# Advanced methods of speech processing and noise reduction for mobile devices

van Khanh Mai

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

UNIVERSITE BRETAGNE LOIRE

**THÈSE / IMT Atlantique**

*sous le sceau de l'Université Bretagne Loire*

pour obtenir le grade de

**DOCTEUR DE IMT Atlantique**

*Mention : Sciences et Technologies de l'Information et de la Communication*

**École Doctorale Sicma**


Présentée par

Van-Khanh Mai

Préparée au département Signal & communication
Laboratoire Labsticc


# Méthodes avancées de traitement de la parole et de réduction du bruit pour les terminaux mobiles

---

# Advanced Methods of speech processing and noise reduction for mobile devices

**Thèse soutenue le 09 mars 2017**

devant le jury composé de :

**Régine Le Bouquin Jeannes**
Professeur, Ecole supérieure d'Ingénieur de Rennes / Présidente

**Jérôme Boudy**
Professeur, Telecom SudParis / Rapporteur

**Abdourrahmane Mahamane Atto**
Maître de conférences (HDR), Polytech Annecy-Chambéry / Rapporteur

**Laurent Navarro**
Maître de conférences, Ecole des mines de Saint-Etienne / Examinateur

**Raphaël Le Bidan**
Maître de conférences, IMT-Atlantique / Examinateur

**Abdeldjalil Aïssa El Bey**
Professeur, IMT-Atlantique / Co-directeur de thèse

**Dominique Pastor**
Professeur, IMT-Atlantique / Directeur de thèse

# Remerciements

Ces travaux de thèse ont été réalisés au sein du département Signal & Communications de IMT-Atlantique. La thèse est financée par la bourse régionale et le PRACOM (Pôle de Recherche Avancée en Communications).

Je tiens tout d'abord à exprimer toute ma gratitude et mes plus vifs remerciements à mes directeurs de thèse, Monsieur Dominique Pastor et Monsieur Abdeldjallil Aïssa-El-Bey, Professeurs à IMT-Atlantique, pour m'avoir donné la possibilité d'entreprendre mes travaux de doctorant, ainsi que pour de leur disponibilité, leur enthousiasme et leur patience malgré leurs nombreuses charges.

Mes remerciements vont également à mon encadrant, Monsieur Raphaël Le Bidan, Maître de Conférence à IMT-Atlantique, pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral et pour ses multiples conseils.

Je tiens aussi à remercier l'ensemble des membres de mon jury : Madame Régine Le Bouquin Jeannes et Messieurs Jerome Boudy, Abdourahmane Atto et Laurent Navarro, pour accepter de participer mon jury de thèse, pour leur lecture de ce manuscrit ainsi que pour les remarques pertinentes sur mes travaux de thèse.

De plus, je remercie tous les membres et ex-membres des équipes COM et TOM du département Signal & Communications de IMT-Atlantique, pour le climat sympathique dans lequel ils m'ont permis de faire de la recherche. Je voudrais exprimer particulièrement toute mon amitié à Thomas Guilment, Budhi Guanadharma et Nicolas Alibert pour leur gentillesse, leurs compétences, leurs conseils, leur amitié et leur humour.

Un grand merci va à mes amis vietnamiens qui m'ont soutenu et avec qui j'ai pu partager des moments inoubliables.

Pour finir, les mots les plus simples étant les plus forts, je tiens à adresser toute mon affection à ma famille et en particulier à ma femme Quynh Trang et ma fille Vivi. Leur confiance, leur amour, leur soutient ont toujours su me porter dans la réussite de ma thèse et continuent de me guider dans la vie.

Merci à vous !

# Table des matières

# Résumé en Français

## Introduction

En traitement du signal une des tâches les plus importantes et fondamentales est l'élimination ou la réduction du bruit de fond. Cette thématique est connue sous le nom de débruitage, suppression du bruit ou rehaussement de la parole dans le cas particulier du traitement de la parole. Cette thèse est consacrée au traitement de la parole, et plus particulièrement à son débruitage. Ces dernières années, l'exploitation du traitement du signal dans les applications mobiles tel que les systèmes de commandes vocales ou les applications dans les smartphones, a connu un intérêt croissant. Dans le cadre de ces applications mobile le rehaussement de la parole a une place centrale. Dans les systèmes de télécommunication, les transmissions ont généralement lieu dans un environnement bruité non-stationnaire ; à l'intérieur d'une voiture, dans la rue ou à l'intérieur d'un aéroport. Le traitement de la parole joue alors un rôle important aux récepteurs pour améliorer la qualité de la parole. Les méthodes du réhaussement de la parole sont également utilisées comme pré-traitement dans les systèmes de codage et de reconnaissance de la parole [1]. Aussi, les algorithmes de rehaussement de la parole peuvent également être appliqués aux prothèses auditives ou aux implants cochléaires pour réduire le bruit ambiant.

Le rehaussement de la parole a pour objectif, d'augmenter le confort auditif d'une part et de diminuer la fatigue de l'auditeur d'autre part . Dans ce contexte, ce rehaussement de la parole vise idéalement à améliorer, non seulement la qualité, mais aussi l'intelligibilité de la parole. Dans la littérature actuelle, les solutions proposées consistent en général à réduire le bruit de font afin d'améliorer la qualité de la parole. Cependant, ces méthodes peuvent générer une distorsion de la parole. C'est la raison pour laquelle, le défi principal du rehaussement de parole est de trouver le meilleur compromis entre la réduction du bruit de fond et la conservation de la qualité de la parole d'origine. De plus, la conception des techniques de rehaussement de la parole dépend aussi de l'application visée, de la bases de données, du type de bruit, de la relation entre le bruit et le signal intérêt et du nombre de capteurs utilisés. En fonction du nombre de capteurs disponibles, les techniques de rehaussement de la parole peuvent être classées en deux catégories : i) les techniques mono-capteur et ii) multi-capteurs. Théoriquement, une amélioration des performances est possible par l'utilisation d'un système multi-capteur au lieu d'un système mono-capteur. Par exemple, un capteur placé près de la source du bruit nous permet d'estimer au mieux ce bruit. Cependant, la complexité de la mise en œuvre, la consommation d'énergie et la taille de l'appareil peuvent être un frein important à la réalisation du rehaussement de la parole dans des applications réelles. De plus, les méthodes utilisant un système mono-capteur peuvent directement être exploitées après un beamforming sur le signal reçu par un système multi-capteur. Par conséquent, nous avons décidé de restreindre notre attention aux méthodes mono-capteur qui sont, non seulement un véritable défi, mais aussi jouent un rôle essentiel dans le traitement de la parole.

De nombreuses méthodes mono-capteur ont été proposées dans la littérature pour le rehaussement de la parole. En général, ces méthodes peuvent être classées en deux catégories :

les méthodes supervisées et non-supervisées. Malgré les bonnes performances obtenues par les méthodes supervisées, les méthodes non-supervisées sont toujours nécessaires. En effet, les méthodes non-supervisée permettent de compenser les lacunes des bases de données qui ne sont pas toujours suffisamment représentatives de l'ensemble des cas d'applications réelles. Dans ces applications, les techniques non-supervisées doivent répondre à tous les critères suivants sans devoir recourir à la aucun apprentissage (training en anglais), ni du bruit ni du signal d'intérêt :

- avoir une bonne performance pour les signaux audio (parole, musique ou autres),
- garantir un bon compromis entre la qualité et l'intelligibilité de la parole
- être robuste aux différents types de bruit stationnaire et non-stationnaire.

Ainsi, la motivation principale de cette thèse est de construire un système complet de débruitage avec des techniques innovantes pour le problème de débruitage de la parole et du signal audio corrompus par un bruit additif. Tout d'abord, une vue d'ensemble de l'architecture générale du système de débruitage mono-capteur en bloc est attentivement étudiée. Cette étude nous permet d'extraire les point clés de chaque bloc et d'identifier les améliorations possibles. Cette thèse est donc divisée en deux parties. Dans la première partie, notre travail consiste à développer une méthode robuste d'estimation du bruit ce qui est un des problème principal dans les systèmes de débruitage monocapteur. Pour ce faire, nous présentons une vue d'ensemble des principales méthodes d'estimation du bruit avec leurs avantages et leurs inconvénients. On nous basant sur cette analyse, nous proposons ensuite une méthode robuste d'estimation du bruit pour les environnements non-stationnaires. Cette méthode repose sur le fait que la transformée de Fourier à courte terme des signaux bruités est parcimonieux dans le sens où les signaux de parole transformés peuvent être représentés par un nombre relativement petit de coefficients avec de grandes amplitudes dans le domaine temps-fréquence. Cette méthode est robuste car elle ne nécessite pas d'information à priori sur la distribution de probabilité du signal intérêt. Ainsi, cette méthode peut améliorer les performances du rehaussement de la parole dans n'importe quel scénario où les signaux bruités peuvent avoir une représentation parcimonieuse faible.

Dans la deuxième partie de cette thèse, nous considérons le cas où nous disposons d'une estimation précise de la densité spectrale de puissance du bruit. Dans ce contexte, nous avons proposé des méthodes de débruitage paramétrique et aussi non-paramétrique. La première famille de méthodes sont des approches paramétriques qui nous permettent d'améliorer non seulement la qualité mais aussi de réduire l'impact négatif sur l'intelligibilité de la parole. Les méthodes proposées sont basées sur la combinaison de la détection et de l'estimation ce qui améliore les performances par rapport aux algorithmes d'estimations paramétriques uniques. Ainsi, deux modèles de la parole bruitées sont pris en compte. Dans le premier modèle, la parole est soit présente, soit absente, alors que dans le deuxième modèle, la parole est toujours présente mais avec différents niveaux d'énergies. La deuxième famille de méthodes sont des approches non-paramétriques. Ces méthodes sont basées sur la fonction SSBS (pour smoothed sigmoid-based shrinkage) dans le domaine de la transformée discrète en cosinus (DCT). Aussi, nous proposons une méthode hybride capable de capter des avantages des méthodes paramétriques et non-paramétriques.

## Architecture générale

Comme nous l'avons introduit précédemment, l'objectif principal de ce travail de thèse est l'étude et le développement d'approches non-supervisées de rehaussement de la parole dans le contexte mono-capteur. Les challenges principaux dans ce contexte pour l'amélioration de la qualité de la

parole sont le manque de ressources (un microphone disponible) et l'absence de bases de données (seul le signal bruité est disponible). Dans ce qui suit (Chapitre 2) nous présentons un aperçu de l'architecture générale des systèmes de débruitage mono-capteur.

Un système de débruitage mono-capteur se compose de quatre blocs principaux : décomposition du signal, estimation du bruit, débruitage et reconstruction du signal (voir Figure 2.1). Le signal bruité observé est segmenté, fenêtré et transformé par une transformée harmonique à court terme dans le bloc de décomposition. En effet, la plupart, des algorithmes de rehaussement de la parole sont appliqués dans un domaine transformé (Discrete Fourier Transform (DFT), Discrete Cosinus Transform (DCT),...) où la séparation entre le signal propre et le bruit est accentuée. La sortie du bloc de décomposition sont donc les coefficients de la transformée à court terme du signal bruité. Ces coefficients sont mis à l'entrée du bloc d'estimation du bruit et du bloc de réduction du bruit. Le bloc d'estimation du bruit à pour objectif d'estimer la densité spectrale de puissance du bruit. L'estimation du bruit est le bloc principal où diverses techniques ont été proposées. Après avoir obtenu une estimation de la densité spectrale de puissance du bruit, un algorithme de réduction du bruit est utilisé pour estimer les coefficients du signal débruités dans le domaine transformé en appliquant une fonction de gain. Cette fonction de gain est généralement calculée à partir de l'amplitude du signal bruité à la sortie du bloc de décomposition et de la densité spectrale de puissance du bruit estimé au bloc d'estimation du bruit. Enfin, le bloc de reconstruction permet de transformer les coefficients estimés dans le domaine temporel. Notez qu'il est possible de récupérer exactement le signal dans le domaine temporel à partir de ses coefficients de la transformée à courte terme (transformations réversibles).

Afin d'évaluer les performances du système de débruitage, un bloc d'évaluation supplémentaire est ajouté (voir Figure 2.7). Dans cette partie, nous présentons certains critères qui sont fréquemment utilisés pour évaluer les performances des méthodes de rehaussement de la parole. Ces critères seront également utilisés dans cette thèse. Ces critères peuvent être divisés en deux catégories ; tests objectifs (SSNR– pour Segmental Signal to noise ratio, SNRI – pour SNR improvement, MARSovrl – pour Multivariate adaptive resgression splines overall speech quality, STOI – pour Short Time Objective Intelligibility) et tests subjectifs (MOS-Mean opinion score). Les tests d'écoute subjectifs sont les critères les plus fiables, mais ils nécessitent plus de temps pour l'évaluation. Certains tests objectifs ont été fortement corrélés avec des tests subjectifs. Par conséquent, ces tests objectifs sont fréquemment utilisés pour évaluer la qualité et l'intelligibilité de la parole.

Comme mentionné précédemment, l'architecture générique des systèmes de rehaussement de la parole est composée de quatre blocs principaux. Par conséquent, une amélioration ou une modification de l'un de ces blocs peut se traduire par une amélioration des performances pour l'ensemble du système. C'est l'objectif des chapitres suivants. Dans le chapitre 3 le bloc d'estimation du bruit sera revisité. Dans le chapitre 4 nous développerons une nouvelle approche pour le bloc de réduction de bruit alors que dans le chapitre 5 nous présenterons une méthode basée sur l'optimisation conjointe des blocs de décomposition du signal et de réduction du bruit.

## Estimation du bruit

Comme nous l'avons présenté précédemment, nous avons motivé l'intérêt d'une approche non-supervisée pour les systèmes de débruitage mono-capteur. Un aperçu général des systèmes a ensuite été présenté. Dans ces systèmes, l'estimation de la densité spectrale de puissance du bruit est une question clé dans la conception des méthodes robustes de réduction du bruit pour le rehaussement de la parole. La question est de savoir comment estimer la densité spectrale de puissance du bruit à partir du signal bruité capturé par un seul capteur. Dans les systèmes

mono-capteur de rehaussement de la parole le principal défi consiste à traiter le cas d'un bruit non-stationnaire. En notant que le signal d'intérêt a une parcimonie faible dans un domaine transformé, une nouvelle méthode d'estimation de la densité spectrale de puissance de bruit non-paramétrique est introduite dans le chapitre 3. Cet algorithme est capable d'estimer efficacement la densité spectrale de puissance du bruit non-stationnaire.

Cette nouvelle méthode ne nécessite pas de modèle ou de connaissances *a priori* des distributions de probabilité des signaux de parole. Fondamentalement, nous ne prenons même pas en considération le fait que le signal d'intérêt ici est la parole. L'approche est appelée extended-DATE (E-DATE) puisqu'elle étend essentiellement le DATE (d-dimensional amplitude trimmed estimator) pour le bruit blanc Gaussien, le bruit stationnaire et le bruit non-stationnaire coloré. Le principe général de l'algorithme E-DATE est la propriété de parcimonie faible de la STFT (pour Short Time Fourier Transform) des signaux bruités. Aussi, la séquence de valeur complexe renvoyée par le STFT dans le domaine temps-fréquence peut être modélisée comme un signal aléatoire complexe avec une distribution inconnue et dont la probabilité inconnue d'occurrence dans le bruit de fond ne dépasse pas la $\frac{1}{2}$. Ainsi, l'estimation du bruit à pour objectif l'estimation de la variance du bruit dans chaque bande fréquentielle ce qui est fourni par le DATE.

L'algorithme E-DATE consiste à réaliser l'estimation de la densité spectrale de puissance du bruit en exécutant l'algorithme DATE pour chaque bande de fréquence sur des périodes de $D$ trames consécutives sans chevauchements, où $D$ est choisi de sorte que le bruit peut être considéré comme approximativement stationnaire dans cet intervalle de temps. Une fois l'estimation de la densité spectrale de puissance du bruit obtenue, elle peut être utilisée pour le débruitage par exemple. Bien que l'algorithme E-DATE ait été spécifiquement conçu pour l'estimation de la densité spectrale de puissance du bruit non-stationnaire, il peut être utilisé sans modification pour l'estimation de la densité spectrale de puissance du bruit blanc ou du bruit stationnaire coloré, ainsi il offre un estimateur de la densité spectrale de puissance du bruit robuste et universel dont les paramètres sont fixés une fois pour tous les types de bruit.

Deux implémentations différentes de l'algorithme E-DATE sont mises en œuvre dans ce chapitre. La première approche est une implémentation simple par blocs de l'algorithme présentée dans Figure 3.3. Il s'agit d'estimer la densité spectrale de puissance du bruit sur chaque période de $D$ trames successives et sans chevauchement. Cela nécessite d'enregistrer $D$ trames, de calculer la densité spectrale de puissance du bruit en utilisant les observations dans ces $D$ trames, puis d'attendre $D$ nouvelles trames sans chevauchement. Cet algorithme s'appelle Bloc-E-DATE (B-E-DATE). L'estimation de la densité spectrale de puissance du bruit sur des périodes séparées de $D$ trames réduit la complexité globale de l'algorithme. Cependant, cela implique une latence de $D$ trames, qui doit être considéré dans les applications à temps réelles. Cette latence peut être contournée comme suit. Tout d'abord, une méthode standard d'estimation est utilisée pour estimer la densité spectrale de puissance du bruit pendant les $D-1$ premières trames. Par la suite, en commençant par la trame $D^{\text{ème}}$ et en faisant glisser une fenêtre d'observation, une version de l'algorithme E-DATE est utilisée pour estimer la densité spectrale de puissance du bruit trame par trame. Cette implémentation alternative s'appelle SW-E-DATE (pour Sliding-Window-E-DATE) montrée dans Figure 3.4.

Les algorithmes B-E-DATE et SW-E-DATE peuvent être considérés comme deux exemples particuliers d'un algorithme général utilisant une fenêtre d'observation. Plus précisément, l'algorithme B-E-DATE correspond au cas extrême où la fenêtre d'observation est totalement vidée et mise à jour une fois toutes les $D$ trames. En revanche, l'algorithme SW-E-DATE correspond à l'autre cas extrême où seul la trame la plus ancienne est enlevée pour stocker la nouvelle, en mode First-In First-Out (FIFO). De toute évidence, une approche plus générale entre ces

deux extrêmes consiste à une mise à jour partielle de la fenêtre d'observation en renouvelant seulement $L$ trames parmi $D$.

Ces algorithmes ont été évaluer en utilisant la base de donnée NOIZEUS. Les résultat sont rapportés dans le tableau 3.1 (pour le nombre de paramètres), dans les figure 3.5, 3.6 (pour l'erreur de l'estimation la densité spectrale de puissance du bruit), la figure 3.7 (pour SNRI), les figures 3.8, 3.9 (pour SSNR), les figures 3.10 et 3.11 (pour MARSovrl). Les résultats expérimentaux montrent que l'algorithme E-DATE fournit généralement l'estimation de la densité spectrale de puissance du bruit la plus précise et qu'il surpasse d'autres méthodes en étant utilisé dans des système de débruitage de la parole en présence de différents types et niveaux de bruit. En raison de ses bonnes performances et sa faible complexité, l'algorithme B-E-DATE devrait être préféré dans la pratique lorsque les fréquences de traitement des données sont suffisamment élevées pour induire des délais acceptables ou même négligeables.

## Débruitage

Dans cette partie, nous proposons deux approches pour estimer l'amplitude spectrale à court terme (STSA). L'objectif principal de cette partie est de prendre en compte les résultats récents de la théorie statistique paramétrique et non paramétrique pour améliorer les performances des systèmes mono-capteur de débruitage de la parole. Le Chapitre 4 prend en considération la théorie statistique en combinant l'estimation et la détection basée sur l'approche paramétrique. Chapitre 5 les performances du rehaussement de la parole sont améliorées en utilisant une approche semi-paramétrique.

### Approche paramétrique

L'objectif de cette partie (chapitre 4) est de de suivre une approche bayésienne visant à optimiser conjointement la détection et l'estimation des signaux de la parole afin d'améliorer l'intelligibilité de la parole. Pour ce faire, nous nous concentrons sur l'estimateur de l'amplitude spectrale basé sur la combinaison de la détection et de l'estimation. En définissant la fonction de coût sur l'erreur d'amplitude spectrale, notre stratégie est de déterminer une fonction de gain sous la forme d'un masque binaire généralisé.

Ainsi, deux modèles d'hypothèse binaire sont utilisés pour déterminer la fonction de gain discontinue. Tout d'abord, on considère les hypothèses binaires où l'absence de la parole est stricte (Strict Model - SM). Dans ce modèle, nous supposons que le signal observé contient du bruit et du signal de parole dans certains atomes temps-fréquence, alors que dans d'autres atomes, l'observation contient uniquement du bruit. La présence de la parole est détectée en contraignant la probabilité de fausse d'alarme comme dans l'approche Neyman-Pearson.

En résumé, pour chaque atome temps-fréquence, la méthode conjointe proposée estime d'abord la STSA de la parole en utilisant l'estimateur bayésien (STSA-MMSE), ainsi le détecteur se base sur cette estimation pour détecter la présence ou l'absence de parole à chaque atome. Si la parole est absent, cette méthode fixe la STSA de la parole à 0. En se concentrant uniquement sur l'estimateur, l'estimation STSA peut être écrite comme un masque binaire. Cette méthode s'appelle SM-STSA. Dans cette méthode, le détecteur dépend de l'estimateur. À son tour, l'estimateur dépend du détecteurs. Cette double dépendance est censée améliorer les performances du détecteur et de l'estimateur.

Deuxièmement, nous supposons que la parole est toujours présente avec différents niveaux d'énergie (Uncertain Model - UM). Plus précisément, sous l'hypothèse nulle, le signal observé est composé du bruit et d'une part négligeable du signal de parole alors que, dans l'hypothèse

alternative, le signal observé est la somme du bruit et de la parole d'intérêt. Comme dans le premier modèle, le détecteur est déterminé par la stratégie de Neyman-Pearson. La différence principale entre les deux modèles est que le premier ne fournit aucune amplitude estimée sous l'hypothèse nulle (la parole est absente) tandis que le dernier introduit une estimation même sous l'hypothèse nulle (le signal de la parole de peu d'intérêt est présent). Ce modèle nous permet de réduire le bruit musical. En effet, les méthodes basées sur le modèle strict de présence/absence de la parole peuvent introduire un bruit musical puisque ces estimateurs peuvent générer au hasard des pics isolés dans le domaine temps-fréquence. Ainsi, sous l'hypothèse nulle, l'estimateur proposé devrait permettre de réduire l'impact de l'erreur des détections manquées. Pour ce modèle, on considère la même fonction de coût pour toutes les situation, nous obtenons ainsi le même estimateur STSA sous les deux hypothèse. C'est la raison pour la quelle nous l'appelons estimateur STSA indépendant (IUM-STSA). Le détecteur influe uniquement sur l'estimateur via un paramètre pondéré (cf. Eq. 4.84).

Pour prendre en compte le rôle de la présence et de l'absence de la parole, nous considérons ensuite la fonction de coût qui nous permet de mettre davantage l'accent sur les détections maquées. L'erreur de détection dépend alors uniquement de la vraie amplitude au lieu de la différence entre la vraie amplitude son estimation. En particulier, lorsque une détection est manquée, le fonction de coût pénalise implicitement non seulement l'erreur estimée mais aussi l'erreur détectée. L'estimation JUM-STSA (c'est-à-dire Joint estimation in the Uncertain Model) peut être écrite comme un masquage binaire généralisé (cf. Eq. 4.93).

Nous avons aussi évalué les performances de nos méthodes proposées sur la base de données NOIZEUS et 11 types de bruit provenant de la base de données AURORA. Les performances de toutes les méthodes proposées et les méthode de référence ont été évaluées dans deux scénarios. Dans le premier scénario, le débruitage est effectué en utilisant la densité spectrale de puissance du bruit de référence. Dans le deuxième scénario, la densité spectrale de puissance du bruit est estimée par la méthode B-E-DATE. Les résultats expérimentaux sont présentés dans les figures 4.2 (pour SSNR), 4.3 (pour SNRI), 4.4 (pour MARSovrl) et 4.5 (pour STOI). Les résultats expérimentaux ont montré la pertinence de l'approche proposée. En d'autre termes, ces résultats expérimentaux confirme l'intérêt de combiner la détection et l'estimation pour l'amélioration de la parole. En effet ces résultats expérimentaux de l'estimateur basé sur la combinaison de la détection et de l'estimation sont généralement meilleurs que ceux de la méthode STSA-MMSE, qui est reconnue comme une approche de référence. Par conséquent, en pratique, nous recommandons l'utilisation de tels détecteurs/estimateurs. Le choix entre eux peut être régi par le type de critère que nous souhaitons optimiser.

**Extension semi-paramétrique**

Dans la partie précédente (Chapitre 4) nous nous sommes concentrés uniquement sur les méthodes paramétriques. Il s'avère que de nombreux résultats dans l'estimation statistique non-paramétrique et robuste établis au cours des deux dernières décennies et basés sur les techniques de seuillage sont suffisamment prometteurs pour suggérer leur utilisation dans le traitement de signal audio non-supervisé afin d'améliorer la robustesse des méthodes de débruitage. De manière générale et comme rappelé ci-dessous, l'intérêt du débruitgae non-paramétrique est double. Tout d'abord, le débruitgae non-paramétrique ne nécessite pas de connaissance a priori de la distribution du signal. Deuxièmement, il permet d'avoir un gain d'intelligibilité de la parole. Étant donné que les approches bayésiennes sont connues pour améliorer la qualité de la parole, l'idée est de combiner ces deux approches. Néanmoins, cette combinaison nécessite d'être mise en place avec soin. En effet, la plupart des estimateurs non-paramétriques forcent à 0 des coefficients de petite amplitude obtenus après une transformation dans un certain domaine. Bien que de

nombreux bruits de fond soient annulés, en éliminant les petits coefficients cela génère du bruit musical et réduit la qualité du signal audio an général et du signal de parole en particulier. Ce problème est bien connu dans le traitement d'image où le forçage à zéro des petits coefficients induit des artefacts.

Par conséquent, si nous voulons améliorer la qualité de la parole en éliminant le bruit musical résiduel, le débruitage non-paramétrique devrait être une bonne alternative dont le principe est d'atténuer les petits coefficients. Un estimateur bayésien peut ensuite être utilisé en aval du débruitage non-paramétrique pour récupérer les informations dans les petits coefficients et ainsi améliorer la qualité globale du signal audio. Une façon de procéder est d'estimer les amplitudes spectrale des coefficients du signal propre dans le domaine temps-fréquence. L'estimation est basée sur le critère MMSE. Cependant, au lieu d'utiliser une DFT, nous nous proposons d'utiliser une transformée en cosinus discrète (DCT), qui évite d'estimer la phase du coefficients et peut réduire la complexité.

Nous commençons par l'amélioration de l'intelligibilité de la parole et de l'audio par une approche non-paramétrique basée sur le SSBS [2], initialement introduit pour le débruitage de l'image. Deux caractéristiques principales de l'approche sont : 1) elle atténue les coefficients DCT qui sont très susceptibles de concerner uniquement le bruit ou la parole avec une faible amplitude dans le bruit ; 2) il tend à maintenir des coefficients DCT de grande amplitude. Cependant, une telle approche non-paramétrique peut être considérée comme un filtrage de Wiener et, en tant que telle, introduit du bruit musical. Nous modifions ensuite l'approche SSBS initiale et proposons l'estimateur de bloc SSBS, ci-après nommé Bloc-SSBS. Bloc-SSBS est pertinent pour éliminer les points isolés dans le domaine temps-fréquence qui peuvent générer du bruit musical. Fondamentalement, Bloc-SSBS applique la même fonction de gain SSBS aux blocs temps-fréquence. La taille de ces blocs est déterminée par le théorème SURE (pour Stein's Unbiased Risk Estimate) [3] afin de minimiser l'estimation impartialle de l'erreur quadratique moyenne sur une régions temps-fréquence. En outre, d'autres paramètres de Bloc-SSBS peuvent être optimisés en se basant sur des résultats récents de traitement du signal statistique non-paramétrique [4] (méthode RDT). Une bonne caractéristique de la procédure d'optimisation des paramètres proposée est le niveau de contrôle offert sur les performances de débruitage qui permet de faire un compromis entre la qualité et l'intelligibilité de la parole. Ceci est rendu possible en distinguant les composants de la parole (ou audio) significatifs et les composants de la parole (resp. audio) avec un intérêt faible.

Les coefficients en sorti de Bloc-SSBS sont supposés satisfaire les mêmes hypothèses que celles généralement utilisées pour l'estimation bayésienne. Par conséquent, dans une deuxième étape, afin de réduire le bruit musical et, surtout, pour améliorer la qualité de la parole, un estimateur statistique bayésien est proposé dans le domaine DCT pour une application à la STSA lissée après Bloc-SSBS. Cette stratégie est nommée BSSBS-MMSE et présentée dans la figure 5.4.

L'évaluation des performances des méthodes proposées ont été effectuées sur la base de données NOIZEUS, avec et sans connaissance de la densité spectrale de puissance du bruit de référence. Différents types de bruits stationnaires et non-stationnaires ont été considérés. Dans le cas où la densité spectrale de puissance du bruit est inconnue, elle est estimés par l'algorithme E-DATE. En outre, des tests objectifs et subjectifs ont été utilisés pour évaluer les performances des estimateurs de la parole. Les tests subjectifs impliquaient un nombre statistiquement significatif d'évaluateurs. Les résultats expérimentaux montrent que BSSBS-MMSE donne de meilleures résultats que les autres méthodes dans la plupart des situations. Ces expériences confirment également la pertinence du choix de la transformée dans le domaine DCT.

## Conclusions

L'objectif de cette thèse était de proposer un système mono-capteur complet d'amélioration de la parole avec des techniques innovantes de traitement du signal pour des applications telles que l'écoute assistée pour les prothèses auditives, les implants cochléaires et les applications de communication vocale avec manque de ressources. Dans ces domaines d'applications, le système complet d'amélioration de la parole devrait non seulement améliorer la qualité de la parole, mais aussi son intelligibilité. En outre, ce système devrait avoir un faible coût de calcul, une faible consommation d'énergie et fonctionner sans aide des bases de données. Afin de surmonter ces contraintes, l'objectif de ce travail est d'évaluer possibilité d'utiliser uniquement des méthodes statistiques non-supervisées, sans recourir à une approche psycho-acoustique ou à de l'apprentissage (supervisé). À cet égard et en tenant compte de la grande quantité de résultats fournis dans la littérature sur le sujet, cette recherche impliquait à la fois des statistiques paramétriques et non paramétriques pour le débruitage audio, lorsque le signal d'intérêt est dégradé par un bruit additif non corrélé et indépendant.

Dans la première partie, l'estimation de la densité spectrale de puissance du bruit a été considérée. Nous avons proposé une nouvelle méthode pour l'estimation de la densité spectrale de puissance du bruit, appelée Étendu-DATE (E-DATE). Cette méthode étend l'algorithme DATE (pour D-dimensional Amplitude Trimmed Estimator), initialement introduit pour l'estimation de la densité spectrale de puissance de bruit Gaussien blanc additif, au cas plus difficile du bruit non-stationnaire. L'idée clé est que, dans chaque bande de fréquence et dans une période de temps suffisamment court, la densité spectrale de puissance instantanée du bruit peut être considérée comme approximativement constante et ainsi estimée comme la variance du bruit gaussien complexe observé en présence du signal d'intérêt. La méthode proposée repose sur le fait que la transformée de Fourier à courte terme des signaux de la parole bruitée est parcimonieuse dans le sens où les coefficients transformés des signaux du signal de parole peuvent être représentés par un nombre relativement petit de coefficients avec de grandes amplitudes dans le domaine temps-fréquence.

L'estimateur E-DATE est robuste car il ne nécessite pas d'informations *a priori* sur la distribution de la probabilité du signal d'intérêt, à l'exception de la propriété de parcimonie faible. Par rapport à d'autres méthodes de l'état de l'art, on constate que l'E-DATE nécessite le plus petit nombre de paramètres (seulement deux). Deux implémentations pratiques de l'algorithme E-DATE ; B-E-DATE et SW-E-DATE, permettent d'obtenir de bonnes performances. En général, l'algorithme E-DATE nous permet d'estimer la densité spectrale de puissance de bruit la plus précise pour différents types et niveaux de bruit. Cet estimateur a également montré sa pertinence pour améliorer la qualité et l'intelligibilité de la parole lorsqu'il est intégré dans un système complet basé sur la méthode STSA-MMSE. Bien que l'algorithme B-E-DATE soit une version simple par blocs de l'algorithme E-DATE, mais il implique un délai d'estimation dû à la latence du traitement. Ceci peut être contourné en recourant à la version SW-E-DATE, basée sur une méthode de fenêtre glissante.

Après l'estimation de la densité spectrale de puissance du bruit par la méthode E-DATE, nous nous sommes concentrés dans la deuxième partie sur les techniques de réduction du bruit. Nous avons considéré deux approches différentes pour récupérer le signal d'intérêt : l'approches paramétrique et non-paramétrique. Dans les deux approches, nous avons exploité une stratégie de combinaison de la détection et de l'estimation pour supprimer ou réduire le bruit de fond, sans augmenter la distorsion du signal. Cette stratégie a été motivée par le fait que, le signal d'intérêt dans le bruit a une représentation parcimonieuse faible qui peut souvent être trouvée

sur une base orthogonale appropriée. Ainsi, nous pouvons supposé raisonnablement que le signal d'intérêt n'est pas être toujours présent dans le domaine temps-fréquence.

Plus précisément, nous avons proposé de nouvelles méthodes pour estimer la STSA de la parole. Ces méthodes sont basées sur la combinaison paramétrique de la détection et de l'estimation. L'idée principale est de prendre en compte la présence et l'absence de la parole dans chaque atome temps-fréquence afin d'améliorer les performances des estimateurs. Les détecteurs optimaux ont été dérivés où ils nous permettent de déterminer l'absence ou la présence du signal de parole dans chaque atome temps-fréquence en fonction de ces estimateurs. Les estimateurs prennent en compte les informations issues de ces détecteurs pour améliorer leurs performances. Deux modèles de signaux incluant une présence et une absence de la parole strictes et incertaines ont été pris en considération. Selon le modèle de signal, la STSA a été forcée à zéro ou remplacé par un petite plancher spectral pour réduire le bruit musical lorsque l'absence de parole a été détectée. Ces méthodes ont été évaluées dans deux scénarios, c'est-à-dire avec et sans connaissance de la densité spectrale de puissance du bruit de référence. Les tests objectifs ont confirmé la pertinence de ces approches en termes de qualité et d'intelligibilité de la parole.

La combinaison de la détection et de l'estimation peuvent être considérées comme une fonction de SSBS. Afin d'améliorer les performances et la robustesse des méthodes de débruitage audio précédemment présentées, une approche semi-paramétrique a été proposée. Il est bien connu que la transformée de Fourier à court terme possède une bonne résolution fréquentielle. Ainsi, la plupart des algorithmes de rehaussement de la parole se base sur cette transformée pour représenter le signal observé dans le domaine temps-fréquence. Cependant, les coefficients de Fourier sont complexes ce qui nécessite une estimation ou une connaissance de la phase de ces coefficients. Pour contourner ce problème, nous avons présenté une nouvelle méthode pour estimer l'amplitude des coefficients du signal de parole dans le domaine temps-fréquence utilisant la transformée cosinus discrète (DCT). Cet estimateur vise à minimiser l'erreur quadratique moyenne de la valeur absolue des coefficients DCT du signal de parole. Afin de tirer des avantages des approches paramétriques et non-paramétriques, on étudie également la combinaison du shrinkage par blocs et de l'estimation bayésienne statistique. Ainsi, la valeur absolue des coefficients du signal d'intérêt est d'abord estimée par Bloc-SSBS. La taille du bloc requise par Bloc-SSBS est obtenue par l'optimisation statistique via l'application du théorème SURE. Cette étape nous permet d'améliorer l'intelligibilité de la parole grâce à un masque binaire lissé. Afin d'évaluer les performances des méthodes proposées, nous avons utilisé des tests subjectif et subjectif informel. Les expériences réalisées démontrent que les méthodes proposées présentent des résultats prometteurs, en termes de qualité et d'intelligibilité de la parole.

En résumé, nous avons proposé plusieurs algorithmes de rehaussement de la parole qui sont tous basés sur une stratégie de combinaison de la détection et de l'estimation. Ceux-ci nous permettent d'améliorer la qualité et l'intelligibilité des signaux vocaux et audio, par rapport aux estimateurs standard. Il est à noter que les approches paramétriques et semi-paramétriques ont été exploitées et que chacune d'entre elles ont montré leur propre pertinence. Par conséquent, selon l'application considérée, un estimateur approprié devrait être choisi. Les estimateurs paramétriques proposés ci-dessus sont plus efficaces pour réduire le bruit musical dans le rehaussement de la parole, alors que les estimateurs non-paramétriques se sont révélés plus pertinents pour le débruitage d'autres types de signaux audio, comme la musique.

## Perspectives

Suite aux travaux réalisés dans le cadre de cette thèse, nous proposons les perspectives suivantes :

1. Bien que notre travail ait porté sur la réduction du bruit dans les systèmes de rehaussement de la parole utilisant la DFT, il faut souligner que l'estimateur E-DATE n'est restreint ni au domaine DFT ni aux signaux de parole. Par conséquent, il pourrait trouver d'autres applications dans n'importe quel scénario où les signaux bruités ont une représentation de parcimonie faible. Par exemple, nous avons réussi à considérer l'utilisation de l'E-DATE dans le domaine DCT. Pour de nombreux signaux d'intérêt, non limités à la parole, une telle représentation de parcimonie faible peut être fournie par une transformation d'ondelettes appropriée. A cet égard, l'application de l'algorithme E-DATE à la séparation de source audio pourrait être considérée. L'estimateur E-DATE repose fondamentalement sur l'estimateur DATE qui peut être considéré comme un détecteur d'anomalie. Par conséquent, l'E-DATE peut également être utilisé comme détecteur d'anomalie dans chaque bande de fréquence. Cela ouvre des perspectives intéressantes dans la détection d'activité vocale basée sur l'analyse de fréquence ainsi que dans la détection et l'estimation de signaux de chirp dans différents types de bruit.

2. Pour tenir compte de la présence ou de l'absence de parole, de nouveaux estimateurs paramétriques ont été proposés en s'appuyant sur la combinaison de la détection et de l'estimation. Ces estimateurs sont basés sur la STSA et la LSA où les hypothèses gaussiennes pour les coefficients DFT sont considérées. Cependant, d'autres distributions pour les coefficients DFT pourraient être étudiées. En outre, plusieurs stratégies qui combinent la détection et l'estimation pour améliorer la performance des estimateurs bayésiens du rehaussement de la parole ont été proposées. L'efficacité de toutes ces approches dépend fortement de la qualité du détecteur. En outre, tous les détecteurs sont basés sur l'hypothèse gaussienne pour les signaux de parole. Étant donné que cette hypothèse peut ne pas être satisfaite, d'autres types de détecteurs de parole dans chaque atome temps-fréquence pourraient être considérés. Une approche prometteuse à cet égard est le détecteur basé sur l'algorithme RDT qui pourra fournir de bonnes performances sans connaissance a priori de la distribution du signal d'intérêt.

3. On a aussi étudié les méthodes de débruitage en utilisant le DCT. Étant donné qu'il ne prend aucune hypothèse sur le signal d'intérêt, Bloc-SSBS peut être appliqué à d'autres applications comme le débruitage de l'image. Nous avons également dérivé un STSA-MMSE dans le domaine DCT en faisant une hypothèse gaussienne sur les coefficients DCT. Il est donc naturel de se demander si d'autres distributions pourraient être plus pertinentes pour la modélisation des coefficients DCT. En outre, il a été observé que bien que la DCT ait une représentation réelle et plus compacte que la DFT, l'application du Bloc-SSBS et du STSA-MMSE dans le domaine DCT sont plus sensibles aux erreurs d'estimation de la densité spectral de puissance du bruit que dans le domaine DFT. Ce point nécessite une étude approfondie.

4. Pour conclure, il convient de noter que toutes les méthodes de rehaussement de la parole exposées dans cette thèse ont été proposées dans le cadre d'un seul microphone disponible et étaient basées uniquement sur des approches statistiques. En tant que tel, quelques perspectives prometteuses apparaissent comme une généralisation naturelle de nos résultats. Tout d'abord et comme discuté dans l'introduction, les systèmes d'amélioration de la parole de multi-microphones peuvent immédiatement s'appliquer et bénéficier des méthodes proposées à la sortie d'une formation de voies. Deuxièmement, les performances de nos algorithmes de rehaussement de la parole peuvent être améliorées en incorporant des informations perceptuelles. Enfin, bien que nous ayons limité l'attention aux approches

non-supervisées, les méthodes proposées peuvent être utilisées comme un post-traitement dans des approches supervisées.

# Abstract

**Abstract:** This PhD thesis deals with one of the most challenging problem in speech enhancement for assisted listening where only one micro is available with the low computational cost, the low power usage and the lack out of the database. Based on the novel and recent results both in non-parametric and parametric statistical estimation and sparse representation, this thesis work proposes several techniques for not only improving speech quality and intelligibility and but also tackling the denoising problem of the other audio signal. In the first major part, our work addresses the problem of the noise power spectrum estimation, especially for non-stationary noise, that is the key part in the single channel speech enhancement. The proposed approach takes into account the weak-sparseness model of speech in the transformed model. Once the noise power spectrum has been estimated, a semantic road is exploited to take into consideration the presence or absence of speech in the second major part. By applying the joint of the Bayesian estimator and the Neyman-Pearson detection, some parametric estimators were developed and tested in the discrete Fourier transform domain. For further improve performance and robustness in audio denoising, a semi-parametric approach is considered. The joint detection and estimation can be interpreted by Smoothed Sigmoid-Based Shrinkage (SSBS). Thus, Block-SSBS is proposed to take into additionally account the neighborhood bins in the time-frequency domain. Moreover, in order to enhance fruitfully speech and audio, a Bayesian estimator is also derived and combined with Block-SSBS. The effectiveness and relevance of this strategy in the discrete Cosine transform for both speech and audio denoising are confirmed by experimental results.

**Keywords:** *speech and audio enhancement, noise reduction, spare representation, parametric estimator, joint detection and estimation, sparse thresholding, non-parametric estimator.*

# Résumé

Cette thèse traite d'un des plus problème stimulant dans traitement de la parole pour la prothèse auditive où un seul capteur est disponible avec les faibles coûts de calcul, la faible l'utilisation d'énergie et l'absence de bases de données. Basée sur les récents nouveaux résultats dans les deux estimation statistiques paramétrique et non-paramétrique et la représentation parcimonieuse, cette étude propose quelques techniques pour non seulement améliorer la qualité et l'intelligibilité de la parole, mais aussi s'attaquer au débruitage du signal audio en général. La thèse est divisé en deux parties. Dans la première partie, on aborde la problème d'estimation de la densité spectrale de puissance du bruit, particulièrement pour le bruit non-stationnaire. Ce problème est un des parties principales du traitement de la parole du mono-capteur. La méthode proposée prend en compte le modèle parcimonieux de la parole dans le domaine transféré. Lors que la densité spectrale de puissance du bruit est estimée, une approche sémantique est exploitée pour tenir en compte la présence ou absence de la parole dans la deuxième partie. En combinant l'estimation Bayésienne et la détection Neyman-Pearson, quelques estimateur paramétriques sont développés et testés dans le domaine Fourier. Pour approfondir la performance et la robustesse de débruitage du signal audio, une approche semi-paramétrique est considérée. La conjointe détection et estimation peut être interprétée par Smoothed Sigmoid-Based Shrinkage (SSBS). Donc, la méthode Bloc-SSBS est proposée pour prendre en compte les atomes voisinages dans le domaine temporel-fréquentiel. De plus, pour améliorer fructueusement la qualité de la parole et du signal audio, un estimateur Bayésien est aussi dérivé et combiné avec la méthode Bloc-SSBS. La efficacité et la pertinence de la stratégie dans le domaine de transformée cosinus pour les débruitages de la parole et de l'audio sont confirmées par les résultats expérimentaux.

**Mots clés :** enrichissement de la parole et de l'audio, débruitage statistique, représentation parcimonieuse, estimation paramétrique, combinaison de détection et estimation, seuillage parcimonieux, estimation non-paramétrique

# Acronyms

| | |
|---|---|
| **AI** | Articulation Index |
| **ANC** | Active Noise Cancellation |
| **AR** | Auto-regressive |
| **B-E-DATE** | Block Extend-d-Dimensional Amplitude Trimmed Estimator |
| **BSSBS** | Block Smoothed Sigmoid-Based Shrinkage |
| **DCT** | Discrete Cosine Transform |
| **DATE** | d-Dimensional Amplitude Trimmed Estimator |
| **DFT** | Discrete Fourier Transform |
| **DNN** | Deep Neutral Network |
| **DTFT** | Discrete Time Fourier Transform |
| **E-DATE** | Extended d-Dimensional Amplitude Trimmed Estimator |
| **FA** | False Alarm |
| **HMM** | Hidden Markov Model |
| **IDCT** | Inverse Discrete Cosine Transform |
| **IMCRA** | Improved Minima-Controlled Recursive-Averaging |
| **IS** | Itakura-Saito distance |
| **ISTCT** | Inverse Short Time Cosine Transform |
| **ISTFT** | Inverse Short Time Fourier Transform |
| **ISTT** | Inverse Short Time Transform |
| **IUM** | Independent Uncertain Model |
| **JUM** | Joint Uncertain Model |
| **KLT** | Karhumen-Loève Transform |
| **LSA** | Log-Spectral Amplitude |
| **MAP** | Maximum A Posteriori |

| | |
|---|---|
| **MARS** | Multivariate Adaptive Regression Spline |
| **MCRA** | Minima-Controlled Recursive-Averaging |
| **ML** | Maximum Likelihood |
| **MMSE** | Minimum Mean Square Error |
| **MOS** | Mean Opinion Score |
| **MS** | Minimum-Statistic |
| **NMF** | Non-negative Matrix Factorization |
| **NR** | Noise Reduction |
| **OSM** | Optimal Strict Model |
| **PESQ** | Perceptual Evaluation of Speech Quality |
| **PDF** | Probability Density Function |
| **RDT** | Random Distortion Threshold |
| **SII** | Speech Intelligibility Index |
| **SM** | Strict Model |
| **SNR** | Signal to Noise Ratio |
| **SNRI** | Signal to Noise Ratio Improvement |
| **SSBS** | Smoothed Sigmoid-Based Shrinkage |
| **SSNR** | Segmental Signal to Noise Ratio |
| **SSM** | Sub-optimal Strict Model |
| **STCT** | Short Time Cosine Transform |
| **STFT** | Short Time Fourier Transform |
| **STHT** | Short Time Harmonic Transform |
| **STI** | Speech Transmission Index |
| **STOI** | Short-Time Objective Intelligibility |
| **STSA** | Short-Time Spectral Amplitude |
| **STT** | Short Time Transform |
| **SURE** | Stein's Unbiased Risk Estimate |
| **SVD** | Singular Value Decomposition |
| **SW-E-DATE** | Sliding-Window Extended d-Dimensional Amplitude Trimmed Estimator |

| | |
|---|---|
| **UM** | Uncertain Model |
| **UMP** | Uniform Most Powerful |
| **WGN** | White Gaussian Noise |

# List of Figures

# List of Tables

# Part I

# Introduction

Chapter 1

# Introduction

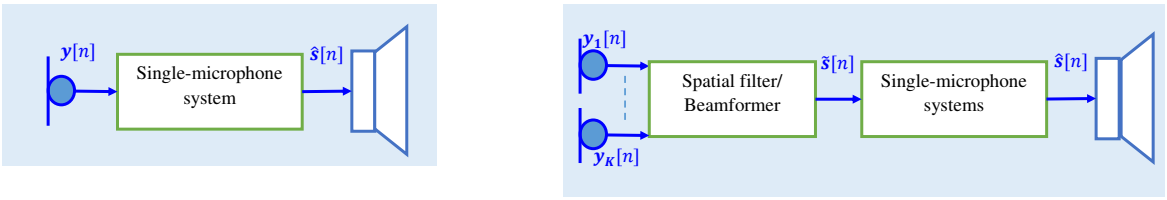
*The important thing is to not stop questioning. Curiosity has its own reason for existence.*

Albert Einstein

## 1.1 Context of the thesis

One of the most fundamental, long-studied and important task in signal processing is the removal or reduction of background noise from a noisy signal, known as denoising, noise suppression or speech enhancement in the particular case of speech signal. This thesis is dedicated to speech enhancement, especially to signal processing techniques for assisted listening. With an increasing interest in mobile speech processing applications such as voice control devices, smart phone application, assisted listening, etc, improving speech quality is a basic requirement in many situations. Communication electronic support, telephone communication, in particular often take place in noisy and non-stationary environments such as the inside of a car, in the street or inside an airport. Speech enhancement methods thus play an important role at the receiving end to improve speech quality. Speech enhancement techniques are also used as pre-processing in speech coding or speech recognition systems, which can be employed in telephone [1]. Speech enhancement algorithms can be also applied to hearing aids like hearing impaired listener or cochlear implant devices for reducing noise before amplification.

Speech enhancement is expected to increase the comfort and also to reduce listener's fatigue. In this respect, speech enhancement ideally aims at improving not only the quality but also the intelligibility of noisy speech. Various solutions make it possible to remove the background noise so as to enhance speech quality. However, they introduce speech distortion. Thus, the main challenge of speech enhancement algorithms is to reduce residual noise without distorting too much the speech signal. Moreover, the design of a speech enhancement technique depends also on the application, the database resource, the nature of noise, the relationship between noise and clean speech, and the number of microphones in the device. Considering the number of microphones or sensors available, speech enhancement technique can be classified into single-microphone and multi-microphones techniques. Technically, the larger the number of microphones, the better the speech quality. For instance, a microphone placed close to the noise source provides a better noise estimate. However, the computational complexity, power consumption, size demands of devices, and etc may impede their usability in real application, for example the invisible in the ear canal hearing aid. Moreover, a technique designed in the single channel case can always be used after beamforming on a microphones array. Indeed, for Gaussian noise model, an optimal method for multi-channel noise reduction is a combination of a minimum-variance distortion-less response multi-microphone beamformer with a single-channel noise reduction algorithm [5]. Figure 1.1 displays the role of single channel technique in the two situations. Therefore, we restrict our attention to single microphone, which is not only the most challenging problem but also play a central role in speech enhancement.




(a) Single channel

(b) Multi-channel

Figure 1.1 – Single channel block function 1.1a and post-processor single channel in block diagram of multi-channel denoising system 1.1b after [6].

The next section will provide a brief review of single channel denoising methods, including supervised and non-supervised approaches, from which we find motivation for taking a closer look at unsupervised techniques. Then, an advanced strategy is going to be proposed by taking into account the constraint of the application. Finally, the main objective, motivation and outline of this thesis will be introduced in Section 1.3.

## 1.2 A brief history of speech enhancement

Many methods have been proposed in the literature for single channel speech enhancement. In general, these methods can be categorized into two broad classes including supervised and unsupervised approaches. Thus, these two types of approaches will be reviewed here by dividing them into some basic sub-classes.

### 1.2.1 Unsupervised methods

Many algorithms have been proposed for speech enhancement with the primary objective to improve speech quality and intelligibility. A detailed review can be found in [1, 7], most of them operating in the Discrete Fourier Transform (DFT) domain in [6]. These methods can be divided into two principal approaches including parametric and non-parametric approaches. In parametric approach, the signal distribution is known. Therefore, possibly up to a certain vector parameter, that makes it possible to resort to standard Bayesian and likelihood theory. In non-parametric approach, the signal distribution is unknown.

**Non-parametric approach:** In this framework, the simplest speech enhancement methods to implement are power spectral subtractions. The methods can be carried out with low computation and without much prior information [8–12]. They are based only on the basic signal model where noise is additive. Another technique is the optimal Wiener algorithm, which assumes a linear relationship between the noisy coefficients and the clean signal coefficients [13–17]. Other non-parametric estimators are based on subspace decomposition. The main idea is that the noisy space can be decomposed into a clean signal space and a noise-only space [18–22]. Recently, some binary masking methods have been proposed in order to improve speech intelligibility [23–25]. In the time-frequency domain, the techniques consist in keeping only some frequency bins from the noisy spectra while forcing to zeros the remaining ones.

**Parametric approach:** By taking the distribution of the clean speech and noise into account, this approach estimates clean speech by formulating denoising as an estimation problem using either maximum likelihood (ML) [26], minimum mean square error (MMSE) [27,28] or maximum *a posteriori* (MAP) [29, 30] estimators. In order to derive MAP and MMSE estimators, the probability density function (PDF) of speech can be assumed to be Gaussian [27], super-Gaussian [31,32], Laplacian [33] or generalized gamma [34]. For MMSE estimators, the cost functions are the mean-square error of magnitude or log-magnitude spectra or the distortion measures, for instance, Itakura-Saito or Cosh measures [35]. In most parametric techniques mentioned above, noise is assumed to be Gaussian. In fact, noise is also supposed to have Laplacian distribution [36]. Some techniques incorporate also the knowledge of speech presence or absence to further improve speech quality [37–39].

### 1.2.2 Supervised methods

For the supervised approach, both the speech and noise model parameters are estimated by learning from the corresponding training samples. Based on these model parameters, a strategy is proposed to combine the signal of interest and the noise models. Then, the denoising problems are tackled with the noisy signal. This broad approach can be divided into four main classes: codebooks-based Wiener approach, Hidden Markov Model (HMM) based approach, dictionary-based approach and Deep Neural Network (DNN) based approach.

**Codebooks-based Wiener approach:** Based on the Wiener filter, this approach uses codebooks of auto-regressive (AR) parameters for linear prediction synthesis of the speech and noise signals. In fact, the Wiener filter is the ratio of the clean signal and the noisy power spectrum. Moreover, the noisy power spectrum is reasonably assumed to be the sum of the clean signal and noise power spectra. These spectra can be determined from the AR parameters. Therefore, this approach first builds codebooks for speech and noise spectra via training the clean signal and noise database. This training can perform offline for both the clean and noise signals [40–42] or offline only for signal and online for noise [43]. The AR parameters (AR coefficients and gain) of the observed signal are then estimated by ML or Bayesian MMSE criteria based on the code-books.

**HMM-based approach:** Here, instead of linear prediction synthesis, the clean speech and noise AR or the other parameters are modeled by HMM. In [44–46], the speech and noise AR parameters are assumed to be Gaussian. More recently, the authors of [47] and [48] work directly with the coefficients in the transformed domain where these signal coefficients are assumed to have complex Gaussian or super Gaussian distribution. The model parameters have been trained from the speech and noise databases via the Expectation Maximization (EM) algorithm. Finally, for estimation of the clean speech, a maximum *a posteriori* (MAP) or a Bayesian MMSE have been proposed to process the noisy signal based on the model parameters. These processes can be done in the discrete Fourier transform (DFT) domain or in the reduced-resolution mel frequency domain.

**DNN-based approach:** DNN has a long history, but was only applied to speech enhancement at the end of year 2013 [49]. Like other supervised approaches, DNN-based approach has two stages: the training and the enhancement stage. The logarithm of clean and noisy amplitude and phase in DFT domain are the parameters of interest here [49–51]. A regression DNN model is used to train these parameters from the signal and noise database. The trained DNN is then fed with the noisy speech to estimate the amplitude of clean speech. In addition, a post-processor can be incorporated to further improve speech quality [50, 51].

**Dictionary-based approach:** This approach can be separated into K-SVD-based methods [52–55] and non-negative matrix factorization (NMF)-based methods [56–59]. The main idea is that, a dictionary or a non-negative matrix for clean speech and/or for noise is trained offline from the database based on K-SVD [60] or on NMF [61]. An over-complete matrix is frequently constructed by concatenating the trained matrix of clean speech with one of noise. In the enhancement stage, a Wiener filter-type, or MMSE estimators are derived from the noisy signal and the over-complete matrix.

## 1.3 Thesis motivation and outline

Despite good results obtained by machine learning (supervised) based approach, there is still room for unsupervised techniques, especially in applications where large enough databases are hardly available for all the types of noise, speech and audio signals that can actually be encountered. This is the case in assisted listening for hearing aids, cochlear implants and voice communication applications with lack of resources. That is the reason why we decided to further investigate unsupervised approach.

In such applications, unsupervised techniques are then expected to fulfill the following criteria, without resorting to any prior training, either for noise or for the signal of interest. Any such method is asked to perform well on both speech and audio signals in noise. It should achieve a good trade-off between intelligibility and quality, for both audio and speech. It must be robust to various stationary and non-stationary types of noise. Its complexity must be low so as to limit computational cost in real-time applications.

Therefore, the main motivation in this thesis work is to construct a complete denoising system with innovative techniques for audio denoising problem where the signal of interest is degraded by uncorrelated and independent additive noise. This system should have low computational cost and low power usage without the help of any database. We also assume that a single noisy observation is available at the system input. It now turns out that many results both in non-parametric and parametric statistical estimation established in the last two decades [4, 62–68] and based on sparse thresholding and shrinkage, are general enough to suggest their use in unsupervised speech and audio denoising. It is worth noting that the parametric approach provides some statistical optimality in terms of MMSE or MAP criteria whereas the non-parametric leads to robustness. Therefore, this work investigates both parametric and non-parametric statistical approaches for single channel speech enhancement. This suggests the use of a semi-parametric method to take advantage of both approaches.

With respect to the content of our research and for the sake of clarity, the thesis will be organized in four main parts : "Introduction", "Noise", "Speech" and "Conclusion and Future Work".

**Part I** consists of Chapter 1 and 2. A general introduction of the thesis is already presented in this chapter. Then, Chapter 2 provides a brief overview of the main single channel speech enhancement methods, which allows us to clearly identify the main foundations of speech enhancement systems and to point out areas for potential improvement.

The next two parts are the main contribution of this PhD research work.

**Part II** reduces to Chapters 3. The main problem of all supervised approach is noise estimation accuracy. Hence, this part focuses on the problem of noise estimation. Our work looks first for a robust noise estimation solution. Many noise estimation methods have been proposed in the literature. Is there still room for further improvement? To answer this question, we present a review of the major noise estimation methods with their advantages and drawbacks. We then propose a robust noise power spectra estimator for non-stationary environments that relies on the fact that the Short-Time Fourier Transform (STFT) of noisy speech signals is sparse in the sense that transformed speech signals can be represented by a relatively small number of coefficients with large amplitudes in the time-frequency domain. The proposed estimation method is robust in that it does not require prior information about the signal probability distribution. Thus, this method can improve performance of speech enhancement in any scenario where noisy signal have weak-sparseness representations.

**Part III** is the core of this research work and consists of Chapter 4 and Chapter 5. All proposed speech enhancement algorithms are developed in this part.

- Chapter 4 proposes a parametric approach for enhancing not only speech quality but also reducing the negative impact on speech intelligibility. The proposed methods are based on joint detection and estimation that improve upon previous parametric algorithms. To this end, two models of noisy speech are taken into account. In the first model, speech is either present, or absent, whereas in the second model the speech is always present but with various levels of energy. The later is also called uncertain speech absence/presence.
- Chapter 5 describes a non-parametric algorithm for audio enhancement. The new non-parametric approach is based on sparse coding and smoothed sigmoid-based shrinkage (SSBS) in the discrete cosine transform (DCT) domain for dealing with speech and other audio signal like music. Moreover, we propose a combined method that captures the advantages of both parametric and non-parametric solution.

**Part IV** is Chapter 6 concludes the PhD thesis and provides some perspectives for further work.

Chapter 2

# Single microphone speech enhancement techniques

*What we know is a drop, what we don't know is an ocean.*

Isaac Newton

## 2.1 Introduction

As introduced in the first chapter, our primary goal in this thesis work is to pursue and improve upon the unsupervised approach in the single microphone situation. The problem in this situation is one of the most difficult problems in speech enhancement because of low resource (one microphone available), lack of database (only noisy signal is presented). This chapter will present an overview of single microphone systems for speech enhancement. In Section 2.2, we first describe the general structure of single microphone system and detail each block in this system. Section 2.3 then introduces several metrics which will be used to evaluate and validate the performance of speech enhancement algorithms in this thesis. Section 2.4 concludes the chapter.

## 2.2 Overview of single microphone speech enhancement system

In audio and speech enhancement, one of the most important tasks is the removal or reduction of background noise from a noisy signal. The observed signal is frequently segmented, windowed and transformed into a representation domain. Then, the clean signal coefficients are usually retrieved by applying an enhancement algorithm to the noisy observations in this domain. Figure 2.1 shows a basic single channel speech enhancement system block diagram. A single microphone system consists of four blocks: Decomposition, Noise Estimation, Noise Reduction Algorithm and Reconstruction Blocks, respectively. In short, the process is performed as follows. First, the noisy signal $\mathbf{y}[n]$ is decomposed using a short time harmonic transform (STHT) in the decomposition block. Second, the time-frequency noisy coefficient $Y[m,k]$ is modified to obtain the enhanced coefficient $\widehat{S}[m,k]$ in the noise reduction block. Note that the noise estimation block provides the noise power spectrum $\widehat{\sigma}_X^2[m,k]$, which is an important input of the noise reduction block. Finally, the enhanced signal $\widehat{\mathbf{s}}[n]$ is synthesized from the enhanced time-frequency coefficient $\widehat{S}[m,k]$ in the reconstruction block. Specially, we used the Hamming window and 50% overlap-add method in implementation for all the algorithms in this thesis. We now describe the role of each block in detail in the following sub-sections.

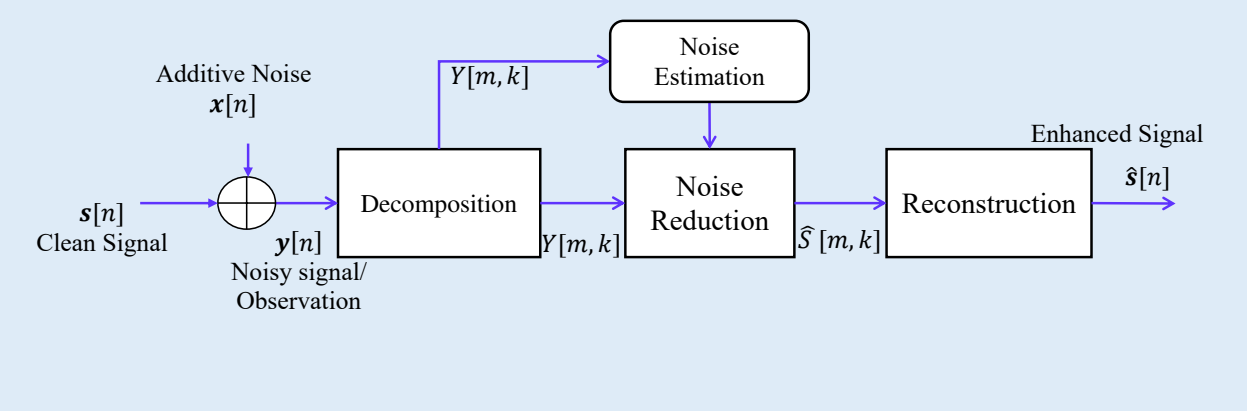



Figure 2.1 – General principle of the classical audio enhancement system [1].

### 2.2.1 Decomposition block

The noisy observed signal is segmented, windowed and transformed by a computational harmonic transform in the decomposition block. In fact, most but not all of speech enhancement algorithms

proceed in the time-domain but rather in a transformed domain where the separate between clean signal and noise is made easier. As mentioned above, we concentrate on speech enhancement scenario where noise is uncorrelated and additive. Therefore, the noisy signal is modeled by $\mathbf{y}[n] = \mathbf{s}[n] + \mathbf{x}[n]$, where $\mathbf{s}$ and $\mathbf{x}$ are respectively the clean signal and independent noise in the time domain and $n = 0, 1, \ldots, N-1$ is the sampling time index. Most enhancement algorithms operate on frame-by-frame where only a finite collection of observation $\mathbf{y}[n]$ is available. A time-domain window $w[n]$ is usually applied to the noisy signal, yielding the windowed signal as:

$$\mathbf{y}_W[n] = \mathbf{y}[n]w[n]. \tag{2.1}$$

In frame-based signal processing, the shape of window is obtained by trading-off between smearing and leakage effects [69]. The second parameter is the window length, which allows to trade-off between spectral resolution and statistical variance. In speech enhancement, if the length of window is too large, can no longer speech be considered stationary within a frame. On the other hand, if the length is too small, the spectral solution may not be accurate enough. Based on previous consideration, Hanning and Hamming window functions are often chosen to truncate the signal of interest in the considered frame. The shape of these windows is illustrated in Figure 2.2. In this thesis, we prefer the Hamming window function, which does not vanish to zero at the end. The Hamming window function is defined as follow:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{K-1}\right) & 0 \leq n \leq K-1, \\ 0 & \text{Otherwise} \end{cases} \tag{2.2}$$

where $K$ is the length of the window.

Once the truncated signal $\mathbf{y}_W[n]$ has been obtained, a short time transform is applied. Common short time transforms include wavelet, Fourier and cosines transform. Let us denote the noisy signal in the transformed domain by:

$$Y[m,k] = S[m,k] + X[m,k], \tag{2.3}$$

where $m$ and $k \in \{0, 1, \ldots, K-1\}$ are the time and frequency-bin indices, respectively. The transformed coefficients can be obtained as:

$$Y[m,k] = \sum_{n=0}^{K-1} \alpha_k[n]w[n]\mathbf{y}[mK^* + n], \tag{2.4}$$

where $K^*$ is the number of shifted samples of the two consecutive frames and $\alpha_k[n]$ is a scaling coefficient dependent on the transform. For instance, in the short time Fourier transform (STFT), the value $\alpha_k^{\mathrm{DFT}}[n]$ is defined following:

$$\alpha_k^{\mathrm{DFT}}[n] = \exp\left(-j\frac{2\pi}{K}kn\right), \tag{2.5}$$

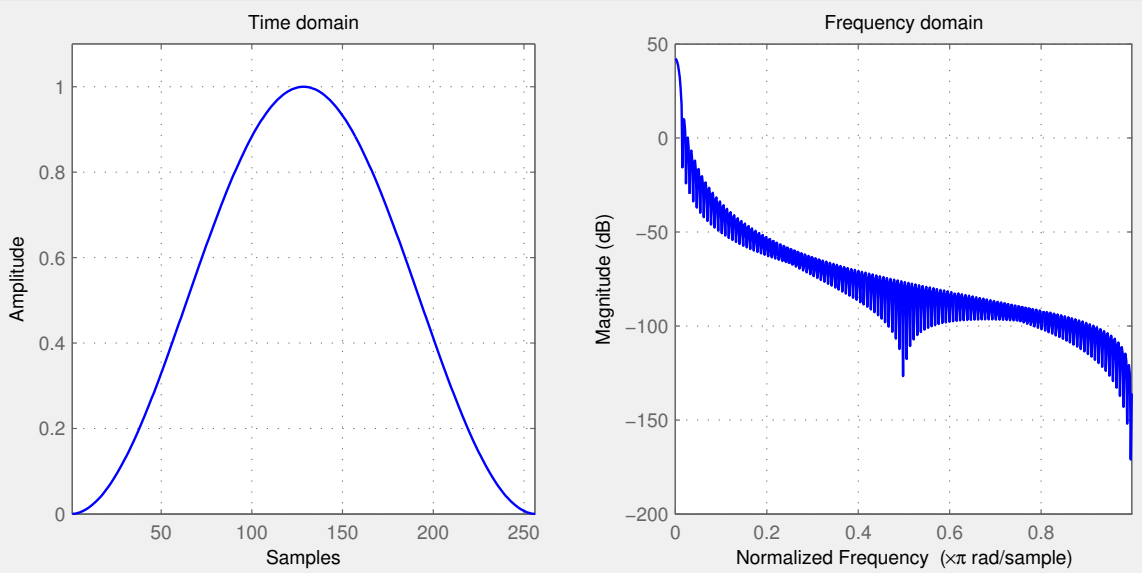
whereas, for the short time cosines transform (STCT):

$$\alpha_k^{\mathrm{DCT}}[n] = \sqrt{\frac{1 + \mathbb{1}_{(0,\infty)}(n)}{K}}\cos\left(\frac{\pi}{2K}k(2n+1)\right), \tag{2.6}$$

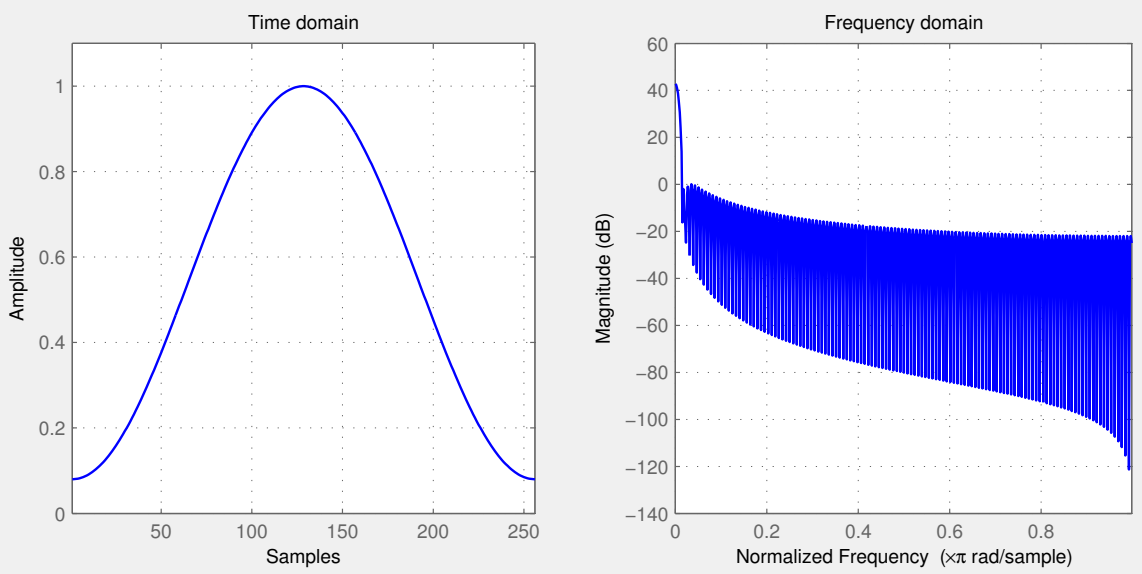
where $\mathbb{1}_{(0,\infty)}(\theta)$ is the indicator function $\mathbb{1}_{(0,\infty)}(\theta) = 1$ if $\theta > 0$ and $\mathbb{1}_{(0,\infty)}(\theta) = 0$ otherwise.

The output of the decomposition block, the noisy signal in the transformed domain $Y[m,k]$, can be written in polar form as:

$$Y[m,k] = A_Y[m,k]\exp\left(j\phi_Y[m,k]\right), \tag{2.7}$$




(a) Hanning window




(b) Hamming window

Figure 2.2 – Two window functions are frequently used in speech enhancement system: the Hanning window shown in Figure 2.2a and the Hamming window shown in Figure 2.2b.

where $A_Y[m,k]$ and $\phi_Y[m,k]$ denote the amplitude and the phase of the noisy signal in the $[m,k]$ time frequency-bin, respectively. Therefore, the signal model is now given by:

$$A_Y[m,k]\exp\left(j\phi_Y[m,k]\right) = A_S[m,k]\exp\left(j\phi_S[m,k]\right) + A_X[m,k]\exp\left(j\phi_X[m,k]\right), \tag{2.8}$$

where $\{A_S[m,k], A_X[m,k]\}$ are also the amplitudes and $\{\phi_S[m,k], \phi_X[m,k]\}$ are the phases of the clean speech, and the noise signal, at the frame $m$ and frequency bin $k$. The well-known quantity periodogram, handled the spectrographic analysis of speech signal, is specified as:

$$\|S[m,k]\|^2 = A_S^2[m,k], \tag{2.9}$$

where $\|\cdot\|$ denotes the $\ell_2$ norm. Moreover, the same definitions $\|Y[m,k]\|^2$ and $\|X[m,k]\|^2$ are used for the periodogram of the noisy speech and the noise signal, respectively.

To summarize, the output $Y[m,k]$ of the decomposition block is the short time transform coefficient of the truncated noisy signal frame $\mathbf{y}[n]$ where $n \in \{mK^*, mK^*+1, \ldots, mK^*+K-1\}$. This coefficient is the input of the noise estimation and noise reduction blocks. The noise estimation block often solely uses the noisy signal periodograms $\|Y[m,k]\|^2$ whereas noise reduction block takes into consideration both the noisy amplitude $A_Y[m,k]$ and the noisy phase $\phi_Y[m,k]$.

### 2.2.2 Noise estimation block

The noise estimation block aims at estimating the power spectrum $\sigma_X^2[m,k] = \mathbf{E}[\|X[m,k]\|^2]$. Therefore, the noise estimation is the main block where various techniques have been proposed. In this section, we discuss only some general points for completeness. For further detail about noise estimation, readers are invited to consult Chapter 3 in Part II. Most noise estimation algorithms are based on the following assumptions [1, Chapter 9]:

(**A1**) As mentioned above, the speech signal is degraded by a statistically independent additive noise.

(**A2**) Speech is not always present. Thus, we can always find an analysis segment, formed by some consecutive frames, that contains speech-pause or noise-only.

(**A3**) Noise is more stationary than clean speech so that we can assume that noise remains stationary within a given analysis segment.

As an example, we will detail one of the first noise power spectrum estimation based on *minimum-statistic* (MS) [70]. This algorithm tracks the minimum value of the noisy speech power spectrum within an analysis segment. For the reason that noise and speech are statistically independent (**A1**), the periodogram of noisy speech is approximated as:

$$\|Y[m,k]\|^2 \approx \|X[m,k]\|^2 + \|S[m,k]\|^2. \tag{2.10}$$

Based on this approximation and the assumption (**A2**), when speech is paused or absent, the periodogram $\|S[m,k]\|^2 = 0$ then $\|Y[m,k]\|^2 \approx \|X[m,k]\|^2$. Moreover $\|S[m,k]\|^2 \geq 0$ so that $\|Y[m,k]\|^2 \geq \|X[m,k]\|^2$. Therefore, the minimum of the periodogram $\|Y[m,k]\|^2$ over a given analysis segment is the estimated noise power spectrum. The periodogram $\|Y[m,k]\|^2$ varies quickly over time. Thus, in order to estimate the noise power spectrum $\sigma_X^2$, instead of the periodogram, a recursive smoothed periodogram is used:

$$P[m,k] = \alpha P[m-1,k] + (1-\alpha)\|Y[m,k]\|^2, \tag{2.11}$$

where $P[m,k]$ is a first-order recursive version of the periodogram, or smoothed periodogram and $\alpha$ is the smoothing constant, which was recommended be equal to 0.95 in [70]. The noise power spectrum is now estimated by tracking the minimum of the smoothed periodogram $P[m,k]$ over an analysis segment. The length of the analysis segment should be long enough to include speech pause but should remain small enough at the same time to track accurately and to adapt to non-stationary noise. Let us denote the minimum of the smoothed periodogram at the frame $m$ and the frequency bin $k$ by $P_{\min}[m,k]$, determined over an analysis segment of $D$ consecutive frames. The minimum $P_{\min}[m,k]$ is updated only after a given analysis segment as show by Algorithm 1 [1]. In this algorithm, $P_{\text{tmp}}[m,k]$ is the temporary minimum periodogram where $P_{\text{tmp}}[0,k] = P[0,k]$ and $mod(\cdot)$ is the modulus operator. Effectively, the temporary minimum periodogram makes it possible to update the minimum of the smoothed periodogram over every $D$ consecutive frames. Once $P_{\min}[m,k]$ is tracked, the estimated noise power spectrum $\widehat{\sigma}_X^2[m,k]$ is given as:

$$\widehat{\sigma}_X^2[m,k] = B_{\min} P_{\min}[m,k], \tag{2.12}$$

where $B_{\min}$ is a factor, which enables to compensate for the bias of the minimum estimate. This factor was found to depend only on the $D$ parameter [71]. Figure 2.3 illustrates an example of minimum tracking at frequency of 500 Hz. We used in this example the parameters recommended in [71], namely $B_{\min} = 1.5$, $\alpha = 0.95$ and $D = 50$.

**Algorithm 1:** Simple MS algorithm for tracking the minimum of the smoothed periodogram and updating it.

**for** $m = 1$ to the end of signal **do**
  **if** $\text{mod}(m/D) = 0$
   $P_{\min}[m,k] = \min\{P_{\text{tmp}}[m-1,k], P[m,k]\}$
   $P_{\text{tmp}}[m,k] = P[m,k]$
  **else**
   $P_{\min}[m,k] = \min\{P_{\min}[m-1,k], P[m,k]\}$
   $P_{\text{tmp}}[m,k] = \min\{P_{\text{tmp}}[m,k], P[m,k]\}$
  **end if**
**end for**

### 2.2.3 Noise reduction block

Once the noise power spectrum estimation is obtained, in single microphone system, a noise reduction algorithm is used for retrieving the enhanced signal $\widehat{S}[m,k]$. Like the noise estimation block, in this section, for the sake of self-completeness, we chose to present one of the first noise reduction method, which is computationally efficient [9] and called the power spectral subtraction algorithm. Further details will be given in the following chapters.

For most noise reduction algorithms, we can define a gain function $G[m,k]$ for which the enhanced amplitude of the signal of interest $\widehat{A}_S[m,k]$ is obtained as follows:

$$\widehat{A}_S[m,k] = G[m,k] A_Y[m,k] \tag{2.13}$$

whereas the enhanced phase $\widehat{\phi}_S[m,k]$ is made equal to the noisy phase $\phi_Y[m,k]$. Therefore, the estimated coefficient in the transformed domain $\widehat{S}[m,k]$ is :

$$\widehat{S}[m,k] = \widehat{A}_S[m,k] \exp\left(j\widehat{\phi}_S[m,k]\right) = G[m,k] Y[m,k]. \tag{2.14}$$

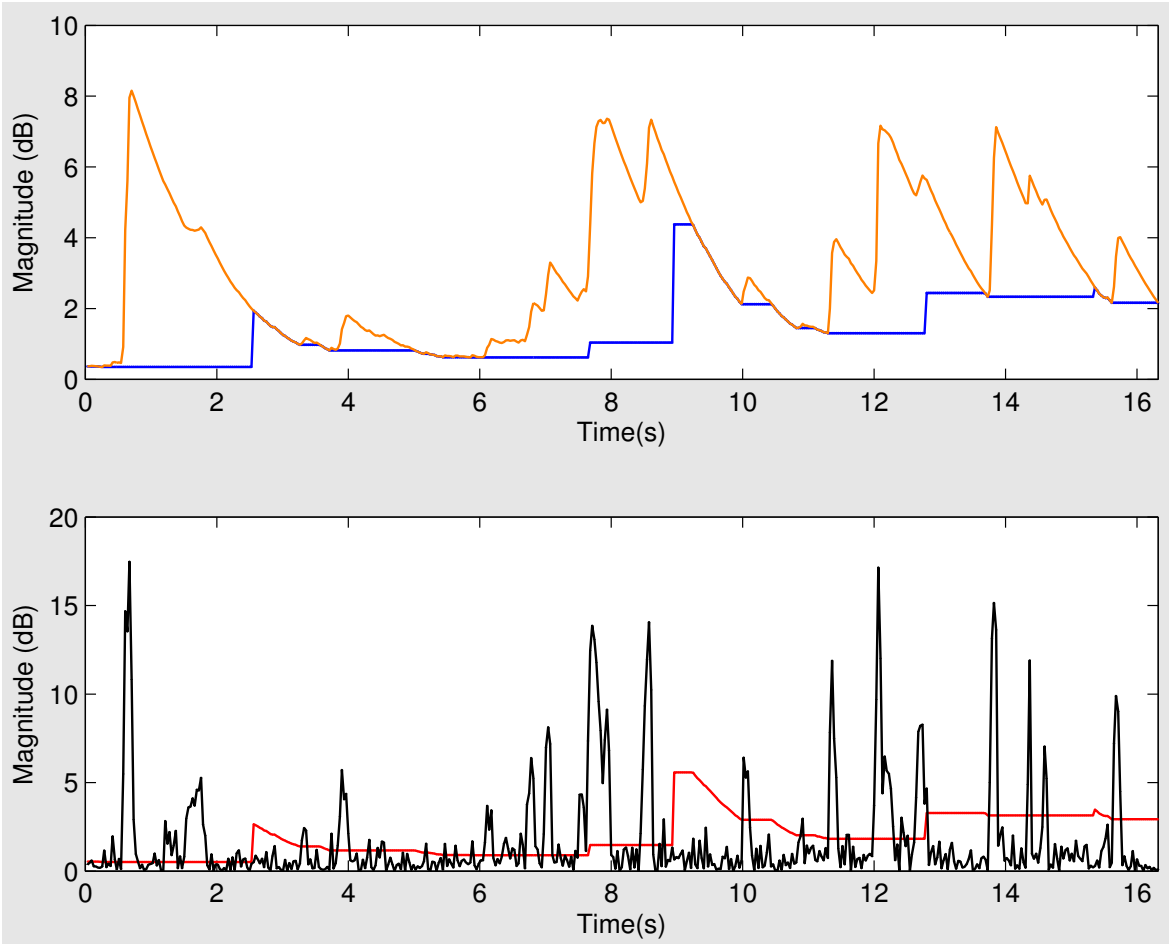



Figure 2.3 – Example of the minimum statistic tracking for the noise power spectrum estimation [71]. Above sub-figure shows the smoothed periodogram $P[m,k]$ (orange line) and its minimum $P_{\min}[m,k]$ (blue line). Below sub-figure displays the periodogram of the car noisy signal at 5 dB SNR (black line) and the noise power estimation (red line).

The gain function of the power spectral subtraction method is given by [9]:

$$G(\gamma[m,k]) = \begin{cases} \sqrt{(\gamma[m,k]-\alpha)/\gamma[m,k]} & \gamma[m,k] > \alpha+\beta, \\ \sqrt{\beta/\gamma[m,k]} & \text{Otherwise} \end{cases} \tag{2.15}$$

where $\alpha \geq 1$ and $\beta$ $(0 < \beta << 1)$ are the over subtraction factor and the spectral floor parameter, respectively. The *a posteriori* signal to noise ratio (SNR) $\gamma[m,k]$ is defined as follows:

$$\gamma[m,k] = \frac{\|Y[m,k]\|^2}{\widehat{\sigma}_X^2[m,k]}. \tag{2.16}$$

The gain function of the power spectral subtraction method is a function of the *a posteriori* SNR only with two parameters $\alpha$ and $\beta$. In general, for other methods, this gain can be dependent on other variables, which can be estimated from the noisy signal. Going back to the power spectral subtraction, parameter $\alpha$ controls the trade-off between the speech distortion and the residual noise whereas $\beta$ is determined by trading-off between *musical noise* and the remaining residual noise. Note that *musical noise* is the noise generated by the isolated point or peak in the

transformed domain or in the spectrum. In [9], $\beta$ is in the range of $[0.005, 0.1]$ and $\alpha$ is obtained from the estimated SNR $\widehat{\gamma}$ in each frame as:

$$\alpha = \alpha_0 - (3\widehat{\gamma})/20, \tag{2.17}$$

where $\alpha_0$ is the over subtraction factor at 0 dB. In addition, for $\widehat{\gamma} \geq 20$, $\alpha = 1$. Figure 2.4 shows the gain function of the power spectral subtraction algorithm as a function of $\gamma$ for fixed floor parameter $\beta = 0.1$ and with various values $\alpha$. At low SNR, these gain functions are the same since they depend only on the floor parameter $\beta$. At high SNR, these gain functions tend to 1. Clearly, the choice of the two parameters $\alpha$ and $\beta$ dictates the performance of this algorithm.

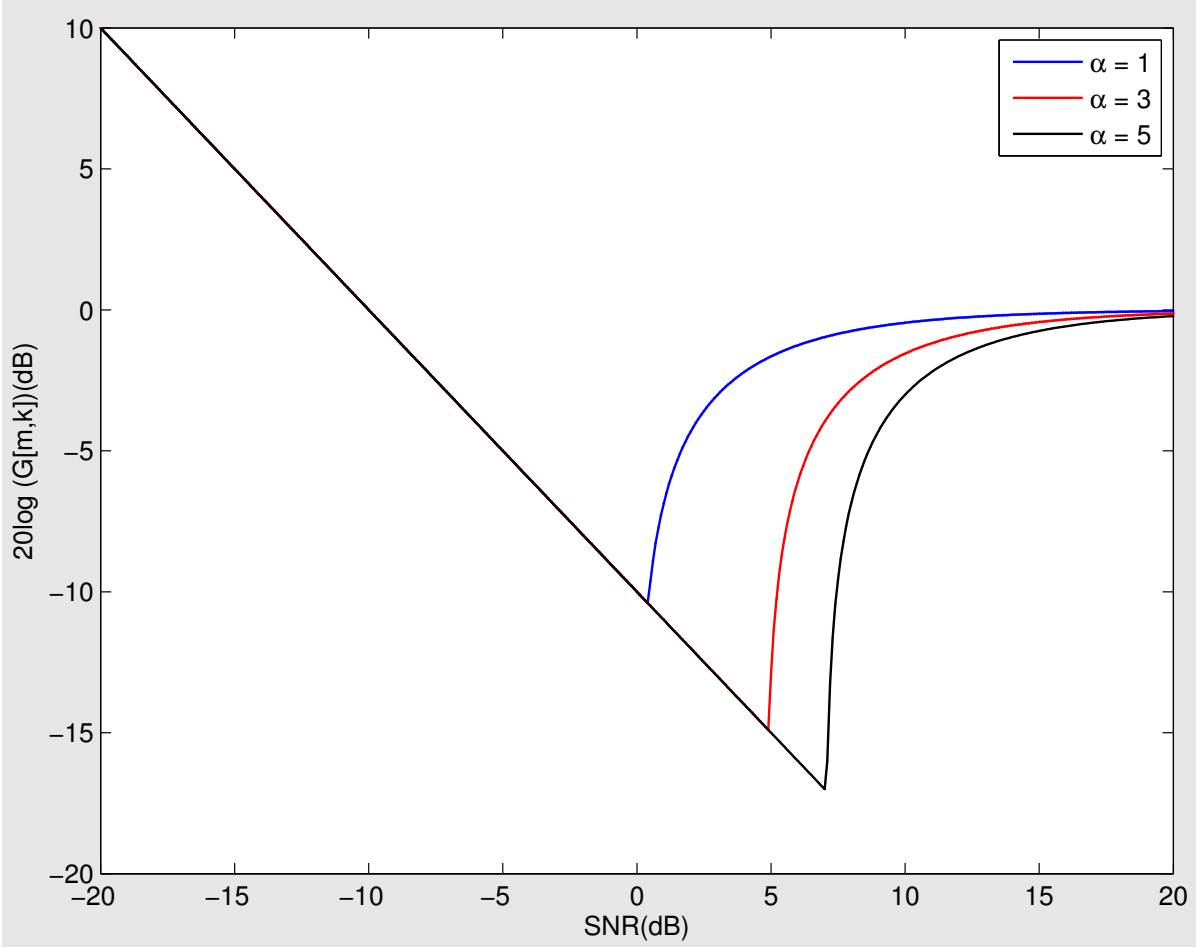



Figure 2.4 – The gain function of the power spectral subtraction for over subtraction factors $\alpha = \{1, 3, 5\}$.

In summary, the noise reduction block estimates the enhanced coefficient $\widehat{S}[m, k]$ in the transformed domain by applying a gain function $G[m, k]$ to the noisy coefficient $Y[m, k]$. This gain function is usually calculated from the noisy amplitude $A_Y[m, k]$ at the output of the decomposition block and from the estimated noise power spectrum $\widehat{\sigma}_X^2[m, k]$ at the downstream of the noise estimation block.

### 2.2.4 Reconstruction block

This block dedicates to transform the estimated clean speech back into the time-domain. Note that it is possible to recover the signal in time domain exactly from its short time transform

coefficients. Several methods have been proposed in the literature [72–77]. In this section, we only present the implementation of the overlap-add method, which is frequently used in speech enhancement. The mathematical framework is detailed in [1, Chap. 2] and [73]. The inverse short time transform is applied to each frame of enhanced coefficients $\{\widehat{S}[m,0], \widehat{S}[m,1], \ldots, \widehat{S}[m,K-1]\}$. The time enhanced signal $\widehat{\mathbf{s}}_m[n]$ in the given frame $m$ is written as:

$$\widehat{\mathbf{s}}_m[n] = \sum_{k=0}^{K-1} \beta_n[k]\widehat{S}[m,k], \tag{2.18}$$

where $\beta_n[k]$ depends on the used transform in the decomposition block. For example, when STFT is used, the value $\beta_n^{\mathrm{DFT}}[k]$ is:

$$\beta_n^{\mathrm{DFT}}[k] = \exp\left(j\frac{2\pi}{K}nk\right) \tag{2.19}$$

while for STCT the value $\beta_n^{\mathrm{DFT}}[k]$ is:

$$\beta_n^{\mathrm{DCT}}[k] = \sqrt{\frac{1+\mathbb{1}_{(0,\infty)}(k)}{K}}\cos\left(\frac{\pi}{2K}(2n+1)k\right). \tag{2.20}$$

Once we have obtained the time enhanced signal $\widehat{\mathbf{s}}_m[n]$ in the relevant overlapped frames, the enhanced signal is calculated as:

$$\widehat{\mathbf{s}}[n] = \sum_{m,k} \widehat{\mathbf{s}}_m[k], \tag{2.21}$$

where $1 \leq m$ and $0 \leq k \leq K-1$ are chosen to satisfy $n = (m-1)K^* + k$. Figure 2.5 shows the reconstruction of the enhanced signal $\widehat{\mathbf{s}}[n]$ from the time enhanced $\widehat{\mathbf{s}}_m[n]$ by using 75% and 50% overlap-add methods. For the 75% overlap-add method, to recover the enhanced signal, we need to know the three consecutive previous frames whereas, for the 50% overlap-add method, only the previous frame $m-1$ is required.

Figure 2.6 displays an example of reconstruction of a single sinusoidal signal $\mathbf{s}[n]$ using 50% overlap-add method, where the sampling rate of the sinusoidal signal is equal to 8 kHz. The sinusoidal signal was Hamming-windowed into 32-ms frames with 50% overlap, and then transformed by DFT. The obtained coefficients $S[m,k]$ is then synthesized by 50% overlap-add to obtain $\widehat{\mathbf{s}}[n]$. In Figure 2.6, the original and reconstructed signals are shown respectively by the red and blue lines. A slight difference is appeared in maximal amplitude between the two signals.

## 2.3 Performance evaluation of speech enhancement algorithms

Practical audio enhancement systems include the four main blocks of Figure 2.1. In research, for evaluating the performance, an additional evaluation block is added as illustrated in Figure 2.7. In this section, we will present some criteria that are frequently selected to evaluate the performance of speech enhancement methods. These criteria will also be used in this thesis. All the criteria can be divided into objective and subjective tests. The subjective listening tests are the most reliable criteria but they require more time for evaluation. Certain objective tests were shown to be highly correlated with subjective tests. Therefore, these objective tests can be often selected to assess the quality and intelligibility of speech. Let us first introduce the objective criteria used in this thesis.

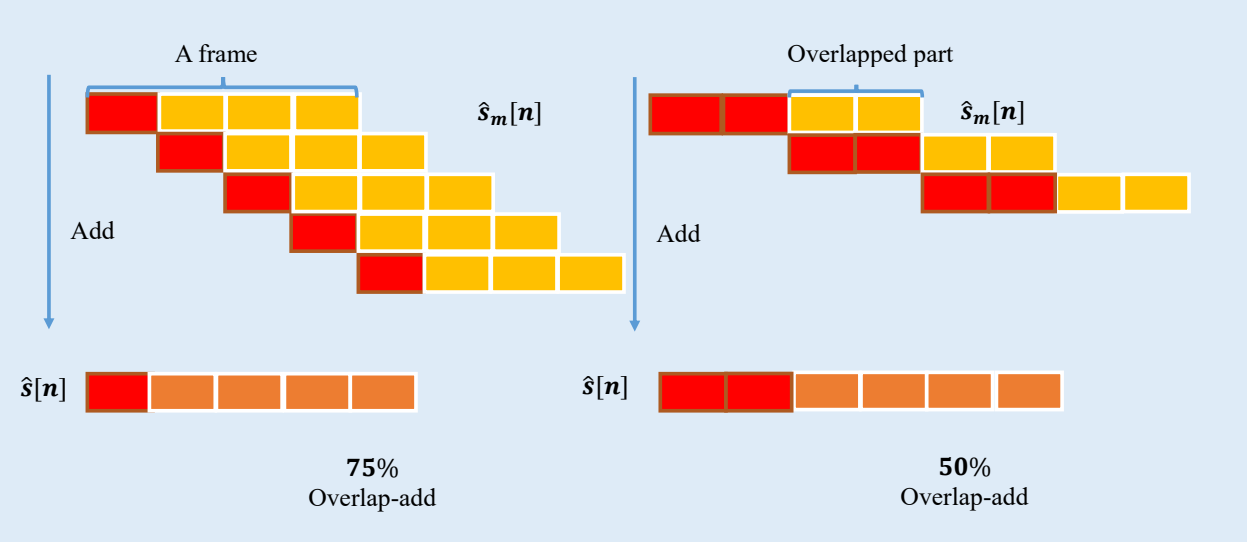



Figure 2.5 – Schema to synthesize the enhanced signal $\hat{\mathbf{s}}[n]$ by 75% (left side) or 50% (right side) overlap-add method. Note that the percent of the overlapped part is the same for decomposition and reconstruction blocks.

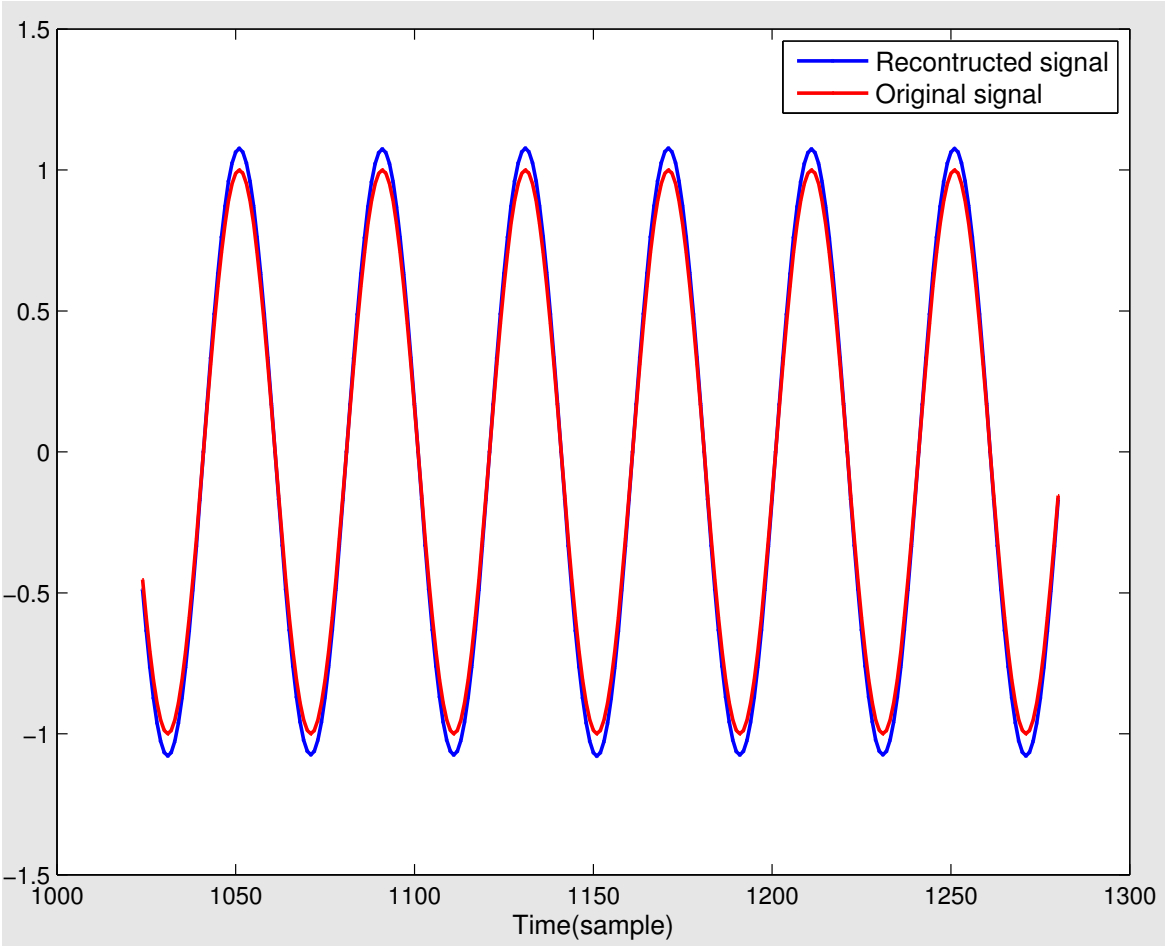



Figure 2.6 – Example of reconstructing a sinusoidal signal using 50% overlap-add method.

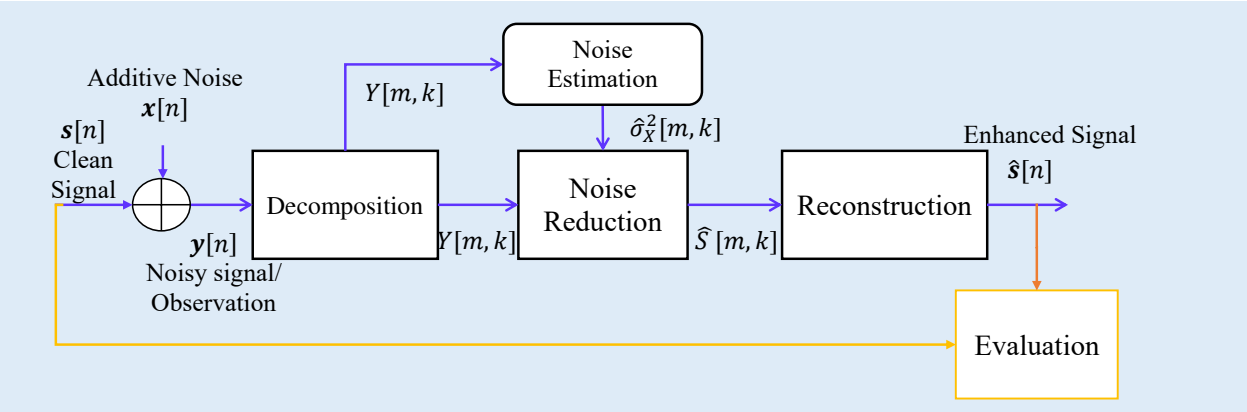



Figure 2.7 – Full audio enhancement system under consideration in the present work.

### 2.3.1 Objective tests

#### 2.3.1.1 Segmental signal to noise ratio measure

The segmental signal to noise ratio (SSNR) measure is one of the simplest and well-known criteria. This measure is the geometric mean of the SNR over all frames of the speech signal [1]:

$$\text{SSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Km}^{Km+K-1}\mathbf{s}^2[n]}{\sum_{n=Km}^{Km+K-1}(\mathbf{s}[n]-\widehat{\mathbf{s}}[n])^2}, \tag{2.22}$$

where $K$ is the frame length in samples and $M$ is the total frames number. The evaluated signal $\widehat{\mathbf{s}}[n]$ and the clean signal $\mathbf{s}[n]$ have the same length and must be synchronized in time. Note that the frames, with SNRs above 35 dB, do not provide a large perceptual difference so that these SNRs are clipped to 35 dB. Moreover, in the noise-only frames, the speech energies are small in that the frame-based SNRs are very low. These frames do not also contribute to the perception of the signal. Therefore, the frame-based SNRs values were trimmed so as to remain within the range [−10,35 dB] instead of using a silence/speech detector [78].

#### 2.3.1.2 Spectral distance measure

The SSNR criterion was based on the frame-based SNRs across all frames of the speech signal in the time domain. We introduce now the second measure called Itakura-Saito (IS) distance based on the dissemblance between all-pole model of the clean signal and the evaluated signal [79]. This distance is defined as:

$$d_{\text{IS}}(\mathbf{a}_\mathbf{s},\mathbf{a}_{\widehat{\mathbf{s}}}) = \frac{\sigma_\mathbf{s}^2}{\sigma_{\widehat{\mathbf{s}}}^2}\left(\frac{\mathbf{a}_{\widehat{\mathbf{s}}}^T\mathbf{R}_\mathbf{s}\mathbf{a}_{\widehat{\mathbf{s}}}}{\mathbf{a}_\mathbf{s}^T\mathbf{R}_\mathbf{s}\mathbf{a}_\mathbf{s}}\right)+\log\left(\frac{\sigma_{\widehat{\mathbf{s}}}^2}{\sigma_\mathbf{s}^2}\right)-1, \tag{2.23}$$

where $\mathbf{R}_\mathbf{s}$ is the auto correlation matrix of the clean signal, $\sigma_\mathbf{s}^2$ and $\sigma_{\widehat{\mathbf{s}}}^2$ denote the LPC gains of the clean speech and evaluated or enhanced speech, respectively. $\mathbf{a}_\mathbf{s}$ and $\mathbf{a}_{\widehat{\mathbf{s}}}$ are the LPC vectors of the clean signal frame and the evaluated signal frame. $\mathbf{R}^T$ is the transpose operator of the matrix $\mathbf{R}$. The LPC vectors of the clean and evaluated signal frame are estimated by assuming that the speech signal over an interval can be modeled as:

$$\mathbf{s}[n] = \sum_{i=1}^{p} a_\mathbf{s}(i)\mathbf{s}[n-i] + \sigma_\mathbf{s}u[n]^T \tag{2.24}$$

where $p$ is the order of the all-pole model, $a_{\mathbf{s}}(i)$ denotes the coefficient of the all-pole filter and $u[n]$ is the white Gaussian noise with unit variance. Thus, the LPC vector of the clean signal $\mathbf{a_s}$ are formed as $\mathbf{a_s} = [1, -a_{\mathbf{s}}(1), -a_{\mathbf{s}}(2), \ldots, -a_{\mathbf{s}}(p)]$ where $a_{\mathbf{s}}(i)$ can be estimated by linear prediction method. A similar way is used for determining the LPC vector $\mathbf{a}_{\widehat{\mathbf{s}}}$ of the evaluated signal.

#### 2.3.1.3 SNR improvement measure

The SNR improvement (SNRI) measure is an objective criterion standardized in the ITU-T G.160 recommendation for evaluating noise reduction algorithms in transmission systems [80]. This measure requires various types of noises and at different SNR levels. The clean speech utterance $\mathbf{s}_i$ is degraded by noise $\mathbf{x}_j$, yielding the noisy speech $\mathbf{y}_{ij}$:

$$\mathbf{y}_{ij} = \mathbf{s}_i + \beta_{ij}\mathbf{x}_j, \tag{2.25}$$

where $\beta_{ij}$ depends on the SNR levels. The output of the speech enhancement system is the corresponding enhanced signal $\widehat{\mathbf{s}}_{ij}$. Like SSNR above, for evaluating the frame-based SNR, the noisy and enhanced signals are segmented into 10-ms frame. We denote the noisy and enhanced framed signal as $\mathbf{y}_{ij}[m,n]$ and $\widehat{\mathbf{s}}_{ij}[m,n]$, where $m$ is the frame indice and $n$ is the sample indice within a given frame $m$. These frames are then divided into the four frame-energy classes including: three speech classes (high, medium and low power of the speech presence) and one noise-only class (the speech absence). For the three speech classes, the output and input speech SNRs are determined in the same way. For instance, the output SNRs and the input SNRs of high power speech class are calculated as follows:

$$\text{SNRout\_}h_{ij} = 10\log\left\{\max\left[\epsilon, \frac{10^{\frac{1}{M_{\text{sph}}}\sum_{m=1}^{M_{\text{sph}}}\log\left\{\max\left[\xi, \sum_n \widehat{\mathbf{s}}_{ij}^2[m,n]\right]\right\}}}{10^{\frac{1}{M_{\text{nse}}}\sum_{m=1}^{M_{\text{nse}}}\log\left\{\max\left[\xi, \sum_n \widehat{\mathbf{s}}_{ij}^2[m,n]\right]\right\}}} - 1\right]\right\} \tag{2.26}$$

$$\text{SNRin\_}h_{ij} = 10\log\left\{\max\left[\epsilon, \frac{10^{\frac{1}{M_{\text{sph}}}\sum_{m=1}^{M_{\text{sph}}}\log\left\{\max\left[\xi, \sum_n \mathbf{y}_{ij}^2[m,n]\right]\right\}}}{10^{\frac{1}{M_{\text{nse}}}\sum_{m=1}^{M_{\text{nse}}}\log\left\{\max\left[\xi, \sum_n \mathbf{y}_{ij}^2[m,n]\right]\right\}}} - 1\right]\right\}, \tag{2.27}$$

where $M_{\text{sph}}$ and $M_{\text{nse}}$ are the total number of high power speech classes and speech absence classes in the considered signal, respectively. $\epsilon$ and $\xi$ are constants that are set equal to $-12$ dB and $-71$ dB. The SNRI$_h_{ij}$ of the high power speech frames is defined as:

$$\text{SNRI\_}h_{ij} = \text{SNRout\_}h_{ij} - \text{SNRin\_}h_{ij}. \tag{2.28}$$

Therefore, the SNRI of the signal $\widehat{\mathbf{s}}_{ij}$ is given by:

$$\text{SNRI}_{ij} = \frac{1}{M_{\text{sph}} + M_{\text{spm}} + M_{\text{spl}}}\left(M_{\text{sph}}\text{SNRI\_}h_{ij} + M_{\text{spm}}\text{SNRI\_}m_{ij} + M_{\text{spl}}\text{SNRI\_}l_{ij}\right), \tag{2.29}$$

where $M_{\text{spm}}$ and $M_{\text{spm}}$ are the total numbers of the medium power and the low power speech frames and SNRI$_m_{ij}$ and SNRI$_l_{ij}$ denote the SNRI of the medium power and the low power speech frames. Finally, the SNRI is obtained by the mean measure over all types and levels of noise:

$$\text{SNRI} = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{I}\sum_{i=1}^{I}\text{SNRI}_{ij}\right), \tag{2.30}$$

where $I$ is the total number of clean speech utterances and $J$ is the number of different background noises.

#### 2.3.1.4 Perceptual motivated Measure

The perceptual evaluation of speech quality (PESQ) criterion is widely used for evaluating the performance of the noise reduction algorithms in telephone handset application, which was recommended by IUT-T P.862 [81]. A semantic description of the PESQ measure is presented in [1, Chap 11. Sec 11.1.3.3].

In brief, the structure of the PESQ measure contains five main blocks, namely, pre-processing, time alignment, auditory transforming, disturbance processing and time and frequency averaging blocks. The clean and the enhanced speech are firstly passed through the pre-processing block to have the same listening level and to adapt to a standard telephone handset. Then, the time alignment block determines the time delay value between the clean and the enhanced signals. This block provides also a delay confidence. Next, the auditory transforming block codifies the clean and the enhanced signal into a perceptual representation of the perceived loudness, where we can point out the loudness spectra of the two signals. Latter, the disturbance processing block measures the dissimilarity between the enhanced and the clean speech representations. Finally, the time and frequency averaging block evaluates the PESQ measure from the dissimilarity determined in the previous block.

#### 2.3.1.5 Composite measures

For capturing different dissimilarities between original and enhanced signals, several composite measures formed by combining multiple objective measures have been proposed. These combinations, either linear or non-linear, make it possible to achieve high correlation with subjective listening tests [78, 82–84]. In this thesis, we use the composite measures that are based on multivariate adaptive regression splines (MARS) and have been found to yield a good correlation with listening tests [83]. The MARSovrl, predicting overall speech quality (OVRL) is the combination of the IS and PESQ criteria and is defined as:

$$\text{MARSovrl} = 1.757 + 1.740\mathbf{BF1} + 0.047\mathbf{BF2} - 0.049\mathbf{BF3} - 2.593\mathbf{BF4} + 11.549\mathbf{BF5}, \quad (2.31)$$

where

$$\mathbf{BF1} = \max(0, \text{PESQ} - 1.696) \qquad \mathbf{BF2} = \max(0, \text{IS} - 11.708) \quad (2.32)$$

$$\mathbf{BF3} = \max(0, \text{IS} - 3.559) \qquad \mathbf{BF4} = \max(0, \text{PESQ} - 2.431) \quad (2.33)$$

$$\mathbf{BF5} = \max(0, \text{PESQ} - 2.564). \quad (2.34)$$

This measure is found to have a high correlation with the mean opinion score (MOS) of subjective listening test [83]. In this paper, two other metrics are introduced, MARSsig and MARSbak, that are designed to provide a high correlation with the two usual corresponding subjective measures that are the signal distortion (SIG) and the background intrusiveness (BAK).

#### 2.3.1.6 Short-time objective intelligibility measure

All criteria mentioned above enable us to estimate speech quality. In this section, we briefly present a widely used criterion for predicting speech intelligibility. Many intelligibility measures have been proposed in the literature. Most of them are based on the articulation index (AI)

with the speech intelligibility index (SII) standardized as S3. 5-1997 [85] or on the speech transmission index (STI) [86]. Recently, a short-time objective intelligibility (STOI) measure has been presented in [87]. This criterion has a high correlation with subjective speech intelligibility test. Therefore, we use it for evaluating all the algorithms considered in this thesis. Figure 2.8 illustrates the structure for evaluating STOI in block.

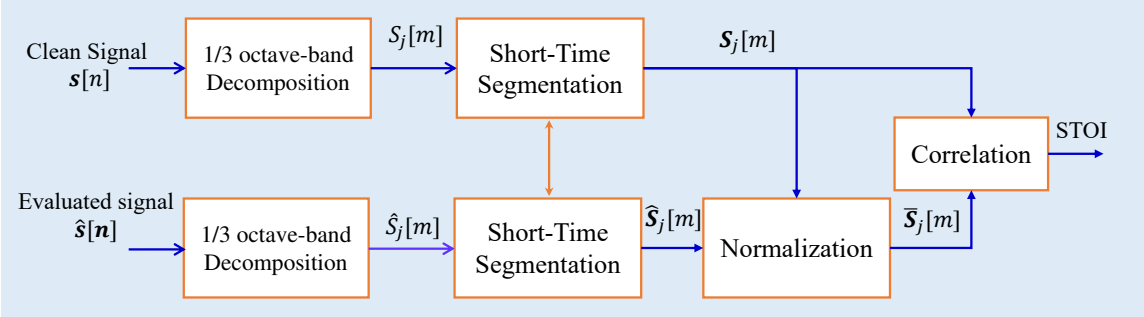



Figure 2.8 – Principle of STOI evaluation [87].

In the one-third octave band decomposition block, both the clean and the enhanced signal are first Hanning-windowed into 256-sample frames with 50% overlap, and then transformed by DFT to obtain $S[m,k]$ and $\widehat{S}[m,k]$, respectively. Then, the norm of the clean speech $S_j[m]$ at the $j^{\text{th}}$ one-third octave band is computed as:

$$S_j[m] = \sqrt{\sum_k \|S[m,k]\|^2}, \tag{2.35}$$

where $k$ is the frequency index pertaining to the $j^{\text{th}}$ one-third octave band. In the same way, we also obtain the norm of the evaluated speech $\widehat{S}_j[m]$ at the one-third octave band $j^{\text{th}}$.

In the short-time segmentation block, the short-time temporal envelope vector $\mathbf{S}_j[m]$ of the clean speech at the $j^{\text{th}}$ one-third octave band is formed from the $N$ consecutive norms $S_j[m]$ as:

$$\mathbf{S}_{jm} = [\mathbf{S}_{jm}(1), \mathbf{S}_{jm}(2), \ldots, \mathbf{S}_{jm}(N)]^T = [S_j[m-N+1], S_j[m-N+2], \ldots, S_j[m]]^T, \tag{2.36}$$

where $N$ is recommended to be equal to 30 frames. Similarly, the same $N$ is used for grouping $\widehat{S}_j[m]$ in $\widehat{\mathbf{S}}_{jm}$. The short-time temporal vector $\widehat{\mathbf{S}}_{jm}$ of the evaluated signal are then normalized and trimmed in the normalization block via:

$$\overline{\mathbf{S}}_{jm}(n) = \min\left(\frac{\|\mathbf{S}_{jm}\|}{\|\widehat{\mathbf{S}}_{jm}\|}\widehat{\mathbf{S}}_{jm}(n), (1+10^{-\beta/20})\mathbf{S}_{jm}(n)\right), \tag{2.37}$$

where $\beta = -15$ dB is the signal to distortion ratio lower bound, $\overline{\mathbf{S}}_{jm}$ is the normalized and trimmed vector of $\widehat{\mathbf{S}}_{jm}$ and $n \in \{1, 2, \ldots, N\}$.

The intelligibility in the $j^{\text{th}}$ one-third octave band is the sample correlation between $\mathbf{S}_{jm}$ and $\overline{\mathbf{S}}_{jm}$:

$$\text{STOI}_{jm} = \frac{\left(\mathbf{S}_{jm} - \mu_{\mathbf{S}_{jm}}\right)^T \left(\widehat{\mathbf{S}}_{jm} - \mu_{\widehat{\mathbf{S}}_{jm}}\right)}{\|\mathbf{S}_{jm} - \mu_{\mathbf{S}_{jm}}\| \quad \|\widehat{\mathbf{S}}_{jm} - \mu_{\widehat{\mathbf{S}}_{jm}}\|}, \tag{2.38}$$

where $\mu_{\mathbf{X}}$ denotes the empirical mean of vector $\mathbf{X}$. The final STOI measure is obtained by averaging over all bands and time-segmentation:

$$\text{STOI} = \frac{1}{JM}\sum_{j,m}\text{STOI}_{jm}, \tag{2.39}$$

where $M$ and $J$ refer to the total numbers of the frames and bands in the considered signal, respectively. In addition, a logistic function is applied to map the STOI measure to a meaningful intelligibility score. This function is defined by:

$$f(\text{STOI}) = \frac{100}{1 + \exp(a \times \text{STOI} + b)}, \tag{2.40}$$

where $a = -17.4906$ and $b = 9.6921$ for fitting with the IEEE sentences in the NOIZEUS database [1].

### 2.3.2 Mean opinion scores subjective listening test

All the criteria mentioned above are objective measures which enable us to save time-consuming tests. For further reliable evaluation, we can not avoid the subjective listening test. In this section, we describe the widely used directed method called mean opinion scores (MOS). This method was selected by the IEEE Subcommittee on Subjective Methods [88]. Raters evaluate the speech quality of the test signal using five numerical scores shown by Table 2.1. Overall speech quality is then determined by averaging all the scores obtained by all raters so that this subjective listening test, hence the name of the test. For further detail, the interested readers are invited to consult paper [88].

Table 2.1 – MOS rating score

| **Score** | **Speech quality** | **Level distortion** |
|---|---|---|
| 5 | *Excellent* | Imperceptible |
| 4 | *Good* | Just perceptible, but not annoying |
| 3 | *Fair* | Perceptible and slightly annoying |
| 2 | *Poor* | Annoying, but not objectionable |
| 1 | *Bad* | Very annoying and objectionable |

In brief, the MOS listening test consists of two steps: training and evaluation. In the training step, the raters listen to a group of reference sentences that are clear representative of each of the five point rates. In the evaluation step, listeners are invited to rate the test signal according to the MOS score table (see Table 2.1). Note that some constraints must be respected in the MOS listening test. First, there must be at least 10 listeners. Second, the test duration of each rater should not exceed 20 minutes because of listening fatigue. Third, headphones should be used for listening to avoid external distortions due to the use of loudspeakers.

## 2.4 Conclusion

In this chapter, we have described the general structure of speech enhancement system where only a single microphone is available to capture or record noisy speech. Objective and subjective criteria for evaluating the performance of speech enhancement algorithms have also been introduced and discussed. From a high-level perspective, the generic structure of speech enhancement systems was shown to include four main signal processing blocks. An improvement or modification in any of these blocks may translate into better performance for the whole systems. This is the purpose of the following chapters. Chapter 3 will revisit the noise estimation block. Chapter 4 will investigate a new approach for noise reduction block whereas Chapter 5 will take a broad perspective and jointly optimize the signal decomposition and noise reduction blocks.

# Part II

# Noise: Understanding the Enemy

In the first chapter, we have motivated the need to carry out an unsupervised approach for single microphone speech enhancement. A general overview of the speech enhancement systems was then given in the second chapter. In these systems, noise power spectrum estimation is a key issue in designing robust noise reduction methods for speech enhancement. The question is how to estimate the noise power spectrum from only the noisy signal captured by only one microphone. In the single speech enhancement, this is the main challenge, especially when background noise is non-stationary. By noting that the signal of interest is weak-sparse in a transformed domain, a novel non-parametric noise power spectrum estimation algorithm is introduced in this chapter. This algorithm can deal efficiently with non-stationary noise. The results described in this chapter have been published in [89].

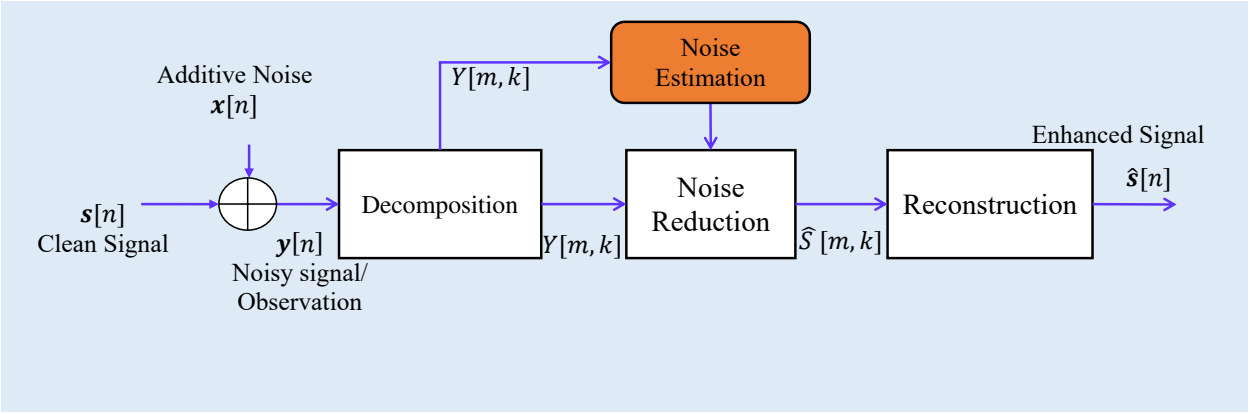
Chapter 3

# Noise estimation block

*Problems are not stop signs, they are guidelines.*

Robert H. Schuller





## 3.1 Introduction

Most noise power spectrum estimation algorithms found in the literature can be classified into four main categories [1], namely histogram-based methods, minimal-tracking algorithms, time-recursive averaging algorithms, and other techniques derived from Maximum-Likelihood (ML) or Bayesian estimation principles, *e.g.* minimum mean square error (MMSE) methods.

In the first category of algorithms, the noise power spectrum is estimated from the maximum of the histogram in the time-frequency domain of the observed signal power spectrum, the latter being determined by using a first-order smoothing recursion [90]. An improvement of this method involves updating the noise power spectrum solely on the frames detected as noise-only by a chi-square test [91]. However, most of the histogram-based algorithms have the drawback of being relatively complex in terms of computational cost and memory resources [92].

In the second family of methods, the noise power spectrum is tracked via minimum statistics, according to the reasonable hypothesis that the noise power spectrum level is below that of noisy speech [71, 93]. First, the smoothed noisy speech power spectrum is evaluated by a first-order recursive operation. Then, the noise variance is computed as the statistical minimum of the smoothed power spectrum with a factor of correction. The main difference between the two methods in [71] and [93] lies in the computation of the smoothing parameter used in the first order recursion. In [71], the smoothing parameter is chosen empirically, whereas this parameter is derived by minimizing the mean square error between the noise and the smoothed noisy speech power spectrum in [93]. Minimum-statistics methods require observing the noisy signals on a sufficiently long time interval so as to track speech power instead of noise power. On the other hand, a long time interval is detrimental to the quality of the estimate in case of non stationary noise. A trade-off is thus necessary, leading to a typical time-delay of 1 to 3 seconds in practice. This causes underestimation which decreases in turn the performance of noise reduction algorithms.

Famous methods in the third category include the Minima-Controlled Recursive-Averaging (MCRA) algorithm [94] and its many modifications such as the Improved-MCRA (IMCRA) [92] or the MCRA2 [95] methods. In this class of algorithms, the noise power spectrum in a given frequency bin is estimated by first-order recursive operations where smoothing parameters depend on the conditional speech presence probability inside the bin. The main difference between MCRA, MCRA2 and IMCRA lies in the way the speech-presence probability is estimated. MCRA and MCRA2 directly estimate the speech-presence probability frame-by-frame via a smoothing operation whereby, for a given frame, the probability of speech presence is increased when this frame is detected as noisy speech and decreased otherwise. A frame is detected as noisy speech if the ratio of the smoothed noisy speech power spectrum to its local minimum is above a certain threshold, the local minimum being computed by using the minimum-statistics technique proposed in [93]. Fixed and frequency-dependent thresholds are used in MCRA and MCRA2, respectively. On the other hand, IMCRA derives the speech-presence probability in each bin by a two-step estimation of the speech-absence probability. The first iteration aims at detecting the absence of speech in a given frame, while the second iteration actually estimates the speech-absence probability from the power spectral components in the speech-absence frame. The main disadvantage of these methods is the estimation delay in case of sudden rising noise, this delay being mainly due to the use of the minimum-statistics methods of [93].

Techniques derived from ML or Bayesian estimation principles overcome the problem of sudden rising noise by estimating the noise power spectrum from the noise periodogram via a statistical criterion. In [96], [97] called MMSE1 and MMSE2, respectively, the noise instantaneous power is evaluated by MMSE and then incorporated in a recursive noise power spectrum estima-

tion technique. [96] proposes a simple bias compensation of the noise instantaneous power before estimating the noise power spectrum via the same recursive smoothing and under the same hypotheses as in [97]. However, the noise instantaneous power estimate in [96] remains biased. In contrast, an unbiased estimator for the noise instantaneous spectrum is obtained in [97] by soft-weighting the noisy speech instantaneous power and the previous noise power spectrum estimate by the conditional probabilities of speech-absence and speech-presence, respectively. The noise power spectrum estimation can also be carried out by recursive ML-Expectation-Maximization (ML-EM) [98], similar to MCRA and IMCRA. This approach allows for rapid noise power spectrum estimation and tracking by avoiding the use of minimum-statistics methods.

In this chapter, we propose a new approach for noise power spectrum estimation, without requiring any model or any prior knowledge for the probability distributions of the speech signals. Fundamentally, we do not even take into consideration the fact that the signal of interest here is speech. The approach is henceforth called extended-DATE (E-DATE) since it basically extends the $d$-dimensional amplitude trimmed estimator (DATE), initially proposed in [65] for white Gaussian noise (WGN), to colored stationary and non-stationary noise. The main principle at the heart of the E-DATE algorithm is the weak-sparseness property of the STFT of noisy signals, according to which the sequence of complex values returned by the STFT in a given time-frequency bin can be modeled as a complex random signal with unknown distribution and whose unknown probability of occurrence in noise does not exceed one half. Noise in each bin is assumed to follow a zero-mean complex gaussian distribution [1, p. 210], so that estimating the noise power spectrum amounts to estimating the noise variance in each bin, the latter being provided by the DATE. The DATE trims the amplitudes in each given bin, after having sorted them by increasing norm. Noise power spectrum estimation by E-DATE is thus similar to and actually extends the quantile-based approach of [99], which relies on assumptions that the weak-sparseness model embraces. More generally, the reader will notice similarities between the proposed method and the state-of-the-art techniques mentioned above. A main difference between the E-DATE approach and standard ones is actually the mathematical justification of the former via the weak-sparseness model, which formalizes more or less standard heuristics in speech processing and yields a reduced number of parameters for more robustness. Although the E-DATE does not rely on minimum-statistics principles or methods, it does however require a time buffer having the same length — typically 80 frames for a sampling rate frequency of 8 kHz — as other popular algorithms.

The chapter is organized as follows. In Section 3.2, the main features of the DATE are reviewed. Section 3.3 develops the weak-sparseness model for noisy speech. The E-DATE is then introduced in Section 3.4, following a step-by-step methodology where we successively deal with WGN, stationary noise and non-stationary noise. Two practical implementations of the E-DATE algorithm are then described. The performance of the E-DATE algorithm is evaluated in Section 3.5 and compared to state-of-the-art methods in terms of number of parameters and estimation errors. Speech enhancement experimental comparisons using objective as well as pseudo-subjective criteria are also conducted by combining the noise power spectrum estimation methods with a noise reduction system. Conclusions are finally given in Section 3.6.

## 3.2 DATE algorithm

For the sake of self-completeness, this section presents the DATE in its full generality. Given $d$-dimensional observations of random signals that are themselves randomly absent or present in independent and additive WGN, the purpose of the DATE is to estimate the noise standard deviation. Such an estimation may serve to detect the signals or to estimate them as in speech

denoising. As in [100], the DATE addresses the frequently-encountered case where 1) most observations follow the same zero-mean normal distribution with unknown variance, 2) signals of interest have unknown distributions and occurrences in noise. Standard robust scale estimators such as the very popular median absolute deviation (MAD) estimator and the trimmed estimator (T-estimator) have performance that degrades significantly when the proportion of signal increases. In contrast, the DATE can still estimate the noise standard deviation when the signals of interest occur with a probability too large for usual scale estimators to perform well. As indicated by its name, the DATE basically trims the norms of the $d$-dimensional observations. However, in contrast to the conventional T-estimator, which applies to one-dimensional data and fixes the number of outliers to remove, the DATE applies to any dimension and chooses adaptively the number of outliers to discard. It performs the trimming by assuming that the signal norms are above some known lower-bound and that the signal probabilities of occurrence are less than one half. These assumptions bound our lack of prior knowledge about the signals and make it possible to separate signals from noise. Moreover, these assumptions are suitable for signal processing applications where noisy signals are considered as outliers with respect to the noise distribution. They are particularly suitable for observations obtained after sparse transforms capable of representing signals by coefficients that are small for the most part except a few ones whose norms are relatively big. In particular, the sequel will exhaustively use the fact that the Fourier transform of speech signals is sparse in a weak sense detailed hereafter.

The DATE basically relies on [65, Theorem 1], which is asymptotic and can be viewed as a method of moments. A detailed presentation of the theoretical background of the DATE is beyond the scope of this chapter and the reader is referred to [65] for details. However, the following brief heuristic presentation may be convenient for the reader. This heuristic exposure departs from that proposed in [65, Theorem 1], so as to shed different light on the theory behind the DATE.

*Notation:* In what follows, $\|\cdot\|$ is the usual euclidean norm in the space of all $d$-dimensional real vectors, $\mathbf{I}_d$ stands for the $d \times d$ identity matrix, $\mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ designates the $d$-dimensional Gaussian distribution with zero-mean and covariance matrix $\sigma^2 \mathbf{I}_d$ and $\mathbb{1}[X \in B]$ stands for the indicator function of the event $[U \in B]$, where $U$ is any random variable and $B$ is any borel set of the real line: $\mathbb{1}[U \in B] = 1$ if $U \in B$ and $\mathbb{1}[U \in B] = 0$, otherwise. In addition, $\Gamma$ is the standard Gamma function and ${}_0\mathrm{F}_1$ is the generalized hyper-geometric function [101, p. 275]. All the random vectors and variables are henceforth assumed to be defined on the same probability space $(\Omega, \mathbb{P}, \mathbb{E})$.

Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of $d$-dimensional random observations such that:

**(A0)** The observations $Y_1, Y_2, \ldots, Y_n, \ldots$ are mutually independent, $Y_n = \varepsilon_n \Lambda_n + X_n$ where $X_n \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_d)$ and $\varepsilon_n$ is Bernoulli distributed with values in $\{0, 1\}$ for each $n \in \mathbb{N}$.

In this model, each observation is either noise alone or the sum of some signal and noise. The probability distributions of the signals $\Lambda_n$ are supposed to be unknown. Our purpose is then to estimate $\sigma_0$.

If all the ratios $\|\Lambda_n\|/\sigma_0$ are known to be above some sufficiently large signal to noise ratio (SNR) $\rho$, it can be expected that some threshold height $\sigma_0 \xi(\rho)$ can suitably be chosen to decide with small error probability that $\Lambda_n$ is present (resp. absent) whenever $\|Y_n\|$ is above (resp. less) $\sigma_0 \xi(\rho)$. Therefore, most of the non-zero terms in the sum $\sum_{n=1}^{N} \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]$ should pertain to noise alone. If the number $\sum_{n=1}^{N} \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]$ of these non-zero terms is itself large enough, we should have an approximation of the form $\frac{\sum_{n=1}^{N} \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]}{\sum_{n=1}^{N} \mathbb{1}[\|Y_n\| \leqslant \sigma_0 \xi(\rho)]} \approx \lambda \sigma_0$.

Such an approximation can actually be proved asymptotically with the help of some additional assumptions. More precisely, suppose that:

**(A1)** $\Lambda_n$, $X_n$ and $\varepsilon_n$ are independent for every $n \in \mathbb{N}$;

**(A2)** the set of priors $\{\mathbb{P}[\varepsilon_n = 1] : n \in \mathbb{N}\}$ is upper-bounded by $1/2$ and the random variables $\varepsilon_n$, $n \in \mathbb{N}$, are independent;

**(A3)** $\sup_{n\in\mathbb{N}} \mathbb{E}[\|\Lambda_n\|^2] < \infty$.

These assumptions including **(A0)** deserve some comments. To begin with, the independence assumption in **(A0)** is mainly technical to prove the results stated in [65]. In fact, our experimental results below suggest that this assumption is not so constraining in speech processing, where we deal with non-overlapping but not necessarily independent time frames. Assumption **(A1)** simply means that the two hypotheses for the observation occur independently and that the noise and signal are independent. The model thus assumes prior probabilities of presence and absence through the random variables $\varepsilon_n$. However, the impact of these priors is reduced by assuming that the probabilities of presence and absence are actually unknown. The role of Assumption **(A2)** is then to bound this lack of prior knowledge about the occurrences of the two possible hypotheses that any $Y_n$ is supposed to satisfy. Assumption **(A3)** simply means that the signals $\Lambda_n$ have finite power.

Under assumptions **(A0)**-**(A3)** and with the help of [102, Theorem 1], [65, Theorem 1] then guarantees that $\sigma_0$ is the unique positive real number $\sigma$ such that:

$$\lim_{\rho\to\infty} \left\| \limsup_{N\to\infty} \left| \frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]} - \lambda\sigma \right| \right\|_\infty = 0 \tag{3.1}$$

where $\lambda = \sqrt{2}\,\Gamma\left(\frac{d+1}{2}\right)/\Gamma\left(\frac{d}{2}\right)$ and $\xi(\rho)$ is the unique positive solution in $x$ to the equality ${}_0\mathrm{F}_1(d/2; \rho^2 x^2/4) = e^{\rho^2/2}$. It is thus natural to estimate the noise standard deviation $\sigma_0$ by seeking for a possibly local minimum with respect to $N$ of:

$$\left| \frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]} - \lambda\,\sigma \right|, \tag{3.2}$$

where $\sigma$ ranges over some search interval $[\sigma_{\min}, \sigma_{\max}]$. Given a lower bound $\rho$ for the ratios $\|\Lambda_n\|/\sigma_0$, the DATE computes the solution in $\sigma$ to the equality:

$$\frac{\sum_{n=1}^N \|Y_n\| \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]}{\sum_{n=1}^N \mathbb{1}[\|Y_n\| \leqslant \sigma\xi(\rho)]} = \lambda\,\sigma. \tag{3.3}$$

Indeed, such a solution trivially minimizes (3.2).

In addition, an application of Bienaymé-Chebyshev's inequality makes it possible to determine the value $n_{\min} \in \{1, 2, \ldots, N\}$ such that the probability that the number of observations due to noise alone be above $n_{\min}$ is larger than or equal to some given probability value $Q$. The main steps of the DATE are summarized in Algorithm 2, where $Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}$ is the sequence $Y_1, Y_2, \ldots, Y_N$ sorted by increasing norm so that $\|Y_{(1)}\| \leqslant \|Y_{(2)}\| \leqslant \ldots \|Y_{(N)}\|$, and where we have defined

$$\mathrm{M}^*_{\{\|Y_1\|,\|Y_2\|,\ldots,\|Y_N\|\}}(n) = \begin{cases} \frac{1}{n}\sum_{k=1}^n \|Y_{(k)}\| & \text{if } n \neq 0 \\ 0 & \text{if } n = 0, \end{cases} \tag{3.4}$$

**Algorithm 2:** DATE algorithm for estimation of noise standard deviation

**Input:**

- A finite subsequence $\{Y_1, Y_2, \ldots, Y_N\}$ of a sequence $Y = (Y_n)_{n\in\mathbb{N}}$ of $d$-dimensional real random vectors satisfying assumptions **(A0-A3)** above
- A lower bound $\rho$ for the SNRs $\|\Lambda_n\|/\sigma_0$, $n \in \mathbb{N}$
- A probability value $Q \leqslant 1 - \frac{N}{4(N/2-1)^2}$

**Constants:** $n_{\min} = N/2 - \sqrt{N/4(1-Q)}$, $\xi(\rho)$, $\lambda$

**Output:** The estimate $\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}}$ of $\sigma_0$

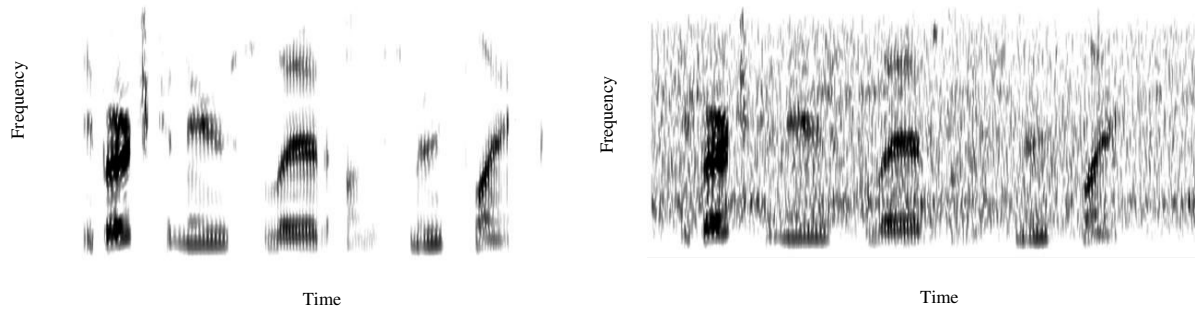
**Computation of $\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}}$:**

**Sort** $Y_1, Y_2, \ldots, Y_N$ by increasing norm so that $\|Y_{(1)}\| \leqslant \|Y_{(2)}\| \leqslant \ldots \leqslant \|Y_{(N)}\|$
**if** there exists a smallest integer $n$ in $\{n_{\min}, \ldots, N\}$ such that:
$\|Y_{(n)}\| \leqslant (\mathrm{M}^*_{\{\|Y_1\|,\|Y_2\|,\ldots,\|Y_N\|\}}(n)/\lambda)\,\xi(\rho) < \|Y_{(n+1)}\|$
  $n^* = n$
**else**
  $n^* = n_{\min}$
**end if**
$\sigma^*_{\{Y_1,Y_2,\ldots,Y_N\}} = \mathrm{M}^*_{\{\|Y_1\|,\|Y_2\|,\ldots,\|Y_N\|\}}(n^*)/\lambda$

The parameters on which the DATE relies are thus: the dimension $d$ of the observations, the number $N$ of observations and the lower bound $\rho$ for the possible SNRs. The two parameters that directly influence the DATE performance are $N$ and $\rho$. As recommended in [65, Remark 4], we can use $\rho = 4$ in practice. Theoretically, $N$ should be large since the theoretical result on which the DATE relies is asymptotic by nature. However, experimental results show that the DATE performance is acceptable when $N$ is above 200. This will be confirmed by the application to speech processing in Sections 3.4 and 3.5.

Another means to choose the minimal SNR required by the DATE is to resort to the notion of universal threshold [103], as proposed in [104]. Indeed, the coordinates of all the $N$ observations $Y_1, Y_2, \ldots, Y_N$ form a set of $N \times d$ random variables. If no signals were present, these $N \times d$ random variables would be i.i.d (independent and identically distributed) Gaussian with zero-mean and variance equal to $\sigma_0^2$. According to [105, Equations (9.2.1), (9.2.2), Section 9.2, p. 187] [106, p. 454] [107, Section 2.4.4, p. 91], the universal threshold $\lambda_u(N \times d) = \sigma_0\sqrt{2\ln(N \times d)}$ could then be regarded as the maximum absolute value of these Gaussian random variables when $N \times d$ is large. Instead of proceeding as in wavelet shrinkage [103] where the universal threshold is utilized to discriminate noisy signal wavelet coefficients from wavelet coefficients of noise alone, the trick proposed in [108] and [104] is to consider $\lambda_u(N \times d)$ as the minimum amplitude that a signal must have to be distinguishable from noise. The minimal SNR can then be defined as $\rho = \rho(N \times d) = \lambda_u(N \times d)/\sigma_0 = \sqrt{2\ln(N \times d)}$. It is an interesting fact that the value of $\rho(N \times d)$ grows rapidly to 4 with $N \times d$.

In the sequel, we will consider values returned by STFT. The DATE will therefore be applied to sequences of real and complex values, that is, one- and two-dimensional data since complex values can be regarded as 2-dimensional real vectors. It is thus worth recalling the specific values of $\xi(\rho)$ and $\lambda$ for $d = 1$ and $d = 2$. If $d = 1$, $\xi(\rho) = \cosh^{-1}(e^{\rho^2/2}) = \frac{1}{2}\rho + \frac{1}{\rho}\log(1 + \sqrt{1 - e^{-\rho^2}})$




(a) Clean speech (b) Noisy speech

Figure 3.1 – Spectrograms of clean and noisy speech signals from the NOIZEUS database. The noise source is car noise. No weighting function was used to calculate the STFT.

and $\lambda = 0.7979$. If $d = 2$, $\xi(\rho) = \mathrm{I}_0^{-1}(e^{\rho^2/2})/\rho$ where $\mathrm{I}_0$ is the zeroth order modified Bessel function of the first kind and $\lambda = 1.2533$. Note that $1/\lambda$ can be regarded as a bias correction factor, similar to those employed by minimum-statistics approaches.

## 3.3 Weak-sparseness model for noisy speech

The main motivation for utilizing the DATE is that noisy speech signals in the time-frequency domain after STFT reasonably satisfy the same type of weak-sparseness model as used to establish [65, Theorem 1]. This weak-sparseness model essentially assumes that the noisy speech signal can be represented by a relatively small number of coefficients with large amplitudes. Indeed, let us consider the spectrograms of Figure 3.1 obtained by STFT of typical examples of clean and noisy speech signals. In the time-frequency domain, speech is composed of a set of time-frequency components or atoms. Most atoms with small amplitudes are masked in the presence of noise. Only the few atoms whose amplitudes are above some minimum value remain visible in noise. Clearly, the proportion of these significant atoms does not exceed one half. These remarks lead to the following model for noisy speech STFTs. In the time domain, as mentioned in Chapter 2, the observed signal is given by

$$\mathbf{y}[n] = \mathbf{s}[n] + \mathbf{x}[n], \tag{3.5}$$

where $\mathbf{s}[n]$ and $\mathbf{x}[n]$ denotes clean speech and independent additive noise. Note that both are real-valued signals. The signal in the time domain is transformed into the time-frequency domain by STFT since most noise reduction systems operate in this particular transform domain. Hence, all processing is frame-based. Let $K$ be the frame length, or equivalently, the STFT length. The corresponding system model in the time-frequency domain then reads:

$$Y[m,k] = S[m,k] + X[m,k], \tag{3.6}$$

in which $m$ denotes the frame index, $k$ is the frequency-bin index, and $S[m,k]$ (resp. $X[m,k]$) stands for the STFT component of the speech signal (resp. noise) at time-frequency point $[m,k]$. Following [1, page 210], we model each $X[m,k]$ as a complex Gaussian random variable. Complex values $Y[m,k]$ are manipulated as 2-dimensional real vectors. According to the empirical remarks above, the weak-sparseness model first assumes that an atomic speech audio source is either present or absent at any given time-frequency point $[m,k]$. The presence or the absence of this

source is modeled by a Bernoulli random variable $\varepsilon[m,k]$. This Bernoulli model is tantamount to and justified by the concept of ideal binary masking in the time-frequency domain, as used in audio source separation [104, 109]. The probability of presence is assumed to be less than or equal to 1/2. Thus $\mathbb{P}[\varepsilon[m,k]=1] \leq 1/2$. Second, the atomic audio source must have significant amplitude so as to contribute effectively to the mixture that composes the speech signal. The minimum amplitude that such a source must have will hereafter be denoted by $\rho$. Let us further denote by $\Theta[m,k]$ the underlying atomic audio source. Then, under the previous assumptions, the noisy speech signal at time-frequency point $[m,k]$ can be modeled as:

$$Y[m,k] = \varepsilon[m,k]\Theta[m,k] + X[m,k] \tag{3.7}$$

We recognize here the weak-sparseness model [110] applied to speech processing, in the continuation of [104].

In summary, our model essentially assumes that the STFT of noisy speech signals satisfies the following three key properties in each time-frequency bin $[m,k]$:
**(A'1)**: the presence/absence of speech $\varepsilon[m,k]$ and the atomic speech audio source $\Theta[m,k]$ are independent,
**(A'2)**: the speech-presence probability does not exceed 1/2,
**(A'3)**: the instantaneous power of the random clean speech signal is upper-bounded by a finite value.

Assumptions (**A'1-A'3**) are adaptations of (**A1-A3**) to the particular case of noisy speech signals. Regarding (**A0**), its equivalent form for noisy speech signals is simply Equation (3.7).

Our purpose is then to estimate the noise power spectrum $\sigma_X^2[m,k] = \mathbb{E}[\|X[m,k]\|^2]$ at any given time-frequency point $[m,k]$. This problem is similar to that addressed in [104], where the signal of interest was a mixture of audio signals, possibly including speech signals, and where additive noise was stationary, Gaussian and white. The DATE was used to estimate the noise power spectrum in [104] because this estimator does not make prior assumption on the statistical nature of the signals of interest. In the present chapter and in contrast to [104], we do not restrict our attention to WGN and generalize the approach of [104] to the estimation of colored and possibly non-stationary noise in the presence of speech. The corresponding extension will be called E-DATE in the following.

## 3.4 Noise power spectrum estimation by E-DATE

In this section, we derive the E-DATE algorithm that will be used for noise power spectrum estimation in all the experiments conducted in Section 3.5. The derivation follows a three-step process, which aims at gradually introducing the modifications required to evolve from the academic WGN model to the much more realistic, but also more challenging, practical case of non-stationary noise. More precisely, we first describe the application of the DATE algorithm to noise power spectrum estimation of noisy speech signals in the time-frequency domain. We extend the DATE to the case of colored stationary Gaussian noise, and then discuss the estimation of non-stationary noise. This leads to the E-DATE algorithm, which is specifically designed for noise power spectrum estimation in non-stationary noisy environments, but can be used with stationary noise as well.

In the following, we suppose to be given $M$ noisy speech frames of $K$ samples. The frames are assumed to be non-overlapping so as to satisfy assumption (**A0**). The STFTs are normalized by $1/\sqrt{K}$.

### 3.4.1 Stationary WGN

In this case, the noise power spectrum is constant and equals $\sigma_X^2$ over the whole time-frequency plane. Accordingly, and by properties of the (normalized) STFT, each noise sample $X[m,k]$ in the time-frequency domain is a zero-mean circularly-symmetric Gaussian complex random variable with variance $\sigma_X^2$:

$$X[m,k] \sim \mathcal{N}_c(0,\sigma_X^2).$$

Equivalently, $X[m,k]$ may be viewed as a zero-mean two-dimensional real Gaussian random vector with covariance matrix $(\sigma_X^2/2)\boldsymbol{I}_2$:

$$X[m,k] \sim \mathcal{N}\left(\mathbf{0},(\sigma_X^2/2)\boldsymbol{I}_2\right).$$

Since the STFT of noisy speech signals is weakly-sparse in the sense of Section 3.3, the $M \times (K/2-1)$ values $Y[m,k]$ for $m \in \{1,2,\ldots,M\}$ and $k \in \{1,2,\ldots,K/2-1\}$ can be used as inputs of the two-dimensional ($d=2$) version of the DATE to provide an estimate $\widehat{\sigma}_X^2$ of $\sigma_X^2$. Note that, due to the Hermitian property of the STFT of real input signals, $\|Y[m,k]\| = \|Y(m,K-k)\|$. Therefore, the frequency bins $K/2+1$ to $K$ are not used in the estimation process as they do not bring additional information. Note also that, in principle, another estimate of $\sigma_X^2$ could be obtained by applying a one-dimensional ($d=1$) DATE on the $2 \times M$ real dataset $Y[1,0],Y[2,0],\ldots,Y[M,0],Y[1,K/2],Y[2,K/2],\ldots,Y[M,K/2]$. However, the size of this second dataset is usually much smaller than that of the first one. Thus only the first option is used in practice as it leads to a more reliable estimate.

### 3.4.2 Colored stationary noise

For colored stationary noise, the noise power spectrum is no longer constant over the whole time-frequency plane but may vary as a function of frequency. Consequently, each noise sample $X[m,k]$ in a given frequency bin $k$ will now be modeled as a zero-mean complex Gaussian random variable with variance $\sigma_X^2(k)$:

$$X[m,k] \sim \mathcal{N}_c(0,\sigma_X^2(k)).$$

Here again, the STFT output sequence $Y[m,k]$ for $m=1,2,\ldots,M$ is assumed to be weakly-sparse in the sense of Section 3.3 so that in each frequency bin $k$, only a few of these values will have an SNR above $\rho$ and in a proportion that does not exceed $1/2$. As a result and as illustrated in Figure 3.2, the extension to colored stationary noise involves running concurrently $K/2+1$ independent instances of the DATE to estimate $\sigma_X^2(k)$ in each frequency bin $k=0,1,2,\ldots,K/2$. As discussed earlier, we do not use the DATE to estimate $\sigma_X^2(k)$ for $k > K/2$ because of the Hermitian symmetry. For $k \in \{1,2,K/2-1\}$, the estimate of $\sigma_X^2(k)$ is computed by the two-dimensional ($d=2$) DATE whereas the one dimensional ($d=1$) DATE is used for bins 0 and $K/2$. For colored noise, assumption (**A'1**) may not always rigorously hold, especially at low frequencies. However, as supported by the experimental results of Section 3.5, this deviation with respect to the underlying theoretical model turns out to be no real issue in practice, thanks to the robust behavior of the DATE, even when the signal presence probability may exceed $1/2$ (see [65, Figure 2]).

In contrast to WGN for which the whole time-frequency plane ($\approx MK/2$ observations) is used to estimate the noise variance $\sigma_X^2$, $M$ frames only are available here to estimate $\sigma_X^2(k)$ in each frequency bin. Clearly a more reliable estimate can be obtained by increasing $M$, but this increases in return the overall computational cost and may also entail some time-delay. A

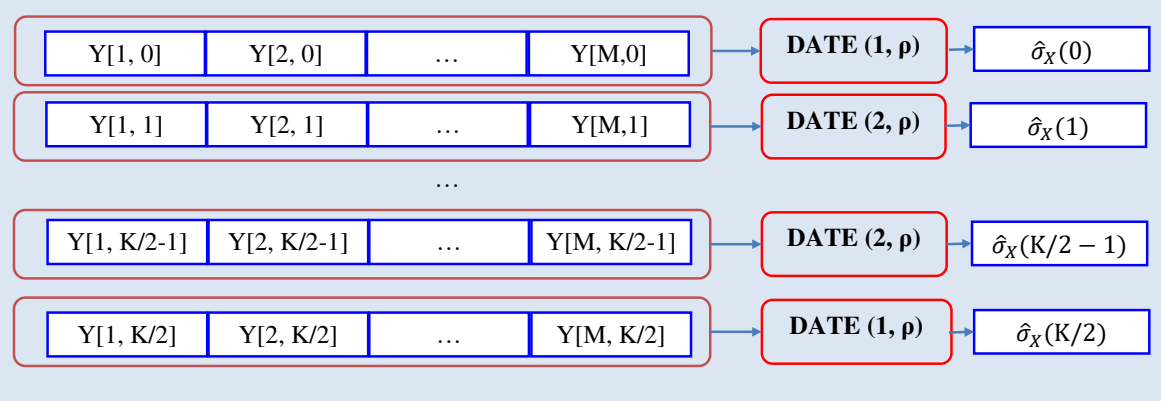



Figure 3.2 – Principle of noise power spectrum estimation based on the DATE in colored stationary noise

possible solution is to begin with a first estimate $\widehat{\sigma}_X^2(k)$ computed over the first $M$ frames, and then to periodically update this estimate as new frames are acquired. For stationary noise, the initial number of frames $M$ does not need to be very high. Even if the first estimate is not very accurate, it is expected to improve rapidly as new frames enter the estimation process.

### 3.4.3 Extension to non-stationary noise: The E-DATE algorithm

Most practical applications including speech denoising usually face a mix of stationary as well as non-stationary noise. Unlike white or colored stationary noise, the power spectrum of non-stationary noise varies over time and frequency, and, as such, proves to be much more challenging to estimate. Interestingly, non-stationary noise models including car noise, babble noise, exhibition noise and others, usually exhibit some form of local stationarity in time and frequency. In such cases, non-stationary noise can be considered as approximately stationary within short time periods of $D$ consecutive frames, where parameter $D$ has to be defined appropriately for each noise model. This amounts to assuming the existence of a noise power spectrum in this time interval, which is a function of frequency only. The DATE algorithm for colored stationary noise introduced in Section 3.4.2 can then be used to estimate the noise power spectrum within this time window of $D$ frames. This is the basis of the E-DATE algorithm.

Parameter $D$ can be preset once for all or could be optimized for applications where prior knowledge about noise is available. The choice for duration $D$ results from a trade-off between estimation accuracy, stationarity and practical constraints such as computational cost and time-delay. A large value for $D$ may violate the local stationary property. On the other hand, the number of frames $D$ should be large enough to produce reliable estimates $\sigma_X^2(k)$. In case $D$ is too small to provide the DATE with a sufficient number of input data, a possible solution consists in grouping several consecutive frequency bins. This is tantamount to assuming that the noise power spectrum is approximately constant over those frequencies. Such a procedure however requires prior knowledge on the noise spectrum properties, which can be irrelevant in practical applications where noise has often unknown type and may evolve across time. For this reason, this solution will not be further studied below.

In summary, the E-DATE algorithm consists in carrying noise power spectrum estimation by running a per-bin instance of the DATE (see Figure 3.2) on periods of $D$ consecutive non-overlapping frames, where $D$ is chosen so that noise can be considered as approximately stationary within this time interval. Once an estimate of the noise power spectrum has been obtained,

it can be used for denoising purpose for instance, but will not be taken into account in the computation of future estimates, as the local power spectrum of non-stationary noise may change significantly from one period of $D$ frames to the next.

Although the E-DATE algorithm was specifically designed for power spectrum estimation of non-stationary noise, it can be used without modification for power spectrum estimation of WGN or colored stationary noise, thereby offering a robust and universal noise power spectrum estimator whose parameters are fixed once for all types of noise considered above. Let us now discuss the practical implementation of the E-DATE algorithm.

### 3.4.4 Practical implementation of the E-DATE algorithm

Two different implementations of the E-DATE algorithm are proposed here. The first approach is a straightforward block-based implementation of the algorithm described in Section 3.4.3. It involves estimating the noise power spectrum on each period of $D$ successive non-overlapping frames. This requires storing $D$ frames, calculating the $K/2+1$ estimates $\widehat{\sigma}_X^2(k)$ using the observations in these $D$ frames, and then waiting for $D$ new non-overlapping frames. The resulting algorithm is called Block-E-DATE (B-E-DATE) and summarized in Algorithm 3, where $\widehat{\sigma} = \mathrm{DATE}_{d,\rho}(y_1, y_2, \ldots, y_n)$ denotes the standard deviation estimate $\widehat{\sigma}$ returned by the $d$-dimensional DATE with minimal SNR $\rho$ and $n$ real $d$-dimensional inputs $y_1, y_2, \ldots, y_n$.

Estimation of the noise power spectrum over separate periods of $D$ non-overlapping frames reduces the overall algorithm complexity. However, this entails a time-delay of $D$ frames, which must be considered in applications. Consider the particular example of speech denoising illustrated in Figure 3.3. Noise reduction is performed on a frame-by-frame basis. A new noise power spectrum estimate is provided to the noise reduction system by the B-E-DATE algorithm once every $D$ non-overlapping frames, and then used to denoise each of those $D$ frames. Clearly, denoising cannot start before the first $D$ non-overlapping frames have been recorded. This results in an overall latency of about 1 or 2 seconds for typical sampling rates of 8 and 16 kHz. This delay can then have some impact for speech applications embedded in current mobile devices. It will naturally be lesser in applications such as Active Noise Cancellation (ANC) where frequency rates are much higher.

**Algorithm 3:** Block-Extended-DATE (B-E-DATE) algorithm for noise power spectrum estimation

**for** $m \geq D$ **do**
  **if** $\mod(m, D) = 0$
    $m^* = m$
    $\widehat{\sigma}_X[m^*, 0] = \mathrm{DATE}_{1,\rho}(Y[m-D+1, 0], Y[m-D+2, 0], \ldots, Y[m, 0])$
    $\widehat{\sigma}_X[m^*, K/2] = \mathrm{DATE}_{1,\rho}(Y[m-D+1, K/2], Y[m-D+2, K/2], \ldots, Y[m, K/2])$
    **for** $k := 1$ to $\frac{K}{2} - 1$ **do**
      $\widehat{\sigma}_X[m^*, k] = \mathrm{DATE}_{2,\rho}(Y[m-D+1, k], Y[m-D+2, k], \ldots, Y[m, k])$
      $\widehat{\sigma}_X[m^*, K-k] = \widehat{\sigma}_X[m^*, k]$
    **end for**
  **else**
    **for** $k := 0$ to $K-1$ **do**
      $\widehat{\sigma}_X[m-D, k] = \widehat{\sigma}_X^*[m^*, k]$
    **end for**
  **end if**
**end for**

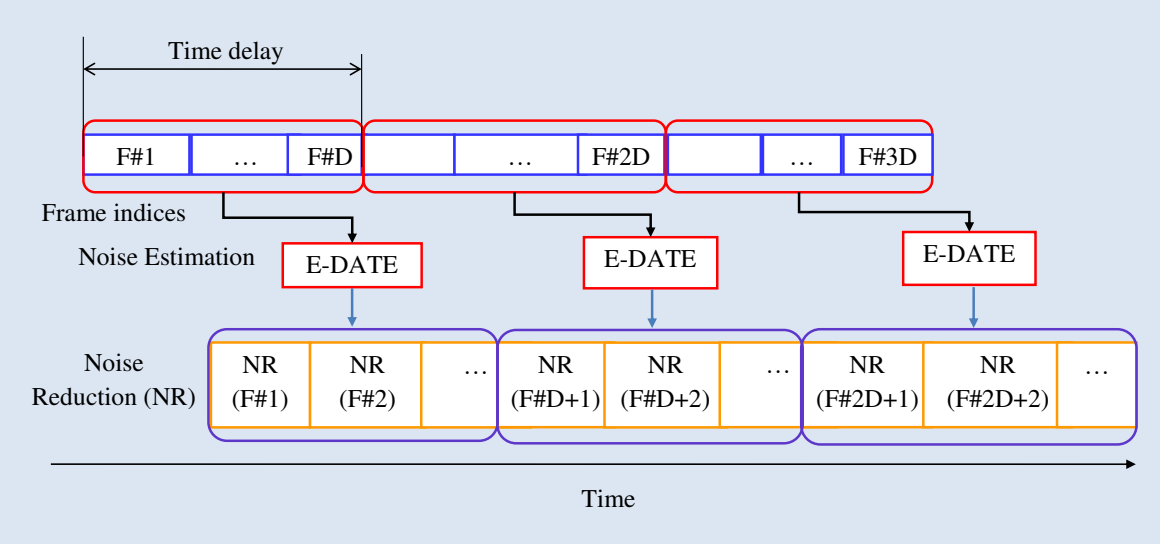



Figure 3.3 – Block E-DATE (B-E-DATE) combined with noise reduction (NR). A single noise power spectrum estimate is calculated every $D$ non-overlapping frames and used to denoise each of these $D$ frames.

The delay limitation can be bypassed as follows. First, a standard noise power spectrum tracking method is used to estimate the noise power spectrum during the first $D-1$ non-overlapping frames. Any of the methods mentioned in the introduction can be used for this purpose. Afterwards, starting from the $D^{\text{th}}$ frame onwards, a sliding-window version of the E-DATE algorithm is used to estimate the noise spectrum on a per-frame basis, using the latest recorded $D$ non-overlapping frames. This alternative implementation called Sliding-Window E-DATE (SW-E-DATE) is summarized in Algorithm 4. Its application to speech denoising is illustrated in Figure 3.4.

**Algorithm 4:** Sliding-Window Extended-DATE (SW-E-DATE) algorithm for noise power spectrum estimation

**for** $m = 1$ to the end of signal **do**
  **if** $m < D$
    Calculate $\widehat{\sigma}_X$ by an alternative method
  **else**
    $\widehat{\sigma}_X[m, 0] = \text{DATE}_{1,\rho}(Y[m-D+1, 0], Y[m-D+2, 0], \ldots, Y[m, 0])$
    $\widehat{\sigma}_X[m, K/2] = \text{DATE}_{1,\rho}(Y[m-D+1, K/2], Y[m-D+2, K/2], \ldots, Y[m, K/2])$
    **for** $k := 1$ to $\frac{K}{2} + 1$ **do**
      $\widehat{\sigma}_X[m, k] = \text{DATE}_{2,\rho}(Y[m-D+1, k], Y[m-D+2, k], \ldots, Y[m, k])$
      $\widehat{\sigma}_X[m, K-k] = \widehat{\sigma}_X[m, k]$
    **end for**
  **end if**
**end for**

The B-E-DATE and the SW-E-DATE algorithms may be viewed as two particular instances of a more general buffer-based algorithm. More precisely, the B-E-DATE algorithm corresponds to the extreme case where the buffer is totally flushed and updated once every $D$ non-overlapping frames. In contrast, the SW-E-DATE algorithm corresponds to the other extreme case where

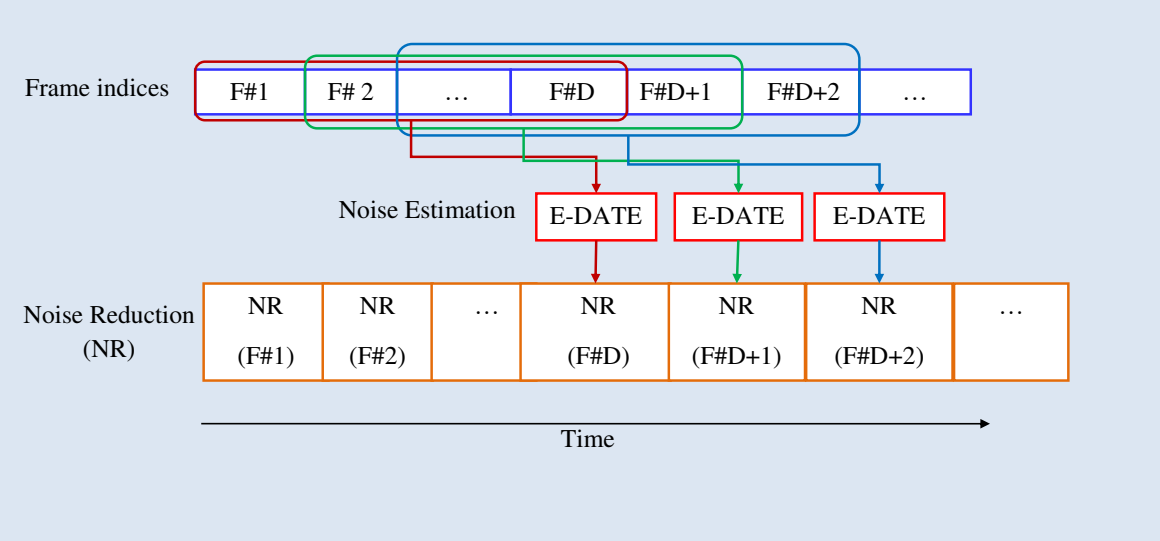



Figure 3.4 – Sliding-Window E-DATE (SW-E-DATE) combined with noise reduction. For the first $D-1$ frames, a surrogate method for noise power spectrum estimation is used in combination with noise reduction. Once $D$ frames are available and upon reception of frame $D+\ell$, $\ell \geq 0$, the SW-E-DATE algorithm provides the NR system with a new estimate of the noise power spectrum computed using the last $D$ frames $F_{\ell+1}, \ldots, F_{\ell+D}$ for denoising of the current frame.

only the oldest frame is discarded in order to store the current one, in a First-In First-Out (FIFO) mode. Clearly, a more general approach between these two extremes consists in partially updating the buffer by renewing only $L$ frames among $D$. This point has not been further investigated in the present work.

Note finally that the proposed implementations of the E-DATE algorithm are not limited to speech denoising but could find use in any application involving signals corrupted by additive and independent non-stationary noise, and to which the weak-sparseness model locally applies.

## 3.5 Performance evaluation

Several comparisons and experiments were conducted in order to assess the performance and benefits of the E-DATE noise power spectrum estimator in comparison with other state-of-the-art algorithms. Both the B-E-DATE and the SW-E-DATE implementations were considered in two different benchmarks. In subsection 3.5.1, we first compare the number of parameters required by the E-DATE and several classical or more recent noise power spectrum estimators. Then, we compare in subsection 3.5.2 the estimation quality of the different algorithms in several distinct noise environments. The combination of the noise power spectrum estimation algorithms with a noise reduction system based on the STSA-MMSE algorithm is investigated using the NOIZEUS speech corpus in subsection 3.5.3. Finally, the time-complexity of the E-DATE algorithm is analyzed in subsection 3.5.4.

### 3.5.1 Number of parameters

Table 3.1 gives the number of parameters required by the E-DATE as well as by the state-of-the-art noise power spectrum estimation algorithms mentioned in the introduction. Derived from robust statistical signal processing concepts, the E-DATE is the simplest algorithm to configure, with only two parameters to specify, namely the SNR lower bound $\rho$ and the number

Table 3.1 – Number of parameters (NP) required by different noise power spectrum estimation algorithms

| Method | MS | IMCRA | MCRA2 | MMSE1 | MMSE2 | E-DATE |
|---|---|---|---|---|---|---|
| NP | 7 | 10 | 7 | 3 | 5 | 2 |

of frames $D$. This stands in sharp contrast with other popular approaches such as Minimum Statistics (MS) [93], which involves 7 parameters. In practice, the minimal SNR $\rho$ can be set as explained at the end of Section 3.2 so that the only crucial parameter is $D$. Working with $D = 80$ non-overlapping frames of $K = 256$ samples was found to yield good performance in all the experiments reported here.

### 3.5.2 Noise estimation quality

The estimation quality of the noise power spectrum estimation algorithms listed in Table 3.1 was evaluated on several noise models using the symmetric segmental logarithmic estimation error measure defined in [111]. The difference between the estimated noise power spectrum $\widehat{\sigma}_X^2[m,k]$ and reference noise power spectrum $\sigma_X^2[m,k]$ is evaluated by

$$LogErr = \frac{1}{MK}\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\left|10\log_{10}\frac{\widehat{\sigma}_X^2[m,k]}{\sigma_X^2[m,k]}\right| \tag{3.8}$$

where $M$ denotes the total number of the available frames. For WGN, the theoretical reference noise power spectrum is known and can be substituted to $\sigma_X^2[m,k]$ in (3.8). This is no longer the case for non-stationary noise involved in the NOIZEUS database. For non stationary noise, the reference noise power spectrum $\sigma_X^2[m,k]$ is estimated as follows [111]:

$$\sigma_X^2[m,k] = \alpha\sigma_X^2[m-1,k] + (1-\alpha)|X[m,k]|^2, \text{with } \alpha = 0.9. \tag{3.9}$$

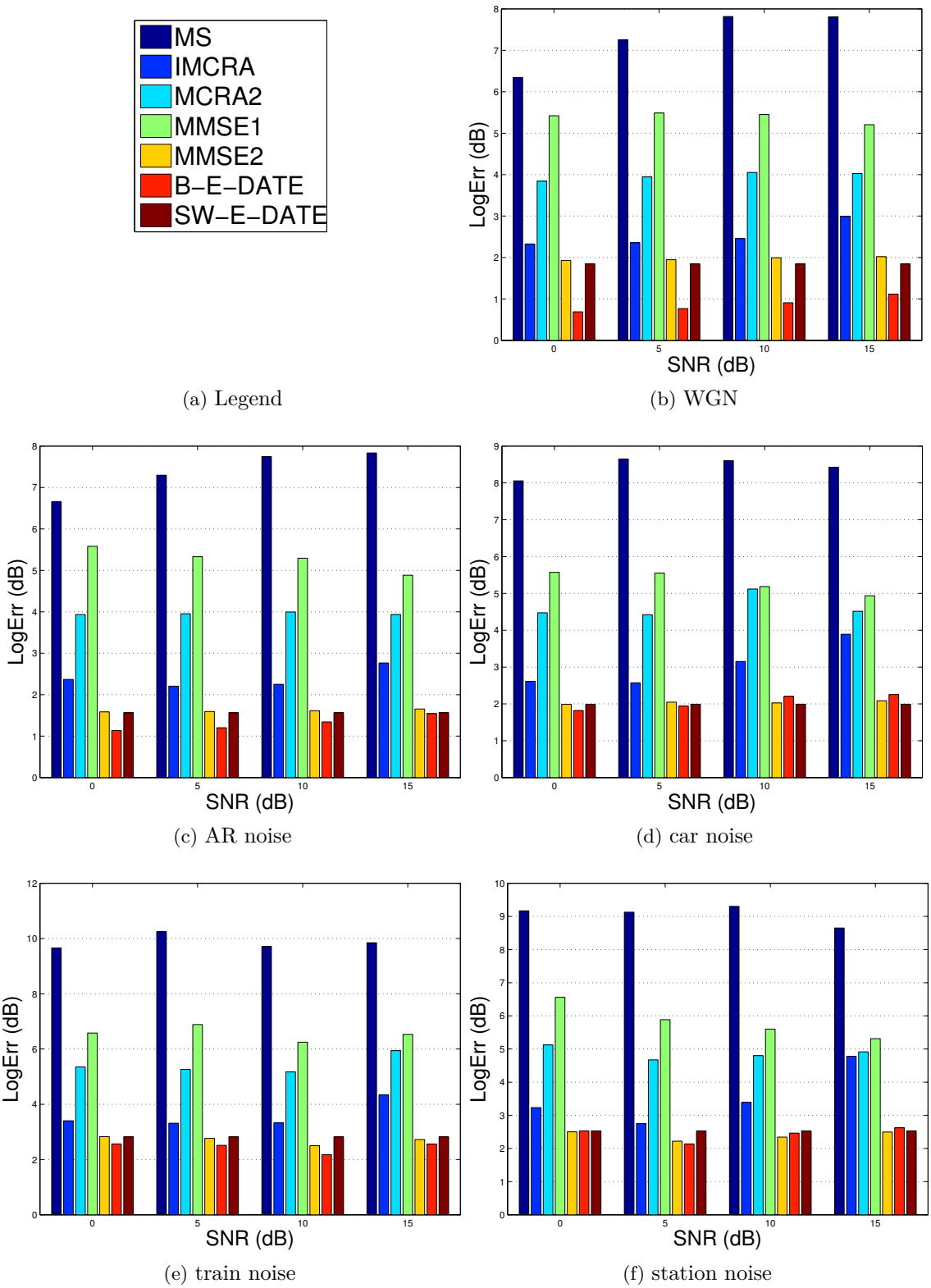
Both the B-E-DATE and the SW-E-DATE implementations of the E-DATE algorithm were evaluated and compared. The SW-E-DATE uses the recently-introduced MMSE2 method [97] as a surrogate algorithm to provide an estimate for the first $D-1$ frames since, as shown below, this algorithm turns out to offer excellent performance among state-of-the-art noise estimators.

The *LogErr* measures obtained with the different noise power spectrum estimators are given in Figures 3.5 and 3.6. All algorithms have been benchmarked at four SNR levels and against various noise models, namely three synthetic noises (WGN, auto-regressive (AR) colored stationary noise and modulated WGN), and 8 typical real non-stationary noise environments.

The results for white and colored stationary noise are given in Figures 3.5b and 3.5c, respectively. The B-E-DATE and SW-E-DATE methods yield the lowest *LogErr* error, the best performance being achieved by the B-E-DATE algorithm in WGN. This is no surprise since the underlying DATE algorithm was originally developed for estimating the standard deviation of additive WGN.

For non-stationary noise with slowly-varying noise spectrum like car, station, train and with speech-like noise including exhibition, restaurant and babble, depending on the noise level, the B-E-DATE algorithm uniformly obtains either the best score, or comes very close to the best score, as shown in Figures 3.5d- 3.5e and 3.6a-3.6c respectively.

Figures 3.6e-3.6f present the results obtained with the least favorable types of non-stationary noise. In the case of modulated WGN (resp. babble noise), the SW-E-DATE algorithm yields the smallest *LogErr* error. As illustrated in Figures 3.6e and 3.6f, the two proposed algorithms are among the best in estimating the very challenging airport noise environment. Their performance closely match those obtained with the state-of-the-art MMSE2 estimator.




(a) Legend

(b) WGN

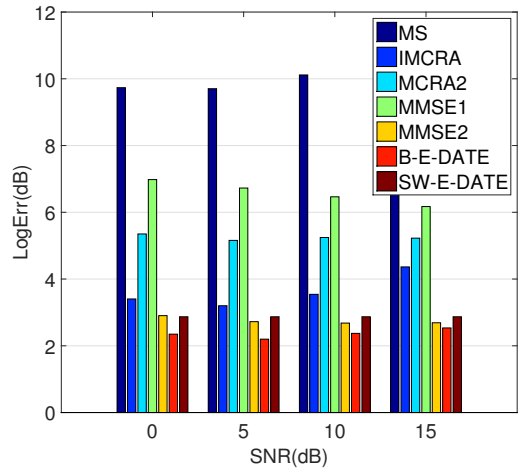
(c) AR noise

(d) car noise

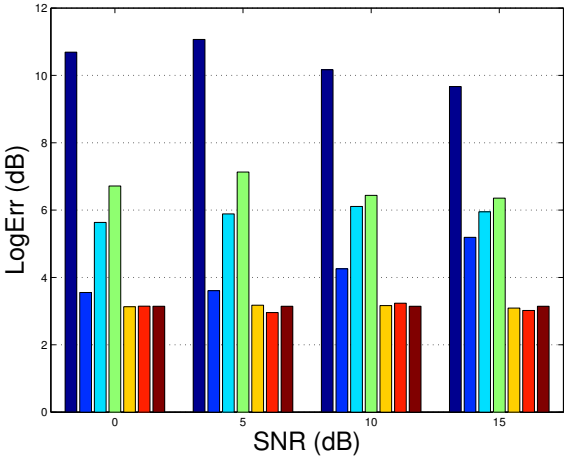
(e) train noise

(f) station noise

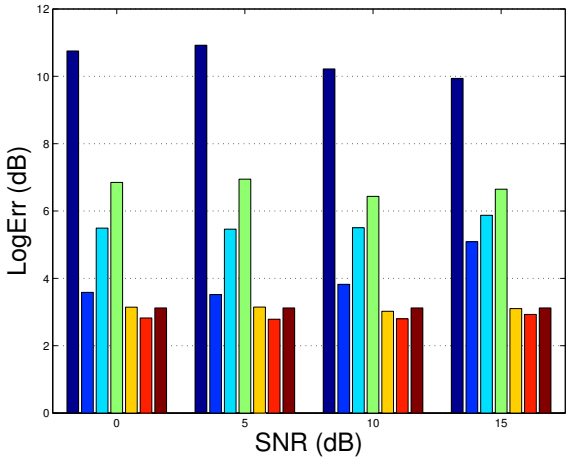
Figure 3.5 – Noise estimation quality comparison of several noise power spectrum estimators at different SNR levels and with different kinds of stationary synthetic noise and slowly varying non-stationary noise. Legend is displayed in Figure 3.5a.



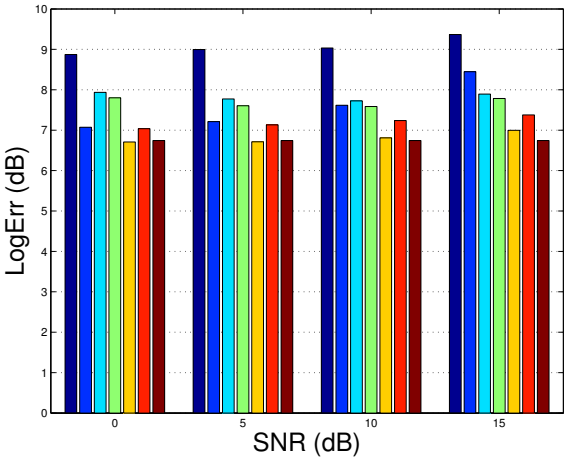

(a) exhibition noise



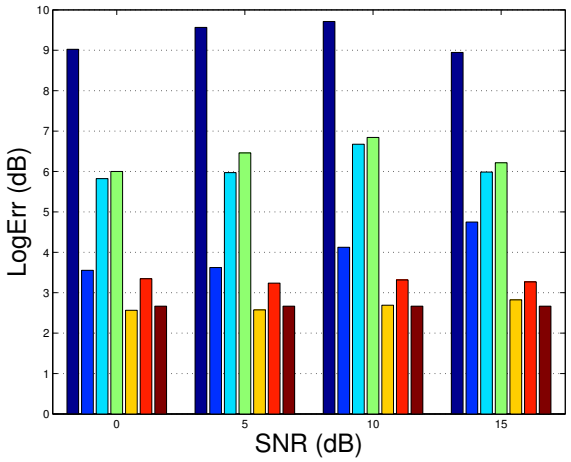

(b) restaurant noise



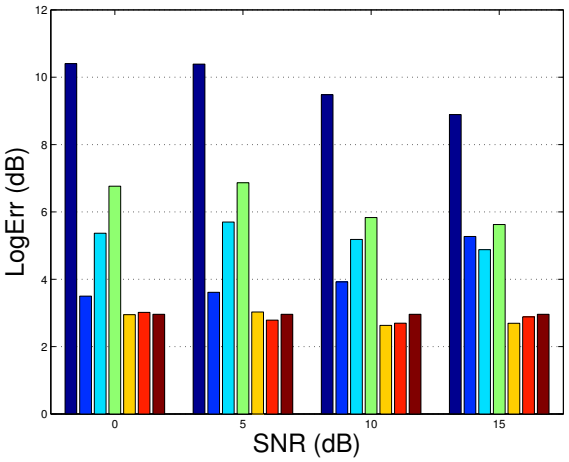

(c) babble noise




(d) modulated WGN




(e) street noise




(f) airport noise

Figure 3.6 – Noise estimation quality comparison of several noise power spectrum estimators at different SNR levels and with different kinds of non-stationary noise where noise power spectra are changing fast. The same legend as in Figure 3.5a is used.

### 3.5.3 Performance evaluation in speech enhancement

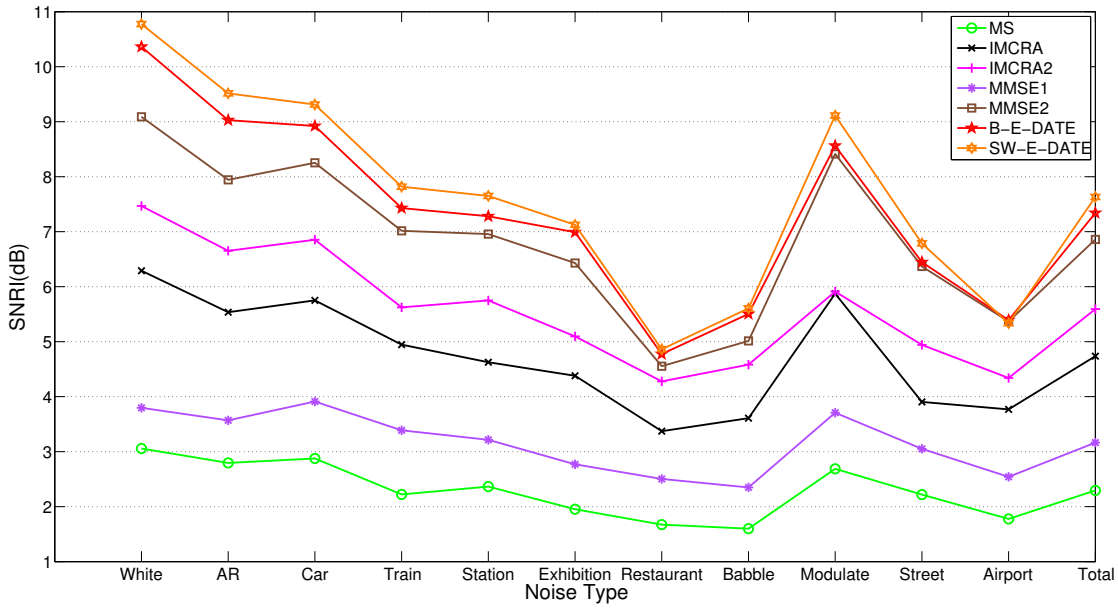



Figure 3.7 – SNRI with various noise types

In complement to the previous study, the performance of the noise power spectrum estimation algorithms listed in Table 3.1 have also been evaluated and compared in combination with a noise reduction system. The speech denoising experiments are based on the NOIZEUS database [1], which contains IEEE sentences corrupted by eight types of noise coming from the AURORA noise database, at four SNR levels, namely 0, 5, 10 and 15 dB. The noise reduction algorithm retained for our experiments is the STSA-MMSE estimator [27]. This method is a standard reference in speech denoising. It can easily be implemented and is known to reduce residual noise without introducing musical noise [1, p. 215, Sec. 7.3].

Three different criteria have been used to compare the different algorithms as mentioned in Section 2.3.1 in Chapter 2. The first one is the SSNR objective criterion. For illustrating the performance of speech enhancement, we evaluated the SSNR improvement (SSNRI) defined as the difference between SSNR of the enhanced signal and SSNR of the noisy signal. Figures 3.8 and 3.9 provide the SSNRI performance for various noise types and SNR levels. B-E-DATE and SW-E-DATE yield good performance in the case of stationary and low-varying non-stationary noise like WGN, AR noise, car noise, train noise and station noise (see Figure 3.8). Even better performances are obtained in Figures 3.9a-3.9c for the exhibition, restaurant and babble speech-like non-stationary noises, respectively. For the fast-changing non-stationary noise shown in Figures 3.9d-3.9f, the two proposed methods are also the best algorithms. Their SSNRI are closed to those achieved with MMSE2 estimator.

The second performance metric is the Signal-to-Noise Ratio Improvement (SNRI) objective criterion standardized in the ITU-T G.160 recommendation for evaluating noise reduction systems [80]. The SNRI performance obtained with the STSA-MMSE combined with the noise power spectrum estimators of Table 3.1 are shown in Figure 3.7 for various noise environments. Note that 4 noise levels were used for each noise type, the final SNRI score being computed as the average score over these 4 levels. We observe that the B-E-DATE and SW-E-DATE yield

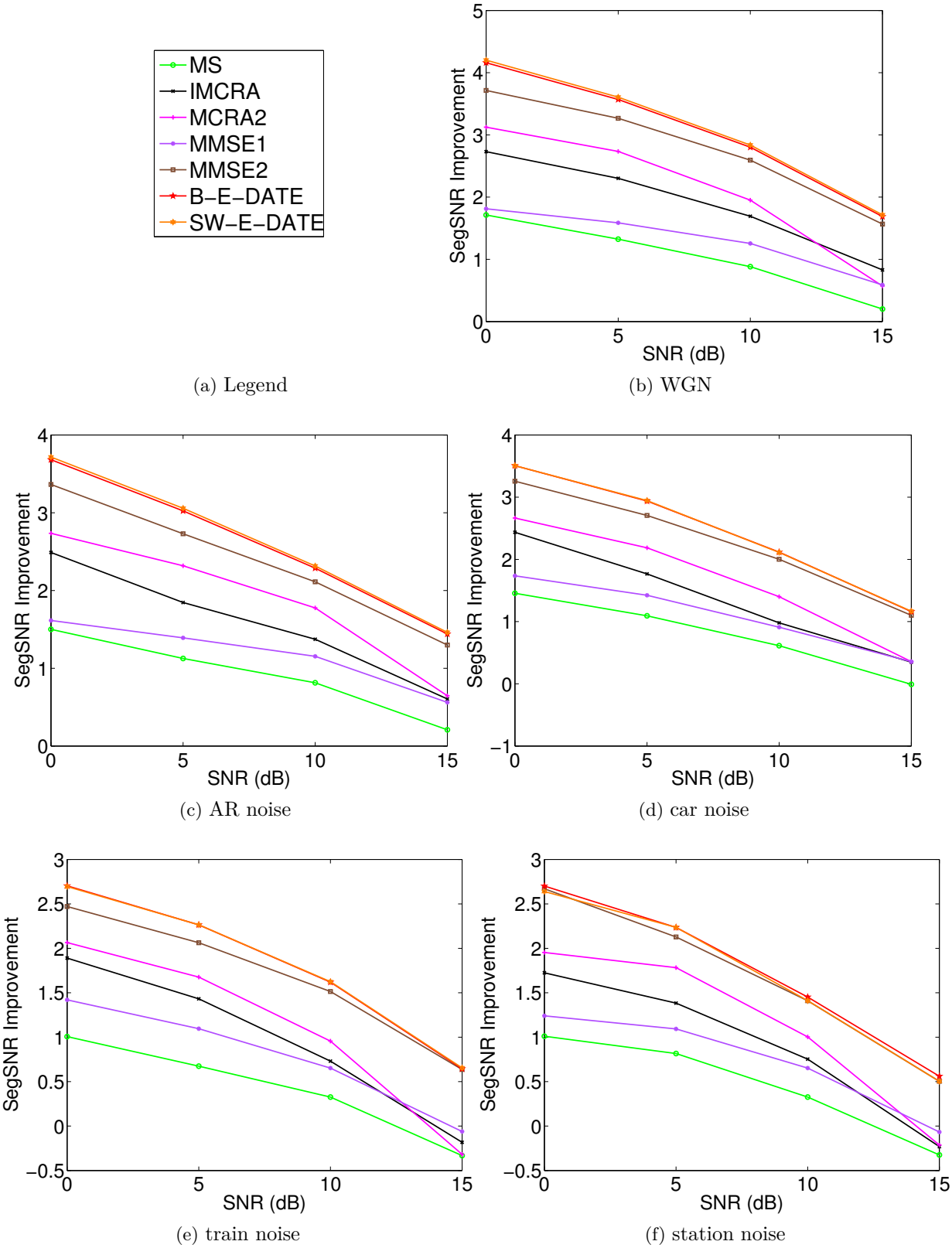
similar performance measurements and that they outperform all other methods for each type of noise. The average SNRI score computed over the 11 noise types and labeled *Total* at the right of Figure 3.7 clearly emphasizes the SNRI gain brought by the E-DATE in comparison to other methods.

The third criterion used to assess noise power spectrum estimation in speech enhancement is the composite objective measures proposed in [83]. As mentioned in Chapter 2, the three measures MARSsig, MARSbak and MARSovrl are designed so as to provide a high correlation with the three widely used corresponding subjective measures that are signal distortion (SIG), background intrusiveness (BAK) and Mean Opinion Score (OVRL). We focus here on the MARSovrl criterion since it has the highest correlation with the real subjective tests. Figures 3.10 and 3.11 show the MARSovrl improvement scores, defined as SSNRI scores, obtained with the different noise power spectrum estimators and noise environments. The good performance of the B-E-DATE and SW-E-DATE are confirmed by the MARSovrl measures obtained in the case of WGN, AR noise, car noise, station noise and train noise. These results allow us to conclude that the E-DATE approach is well-suited for stationary or slowly varying non-stationary noise. Although not shown here for space limitation, we hasten to mention that very similar trends were observed for the other two criteria MARSsig and MARSbak. In the challenging case of fast-changing noise, all the methods in this chapter yield the same result at 0dB. At higher SNR levels, depending on the kind of noise, the E-DATE MARSovrl scores are similar to that obtained by the best method or are the highest scores (see Figure 3.11).

Two final remarks are in order here. First, the B-E-DATE algorithm generally performs better than the SW-E-DATE algorithm. This is particularly evident in Figures 3.9 and 3.9 and can also be noticed in the other experimental results. This is mainly due to the fact that our implementation of the SW-E-DATE initially resorts to a surrogate algorithm to estimate noise power spectrum during the first $D = 80$ frames, which has inferior performance to the B-E-DATE. Since these $D$ frames represent a significant part of the total duration of many of the tested utterances, the performance loss incurred by the use of a worse estimator significantly impacts the overall score. Second, in the previous section was evoked the possibility to partially update the buffer by renewing only $L$ frames among $D$ instead of flushing it completely (B-E-DATE), or renewing it only one frame at a time in a FIFO manner (SW-E-DATE). The difference in performance between these two E-DATE implementations suggests that such a partial renewal should not dramatically modify the results. This means that buffer optimization can be performed in practice whenever required by practical constraints, and without significantly impacting the denoising performance. For instance, additional experimental results with airport, babble, station, car and train noises suggest that $D$ can be chosen in the range $[50, 80]$ without really affecting MARSovrl for SNR $>$ 0dB.

### 3.5.4 Complexity analysis

Tables 3.3 and 3.4 compare the computational costs of the B-E-DATE and SW-E-DATE implementations, respectively. Each table gives the number of real additions, multiplications, divisions and square roots required to perform the estimate. Both the B-E-DATE and the SW-E-DATE use $D$ frames to compute the noise power spectrum estimate. However computation is performed only once every $D$ frames for the B-E-DATE algorithm, whereas it is performed once per frame in the SW-E-DATE implementation. Hence the number of operations in Table 3.3 should be divided by $D$ to allow for a fair per-frame computational cost comparison between the two implementations. For reference purpose, Table 3.2 lists the number of operations required by the MMSE2 estimator of [97]. Inspection of Tables 3.3 and 3.2 shows that the B-E-DATE and




(a) Legend

(b) WGN

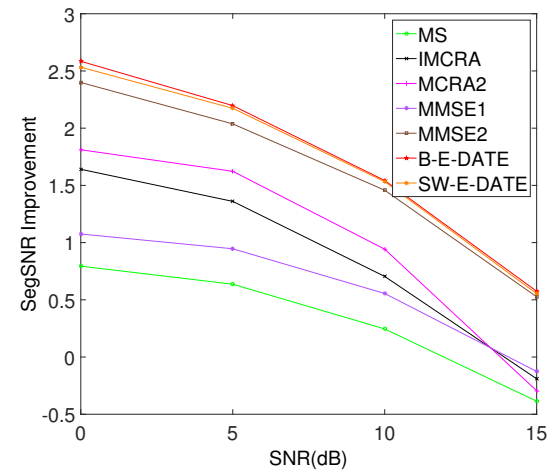
(c) AR noise

(d) car noise

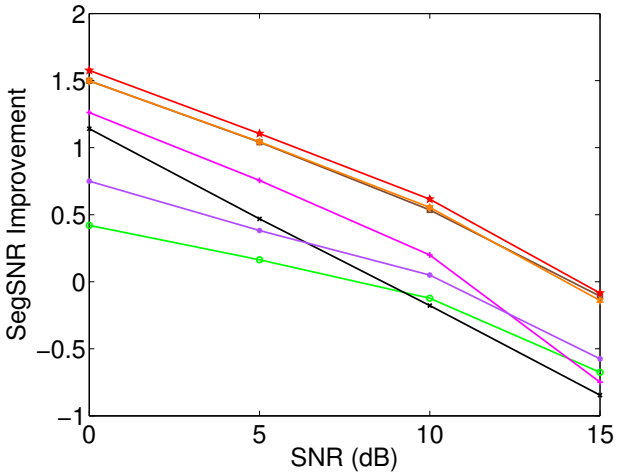
(e) train noise

(f) station noise

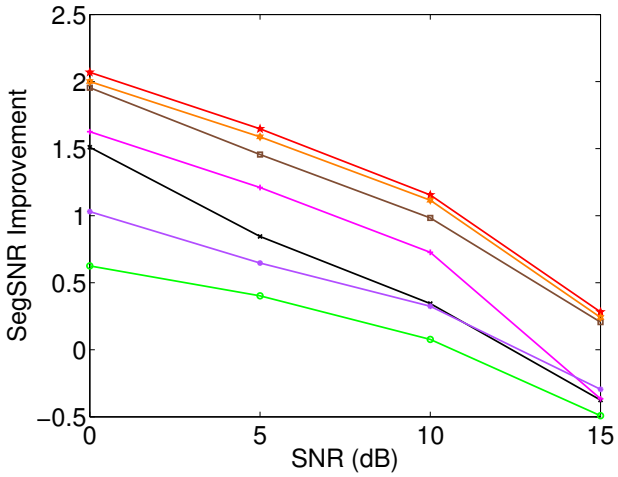
Figure 3.8 – Speech quality evaluation after speech denoising (SSNR) for the stationary and low-varying non-stationary noise. Legend of all sub-figure is illustrated in Figure 3.8a.



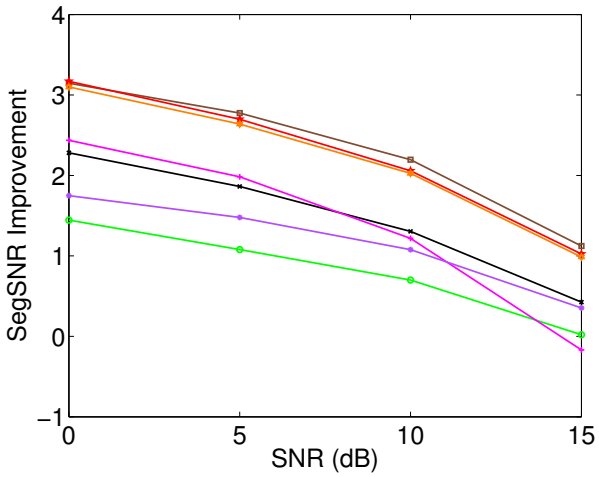

(a) exhibition noise



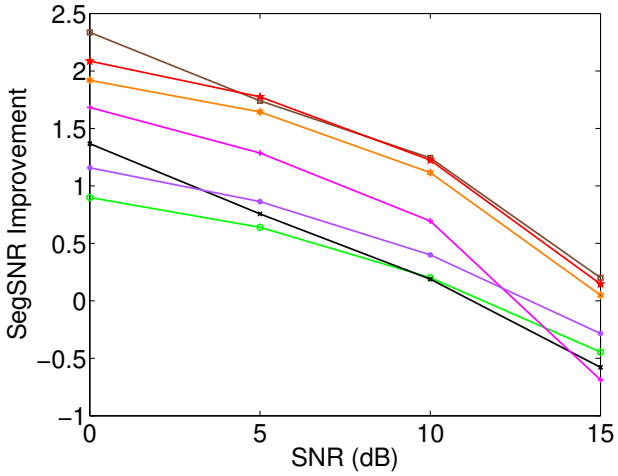

(b) restaurant noise



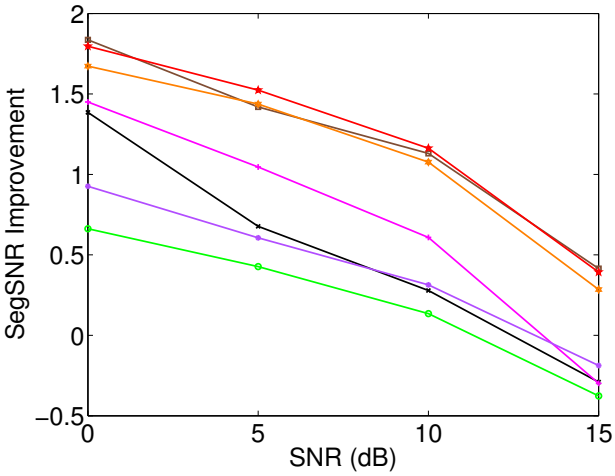

(c) babble noise




(d) modulated WGN

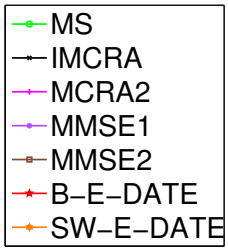



(e) street noise

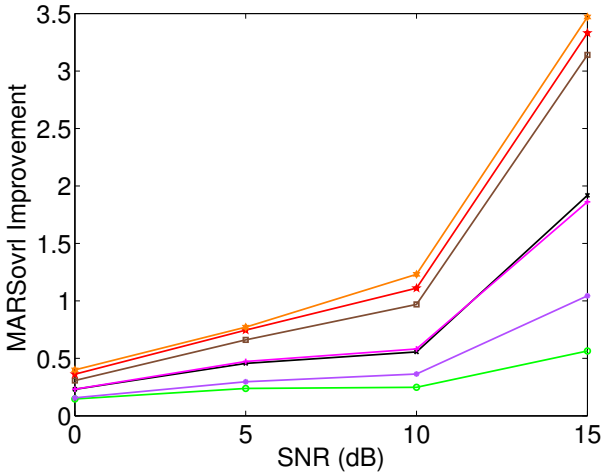



(f) airport noise

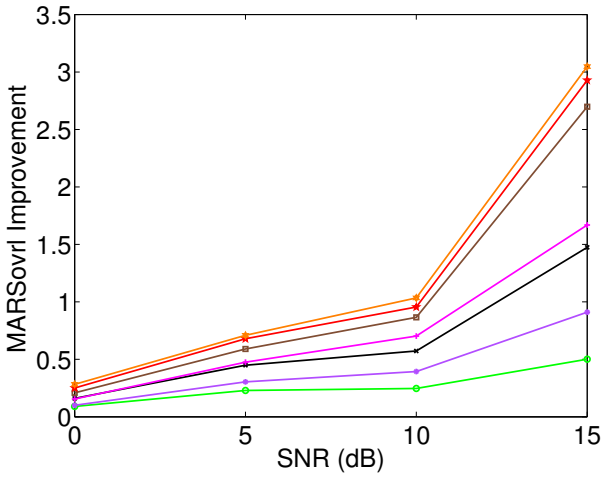
Figure 3.9 – Speech quality evaluation after speech denoising (SSNR)for the fast-changing or speech-like non-stationary noise. Legend is the same as in Figure 3.8a.



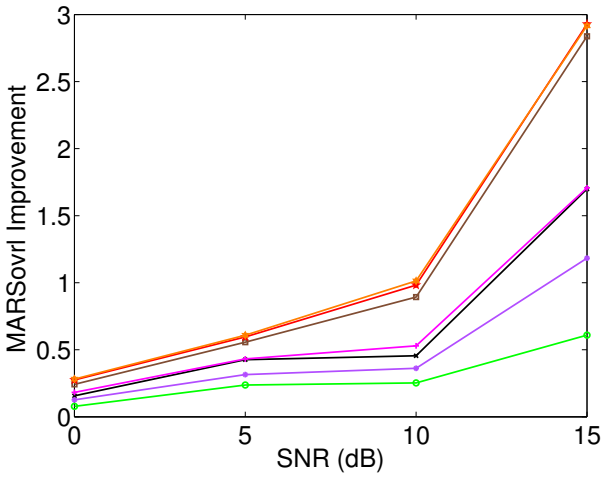

(a) Legend



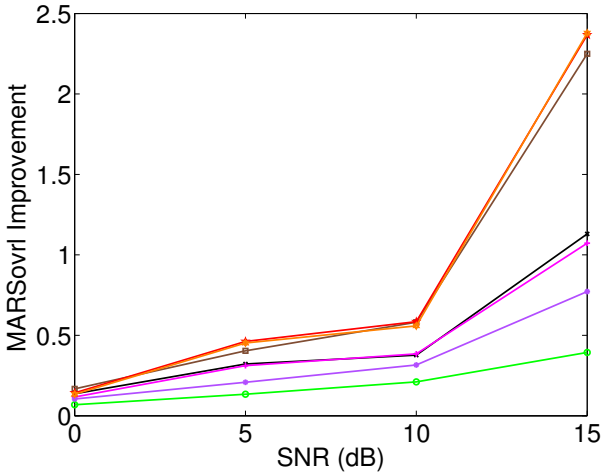

(b) WGN



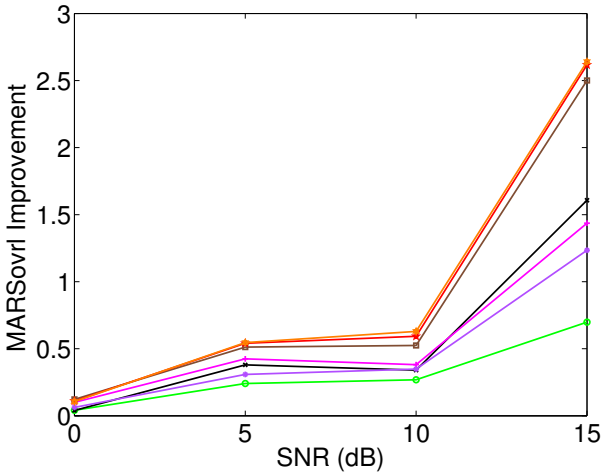

(c) AR noise




(d) car noise

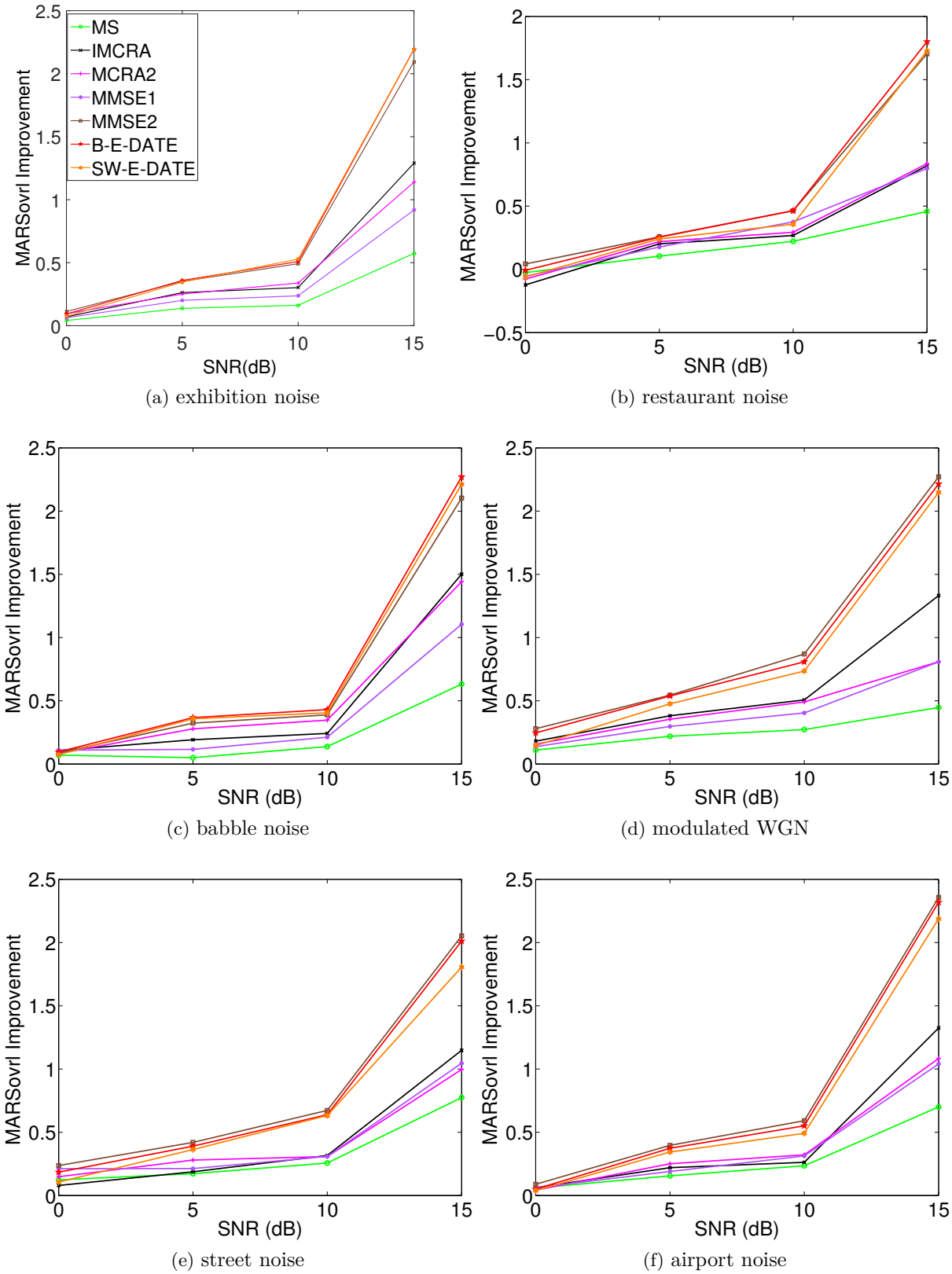



(e) train noise




(f) station noise

Figure 3.10 – Speech quality evaluation after speech denoising ($MARS_{ovrl}$ composite criterion) for stationary or low-varying non-stationary noise. Legend is the same as in Figure 3.10a.




(a) exhibition noise

(b) restaurant noise

(c) babble noise

(d) modulated WGN

(e) street noise

(f) airport noise

Figure 3.11 – Speech quality evaluation after speech denoising ($MARS_{ovrl}$ composite criterion) for fast-changing or speech-like non-stationary noise. Legend is also pointed out in Figure 3.10a.

Table 3.2 – Computational cost of MMSE2 per new frame and per frequency bin

| Addition | Multiplication | Division | Exponent |
|---|---|---|---|
| 12 | 10 | 2 | 1 |

Table 3.3 – Computational cost of B-E-DATE per group of $D$ frames and per frequency bin

| | Addition | Multiplication | Division | Square root |
|---|---|---|---|---|
| Norm | $D$ | $2D$ | 0 | $D$ |
| Sorting | $D \log D$ | 0 | 0 | 0 |
| Search $n^*$(worst case) | $D(D-1)/2$ | $D$ | $D$ | 0 |
| Total | $D\left(\log D + (D+1)/2\right)$ | $3D$ | $D$ | $D$ |

MMSE2 estimators have similar computational complexity. This is confirmed by execution times of Matlab implementations of these algorithms where the B-E-DATE algorithm is found to have a processing time about 1.53 times that of the MMSE2 algorithm. We also note from Tables 3.3 and 3.4 that SW-E-DATE requires approximately $D/3$ times more operations that B-E-DATE. Indeed, B-E-DATE requires $3D$ multiplications to process $D$ frames at once, whereas SW-E-DATE requires $D+2$ multiplications per frame. Execution times of Matlab implementations of these algorithms also confirm this ratio.

## 3.6 Conclusion

In this chapter, we have proposed a novel method to estimate the power spectrum of some non-stationary noise, in applications where a weak-sparse transform makes it possible to represent the signal of interest by a relatively small number of coefficients with significantly large amplitude. The resulting estimator called Extended-DATE (E-DATE) is robust in that it does not use prior knowledge about the signal or the noise except for the weak-sparseness property. Compared to other methods in the literature, the E-DATE algorithm has the remarkable advantage of requiring only two parameters to specify. A straightforward block-based implementation of the E-DATE, called B-E-DATE, has first been introduced. This implementation entails an estimation delay, which diminishes as the frequency rate increases. This delay could be reduced by grouping frequency bins. Another solution to shorten this delay involves resorting to a sliding-window implementation called SW-E-DATE, but at the price of a higher computational cost. The B-E-DATE and SW-E-DATE have been benchmarked against various classical and recent noise power spectrum estimation methods in two situations: with and without noise reduction. The experimental results show that the E-DATE estimator generally provides the most accurate

Table 3.4 – Computational cost of SW-E-DATE per new frame and per frequency bin

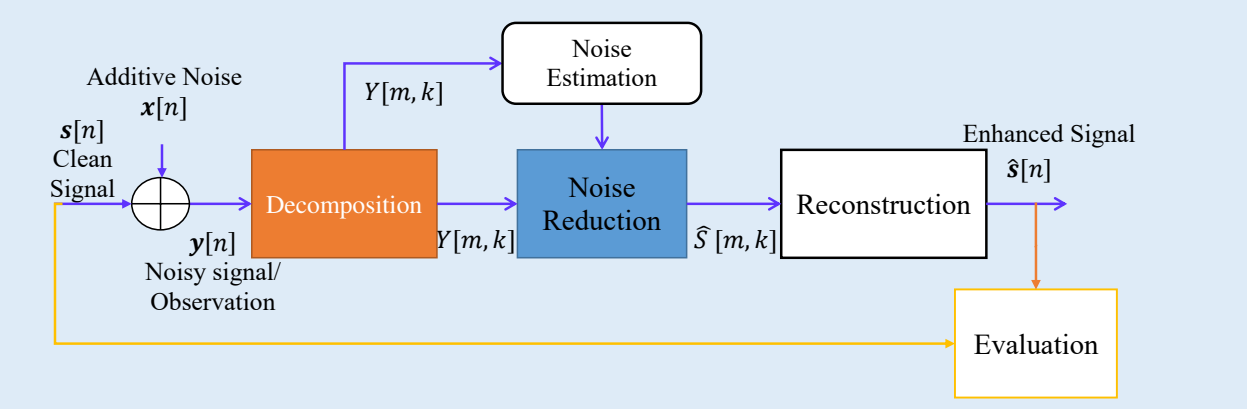
| | Addition | Multiplication | Division | Square root |
|---|---|---|---|---|
| Norm | 1 | 2 | 0 | 1 |
| Sorting | $\log D$ | 0 | 0 | 0 |
| Search $n^*$(worst case) | $D(D-1)/2$ | $D$ | $D$ | 0 |
| Total | $1 + \log D + D(D-1)/2$ | $D+2$ | $D$ | 1 |

noise estimate, and that it outperforms other methods for speech denoising in the presence of various noise types and levels. For its good performance and low complexity, the B-E-DATE should be preferred in practice when frequency rates are high enough to induce acceptable or even negligible time-delay.

# Part III

# Speech: Improving you

In this part, we propose two approaches for estimating speech short-time spectral amplitude (STSA). The main objective of this part is to take into account the recent result in parametric and non-parametric statistical theory to improve the performance of speech enhancement system. Chapter 4 takes into consideration the joint estimation and detection theory based on the parametric approach. Chapter 5 further improves speech quality by resorting to a semi-parametric approach.



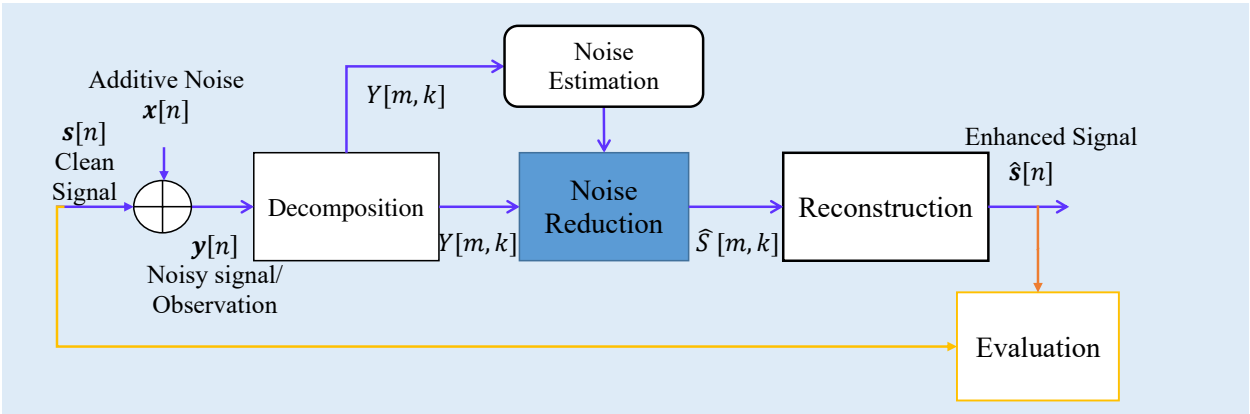
Chapter 4

# Spectral amplitude estimator based on joint detection and estimation

*It does not matter how slowly you go as long as you do not stop.*

Confucius





## 4.1 Introduction

Optimal Bayesian estimator algorithms aimed to remove or to reduce background noise are frequently used in speech enhancement. By assuming a statistical distribution for the signal of interest and the observation in the STFT domain, the estimator of the short-time spectral amplitude (STSA) is obtained by minimizing the statistical expectation of a cost function that measures the difference between the true amplitude and its estimate. These optimal estimators perform better than most unsupervised methods including the spectral-subtractive algorithms, the Wiener filtering and subspace approach [1].

The first original optimal Bayesian STSA estimator was proposed in [27], where the cost function is the square error between the clean signal and its STSA estimate. A general STSA estimator was developed in [112], where the cost function of this method is defined by the square error of the $\beta$ power amplitude. Based on the properties of auditory systems, a number of STSA Bayesian estimators are also derived by defining the cost function as the perceptual distortion metric [28, 35]. Taking advantage of the $\beta$-power and the auditory approaches, a weighted estimator is proposed in [113]. Similarly, instead of the Gaussian assumption as in the above methods, some Bayesian estimators are calculated or approximated by supposing the super-Gaussian or generalized Gamma distribution for the STSA [34, 114, 115].

Nevertheless, these algorithms implicitly suppose that speech is present in all time-frequency bins, which may degrade their performance. Hence, some studies take into account speech presence uncertainty to estimate STSA for improving speech quality [27, 116, 117]. In those approaches, the gain function is simply multiplied by the speech presence probability, which provides much more attenuation. The speech presence probability is calculated by using the *a priori* probability of speech presence, which is assumed to be fixed or to vary with time and frequency [37, 118]. An optimal approach applied to log-short-time spectral amplitude (LSA) is also proposed in [37] but this method does not yield better performance than the original LSA estimator [119]. In addition, most algorithms do not improve speech intelligibility [120].

Recently, some researches try to combine detection and estimation as in the binary masking approach where spectral amplitudes in some time-frequency bins are retained, whereas the other amplitudes are discarded for improving performance [121]. The gain function of these methods is defined as a generalized binary mask function, which enables to recover speech intelligibility [122]. This is the reason why we decided to pursue these approaches.

In this respect, the purpose of this chapter is to follow a Bayesian approach aimed at jointly optimizing detection and estimation of speech signals so as to improve speech intelligibility. To the best of our knowledge, this approach is the first attempt of that kind in speech processing. To this end, we focus on the spectral amplitude estimator based on joint detection and estimation theory. By defining the cost function on the spectral amplitude error, our strategy tries to determine a gain function in the form of a generalized binary masking. Furthermore, two binary hypothesis state-models are used to figure out the discontinuous gain function. First, the well-known strict binary speech and absence hypotheses are considered. In this model, we assume that the observed presence signal contains noise and speech signal in some given time-frequency bins, whereas in other time-frequency bins, the observation is noise-only. The presence of speech is detected by constraining the false alarm probability as in the Neyman-Pearson approach. Second, we assume that speech is always present with variable energies. Specifically, we assume that, under the null hypothesis, the observed signal is composed of noise and negligible speech while, in the alternative hypothesis, the observed signal is the sum of noise and speech of actual interest. As in the first model, the detector is determined by the Neyman-Pearson strategy. The main difference between the two models is that the former provides no estimated amplitude

under the null hypothesis (*i.e* speech is absent) whereas the later introduces a rough estimate even under the null hypothesis (*i.e* some speech of little interest is present).

The remainder of this chapter is organized as follows. Section 4.2 presents notation and assumptions about noise and the signal of interest. In Section 4.3, the combination strategy of detection and estimation for speech enhancement is presented in the strict speech presence and absence model. Based on this, we derive the generalized binary STSA combined estimator by defining different cost functions under each hypothesis. Similarly, Section 4.4 introduces the uncertainty of speech presence/absence and the derivation of the discontinuous STSA estimators are also proposed. Then, in Section 4.5, experimental results conducted on both synthetic and real-word noise emphasizes the gain brought by our methods. Finally, Section 4.6 concludes this chapter.

## 4.2 Signal model in the DFT domain

As mentioned in the second chapter, one most important problem in speech enhancement applications is to estimate the clean speech from noisy speech $\mathbf{y}[n] = \mathbf{s}[n] + \mathbf{x}[n]$, where $\mathbf{s}[n]$ and $\mathbf{x}[n]$ are respectively the clean signal and independent noise in the time domain. The observed signal is frequently segmented, windowed and transformed by a computational harmonic transform as the short-time Fourier, wavelet or discrete cosine transforms. As most methods in the literature, this chapter considers the STFT.

The corrupted speech in the time-frequency domain is denoted by $Y[m,k] = S[m,k] + X[m,k]$, where $m$ and $k$ denote the time frame and frequency-bin indices, respectively and $S[m,k]$ and $X[m,k]$ also denote the STFT coefficients of the clean speech signal and noise, correspondingly. These STFT coefficients are assumed to have complex Gaussian distributions with zero-mean and to be uncorrelated [27]. For convenience, the $m$ and $k$ indices will be omitted in the sequel unless for clarification, and estimates are pointed by a wide hat symbol: *e.g.* $\widehat{\psi}$ is an estimate of $\psi$. The complex noisy coefficients in polar form are also given as $A_Y e^{\Phi_Y} = A_S e^{\Phi_S} + A_X e^{\Phi_X}$, where $\{A_Y,\, A_S,\, A_X\}$ and $\{\Phi_Y,\, \Phi_S,\, \Phi_X\}$ are the amplitudes and phases of the observed signal, clean speech and noise respectively. Clean speech and noise are furthermore supposed to be independent and centered so that $\mathbf{E}(A_Y^2) = \mathbf{E}(A_S^2) + \mathbf{E}(A_X^2) = \sigma_S^2 + \sigma_X^2$, with $\mathbf{E}(A_S^2) = \sigma_S^2$, $\mathbf{E}(A_X^2) = \sigma_X^2$, where $\mathbf{E}$ is the expectation. The *a priori* signal-to-noise ratio (SNR) $\xi$ and the *a posteriori* SNR $\gamma$ are defined as follows $\xi = \sigma_S^2/\sigma_X^2$, $\gamma = A_Y^2/\sigma_X^2$. For the sake of simplicity, we then denote also $A$ for the clean speech amplitude $A_S$.

Concerning the two-state model, the true hypothesis $H$ is valued in $\{H_0, H_1\}$. The decision $\mathbf{D}$ takes its value in $\{0,1\}$ and thus returns the index of the so-called accepted hypothesis. For simplicity sake, let $\mathbb{P}_{H_j}(\mathbf{D} = i)$ denote the probability that $\mathbf{D} = i$ under the true hypothesis $H_j$ and $\mathbb{P}(\mathbf{D} = i | Y = y)$ denote the probability that $\mathbf{D} = i$ given $Y = y$, where $i, j \in \{1, 0\}$.

Generally, for determining the decision rule $\mathbf{D}$, the Neyman-Pearson test maximizes the detection probability $\mathbb{P}_D(\mathbf{D}) = \mathbb{P}_{H_1}(\mathbf{D} = 1)$ or minimizes the miss probability $\mathbb{P}_M(\mathbf{D}) = \mathbb{P}_{H_1}(\mathbf{D} = 0)$ subject to $\mathbb{P}_F(D) = \mathbb{P}_{H_0}(\mathbf{D} = 1) \leq \alpha$, where $\alpha$ is the so-called *level* of the test [123]. On the other hand, for estimating the signal of interest, Bayesian estimators minimize Bayes risks $\mathbf{R}$ that are constructed via a cost function $c(\widehat{A}, A)$, where $A$ is the clean signal amplitude and $\widehat{A}$ is its estimate [1, p.241]. Usually, $\mathbf{R}(\widehat{A}) = \mathbf{E}[c(\widehat{A}, A)]$. The two optimization problems are summarized as:

$$\begin{aligned} \textbf{Detection:} \quad & \min_{\mathbf{D}} \mathbb{P}_M(\mathbf{D}) \text{ subject to } \mathbb{P}_F(\mathbf{D}) \leq \alpha \\ \textbf{Estimation:} \quad & \min_{\widehat{A}} \mathbf{E}\left[c(\widehat{A}, A)\right] \end{aligned} \tag{4.1}$$

In the sequel, we focus our attention on decisions $\mathbf{D}$ for which exists some test $\delta$ such that $\mathbf{D} = \delta(Y)$. We recall that a test $\delta$ is a function defined on $\mathbb{C}$ and valued in $\{0, 1\}$. By taking into account two well-known approaches, we now present several joint detection/estimation of speech.

## 4.3 Strict presence/absence estimators

In the certain two-state modeled by binary hypothesis, the noisy speech signal is given by

$$\begin{array}{lll} H_0: & \text{speech is absent:} & Y = X \\ H_1: & \text{speech is present:} & Y = S + X, \end{array} \tag{4.2}$$

where $H_0$ and $H_1$ are the null and alternative hypotheses denoting speech presence and speech absence in the given time-frequency bin, respectively. Using the assumptions of the above section, we have the probability density function of $Y$ under each hypothesis $H_i$, $i \in \{0, 1\}$ as follows

$$f_Y(y; H_0) = \frac{1}{\pi\sigma_X^2}\exp\left(-\frac{|y|^2}{\sigma_X^2}\right), \tag{4.3}$$

$$f_Y(y; H_1) = \frac{1}{\pi\sigma_X^2(1+\xi)}\exp\left(-\frac{|y|^2}{\sigma_X^2(1+\xi)}\right), \tag{4.4}$$

for any complex value $y$. In order to improve performance of optimal Bayesian estimators, a detector is applied to each time-frequency bin for detecting the presence of speech. Then, an estimator allows us to retrieve the signal of interest. Furthermore, the estimator and the detector are obtained by defining a Bayesian/Neyman-Pearson based risk.

Following [67], in order to combine detection and estimation, we aim to find the couple $(\widehat{A}, \delta)$, where $\widehat{A}$ is the estimate of $A$ and $\delta$ denotes a test valued in $\{0, 1\}$. The decision made by the test when the observation is $Y$ is thus $\mathbf{D} = \delta(Y)$ and is the index of the accepted hypothesis. Basically, when the decision is 0, the absence of speech is accepted and thus, the estimate of $A$ must be $\widehat{A} = 0$ and the cost is then $c(A) := c(0, A)$. Otherwise, the presence of speech is accepted and the estimation cost must then be $c(\widehat{A}, A)$. Therefore, given the observation $Y$, the estimated cost is defined by

$$C(\widehat{A}, A) = c(\widehat{A}, A)\delta(Y) \ + c(A)(1 - \delta(Y)). \tag{4.5}$$

Thus, the average Bayes risk $\mathbf{R}$ under $H_1$ is defined by:

$$\mathbf{R}\left(\widehat{A}, \mathbf{D}\right) = \mathbf{E}_1\left[C(\widehat{A}, A)\right], \tag{4.6}$$

where $\mathbf{E}_1$ stands for the expectation under $H_1$ with respect to $Y$ and $A$.

By taking the constraint on the Neyman-Pearson detector and the generalized cost of the Bayesian estimator, the joint detection and estimation problem becomes the following constrained minimization problem:

$$\begin{array}{c} \min\limits_{\widehat{A}, \mathbf{D}} \mathbf{R}\left(\widehat{A}, \mathbf{D}\right) \\ \text{subject to: } \mathbb{P}_{H_0}(\mathbf{D} = 1) \leq \alpha. \end{array} \tag{4.7}$$

This problem is investigated and solved in [67, Theorem 1] for randomized tests. The proof of this result can be simplified for non-randomized tests, which are sufficient for application to

speech. In fact, the result follows by minimizing the Lagrange multipliers:

$$\begin{aligned} L(\widehat{A}, \mathbf{D}) &= \mathbf{R}\left(\widehat{A}, \mathbf{D}\right) + \tau\left(\mathbb{P}_{H_0}(\mathbf{D}=1) - \alpha\right) \\ &= \tau(1-\alpha) + \left(\mathbf{E}_1\left[C(\widehat{A}, A)\right] - \tau\mathbb{P}_{H_0}(\mathbf{D}=0)\right). \end{aligned} \tag{4.8}$$

Therefore, the problem amounts to minimizing the second term in the right hand side (rhs) term of the second equality in Equation (4.8). This term is henceforth named $L_1(\widehat{A}, \mathbf{D})$. Let us compute it. We have first

$$\mathbf{E}_1\left[C(\widehat{A}, A)\right] = \int C(\psi(y), a) f_{A,Y}(a, y; H_1) da dy, \tag{4.9}$$

where $\psi$ is a map of $\mathbb{C}$ into $[0, \infty)$ and given an observation $Y$, the estimate $\widehat{A}$ of $A$ provided by this map is:

$$\widehat{A} = \psi(Y). \tag{4.10}$$

Using the cost function defined by Equation (4.5) and Bayes's theorem, we obtain

$$\begin{aligned} \mathbf{E}_1\left[C(\widehat{A}, A)\right] &= \int \left[c(\psi(y), a)\delta(y) + c(a)(1-\delta(y))\right] f_{A|Y=y}(a) f_Y(y; H_1) da dy \\ &= \int \left[\delta(y) \int c(\psi(y), a) f_{A|Y=y}(a) da + (1-\delta(y)) \int c(a) f_{A|Y=y}(a) da\right] f_Y(y; H_1) dy. \end{aligned} \tag{4.11}$$

Moreover, let us recall that the conditional expectation $\mathbf{E}\left[g(X, Y)|Y=y\right]$ of a measurable function $g(X, Y)$ is given by:

$$\mathbf{E}\left[g(Y, X)|Y=y\right] \doteq \int g(x, y) f_{X|Y=y}(x) dx, \tag{4.12}$$

so that, the Bayesian risk is :

$$\mathbf{E}_1\left[C(\widehat{A}, A)\right] = \int \left[\mathbf{E}\left[c(\psi(Y), A)|Y=y\right]\delta(y) + \mathbf{E}\left[c(A)|Y=y\right](1-\delta(y))\right] f_Y(y; H_1) dy. \tag{4.13}$$

For the sake of simplicity, we denote $\mathbf{E}\left[c(\psi(Y), A)|Y=y\right]$ and $\mathbf{E}\left[c(A)|Y=y\right]$ by $r(y; \widehat{A})$ and $r(y)$, respectively. Therefore, $L_1(\widehat{A}, \mathbf{D})$ rewrites:

$$\begin{aligned} L_1(\widehat{A}, \mathbf{D}) &= \int \left[r(y; \widehat{A})\delta(y) + r(y)(1-\delta(y))\right] f_Y(y; H_1) dy + \tau \int (1-\delta(y)) f_Y(y; H_0) dy \\ &= \int \left[r(y; \widehat{A}) f_Y(y; H_1)\delta(y) + (r(y) f_Y(y; H_1) - \tau f_Y(y; H_0))(1-\delta(y))\right] dy \end{aligned} \tag{4.14}$$

Since we consider non-randomized tests, $\delta$ is completely specified by its critical region $\mathcal{A}$ so that:

$$\delta = \mathbb{1}_{\mathcal{A}} \quad \text{and} \quad \delta(Y) = \mathbb{1}_{\mathcal{A}}(Y) \tag{4.15}$$

where $\mathbb{1}_{\mathcal{A}}$ is the indicator function of $\mathcal{A}$. It follows that:

$$\begin{aligned} L_1(\widehat{A}, \mathbf{D}) = &\int_{\mathcal{A}} \left[r(y; \widehat{A}) f_Y(y; H_1) - (r(y) f_Y(y; H_1) - \tau f_Y(y; H_0))\right] dy \\ &+ \int \left[r(y) f_Y(y; H_1) - \tau f_Y(y; H_0)\right] dy. \end{aligned} \tag{4.16}$$

The second term in the right-hand side term in Equation (4.16) depends neither on $\widehat{A}$ nor on the decision $\mathbf{D}$. Therefore, minimizing $L_1(\widehat{A}, \mathbf{D})$ with respect to $\widehat{A}$ and $\mathbf{D}$ amounts to minimizing the first integral to the right-hand side of Equation (4.16). Using Lemma 1 (see Appendix A), the optimal critical region that minimizes $L_1(\widehat{A}, \mathbf{D})$ is:

$$\mathcal{A} = \left[\Big(r(y;\widehat{A}) f_Y(y;H_1) - r(y) f_Y(y;H_1) + \tau f_Y(y;H_0)\Big) < 0\right], \tag{4.17}$$

where $[f < 0] = \{x \in \mathbb{C} : f(x) < 0\}$. Furthermore, over this set, we must also minimize $r(y;\widehat{A})$ with respect to $\widehat{A}$, which is the standard Bayesian estimator. Thus, the obtained result is simply given by

$$\begin{cases} \widehat{A} = \psi(Y) \text{ where } \psi(y) = \underset{a}{\operatorname{argmin}}\ r(y;a) \\ \text{Test } H_0 \text{ vs. } H_1 : \dfrac{f_Y(y;H_1)}{f_Y(y;H_0)} \left[r(y) - r(y;\widehat{A})\right] \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau, \\ \text{if } \mathbf{D} = 0, \text{ force } \widehat{A} = 0 \end{cases} \tag{4.18}$$

In addition, $\tau$ is calculated by imposing $\mathbb{P}_{H_0}(\mathbf{D} = 1) = \alpha$. Moreover, using Equations (4.3) and (4.4), the likelihood ratio is given by:

$$\frac{f_Y(y;H_1)}{f_Y(y;H_0)} = \frac{\exp\left(\dfrac{\gamma\xi}{1+\xi}\right)}{1+\xi} = \frac{\exp(\nu)}{1+\xi}, \tag{4.19}$$

where

$$\nu = \frac{\gamma\xi}{1+\xi}$$

This approach is similar to ideal binary masking [122]. When the decision is that noise only is present, the amplitude is set to 0. The difference is that, when the decision is that speech is present, the binary masking keeps simply the noisy amplitude, whereas joint detection/estimation provides a Bayesian estimator. Additionally, in speech enhancement, the square error between the clean STSA (or the clean LSA) and its estimate is widely used as the cost function. Therefore, we propose the following detectors derived from the the STSA and LSA cost functions.

### 4.3.1 Strict joint STSA estimator

The STSA cost function is:

$$c(A) = A^2 \tag{4.20}$$

$$c(\widehat{A}, A) = \left(\widehat{A} - A\right)^2 \tag{4.21}$$

Under hypothesis $H_1$, the Bayesian estimator of the speech STSA when deciding $\mathbf{D} = 1$ is a map $\psi^{\mathrm{SM}}_{\mathrm{STSA}}$ [1] of $\mathbb{C}$ into $[0, \infty)$ aimed at minimizing $r(y;\widehat{A})$. It is known to be derived from the conditional mean and is given for every $y \in \mathbb{C}$ by [124]:

$$\psi^{\mathrm{SM}}_{\mathrm{STSA}}(y) = \int_0^\infty a f_{A|Y=y}(a;H_1)da = \frac{\displaystyle\int_0^\infty a f_{Y|A=a}(y;H_1) f_A(a)da}{\displaystyle\int_0^\infty f_{Y|A=a}(y;,H_1) f_A(a)da}. \tag{4.22}$$

[1] SM: Strict Model

Thus, given the DFT coefficient $Y$, the estimate $\widehat{A}$ of $A$ provided by this estimator is:

$$\widehat{A} = \psi^{\mathrm{SM}}_{\mathrm{STSA}}(Y). \tag{4.23}$$

In order to determine $\psi^{\mathrm{SM}}_{\mathrm{STSA}}$ via Equation (4.22), the DFT coefficients of the clean speech and noise are assumed to be statistically independent and to have complex centered Gaussian distributions. According to this assumption, the probability density function (pdf) of the STSA $A$ and the phase $\Phi_S$ are Rayleigh and uniform in the range $(0, 2\pi)$ [125, Example 6-22, p. 202], respectively. We have:

$$f_{Y|A=a}(y; H_1) f_A(a) = \int_0^{2\pi} f_{Y|A=a,\Phi_S=\phi}(y; H_1) f_{A,\Phi_S}(a, \phi) d\phi, \tag{4.24}$$

where

$$f_{A,\phi_S}(a, \phi) = \frac{a}{\pi\sigma_S^2} \exp\left\{-\frac{a^2}{\sigma_S^2}\right\}, \tag{4.25}$$

and

$$f_{Y|A=a,\Phi_S=\phi}(y; H_1) = \frac{1}{\pi\sigma_X^2} \exp\left\{-\frac{|y - a\exp(-i\phi)|^2}{\sigma_X^2}\right\}. \tag{4.26}$$

The map $\psi^{\mathrm{SM}}_{\mathrm{STSA}}$ is given by [27]:

$$\psi^{\mathrm{SM}}_{\mathrm{STSA}}(y) = G_{\mathrm{STSA}}(\xi, \gamma)|y|, \tag{4.27}$$

where

$$G_{\mathrm{STSA}}(\xi, \gamma) = \frac{\sqrt{\pi\nu}}{2\gamma} \exp\left(\frac{-\nu}{2}\right) \left[(1+\nu) I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right], \tag{4.28}$$

where $I_0(\,\cdot\,)$ and $I_1(\,\cdot\,)$ are the modified Bessel functions of zero and first order, respectively. This gain is a function of two variables: the *a priori* SNR $\xi$ and the *a posteriori* SNR $\gamma$. As mentioned in [27], for high *a posteriori* SNR, this gain function is close to the Wiener gain function. The *a posteriori* SNR is directly given by the observed amplitude $A_Y$. In contrast, the *a priori* SNR is unknown. This variable $\xi$ can be estimated via the decision directed approach [27]:

$$\xi[m, k] = \beta \frac{\widehat{A}^2[m-1, k]}{\sigma_X^2[m-1, k]} + (1-\beta) \max\left((\gamma[m, k] - 1), 0\right), \tag{4.29}$$

where $0 < \beta < 1$ is the smoothing parameter and $\widehat{A}[m-1, k]$ is the estimated STSA at the previous frame. Thus, the STSA estimate under hypothesis $H_1$ is obtained as :

$$\widehat{A} = G_{\mathrm{STSA}}(\xi, \gamma) A_Y. \tag{4.30}$$

The joint detector is determined via two risks $r(y)$ and $r(\widehat{A}, y)$. They are respectively the miss detection risk and the standard Bayesian risk under $H_1$. In this case, the miss detection risk is calculated by:

$$r(y) = \int_0^{\infty} c(a) f_{A|Y=y}(a; H_1) da = \int_0^{\infty} a^2 f_{A|Y=y}(a; H_1) da. \tag{4.31}$$

Similarly, the cost for the Bayesian estimating error is also provided by:

$$r(y; \widehat{A}) = \int_0^{\infty} c(\psi(y), a) f_{A|Y=y}(a; H_1) da = \int_0^{\infty} (\psi^{\mathrm{SM}}_{\mathrm{STSA}}(y) - a)^2 f_{A|Y=y}(a; H_1) da. \tag{4.32}$$

Expanding the square in the rhs of Equation (4.32), the Bayesian risk can be written as the function of the miss detection risk by using Equation (4.22):

$$r(y;\widehat{A}) = \int_0^\infty \left( \left(\psi_{\text{STSA}}^{\text{SM}}(y)\right)^2 - 2a\psi_{\text{STSA}}^{\text{SM}}(y) + a^2 \right) f_{A|Y=y}(a;H_1)da = r(y) - \left(\psi_{\text{STSA}}^{\text{SM}}(y)\right)^2. \quad (4.33)$$

So, in the strict present/absent model considered in this section for STSA estimation, the decision in Equation (4.18) is

$$\mathcal{D}_{\text{STSA}}^{\text{SM}}(y) \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau, \quad (4.34)$$

where:

$$\mathcal{D}_{\text{STSA}}^{\text{SM}}(y) = \frac{\exp(\nu)}{1+\xi}\left(r(y) - r(y;\widehat{A})\right) = \frac{\exp(\nu)}{1+\xi}\left(\psi_{\text{STSA}}^{\text{SM}}(y)\right)^2. \quad (4.35)$$

In short, for each time-frequency bin, the proposed joint method estimates first the speech STSA by using the Bayesian estimator, then the detector is based on this estimate to detect the presence or absence of speech at each bin. If speech is absent, this method sets the speech STSA to 0. Focusing only on the estimator, the STSA estimate can be written as a binary masking:

$$\widehat{A} = G_{\text{STSA}}^{\text{SM}}(\xi,\gamma)A_Y, \quad (4.36)$$

where the gain function $G_{\text{STSA}}^{\text{SM}}(\xi,\gamma)$ is:

$$G_{\text{STSA}}^{\text{SM}}(\xi,\gamma) = \begin{cases} G_{\text{STSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\text{STSA}}^{\text{SM}}(y) \geq \tau_{\text{STSA}}^{\text{SM}} \\ 0 & \text{otherwise,} \end{cases} \quad (4.37)$$

where the threshold $\tau_{\text{STSA}}^{\text{SM}}$ is determined by seeking a solution to $\mathbb{P}_{H_0}(\mathbf{D}=1) = \alpha$ (see Appendix B).

### 4.3.2 Strict joint LSA estimator

#### 4.3.2.1 Optimal joint LSA estimator

We now consider that the cost function is the square of the error between the clean LSA and its estimate:

$$c(A) = (\log(A) - \log(\varepsilon))^2 \quad (4.38)$$

$$c(\widehat{A}, A) = \left(\log(\widehat{A}) - \log(A)\right)^2 \quad (4.39)$$

where $\varepsilon$ satisfies $(0 < \varepsilon \leq A)$ and is a fixed constant that enables us to ensure a monotonic cost function under hypothesis $H_0$. Therefore, similarly to above, the Bayesian estimator under hypothesis $H_1$ of the speech LSA is also a map $\psi_{\text{LSA}}^{\text{OSM}}$ [2] of $\mathbb{C}$ into $(-\infty,\infty)$:

$$\psi_{\text{LSA}}^{\text{OSM}}(y) = \int_0^\infty \log(a) f_{A|Y}(a|y,H_1)da = \frac{\int_0^\infty \log(a) f_{Y|A}(y|a,H_1) f_A(a)da}{\int_0^\infty f_{Y|A}(y|a,H_1) f_A(a)da}. \quad (4.40)$$

Using the moment-generating function of $A$, this estimator is calculated in [28]:

$$\widehat{A} = \exp\left(\psi_{\text{LSA}}^{\text{OSM}}(Y)\right) = G_{\text{LSA}}(\xi,\gamma)A_Y, \quad (4.41)$$

[2] OSM means Optimal estimator in the "Strict Model".

where $G_{\mathrm{LSA}}(\xi,\gamma)$ is the LSA gain function given by :

$$G_{\mathrm{LSA}}(\xi,\gamma)=\frac{\xi}{1+\xi}\exp\left\{\frac{1}{2}\int_{\nu}^{\infty}\frac{e^{-t}}{t}dt\right\}. \tag{4.42}$$

Note that the *a priori* SNR $\xi$ is estimated by decision-directed approach again. The integral in Equation (4.42) can be numerically approximated.

For determining the detector, the two Bayesian risks $r(y)$, $r(\widehat{A},y)$ need to be evaluated. The value of $r(y)$ is calculated by

$$r(y)=\int_{0}^{\infty}(\log(a)-\log(\epsilon))^2 f_{A|Y=y}(a;H_1)da \tag{4.43}$$

and then, the cost value with the optimal estimate $\widehat{A}$ is

$$r(y;\widehat{A})=\int_{0}^{\infty}\left(\log(a)-\psi_{\mathrm{LSA}}^{\mathrm{OSM}}(y)\right)^2 f_{A|Y=y}(a;H_1)da. \tag{4.44}$$

Based on Equation(4.43), Equation (4.44) simplifies to

$$r(y;\widehat{A})=r(y)-\left(\psi_{\mathrm{LSA}}^{\mathrm{OSM}}(y)-\log(\varepsilon)\right)^2. \tag{4.45}$$

Thus, with the results of Eqs. (4.41) and (4.45), the joint optimal LSA estimator under strict speech absence/presence model (OSM) turns into a binary masking function as well:

$$\widehat{A}=G_{\mathrm{LSA}}^{\mathrm{OSM}}(\xi,\gamma)A_Y, \tag{4.46}$$

where the spectral gain function $G_{\mathrm{LSA}}^{\mathrm{OSM}}(\xi,\gamma)$ is

$$G_{\mathrm{LSA}}^{\mathrm{OSM}}(\xi,\gamma)=\begin{cases} G_{\mathrm{LSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\mathrm{LSA}}^{\mathrm{OSM}}(Y)\geq\tau_{\mathrm{LSA}}^{\mathrm{OSM}},\\ 0 & \text{otherwise.}\end{cases} \tag{4.47}$$

where:

$$\mathcal{D}_{\mathrm{LSA}}^{\mathrm{OSM}}(y)=\frac{\exp(\nu)}{1+\xi}\left(\psi_{\mathrm{LSA}}(y)-\log(\varepsilon)\right)^2. \tag{4.48}$$

Note that $\widehat{A}$ is a function of three parameters: the *a priori* SNR $\xi$, the *a posteriori* SNR $\gamma$ and the spectral amplitude $A_Y$. The calculation of the threshold $\tau_{\mathrm{LSA}}^{\mathrm{OSM}}$ is presented in Appendix B.

#### 4.3.2.2 Sub-optimal joint LSA estimator

In the above subsection, because of the logarithmic-based non-decreasing cost function, we had to introduce a fixed constant $\varepsilon$, which cannot be chosen theoretically since $A$ is unknown and which can be fixed in practice after some preliminary experiments. For eliminating this undesired constant $\varepsilon$ and taking advantage from the performance of the LSA approach, an alternative cost function can be defined as:

$$c(A)=(\log(A+1))^2 \tag{4.49}$$

$$c(\widehat{A},A)=\left(\log(\widehat{A}+1)-\log(A+1)\right)^2 \tag{4.50}$$

The choice of $c(A)$ defined by (4.49) is suitable for penalizing the decision in terms of LSA. This cost function is monotonically increasing and equals zero when the true amplitude is zero. In the

same way, the choice of $c(\widehat{A}, A)$ under hypothesis $H_1$ is adapted to the same kind of constraints: this cost function increases with $\widehat{A}$ and equal to 0 when $\widehat{A} = A$.

Following the same process as in the above subsection, the corresponding Bayesian estimator under hypothesis $H_1$ is given by a map $\psi_{\text{LSA}}^{\text{SSM}}$ [3] of $\mathbb{C}$ into $[0, \infty)$:

$$\psi_{\text{LSA}}^{\text{SSM}}(y) = \int_0^{\infty} \log(a+1) f_{A|Y=y}(a; H_1) da. \tag{4.51}$$

Thus, the STSA estimated is obtained as:

$$\widehat{A} = \exp\left(\psi_{\text{LSA}}^{\text{SSM}}(Y)\right) - 1. \tag{4.52}$$

Then, the Bayesian risk for the miss detection $r(y)$ is written as

$$r(y) = \int_0^{\infty} \left(\log(a+1)\right)^2 f_{A|Y=y}(a; H_1) da. \tag{4.53}$$

and so that the standard Bayesian risk under hypothesis $H_1$, $r(\widehat{A}, y)$ is

$$r(y; \widehat{A}) = r(y) - \left(\psi_{\text{LSA}}^{\text{SSM}}(y)\right)^2. \tag{4.54}$$

Even if we use the moment-generating function of $A+1$, the integral in Equation(4.51) is hardly tractable. In addition, the estimator of Equation (4.51) is similar to that of Equation (4.40). The latter will thus be used to approximate the former. We thus propose the sub-optimal spectral gain function in the strict presence/absence model (SSM) as follows:

$$G_{\text{LSA}}^{\text{SSM}}(\xi, \gamma) = \begin{cases} G_{\text{LSA}}(\xi, \gamma) & \text{if } \mathcal{D}_{\text{LSA}}^{\text{SSM}}(y) \geq \tau_{\text{LSA}}^{\text{SSM}}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.55}$$

where $\mathcal{D}_{\text{LSA}}^{\text{SSM}}(y)$ is given by:

$$\mathcal{D}_{\text{LSA}}^{\text{SSM}}(y) = \frac{\exp(\nu)}{1+\xi} \left\{\log\left[\exp\left(\psi_{\text{LSA}}^{\text{OSM}}(y)\right) + 1\right]\right\}^2 \tag{4.56}$$

and the calculation of the threshold $\tau_{\text{LSA}}^{\text{SSM}}$ is detailed in Appendix B.

The detectors $\mathcal{D}_{\text{LSA}}^{\text{OSM}}$ (Equation (4.48)) and $\mathcal{D}_{\text{LSA}}^{\text{SSM}}$ (Equation (4.56)) are slightly different. Both are monotonic increasing and depend on the LSA estimator. In turn, the OSM-LSA and SSM-LSA estimators depend on the detectors. This twofold dependency is expected to improve the performance of the two detectors and estimators. However, in contrast to the optimal estimator ($\mathcal{D}_{\text{LSA}}^{\text{OSM}}$) , the sub-optimal ($\mathcal{D}_{\text{LSA}}^{\text{SSM}}$) does not introduce any auxiliary parameter $\varepsilon$, which should be beneficial.

## 4.4 Uncertain presence/absence estimators

The proposed above methods based on strict presence/absence hypotheses may introduce musical noise since these estimators can randomly generate some isolated peaks in the time frequency domain. Thus, under $H_0$, it should be proposed an estimator that allows us to reduce the impact of miss detection error, since such error may introduce musical noise [1, pp.132]. Normally, under $H_0$, this estimate should be $\widehat{A}_0 = \sqrt{\beta} A_X$ where $\beta$ $(0 < \beta \ll 1)$ is a constant spectral floor

[3] SSM means "Sub-optimal" estimator in the "Strict Model"

parameter [9], which is empirically chosen. In favor of this suggestion, as in [126], we now assume that, under hypothesis $H_0$, the signal of little interest $S_0$ is present but with small amplitude. Under the alternative hypothesis $H_1$, the noisy signal remains the sum of the signal of actual interest $S_1$ and noise. Therefore, with these hypotheses, the two-state model is

$$\begin{aligned} H_0 &: Y = S_0 + X, \\ H_1 &: Y = S_1 + X, \end{aligned} \tag{4.57}$$

where $S_0$ is key to distinguish between the two models summarized by Equations (4.2) and (4.57). Furthermore, supposing that $S_0 = \sqrt{\beta}X$, we similarly get the conditional pdf of the observed signal, so that :

$$f_Y(y; H_0) = \frac{1}{\pi\sigma_X^2(1+\beta)} \exp\left(-\frac{|y|^2}{\sigma_X^2(1+\beta)}\right) \tag{4.58}$$

$$f_Y(y; H_1) = \frac{1}{\pi\sigma_X^2(1+\xi)} \exp\left(-\frac{|y|^2}{\sigma_X^2(1+\xi)}\right) \tag{4.59}$$

The main difference between the conditional pdfs above is that, under hypothesis $H_0$, the *a priori* SNR $\beta$ is identical for all frequency bins since $\beta$ is fixed once for all, whereas, under hypothesis $H_1$, the *a priori* SNR $\xi = \xi[m,k]$ varies in time and frequency.

The true signal $S$ is either $S_0$ or $S_1$, depending on the true hypothesis. Set $A_i = |S_i|$ for $i \in \{0,1\}$ and denote the clean speech amplitude by $A$. Under hypothesis $H_i$, we have $A = A_i$. Let $\widehat{A}_j$ be the estimate of $A$ when the decision is $H_j$, that is, when $\mathbf{D} = j$. As in Section 4.3 and in Bayesian detection, we then define *a priori* cost function $c_{ji}$ for deciding $\mathbf{D} = j$ under the true hypothesis $H_i$. When deciding $\mathbf{D} = j$, the cost of providing the estimate $\widehat{A}_j$ of $A$ under the true hypothesis $H_i$ is thus $c_{ji}(\widehat{A}_j, A_i)$. By involving the decision made by the test $\delta$, the weighted cost function under true hypothesis $H_i$ becomes:

$$C_i(\widehat{A}_1, \hat{A}_0, A_i) = c_{1i}(\widehat{A}_1, A_i)\delta(Y) + c_{0i}(\widehat{A}_0, A_i)(1 - \delta(Y)). \tag{4.60}$$

Thus, the Bayesian risk under hypothesis $H_i$ can now be computed as :

$$\mathbf{R}_i\left(\widehat{A}_1, \hat{A}_0, \mathbf{D}\right) = \mathbf{E}_i\left[C_i(\widehat{A}_1, \widehat{A}_0, A_i)\right], \tag{4.61}$$

where $\mathbf{E}_i$ denotes the statistical expectation under hypothesis $H_i$ with respect to $Y$ and $A_i$ and where $i \in \{0,1\}$. Since a non-null estimate of the clean speech amplitude is provided when the decision is 0, which entails an estimation cost, we follow [127] by tackling the following constrained optimization problem:

$$\begin{aligned} &\min_{\widehat{A}_1, \widehat{A}_0, \mathbf{D}} \mathbf{R}_1\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right) \\ &\text{subject to } \mathbf{R}_0\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right) \leq \alpha, \end{aligned} \tag{4.62}$$

where the false alarm probability of Equation (4.7) is replaced by the cost under $H_0$. In this strategy, we control the cost of erroneously estimating the signal amplitude under $H_0$, that is, when the signal is of little interest and there is no real need to estimate it accurately. So, we can be satisfied by upper-bounding the estimation cost under $H_0$. Of course, the upper-bound must be fixed to a small value. In contrast, under $H_1$, the speech signal must be estimated as well as possible and thus, we want to minimize the estimation cost.

It is worth noticing that this optimization problem (4.62) is a general case of the problem treated in the section above. Indeed, if under $H_0$, we assume that $A = A_0 = 0$ and thus force the estimate under $H_0$ to $\widehat{A}_0 = 0$, we choose $c_{00}(\widehat{A}_0, A_0) = 0$ and the upper-bounding of the cost under $H_0$ amounts to upper-bounding the false alarm probability. Furthermore, if we focus only on the detection cost and set $c_{ii}(\widehat{A}_i, A_i) = 0$, $c_{ji}(\widehat{A}_j, A_i) = 1$ with $j \neq i$, the optimization problem of Equation 4.62 becomes the testing problem solved by the Neyman-Pearson lemma.

The problem (4.62) is considered and solved by using [127, theorem 2.1]. As above, an alternative and much simpler proof of [127] is given in Appendix B by using Lagrange multiplier and seeking estimators and a non-randomized test that solves:

$$\min_{\widehat{A}_1, \widehat{A}_0, \mathbf{D}} L(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) \tag{4.63}$$

with:

$$L(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \mathbf{R}_1(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) + \tau \left( \mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) - \alpha \right). \tag{4.64}$$

For $i = 0, 1$, we have:

$$\begin{aligned} \mathbf{R}_i(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) &= \mathbf{E}_i \left[ C_i(\widehat{A}_1, \widehat{A}_0, A_i) \right] \\ &= \int \mathbf{E}_i \left[ C_i(\widehat{A}_1, \widehat{A}_0, A_i) | Y = y \right] f_Y(y; H_i) dy, \end{aligned} \tag{4.65}$$

by definition of a conditional. Properties of a conditional now induce that:

$$\begin{aligned} \mathbf{E}_i \left[ C_i(\widehat{A}_1, \widehat{A}_0, A_i) | Y = y \right] &= \mathbf{E}_i \left[ c_{1i}(\widehat{A}_1, A_i)\delta(Y) + c_{0i}(\widehat{A}_0, A_i)(1 - \delta(Y)) | Y = y \right] \\ &= \mathbf{E}_i \left[ c_{1i}(\widehat{A}_1, A_i)\delta(Y) | Y = y \right] + \mathbf{E}_i \left[ c_{0i}(\widehat{A}_0, A_i)(1 - \delta(Y)) | Y = y \right] \\ &= \mathbf{E}_i \left[ c(\widehat{A}_1, A_i) | Y = y \right] \delta(y) + \mathbf{E}_i \left[ c(\widehat{A}_0, A_i) | Y = y \right] (1 - \delta(y)) \\ &= \mathbf{E}_i \left[ c_{1i}(\psi_1(Y), A_i) | Y = y \right] \delta(y) + \mathbf{E}_i \left[ c_{0i}(\psi_0(Y), A_i) | Y = y \right] (1 - \delta(y)) \end{aligned}$$

For any $\psi : \mathbb{C} \to [0, \infty)$, we set

$$r_{ji}(y; \psi) = \mathbf{E}_i \left[ c_{ji}(\psi(Y), A_i) | Y = y \right] \tag{4.66}$$

Therefore,

$$\mathbf{E}_i \left[ C_i(\widehat{A}_1, \widehat{A}_0, A_i) | Y = y \right] = r_{1i}(y; \psi_1)\delta(y) + r_{0i}(y; \psi_0)(1 - \delta(y)) \tag{4.67}$$

It follows that:

$$\mathbf{R}_i(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \int r_{1i}(y; \psi_1)\delta(y) f_Y(y; H_i) dy + \int r_{0i}(y; \psi_0)(1 - \delta(y)) f_Y(y; H_i) dy \tag{4.68}$$

Injecting Equation (4.68) into Equation (4.64), the optimization problem of the latter is simplified into the minimization of the function $L_1\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right)$ given by

$$\begin{aligned} L_1\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right) &= \mathbf{R}_1(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) + \tau \mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) \\ &= \int \left[ \mathbb{D}_1(y; \psi_1)\delta(y) + \mathbb{D}_0(y; \psi_0)(1 - \delta(y)) \right] dy, \end{aligned} \tag{4.69}$$

where

$$\mathbb{D}_i(y; \psi_i) = f_Y(y; H_1) r_{i1}(y; \psi_i) + \tau f_Y(y; H_0) r_{i0}(y; \psi_i), (i = 0, 1) \tag{4.70}$$

As in Section 4.3, we are looking for a test defined by Equation (4.15). Equation (4.69) rewrites

$$L_1\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right) = \int_{\mathcal{A}} \left[\mathbb{D}_1(y;\psi_1) - \mathbb{D}_0(y,\psi_0)\right] dy + \int \mathbb{D}_0(y;\psi_0) dy. \tag{4.71}$$

Using Lemma 1 again, we obtain:

$$\mathcal{A} = \left\{y \in \mathbb{C} : \left(\mathbb{D}_1(y;\psi_1) - \mathbb{D}_0(y;\psi_0)\right) \leq 0\right\}. \tag{4.72}$$

This specifies $\delta$ and $\mathbf{D}$. In particular,

$$\forall y \in \mathbb{C}, \delta(y) = \begin{cases} 1 & \text{if } y \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases} \tag{4.73}$$

It remains to calculate optimal $\psi_0$ and $\psi_1$. To this end, with our choice for $\delta$, it follows from Equation (4.69) that:

$$L_1\left(\widehat{A}_1, \widehat{A}_0, \mathbf{D}\right) = \int_{\mathcal{A}} \mathbb{D}_1(y;\psi_1) dy + \int_{\mathcal{A}^c} \mathbb{D}_0(y;\psi_0) dy, \tag{4.74}$$

where $\mathcal{A}^c$ is the complementary set of $\mathcal{A}$. Provided they exist, it suffices to choose

$$\psi_i = \underset{\psi}{\operatorname{argmin}}\, \mathbb{D}_i(y;\psi) = \underset{\psi}{\operatorname{argmin}} \left[f_Y(y;H_1) r_{i1}(y;\psi) + \tau f_Y(y;H_0) r_{i0}(y;\psi)\right] \tag{4.75}$$

Summarizing the foregoing, the estimator and detector are respectively given by

$$\widehat{A}_i = \psi_i(Y) \text{ with } \psi_i = \underset{\psi}{\operatorname{argmin}} \left[f_Y(y;H_1) r_{i1}(y;\psi) + \tau f_Y(y;H_0) r_{i0}(y;\psi)\right], \quad i = 0,1 \tag{4.76}$$

$$\frac{f_Y(y;H_1)}{f_Y(y;H_0)}\left[r_{01}(y;\psi_0) - r_{11}(y;\psi_1)\right] \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau \left[r_{10}(y;\psi_1) - r_{00}(y;\psi_0)\right], \tag{4.77}$$

where $\tau$ is determined by using the constraint with equality (see Appendix B). Moreover, the standard likelihood ratio $\Lambda(\xi,\gamma)$ is directly computed by using Equations (4.58) and (4.59)

$$\Lambda(\xi,\gamma) = \frac{f_Y(y;H_1)}{f_Y(y;H_0)} = \frac{1+\beta}{1+\xi} \exp\left(\frac{\gamma(\xi-\beta)}{(1+\beta)(1+\xi)}\right). \tag{4.78}$$

In the next sections, we consider different cost functions, we propose some approaches to derive several joint detector/estimator with the structure of Equations (4.76) and (4.77).

### 4.4.1 Uncertain joint STSA detector/estimator

#### 4.4.1.1 Independent STSA estimator

In this section, we consider the same cost function for the four different situations where $D = j$ under true hypothesis $H_i$ with $(j,i) \in \{0,1\}^2$. This cost is defined as:

$$c_{ji}(a,b) = c(a,b) = (a-b)^2 \tag{4.79}$$

It follows from Equations (4.12) and (4.66) that:

$$r_{ji}(y;\psi) = \int c(\psi(y), a_i) f_{A_i|Y=y}(a_i) da_i \tag{4.80}$$

which does not depend on $j$ any more. Given $\psi : \mathbb{C} \to [0, \infty)$ and $y \in \mathbb{C}$, set $t = \psi(y)$ and rewrite $\mathbb{D}_i(y; \psi)$ as:

$$\begin{aligned}\mathbb{D}_i(y;\psi) &= f_Y(y;H_1)\int c(t,a_1) f_{A_1|Y=y}(a_1)da_1 + \tau f_Y(y;H_0)\int c(t,a_0) f_{A_0|Y=y}(a_0)da_0 \\ &= f_Y(y;H_1)\int (t-a_1)^2 f_{A_1|Y=y}(a_1)da_1 + \tau f_Y(y;H_0)\int (t-a_0)^2 f_{A_0|Y=y}(a_0)da_0\end{aligned}$$

We have a convex function of $t$ and by derivation with respect to $t$, some routine algebra shows that the value of $t = \psi(y)$ that minimizes $\mathbb{D}_i(y; \psi)$ is [4]

$$\psi^{\text{IUM}}_{\text{STSA}}(y) = \frac{f_Y(y;H_1)G_{\text{STSA}}(\xi,\gamma) + \tau^{\text{IUM}}_{\text{STSA}} f_Y(y;H_0)G_{\text{STSA}}(\beta,\gamma)}{f_Y(y;H_1) + \tau^{\text{IUM}}_{\text{STSA}} f_Y(y;H_0)}|y|, \tag{4.81}$$

where $G_{\text{STSA}}(\xi, \gamma)$ is defined by Equation (4.28) and $\tau^{\text{IUM}}_{\text{STSA}}$ is given in Appendix B. This function can be simplified by using the likelihood ratio $\Lambda(\xi, \gamma)$:

$$\psi^{\text{IUM}}_{\text{STSA}}(y) = \frac{\Lambda(\xi,\gamma)G_{\text{STSA}}(\xi,\gamma) + \tau^{\text{IUM}}_{\text{STSA}} G_{\text{STSA}}(\beta,\gamma)}{\Lambda(\xi,\gamma) + \tau^{\text{IUM}}_{\text{STSA}}}|y|. \tag{4.82}$$

It is important to note that $\psi^{\text{IUM}}_{\text{STSA}}(y)$ of Equation (4.82) does not depend on $i$. Therefore, the estimated STSA $\widehat{A}_1$ and $\widehat{A}_0$ are given as:

$$\widehat{A}_1 = \widehat{A}_0 = \psi^{\text{IUM}}_{\text{STSA}}(Y) = G^{\text{IUM}}_{\text{STSA}}(\xi,\gamma)A_Y, \tag{4.83}$$

where the gain function $G^{\text{IUM}}_{\text{STSA}}(\xi, \gamma)$ is given by:

$$G^{\text{IUM}}_{\text{STSA}}(\xi,\gamma) = \frac{\Lambda(\xi,\gamma)G_{\text{STSA}}(\xi,\gamma) + \tau^{\text{IUM}}_{\text{STSA}} G_{\text{STSA}}(\beta,\gamma)}{\Lambda(\xi,\gamma) + \tau^{\text{IUM}}_{\text{STSA}}}. \tag{4.84}$$

Because under any hypothesis, we get the same STSA estimator, we call it as independent STSA estimator. The detector influences the estimator only via the threshold $\tau^{\text{IUM}}_{\text{STSA}}$ in the gain function $G^{\text{IUM}}_{\text{STSA}}(\xi, \gamma)$.

#### 4.4.1.2 Joint STSA estimator

For further taking into account the role of the presence and absence of speech, we consider the cost function as follows:

$$c_{ji}(\widehat{A}_j, A_i) = \begin{cases} A_i^2, & i \neq j, \\ (\widehat{A}_j - A_i)^2, & i = j, \end{cases} \tag{4.85}$$

where $i, j \in \{0, 1\}$. The cost function enables us to put more emphasis on the miss detection. Thus, the error miss detection depends only on the true amplitude instead of the difference between the true amplitude under true hypothesis and its estimate under deciding the other hypothesis. Particularly, when we make the false-alarm and miss detection, unlike Subsection 4.4.1.1, the cost functions now not only implicitly penalize the estimated error but also the detected error.

Similar to the above subsection, given $\psi : \mathbb{C} \to [0, \infty)$ and $y \in \mathbb{C}$, set $t = \psi(y)$, $\mathbb{D}_i(y; \psi)$ can be now rewritten as:

$$\mathbb{D}_1(y;\psi) = f_Y(y;H_1)\int (t-a_1)^2 f_{A_1|Y=y}(a_1)da_1 + \tau f_Y(y;H_0)\int a_0^2 f_{A_0|Y=y}(a_0)da_0 \tag{4.86}$$

$$\mathbb{D}_0(y;\psi) = f_Y(y;H_1)\int a_1^2 f_{A_1|Y=y}(a_1)da_1 + \tau f_Y(y;H_0)\int (t-a_0)^2 f_{A_0|Y=y}(a_0)da_0 \tag{4.87}$$

[4] IUM means "Independent" estimator in the "Uncertain Model"

By derivation with respect to $t$ of each $\mathbb{D}_i(y;\psi)$, the value of $t=\psi(y)$ that minimizes $\mathbb{D}_i(y;\psi)$ defines the function $\psi_{\text{STSA}}^{\text{JUM}(i)}$ [5] evaluated as:

$$\psi_{\text{STSA}}^{\text{JUM}(i)}(y)=\int_0^\infty a_i f_{A_i|Y=y}(a_i;H_i)da_i = G_{\text{STSA}}(\xi_i,\gamma)|y|. \tag{4.88}$$

where $\xi_1=\xi$ as in the standard gain function $G_{\text{STSA}}(\xi,\gamma)$ whereas $\xi_0=\beta$. Therefore, the estimated $\widehat{A}_i$ is given as

$$\widehat{A}_i=\psi_{\text{STSA}}^{\text{JUM}(i)}(Y)=G_{\text{STSA}}(\xi_i,\gamma)A_Y, \tag{4.89}$$

According to Equation (4.66), the Bayesian risk $r_{ji}$ for $j\neq i$ is given by:

$$r_{ji}(y;\psi_j)=\int_0^\infty a_i^2 f_{A_i|Y=y}(a_i;H_i)da_i \quad (j\neq i) \tag{4.90}$$

Moreover, under correct detection, the Bayesian risk $r_{ii}$ is computed by using Equations (4.66) and (4.88) and equals:

$$r_{ii}(y;\psi_i)=\int_0^\infty \left(\psi_{\text{STSA}}^{\text{JUM}(i)}(y)-a_i\right)^2 f_{A_i|Y=y}(a_i;H_i)da_i = r_{ji}(y;\psi_j)-\left(\psi_{\text{STSA}}^{\text{JUM}(i)}(y)\right)^2, \tag{4.91}$$

with $j\neq i$. Injecting Equations 4.91 and 4.78 into Equation (4.77), we obtain the decision rule as:

$$\Lambda(\xi,\gamma)\left(\psi_{\text{STSA}}^{\text{JUM}(1)}(y)\right)^2 \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau_{\text{STSA}}^{\text{JUM}}\left(\psi_{\text{STSA}}^{\text{JUM}(0)}(y)\right)^2 \tag{4.92}$$

where $\tau_{\text{STSA}}^{\text{JUM}}$ is given in Appendix B. Finally, the gain function in this situation is written as

$$G_{\text{STSA}}^{\text{JUM}}(\xi,\gamma)=\begin{cases} G_{\text{STSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\text{STSA}}^{\text{JUM}}(y)\geq\tau_{\text{STSA}}^{\text{JUM}},\\ G_{\text{STSA}}(\beta,\gamma) & \text{otherwise,}\end{cases} \tag{4.93}$$

where $\mathcal{D}_{\text{STSA}}^{\text{JUM}}$ is given by:

$$\mathcal{D}_{\text{STSA}}^{\text{JUM}}(y)=\Lambda(\xi,\gamma)\left(\frac{\psi_{\text{STSA}}^{\text{JUM}(1)}(y)}{\psi_{\text{STSA}}^{\text{JUM}(0)}(y)}\right)^2. \tag{4.94}$$

### 4.4.2 Uncertain joint LSA estimator

#### 4.4.2.1 Independent LSA estimator

The first method is derived by defining the cost function as follows

$$c_{ji}(\widehat{A}_j,A_i)=\left(\log(\widehat{A}_j)-\log(A_i)\right)^2, \tag{4.95}$$

The estimators are sequentially evaluated following Equation (4.76). We have fist:

$$\mathbb{D}_i(y;\psi)=f_Y(y;H_1)\int(t-\log(a_1))^2 f_{A_1|Y=y}(a_1)da_1+\tau f_Y(y;H_0)\int(t-\log(a_0))^2 f_{A_0|Y=y}(a_0)da_0,$$

where, given $y\in\mathbb{C}$, $t=\log(\psi(y))$ with $\psi:\mathbb{C}\to(0,\infty)$. Therefore, the value of $t=\log(\psi(y))$ that minimize $\mathbb{D}_i(y;\psi)$ is :

$$t_{\text{LSA}}^{\text{IUM}}=\frac{f(y;H_1)\int\log(a_1))f_{A_1|Y=y}(a_1)da_1+\tau_{\text{LSA}}^{\text{IUM}}f_Y(y;H_0)\int\log(a_0))f_{A_0|Y=y}(a_0)da_0}{f_Y(y;H_1)+\tau_{\text{LSA}}^{\text{IUM}}f_Y(y;H_0)}, \tag{4.96}$$

[5] JUM mean "Joint" estimation in the "Uncertain Model".

where $\tau_{\text{LSA}}^{\text{IUM}}$ is calculated in Appendix B. This value can be evaluated by using the standard LSA gain function as follows:

$$t_{\text{LSA}}^{\text{IUM}} = \frac{\Lambda(\xi,\gamma)\log(G_{LSA}(\xi,\gamma)|y|) + \tau_{\text{LSA}}^{\text{IUM}}\log(G_{LSA}(\beta,\gamma)|y|)}{\Lambda(\xi,\gamma) + \tau_{\text{LSA}}^{\text{IUM}}}, \tag{4.97}$$

The corresponding map function $\psi_{\text{LSA}}^{\text{IUM}}$ from $\mathbb{C}$ to $(0,+\infty)$ under the two hypotheses are the same and equal to

$$\psi_{\text{LSA}}^{\text{IUM}}(y) = \exp\left(t_{\text{LSA}}^{\text{IUM}}\right) = \exp\left(\frac{\Lambda(\xi,\gamma)\log(G_{\text{LSA}}(\xi,\gamma)) + \tau_{\text{LSA}}^{\text{IUM}}\log(G_{\text{LSA}}(\beta,\gamma))}{\Lambda(\xi,\gamma) + \tau_{\text{LSA}}^{\text{IUM}}}\right)|y|. \tag{4.98}$$

Thus, the amplitude estimators under two hypotheses are identical $\widehat{A}$, which is determined as follows:

$$\widehat{A} = \psi_{\text{LSA}}^{\text{IUM}}(Y) = G_{\text{LSA}}^{\text{IUM}}(\xi,\gamma)A_Y, \tag{4.99}$$

where the gain function $G_{\text{LSA}}^{\text{IUM}}(\xi,\gamma)$, independent on the decided hypothesis, is a combination of $G_{\text{LSA}}(\beta,\gamma)$ and $G_{\text{LSA}}(\xi,\gamma)$.

$$G_{\text{LSA}}^{\text{IUM}}(\xi,\gamma) = \exp\left(t_{\text{LSA}}^{\text{IUM}}\right) = \exp\left(\frac{\Lambda(\xi,\gamma)\log(G_{\text{LSA}}(\xi,\gamma)) + \tau_{\text{LSA}}^{\text{IUM}}\log(G_{\text{LSA}}(\beta,\gamma))}{\Lambda(\xi,\gamma) + \tau_{\text{LSA}}^{\text{IUM}}}\right). \tag{4.100}$$

Note that the gain function $G_{\text{LSA}}^{\text{IUM}}(\xi,\gamma)$ in Equation (4.100) and $G_{\text{STSA}}^{\text{IUM}}(\xi,\gamma)$ in Equation (4.84) clearly become $G_{\text{LSA}}(\xi,\gamma)$ and respectively (res.) $G_{\text{STSA}}(\xi,\gamma)$ when the threshold $\tau_{\text{LSA}}^{\text{IUM}} = 0$ and res. $\tau_{\text{STSA}}^{\text{IUM}} = 0$.

#### 4.4.2.2 Sub-optimum joint LSA estimator

As Subsection 4.4.1.2, for more emphasizing the role of the detector, the second LSA estimator in the uncertain model is built by basing on the cost function succeeding

$$c_{ji}(\widehat{A}_j, A_i) = \begin{cases} \left(\log(A_i+1)\right)^2, & i \neq j, \\ \left(\log(\widehat{A}_j+1) - \log(A_i+1)\right)^2, & i = j. \end{cases} \tag{4.101}$$

In the same way, as for the above cost function, we firstly calculate the risk $\mathbb{D}_i(y;\psi)$ where $\psi : \mathbb{C} \to [0,\infty)$. Then, we set the derivative of this risk with respect to $t = \log(\psi(y)+1)$ to equal zero for seeking the $\psi_i$. First, the risk $\mathbb{D}_i(y;\psi)$ following the cost defined in Equation (4.101) is given by:

$$\begin{aligned}
\mathbb{D}_1(y;\psi) =& f_Y(y;H_1)\int (t-\log(a_1+1))^2 f_{A_1|Y=y}(a_1)da_1 \\
&+ \tau f_Y(y;H_0)\int (\log(a_0+1))^2 f_{A_0|Y=y}(a_0)da_0, \\
\mathbb{D}_0(y;\psi) =& f_Y(y;H_1)\int (\log(a_1+1))^2 f_{A_1|Y=y}(a_1)da_1 \\
&+ \tau f_Y(y;H_0)\int (t-\log(a_0+1))^2 f_{A_0|Y=y}(a_0)da_0,
\end{aligned}$$

Thus, the value of $t_{\text{LSA}}^{\text{JUM}(i)}$ that minimizes $\mathbb{D}_i(y;\psi)$ is

$$t_{\text{LSA}}^{\text{JUM}(i)} = \int \log(a_i+1) f_{A_i|Y=y}(a_i)da_i. \tag{4.102}$$

As discussed in Subsection 4.4.1.2, $t_{\mathrm{LSA}}^{\mathrm{JUM}(i)}$ can be approximated by :

$$t_{\mathrm{LSA}}^{\mathrm{JUM}(i)} = \log\left[G_{\mathrm{LSA}}(\xi_i,\gamma)|y|+1\right], \tag{4.103}$$

where $\xi_1 = \xi$ and $\xi_0 = \beta$. The corresponding map $\psi_{\mathrm{LSA}}^{\mathrm{JUM}(i)}$ from $\mathbb{C}$ to $[0,+\infty)$ writes:

$$\psi_{\mathrm{LSA}}^{\mathrm{JUM}(i)}(\mathbf{y}) = \exp\left(t_{\mathrm{LSA}}^{\mathrm{JUM}(i)}\right) - 1 = G_{\mathrm{LSA}}(\xi_i,\gamma)|y|. \tag{4.104}$$

In order to determine the decision rule, as in Section 4.4.1.2, with $j \neq i$, we have:

$$\begin{aligned} r_{ji}(y;\psi_j) &= \int (\log(a_i+1))^2 f_{A_i|Y=y}(a_i)da_i, \\ r_{ii}(y;\psi_i) &= \int (t_{\mathrm{LSA}}^{\mathrm{JUM}(i)} - \log(a_i+1))^2 f_{A_i|Y=y}(a_i)da_i = r_{ji} - (t_{\mathrm{LSA}}^{\mathrm{JUM}(i)})^2, \end{aligned} \tag{4.105}$$

Thus, we obtain the decision rule as follows:

$$\mathcal{D}_{\mathrm{LSA}}^{\mathrm{JUM}}(y) \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau_{\mathrm{LSA}}^{\mathrm{JUM}}, \tag{4.106}$$

where $\tau_{\mathrm{LSA}}^{\mathrm{JUM}}$ is given in Appendix B and the lhs of the decision rule is

$$\mathcal{D}_{\mathrm{LSA}}^{\mathrm{JUM}}(y) = \Lambda(\xi,\gamma)\left(\frac{\log\left(\psi_{\mathrm{LSA}}^{\mathrm{JUM}(1)}(y)+1\right)}{\log\left(\psi_{\mathrm{LSA}}^{\mathrm{JUM}(0)}(y)+1\right)}\right)^2. \tag{4.107}$$

The detector $\mathcal{D}_{\mathrm{LSA}}^{\mathrm{JUM}}$ can be simply expressed as a function of two variables $\widehat{A}_0$ and $\widehat{A}_1$ to point out the relation between the detector and the estimator. Additionally, the detector depends on the *a priori* SNR $\xi$ and the *a posteriori* SNR $\gamma$. Finally, the estimator is summarized as

$$\widehat{A} = G_{\mathrm{LSA}}^{\mathrm{JUM}}(\xi,\gamma)A_Y, \tag{4.108}$$

where

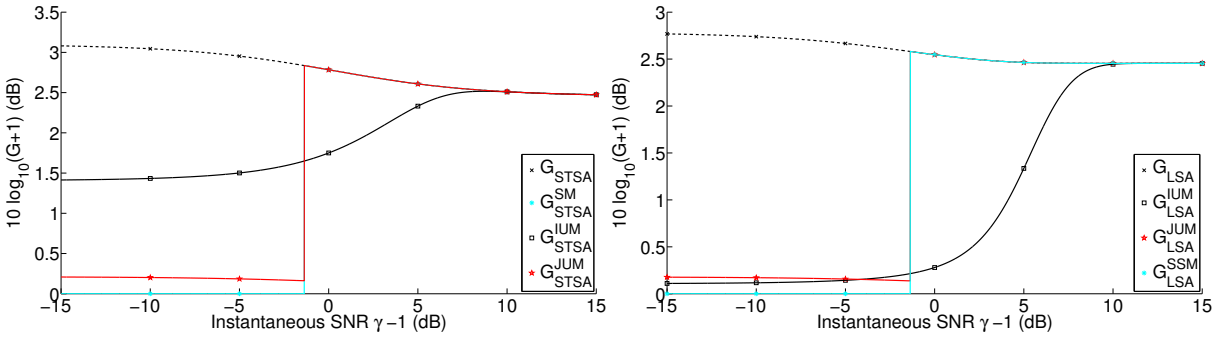
$$G_{\mathrm{LSA}}^{\mathrm{JUM}}(\xi,\gamma) = \begin{cases} G_{\mathrm{LSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\mathrm{LSA}}^{\mathrm{JUM}}(y) \geq \tau_{\mathrm{LSA}}^{\mathrm{JUM}}, \\ G_{\mathrm{LSA}}(\beta,\gamma) & \text{otherwise.} \end{cases} \tag{4.109}$$

The gain functions of all the methods in this chapter are displayed by Figure 4.1. Compared to the standard STSA and LSA methods (Figures 4.1a and 4.1b, respectively), these joint methods provide more impact at low instantaneous SNR. We recall that the instantaneous SNR is defined by $\gamma - 1$ [1].

## 4.5 Experimental results

### 4.5.1 Database and Criteria

We assessed our proposed methods on the NOIZEUS database [1] and 11 types of noise from the AURORA database. We also involved synthetic white noise and auto-regressive noise (AR). These tests were conducted at four SNR levels, namely 0, 5, 10 and 15 dB, as in Chapter 3. In our experiments, speech signals are sampled at 8 kHz, segmented into frames of 256 samples each, transformed by STFT with 50% overlapped Hamming windows. All thresholds are calculated by fixing the false alarm probability $\alpha$ to 0.05 for all noise levels (see Appendix B). The parameter $\beta$ is chosen as $\beta = 0.002$.




(a) STSA (b) LSA

Figure 4.1 – Attenuation curves of all joint detection/estimations in comparison with the standard STSA and LSA methods at *a piori* SNR level $\xi = 5$dB . The detector thresholds were calculated with $\alpha = 0.05$ and $\beta = -25$ dB.

The performance of all the methods were evaluated in two scenarios. In the first scenario, denoising is performed by using the reference noise power spectrum. This one is simply the theoretical power spectrum if noise is stationary. Otherwise, the reference noise power spectrum of the frame $m$ in a given bin $k$ is estimated as in [111] by:

$$\sigma_X^2[m,k] = \mu\sigma_X^2[m-1,k] + (1-\mu)A_X^2[m,k], \tag{4.110}$$

where $\mu = 0.9$. This iterative estimation is initialized by setting $\sigma_X^2[0,k] = A_X^2[0,k]$. The purpose of this scenario is to assess the performance of the denoising in itself, as much as possible. In the second scenario, for all the methods, the noise power spectrum was estimated using the B-E-DATE algorithm introduced in the chapter above [89]. This scenario makes it possible to estimate the performance loss in denoising incurred by integrating an up-to-date noise estimator.

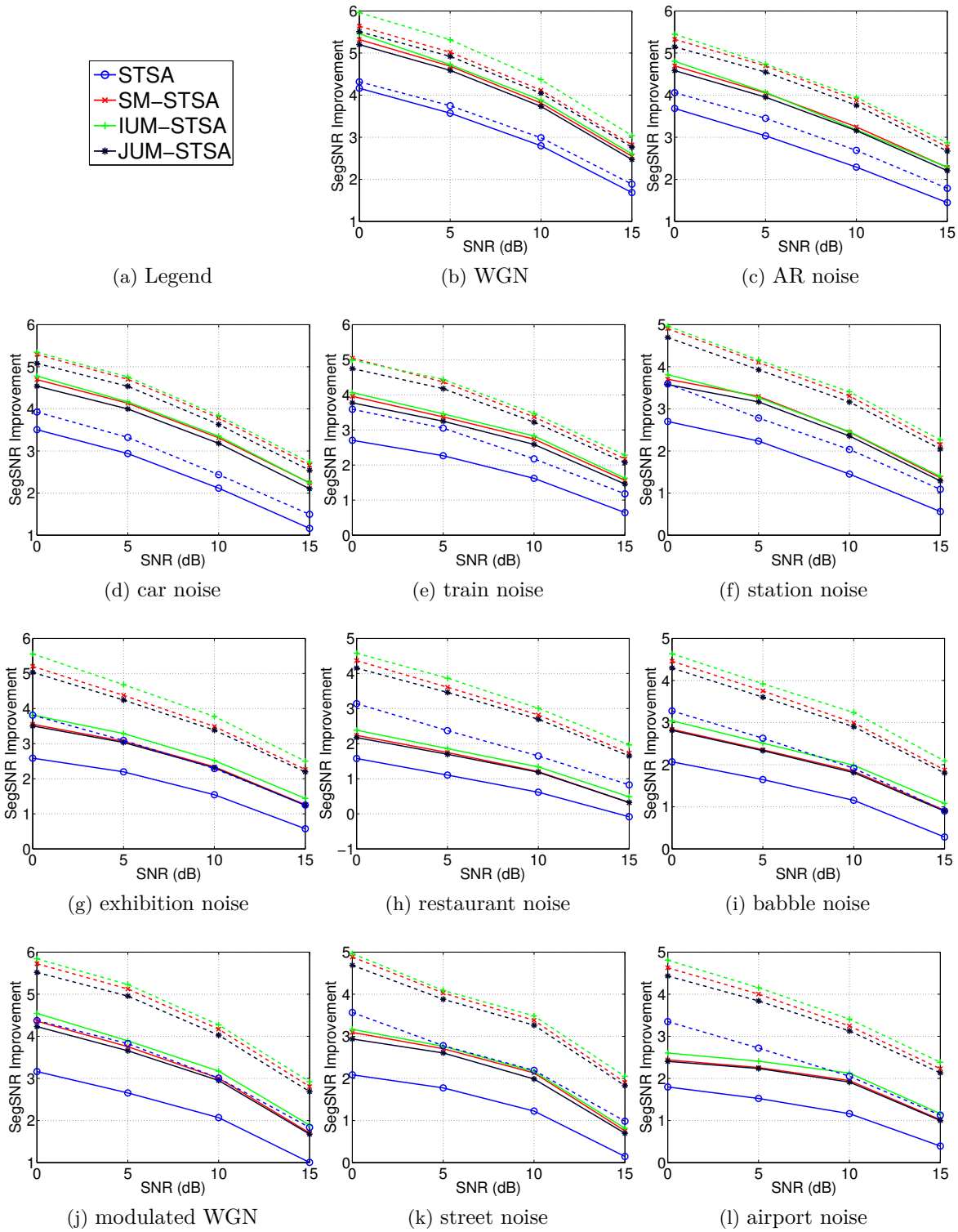
For assessing speech quality and preliminary speech intelligibility after denoising, objective quality and intelligibility criteria have been used. Speech quality is firstly measured by SSNR improvement and then by SNRI. The overall quality of speech was also predicted by MARSovrl measure as in Chapter 3. Secondly, intelligibility of speech was initially evaluated by the short-time objective intelligibility measure (STOI), which highly correlates with intelligibility measured by listening tests.

### 4.5.2 STSA-based results

| Methods | STSA | SM-STSA | IUM-STSA | JUM-STSA |
|---|---|---|---|---|
| Gain | Eq. (4.28) | Eq. (4.37) | Eq. (4.84) | Eq. (4.93) |

Table 4.1 – All jointed STSA methods have been implemented in the simulation

In this section, we consider all methods given by Table 4.1. All algorithms have been benchmarked at four SNR levels and against various noise models, namely white Gaussian noise (White), 2nd-order auto-regressive (AR) noise, 3 usual types of quasi-stationary noise (car, train and station) and 6 kinds of non-stationary noise (airport, exhibition, restaurant, street, modulated WGN and babble). The experimental results of all the methods based on the STSA




(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.2 – Speech quality evaluation by SSNR improvement after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. The common legend to all the sub-figures is that of Figure 4.2a.

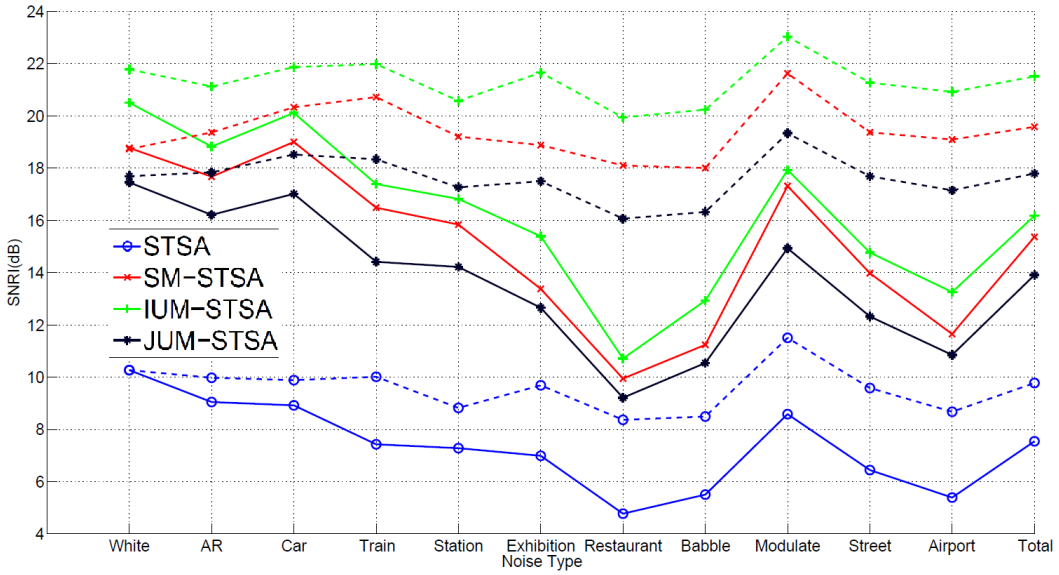



Figure 4.3 – SNRI with various noise types for all STSA-based methods with and without the reference noise power spectrum

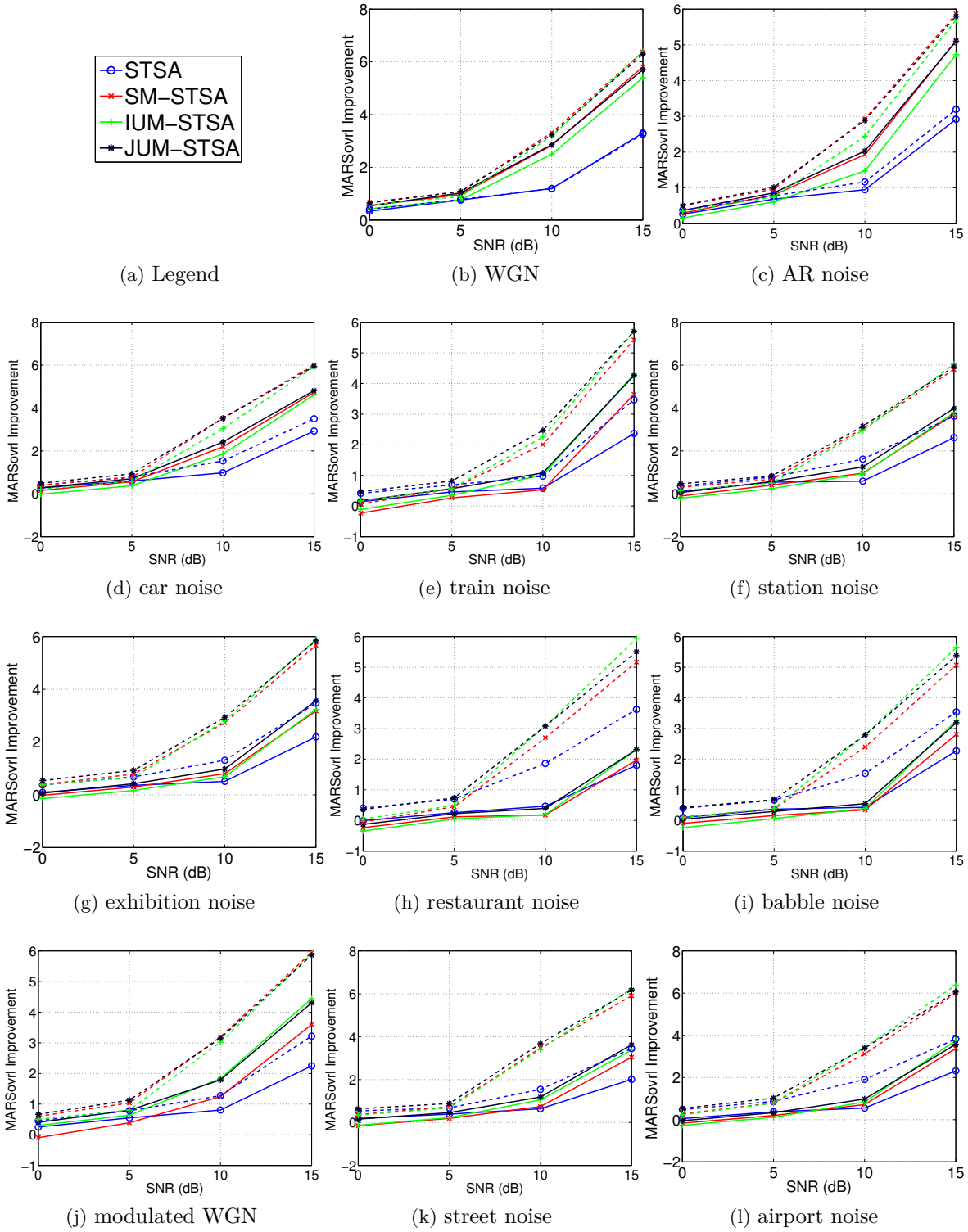
cost function are shown by Figures 4.2-4.5. Our methods are compared to the standard short-time spectral amplitude estimator (STSA) proposed in [27]. This STSA-based method is simple to implement and generally considered as a good reference method.

All measures of the STSA, SM-STSA, IUM-STSA and JUM-STSA methods are designated by the blue, red, green and black lines with the circle, x-mark, plus and star makers, respectively, as displayed in Figure 4.2a. Moreover, all of the measures obtained with the reference noise power spectrum and with B-E-DATE methods are drawn by dashed and solid lines, correspondingly.

Figure 4.2 displays the average results of the objective criterion SSNR improvement for different noise types and SNR levels and with two noise estimators. In the ideal situation where noise is Gaussian and known, IUM-STSA yields the best score at all SNR levels shown by Figure 4.2b. More specifically, in the same situation, between two strict and uncertain models, SM-STSA and JUM-STSA provide almost the same measures, whereas IUM-STSA derived from the uncertain model perform better than SM-STSA derived from the strict model. The gain is about 0.5 dB. Compared to STSA, the gain of the joint estimators is around $1 - 1.8$ dB. In the more realistic case where noise power spectrum is estimated by B-E-DATE, SSNR improvement measures obtained by the joint estimators are not so much different. The gain of the joint estimators regarding to STSA is now around 1 dB. The loss due to the use of noise estimator. The error of noise estimator can generate undesirable effect both in the estimator and the detector of the joint detection and estimation.

For stationary (AR) and slowly-change non-stationary (car, train and station) noise as in Figures 4.2c-4.2f, all joint estimators lead the same measure and outperform STSA with a gain around 1.5 dB in the first scenario where the reference noise power spectrum is used and with a gain around 1 dB in the second scenario.

For fast-changing and speech-like non-stationary (modulated, street, airport, exhibition, restaurant and babble) noise, the SSNR improvement score of IUM-STSA achieves also the best measure (see Figures 4.2g- 4.2l). The gain is also equal around to 1.5 dB in the first scenario and to 1 dB in the second scenario in comparison to standard STSA.



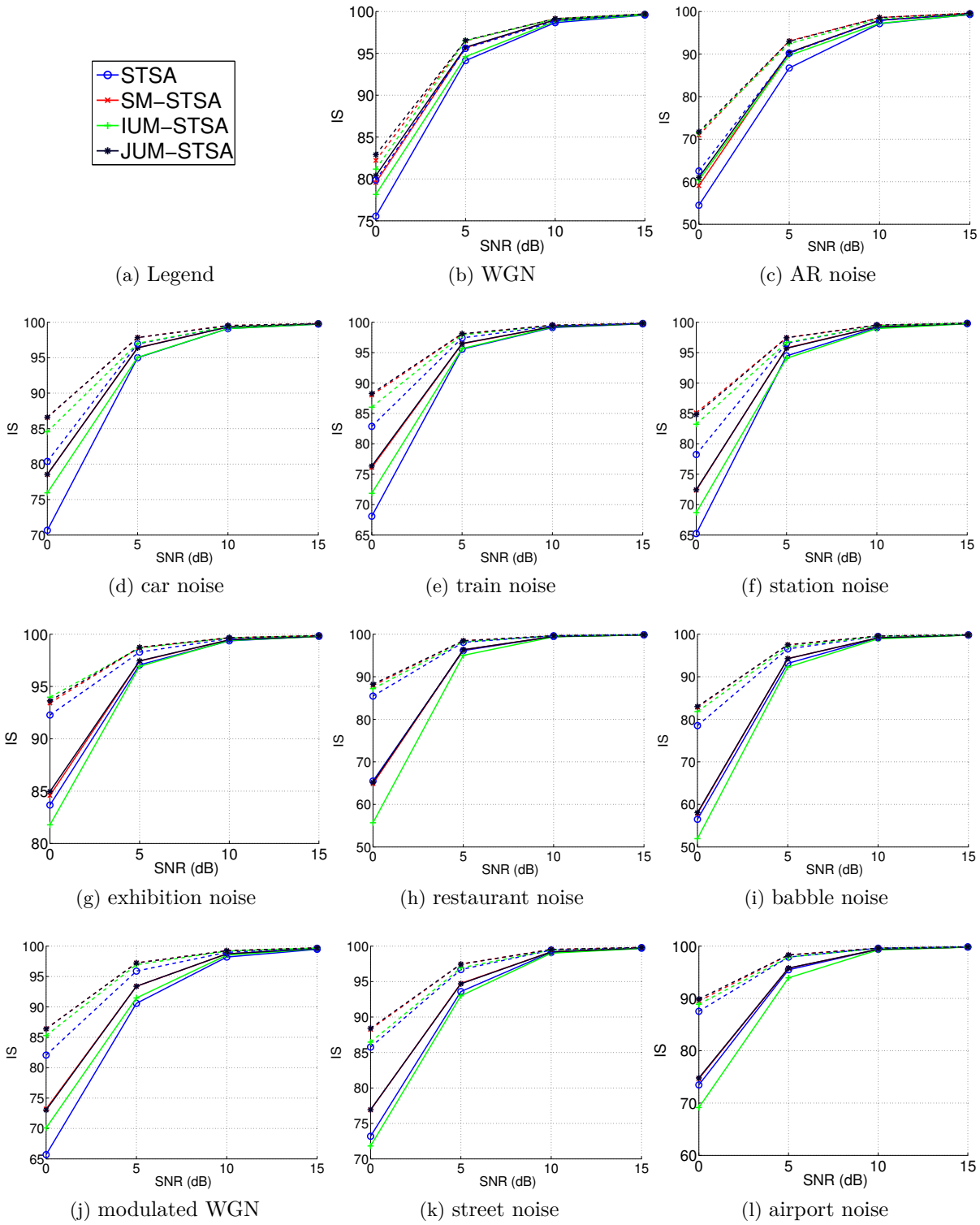

(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.4 – Speech quality evaluation by MARSovrl improvement after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Legend is also pointed out in Figure 4.4a.




(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.5 – Speech intelligibility evaluation by STOI after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Legend of all sub-figure is also illustrated in Figure 4.5a.

The second criterion SNRI measure is displayed in Figure 4.3. The legend is that in Figure 4.2a. This criterion confirms that IUM-STSA gives the best overall SNR improvement in the two scenarios. The gain is around 6 dB when using B-E-DATE and 8 dB when using the reference noise power for fast-changing non-stationary noise). For stationary and slowly-changing non-stationary noise, the gain is around 10 dB (res. 11.5 dB) when using noise power spectrum estimated by B-E-DATE (res. when using the reference noise power spectrum). We can summarize the foregoing by saying that joint estimators generally outperform standard STSA [27] in terms of SSNR improvement and SNRI in all situations. The overall gain is around 6 to 10 dB, which is emphasized by label "Total" in Figure 4.3.

The composite speech quality overall MARSovrl improvement measure results are illustrated by Figure 4.4. The legend is given by Figure 4.4a. For stationary (white and AR) noise, in the two scenarios, at low SNR levels, SM-STSA and JUM-STSA yield the same score and outperform IUM-STSA and STSA (see Figures 4.4b-4.4c). However, the gain is not significant. At high SNR levels, joint estimators outperform STSA as well. For slowly-changing non-stationary noise, in the two scenarios, at low SNR levels, JUM-STSA and STSA lead the same measure and slightly perform better than SM-STSA and IUM-STSA. At high SNR levels, joint estimators perform outperform standard STSA, except at 10 dB, for train noise, where SM-STSA and STSA yield the same score for using the noise power spectrum estimated by B-E-DATE (see Figures 4.4d-4.4f).

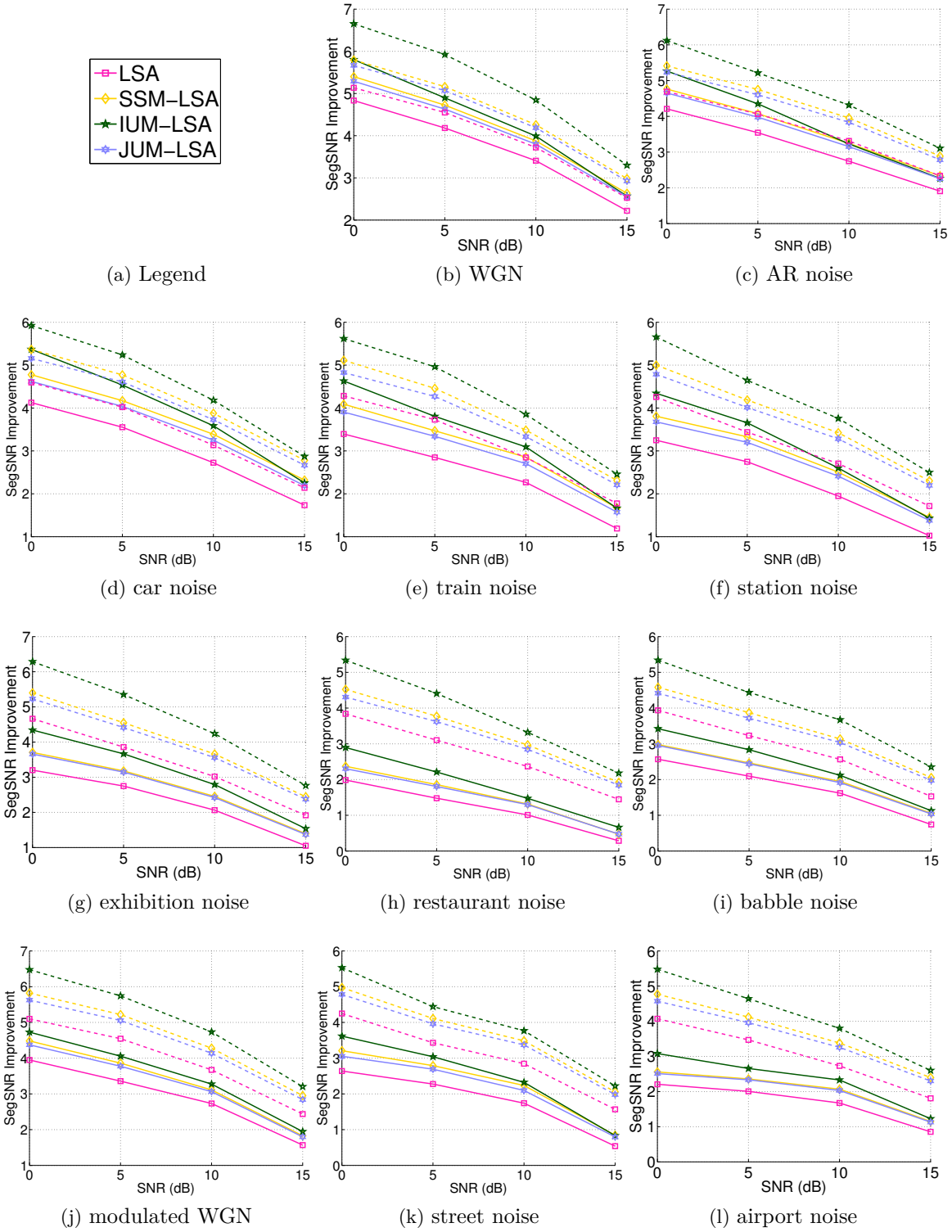
In the case of fast-changing and speech-like non-stationary noise, when all estimators are combined with the B-E-DATE noise power spectrum estimator, all methods provide similar scores at low SNR levels, even at 10 dB except for modulated WGN. The relevance of joint detection/estimation is only confirmed at higher SNR levels (see Figures 4.4g-4.4l). However, when using the reference noise power spectrum, a significant gain is yielded by joint detector/estimators in comparison to STSA at high SNR levels. This emphasizes the impact of noise estimation which seemingly provide undesirable effects in the detection quality. The good performance of the detector is obtained with the reference noise power spectrum and at high SNR levels.

With respect to the foregoing three criteria, we can say that, in a nutshell, in terms of SSNR and SNRI, SM-STSA leads better scores than JUM-STSA but performs lesser than IUM-STSA. In terms of overall speech quality, providing an estimation of the speech signal under $H_0$ hypothesis (JUM-STSA) makes it possible to obtain a better score than forcing the estimated amplitude to 0 under the null hypothesis (SM-STSA).

Finally, the intelligibility score (IS) obtained by mapping the STOI measure is shown by Figure 4.5. At high SNR levels, the methods yield the same results. Therefore, we focus at 0 dB. For stationary (white and AR) and slowly-changing non-stationary (car, train and station) noises, the proposed SM-STSA and JUM-STSA obtain the best score. The IS measure of these methods improves 5 to 10% in comparison to STSA and in the two scenarios. However, for the fast-changing airport and speech-like non-stationary noises, SM-STSA, JUM-STSA and STSA give the same score and are better than IUM-STSA. For modulated WGN and street noises, the gain in IS is around $4-7\%$ when using SM-STSA and JUM-STSA instead of STSA (see Figures 4.5g-4.5l).

### 4.5.3 LSA-based results

We now consider the joint detection and estimation methods mentioned in Table 4.2 and based on the log-spectral amplitude error function. We compare these methods to the standard log-spectral amplitude estimator (LSA) presented in [28]. In order to perform a significant analysis,




(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.6 – Speech quality evaluation by SSNR improvement after speech denoising using LSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Legend of all sub-figure is also given in Figure 4.6a.

Table 4.2 – All jointed LSA methods have been implemented in the simulation

| Methods | LSA | SSM-LSA | IUM-LSA | JUM-LSA |
|---|---|---|---|---|
| Gain | Eq. (4.42) | Eq. (4.55) | Eq. (4.100) | Eq. (4.109) |

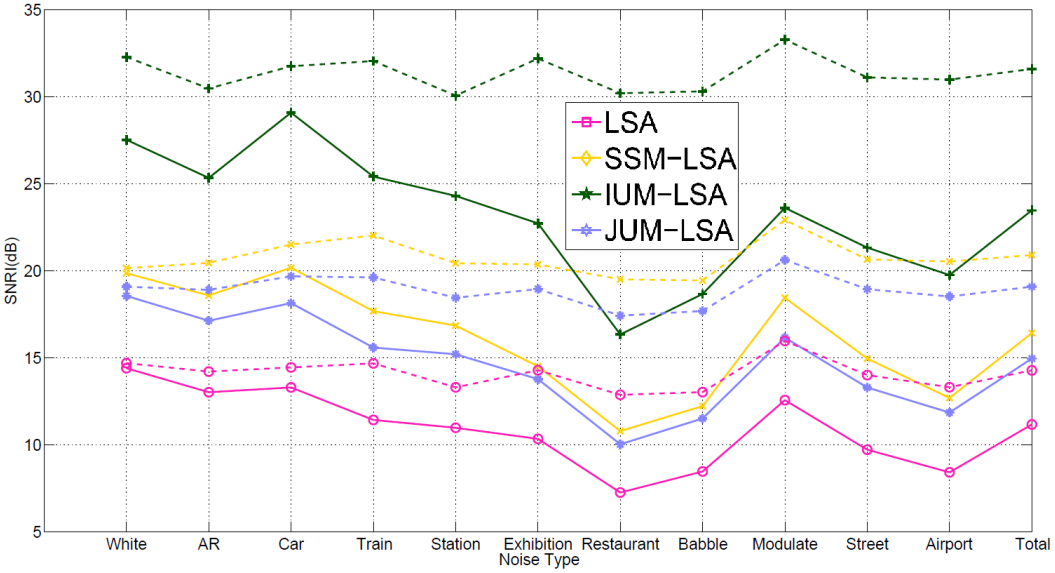


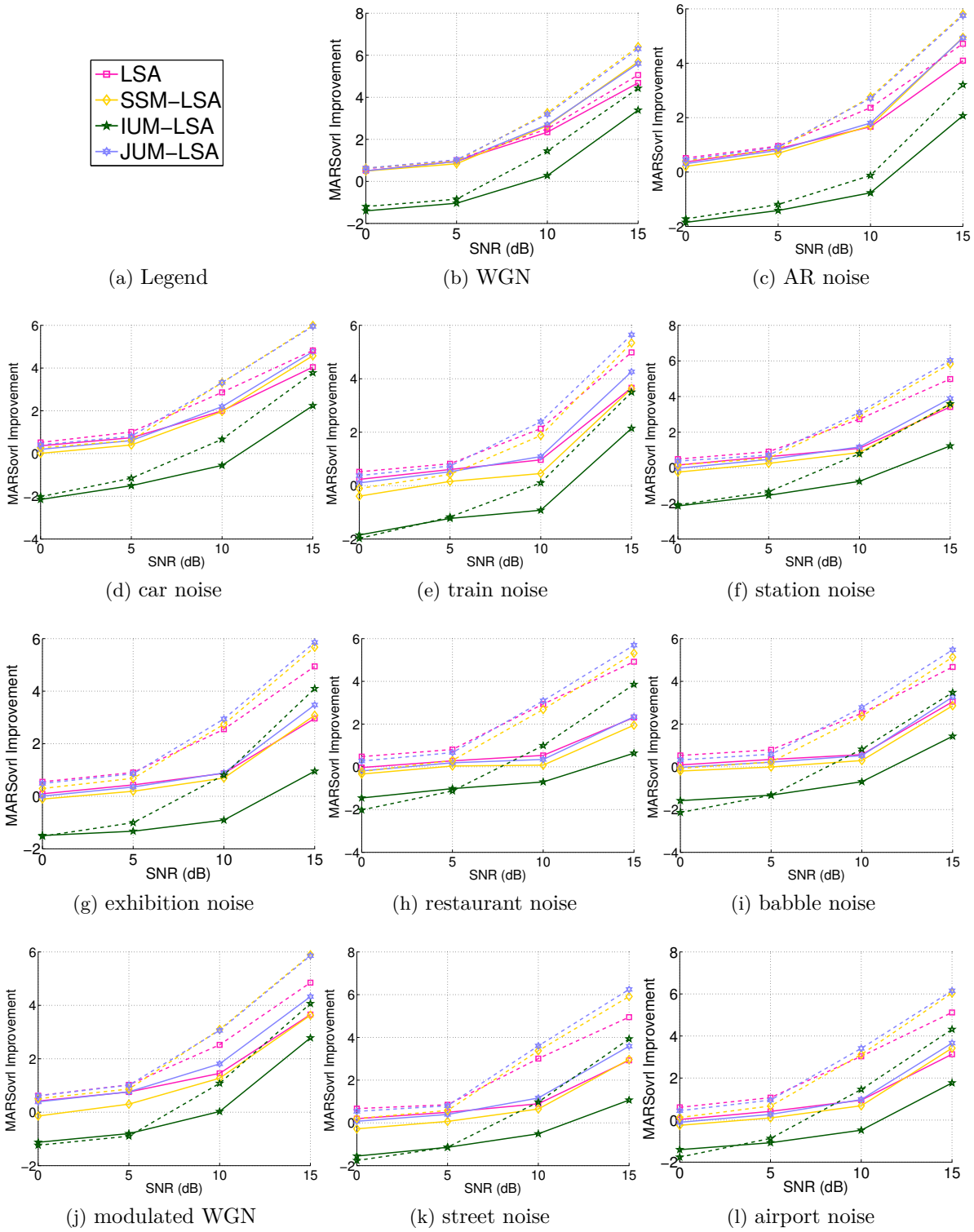

Figure 4.7 – SNRI with various noise types for all LSA-based methods in two scenarios where the reference noise power spectrum is used or not.

all the methods have also been tested at four SNR levels and against 11 kinds of noise, from stationary to slowly-changing non-stationary and fast-changing or speech-like non-stationary noises. All the scores obtained with these methods at all SNR levels and for all types of noise are displayed in Figures 4.6-4.9.

The scores of the LSA, SSM-LSA, IUM-LSA and JUM-LSA are plotted in pink, yellow, dark green, light blue lines with square, diamond, pentagram and hexagram markers, respectively (see legend of Figure 4.6a). As in the above section, the scores obtained when using the reference noise power spectrum are designed by dashed lines with same colors as that used to represent results obtained when the noise power spectrum is estimated by B-E-DATE.

The objective criterion SSNR improvement is shown in Figure 4.6 for all kinds of noise and in the two possible scenarios. We can see that IUM-LSA achieves the best score under all situations, from stationary noise (see Figures 4.6b-4.6c) to slowly-changing non-stationary noise (see Figures 4.6d-4.6f), and up to speech-like and fast-changing non-stationary noise (see Figures 4.6g-4.6l). At 15 dB SNR level, combining with the noise power spectrum estimated by B-E-DATE, the three joint detector/estimators (SSM-LSA, IUM-LSA, JUM-LSA) lead to the same score. By forcing the estimated amplitude $A_0$ under decision $H_0$ to 0, SSM-LSA yields slightly better results than JUM-LSA when the reference noise power spectrum is given. Moreover, the three joint detector/estimators outperform the standard LSA for all noise types and in the two scenarios in terms of SSNR improvement. This gain is more significant at low SNR levels. In this case, the gain is around $0.5 - 1.5$ dB.

Figure 4.7 displays the average ITU criterion SNRI for various noise types, levels, and also in the two scenarios. The legend is the same as in Figure 4.6a. SSNR improvement demonstrates the gain of the joint detector/estimators at each noise SNR levels, SNRI confirms the performance of joint detector/estimators at all SNR levels for each noise and for all considered noise. IUM-LSA



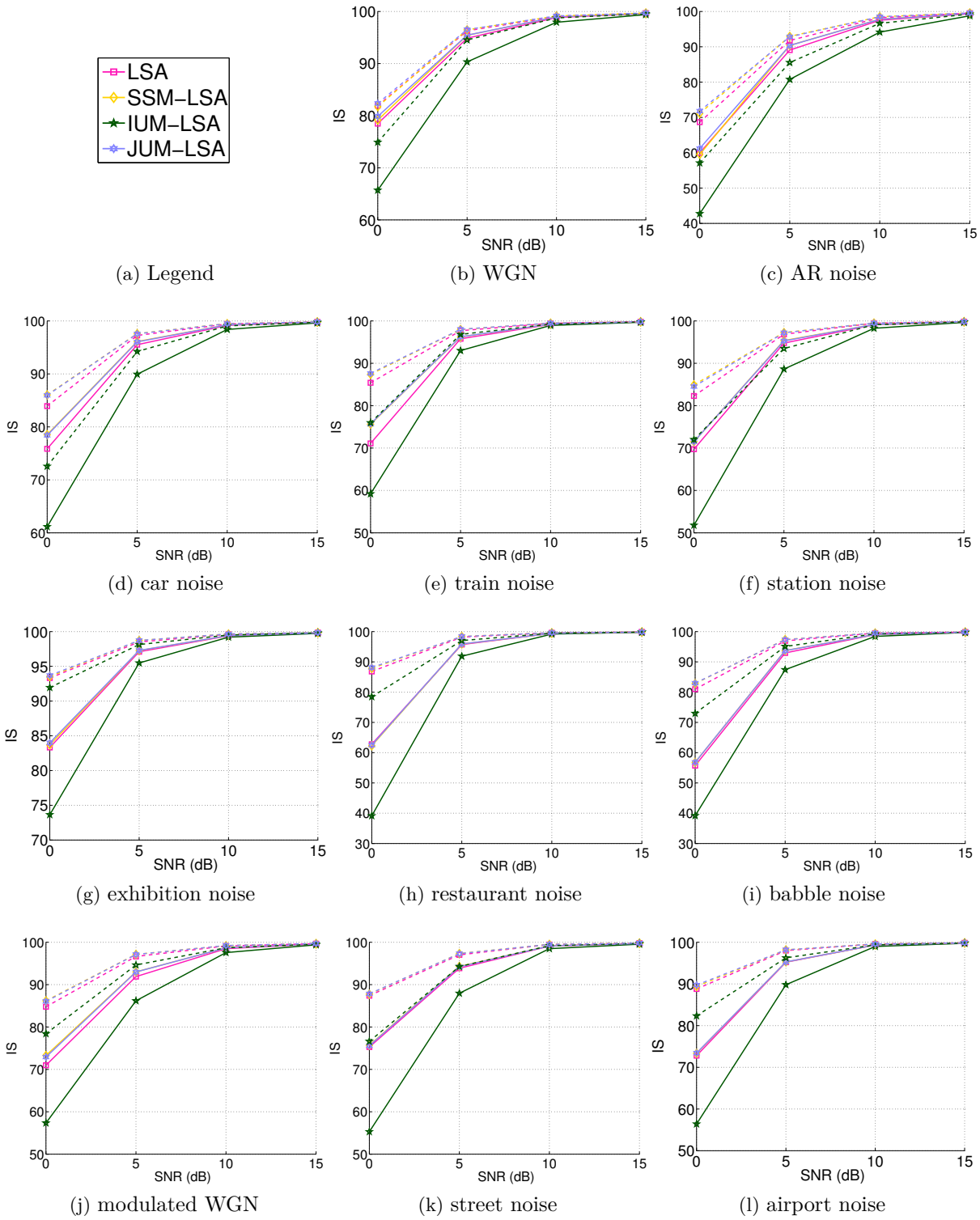

(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.8 – Speech quality evaluation by MARSovrl improvement after speech denoising using LSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Legend of all sub-figure is also illustrated in Figure 4.8a.




(a) Legend (b) WGN (c) AR noise

(d) car noise (e) train noise (f) station noise

(g) exhibition noise (h) restaurant noise (i) babble noise

(j) modulated WGN (k) street noise (l) airport noise

Figure 4.9 – Speech intelligibility evaluation by STOI after speech denoising using LSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Legend is also pointed out in Figure 4.9a.

remains better than the other methods. Compared to LSA, the gain is in the range of $[10-17]$ dB. This gain is higher in the situation where the noise reference is used. Generally, all of the joint detector/estimators give better scores than standard LSA. The gain is at least 3 dB.

In terms of MARSovrl improvement, the averaged scores are illustrated in Figure 4.8. Note that IUM-LSA, which achieves the best SSNR improvement and SNRI scores, gives the smallest MARSovrl score in all situations. IUM-LSA as IUM-STSA removes more the background noise so that they may suppress the signal of interest. Therefore, IUM-LSA leads the smallest MARSovrl measure. An informal listening confirms that IUM-LSA provides a large noise distortion and also a signal distortion.

Considering the other methods, at low SNR levels in the two scenarios, for stationary noise, LSA, SSM-LSA and JUM-LSA obtain similar results (see Figure 4.8b and 4.8c). In the same situation, for non-stationary noise, LSA and JUM-LSA yield better scores than SSM-LSA (see Figures 4.8d-4.8l). At high SNR levels, in the first scenario (*i.e.* using the reference noise power spectrum) JUM-LSA and SSM-LSA achieve better scores than LSA for all noise types. For the second scenario (*i.e.* using the noise power spectrum estimated by B-E-DATE), only at 15 dB JUM-LSA yields a significantly better score than LSA and SSM-LSA, except for white noise and its modulation (see Figures 4.8b and 4.8j ). The fact that JUM-LSA outperforms SSM-LSA in the realistic situation where noise is unknown strengthens the motivation to provide a small estimate under hypothesis $H_0$,

In terms of speech intelligibility, the IS scores are shown by Figure 4.9. In general, we focus also on low SNR levels. JUM-LSA presents the best measure, whereas IUM-LSA returns the smallest measure in all the situations under consideration. In the first scenario, SSM-LSA has the same score than JUM. Compared to LSA the gain is around $1-2\%$. For the second scenario, for stationary and slowly-changing non-stationary noise, the gain is equal to $1-2\%$. It can even reach 5% for train noise (see Figure 4.9e). For speech-like and fast-changing noise, LSA, SSM-LSA and JUM-LSA yield same scores, except for modulated WGN.

## 4.6 Conclusion

In this chapter, for speech enhancement, we have proposed joint detection and estimation methods based on STSA and LSA estimation. The key idea is to take into account the presence and absence of speech in each time-frequency bin. Thus, optimal detectors are derived to improve quality of speech in noisy environments. When the absence of speech is detected, our methods have set the STSA to zero or to a small spectral floor for avoiding musical noise. The performance evaluation was conducted in two scenarios, one where the reference noise power spectrum is used and one where noise is estimated by an up-to-date method. The experimental results have shown the relevance of the approach. In a nutshell, these experimental results enhance the interest to combine speech detection and estimation. Actually, joint detection/estimation generally outperforms the standard STSA, which is still recognized as a reference approach. So, in practice, we would recommend the use of such detector/estimators. the choice between them can be ruled by the type of criterion the practitioner wishes to optimize.

We proposed a unified framework based on the combination of detection and estimation for improving the performance of Bayesian estimators in speech enhancement. The efficiency of the approaches deriving from this framework is totally dependent on the quality of the speech detection on which the estimation is based. Indeed, miss detection induces degradation of the estimation, whereas false alarms may generate musical noise. In addition, all detector/estimators are based on the Gaussian assumption, which may not be respected. Therefore, another type of detector could be considered in each time-frequency bin. Prospects in this respect will be

discussed in the final chapter. In the next chapter, we will introduce a semi-parametric approach, where the Gaussianity of the speech is not required.

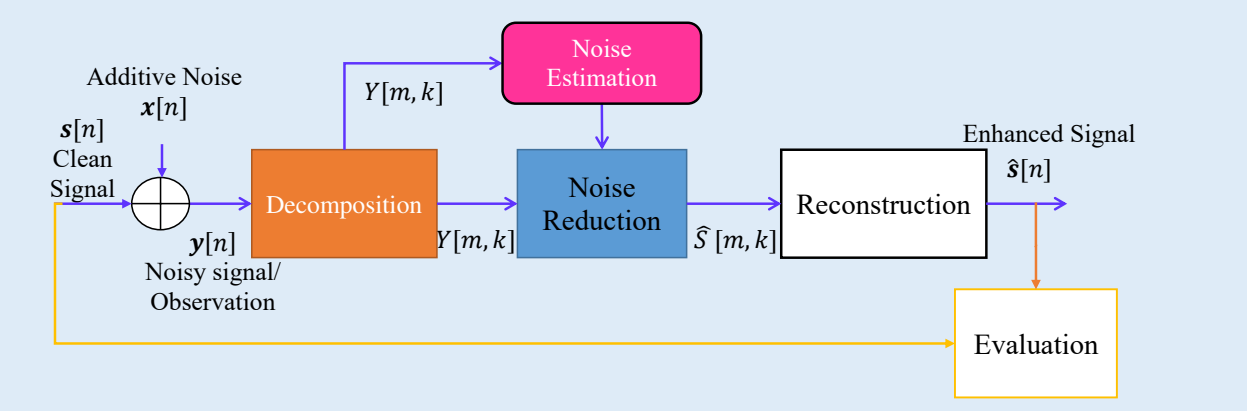
Chapter 5

# Non-diagonal smoothed shrinkage for robust audio denoising

*Believe you can and you're halfway there.*

Theodore Roosevelt





## 5.1 Introduction

### 5.1.1 Motivation and organization

The previous chapter focused solely on the parametric methods. But it turns out that many results in non-parametric and robust statistical estimation established in the last two decades [2,62–64,66,103] and based on sparse thresholding and shrinkage, are generally enough to suggest their use in unsupervised speech and audio denoising for improving the robustness of the denoising methods. Generally speaking and as recalled below, the interest in non-parametric audio and speech denoising is twofold. First, it performs regardless of the signal distribution. Second, it achieves gain in intelligibility [120]. Since Bayesian approaches are known to improve quality [82], the idea is to combine the two approaches. Nonetheless, this combination requires some care. Indeed, most non-parametric estimators force to 0 small amplitude coefficients obtained after transformation into a certain domain. Although much background noise is canceled by doing so, removing small noisy coefficients pertaining to the signal of interest generates musical noise and reduces speech and audio quality [1]. This problem is well known in image processing where zero-forcing of small coefficients induces artifacts [63].

Therefore, if we want to improve quality by eliminating residual musical noise, the non-parametric denoising should be a smooth shrinkage merely aimed at attenuating small coefficients. A Bayesian estimator can then be used downstream the non-parametric one to retrieve speech and audio information in small coefficients and thus improving the overall quality. Note that if the Bayesian estimator were used before the non-parametric one, the latter would tend to shrink small coefficients estimated by the former, which is not desirable because even small coefficients after Bayesian estimation may pertain to relevant speech or audio contents for overall quality.

With respect to the foregoing, the problem addressed in this chapter is the design and combination of non-parametric and Bayesian estimation for speech and audio denoising. In this chapter, as the other methods mentioned above, we estimate the amplitudes of the clean signal coefficients in the time-frequency domain. The estimation is based on the MMSE criterion. However, instead of the DFT, we focus on the discrete cosine transform (DCT), which avoids estimating the phase spectrum and may reduce complexity [128, 129]. To this end, we will consider the following strategy.

We begin by improving speech and audio intelligibility by a non-parametric approach based on smoothed sigmoid-based shrinkage (SSBS) [2], originally introduced for image denoising. Two main features of the approach are: 1) it attenuates DCT coefficients that are very likely to pertain to noise only or to speech with small amplitude in noise; 2) it tends to keep unaltered large-magnitude DCT coefficients. However, such a non-parametric approach can be regarded as an approximated Wiener filtering and, as such, introduces musical noise. We then modify the original SSBS approach and propose the SSBS block estimator, hereafter named Block-SSBS. Block-SSBS is relevant to eliminate isolated points in the time-frequency domain that may induce musical noise. Basically, Block-SSBS applies the same SSBS gain function to time-frequency blocks. The sizes of these blocks are determined by adaptive Stein's Unbiased Risk Estimate (SURE) [3] so as to minimize the unbiased estimate of the mean square error over regularly distributed time-frequency regions. In addition, other parameters of Block-SSBS can be optimized by resorting to recent results in non-parametric statistical signal processing [4]. A nice feature of the proposed parameter optimization procedure is the level of control offered on the denoising performance that allows trading-off speech quality and intelligibility. This is made

possible by discriminating speech (or audio) components with significant contents from speech (resp. audio) components with lesser interest.

For reasons detailed below, the outcome of Block-SSBS is assumed to satisfy the same hypotheses as those generally used for Bayesian estimation. Therefore, in a second step, to further reduce musical noise and, above all, to improve speech quality, a Bayesian statistical estimator is devised for application to smoothed short-time spectral amplitude (STSA) after Block-SSBS. This Bayesian estimator is hereafter called STSA-MMSE.

In a nutshell, the main contributions of this chapter are as follows. To begin with, the whole method is carried out in the DCT domain, so as to get rid of the phase estimation problem. It introduces Block-SSBS in the DCT domain for speech and audio denoising in presence of stationary or non-stationary noise. Block-SSBS is then optimized via automatic and adaptive statistical methods tailored to speech and audio enhancement. The derivation of STSA-MMSE in the DCT domain is another contribution. The chapter also propounds and studies the combination of Block-SSBS and STSA-MMSE and shows that this combination is very promising for speech and audio denoising in presence of various types of noise, via objective and subjective tests. It must also be pointed out that these tests include situations where the noise spectrum is known, as well as cases where this spectrum is estimated via an up-to-date estimator.

The rest of this chapter is organized as follows. Sub-Section 5.1.2 introduces the signal model, the notation and makes some general recalls on the DCT. Sub-Section 5.1.3 reviews the non-parametric thresholding methods originally developed for image denoising, with a particular emphasis on SSBS. In Section 5.2, we present semi-parametric audio and speech enhancement by Block-SSBS, derive the Bayesian STSA-MMSE in the DCT domain and then combine the two. Experimental results, both objective and subjective, are reported and analyzed in Section 5.3. Finally, Section 5.4 concludes this chapter.

### 5.1.2 Signal model and notation in the DCT domain

As announced above, the DCT will hereafter be used for denoising. Therefore, this section reviews the principle of the DCT and the reasons why the DCT can be applied to speech and audio enhancement.

The DCT is analyzed from a general point of view in [130]. Originally developed for pattern recognition and Wiener filtering in image processing, its interest in speech and audio enhancement is more specifically studied in [128, 129]. Basically, given a sequence $\{y[n]\}$ with $0 \leq n \leq K-1$, the DCT coefficients are calculated as:

$$Y[k] = \alpha_k \sum_{n=0}^{K-1} y[n] \cos \frac{(2n+1)k\pi}{2K}, \tag{5.1}$$

with $\alpha_0 = \sqrt{1/K}$ and $\alpha_k = \sqrt{2/K}$ for $1 \leq k \leq K-1$ [131]. The inverse DCT is then given by:

$$y[n] = \sum_{k=0}^{K-1} \alpha_k \, Y[k] \cos \frac{(2n+1)k\pi}{2K}. \tag{5.2}$$

The DCT defined by (5.1) and (5.2) can be effectively used in speech and audio enhancement or noise reduction for the subsequent reasons. As discussed in [128, 130, 131], DCT has higher energy compaction than DFT. The signal of interest can thus have a sparse representation in the DCT domain. That is why the DCT is widely used in image compression [130] and dictionary learning [132]. Second, the DCT coefficients are real whereas DFT coefficients are complex. The DCT coefficients have binary phase, whereas phases of the DFT coefficients are often assumed

to follow the uniform distribution in the range $[-\pi, \pi]$. Therefore, the DCT phase [128] does not need to be estimated because error in the DCT phase has no important role for estimating the signal of interest. Third, DCT is known to be better than DFT for approximating the Karhumen-Lovève transform (KLT), which is optimal in terms of variance distribution, rate distortion function and mean-square estimation error. Moreover, DCT and inverse DCT (IDCT) can be also calculated by fast computation algorithms.

For estimating a clean audio signal from its noisy observation, the latter is often segmented, windowed and transformed by computational harmonic analysis. In the present framework, this harmonic analysis will be performed by DCT.

Let us denote the noisy signal in the DCT domain by:

$$Y[m,k] = S[m,k] + X[m,k], \tag{5.3}$$

where $m$ and $k \in \{0, 1, \ldots, K-1\}$ are the time and frequency-bin indices, respectively. As an extension of (5.1) and similarly to the expressions of DFT coefficients, the DCT coefficients are obtained as [1]:

$$Y[m,k] = \sum_{n=0}^{K-1} \alpha_n w[n] y[mK^* + n] \cos \frac{(2n+1)k\pi}{2K}, \tag{5.4}$$

where $K$ is the frame length, $K^*$ is the number of shifted samples between two consecutive frames and $w[n]$ is a window such as the Hamming or the Hanning windows with length $K$. For the sake of simplicity, the indices $m$ and $k$ will be omitted unless for clarification. Wide hat symbols are henceforth used to denote estimates. Moreover, lower case letters denote realizations of random variable. The absolute value (resp. sign) of the DCT coefficients of the noisy signal, signal of interest and noise are denoted by $A_Y, A_S, A_X$ (resp. $\phi_Y, \phi_S, \phi_X$), correspondingly.

The signal of interest and noise are assumed to be independent and zero mean, so that $\mathbf{E}(Y^2) = \mathbf{E}(S^2) + \mathbf{E}(X^2) = \sigma_S^2 + \sigma_X^2$, where the spectra of the clean signal and noise are denoted by $\mathbf{E}(S^2) = \sigma_S^2$, $\mathbf{E}(X^2) = \sigma_X^2$, respectively, and where $\mathbf{E}(.)$ is the expectation. We also define the *a priori* signal-to-noise ratio (SNR) $\xi$ and the *a posteriori* SNR $\gamma$ as $\xi = \sigma_S^2/\sigma_X^2$, $\gamma = |Y|^2/\sigma_X^2$. As usual [28], the DCT coefficients $Y[m,k]$ with $k \in \{0, 1, \ldots, K-1\}$ are assumed to be uncorrelated. The notation introduced above is used throughout with always the same meaning.

### 5.1.3 Sparse thresholding and shrinkage for detection and estimation

Because we want to study to what extent sparse thresholding and, more precisely, smooth shrinkage can contribute to speech and audio denoising, this section provides recalls on such methods, originally devised for retrieving the transformed coefficients of a clean image observed in noise. These methods perform estimation regardless of the signal distribution. Their implementation is quite simple.

Image denoising can be typically achieved via shrinkage functions, whereby an estimate of the signal of interest is obtained by thresholding the coefficients obtained by projection of the noisy observation onto an orthogonal basis. Given an observation coefficient $Y$ in the wavelet, DCT or DFT domain with $Y = S + X$, the estimate $\widehat{S}$ of the clean signal coefficient is obtained by $\widehat{S} = GY$, where $G$ is a gain or shrinkage function. Such a shrinkage function depends on $Y$. In the sequel, $G$ will be expressed as a function of $\gamma$ or an estimate of $\gamma$. In this respect, hard thresholding is the first shrinkage function introduced in [103] and further developed in [62]. Hard thresholding estimates $S$ by keeping or discarding $Y$ according to:

$$\widehat{S} = \begin{cases} Y & \text{if} \quad \|Y\| \geq \lambda \sigma_X, \\ 0 & \text{otherwise}, \end{cases} \tag{5.5}$$

where $\lambda$ is an appropriate threshold. The hard thresholding gain function is thus:

$$G_\lambda(\gamma) = \begin{cases} 1 & \text{if} \quad \gamma \geq \lambda^2, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

Smooth shrinkage performs estimation of the clean transformed coefficient by the soft thresholding gain function proposed in [62]:

$$G_\lambda(\gamma) = \begin{cases} 1 - \frac{\lambda}{\sqrt{\gamma}} & \text{if} \quad \gamma \geq \lambda^2, \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

It is illuminating to interpret shrinkage by soft thresholding as a combined detection and estimation process. By comparing the *a posteriori* SNR $\gamma$ to a suitable threshold and setting $\widehat{S}$ to zero if the *a posteriori* SNR $\gamma$ falls below the threshold, a kind of speech and audio detection is indeed realized. In addition, soft thresholding provides a transformed coefficient estimate of the desired signal by subtracting the threshold from the noisy coefficients. A similar approach operating in the time domain was proposed in [133].

Another form of smoothed shrinkage is provided by the SSBS approach proposed and analyzed in [2, 63]. The SSBS gain function relies on the sigmoid function, also called logistic function, widely used in machine learning. Based on three desirable properties for any shrinkage function, that is, smoothness, penalized shrinkage and vanishing attenuation at infinity, the SSBS method allows for a trade-off between hard and soft thresholding. The original SSBS gain function in [2] reads:

$$G_{\tau,\lambda}(\gamma) = \frac{1}{1 + e^{-\tau\left(\sqrt{\gamma} - \lambda\right)}}, \tag{5.8}$$

where parameter $\lambda$ influences the detection performance, whereas $\tau$ controls the attenuation provided by the SSBS function. The hard thresholding gain function is a limiting case of SSBS gain function. Furthermore, SSBS functions make it possible to attenuate $\sqrt{\gamma}$ below $\lambda$ in a continuous manner, instead of setting it to zero as conventionally done in hard and soft thresholding. SSBS methods are simple to implement since they only require multiplying the noisy coefficients by the logistic function to obtain the enhanced transformed coefficients.

In the DFT or DCT domain, the attenuation factors or gain functions $G_\lambda(\gamma)$ , $G_{\tau,\lambda}(\gamma)$ of the above methods are independently and singly evaluated for each $[m, k]$ atom. Therefore, in order to incorporate the impact of neighboring time-frequency atoms, the block thresholding approach, originally proposed in [134] for wavelet transform, can be applied to the DFT [135]. In both cases, the method is based on the so-called subtraction gain function or soft thresholding so that the gain function for block $B_i$ with size $L_i$ is:

$$G_\lambda(\gamma_i) = \left(1 - \frac{\lambda}{\gamma_i}\right)_+, \tag{5.9}$$

where

$$\gamma_i = \frac{\sum_{[m,k]\in B_i} |Y[m,k]|^2}{L_i \sigma_X^2}, \tag{5.10}$$

and $(\theta)_+ = \theta$ if $\theta \geq 0$ and $(\theta)_+ = 0$ otherwise.

## 5.2 Non-diagonal audio estimation of Discrete Cosine Coefficients

Non-parametric and parametric estimations are very different. Non-parametric methods can cope with lack of prior knowledge about the signal of interest and its distribution. Thus, they can deal with various signals.

However, because non-parametric methods perform estimation regardless of the signal distribution, the quality of denoised speech can be reduced. Moreover, such methods tend to introduce musical noise. To the contrary, parametric methods take a model for the distribution of the signal of interest into account. Therefore, if the model is reasonably good, they can achieve good performance in speech enhancement applications by noticeably improving speech quality. However, they can fail to improve speech intelligibility [120].

Therefore, our objective is to design a method that takes advantage of both the parametric and non-parametric approaches so as to achieve a good trade-off between intelligibility and quality. To this end, we combine an SSBS-based method with a Bayesian statistical estimator. The rationale for this combination is the following. Bayesian statistical estimators of STSA in the DCT domain can be expected to provide good performance in speech enhancement, especially to improve quality without introducing musical noise. Since an SSBS-based approach merely attenuates small coefficients, the idea is to enhance speech quality thanks to a Bayesian estimator. This one, placed downstream an SSBS-based estimator aimed at canceling most of the background noise, should retrieve information on clean speech.

In this respect, the next subsection introduces the Block-SSBS approach. Based on the SSBS estimator, it is designed for audio denoising. Section 5.2.2 then presents STSA-MMSE, a Bayesian estimation of STSA in the DCT domain. The combination of Block-SSBS and STSA-MMSE is described in Section 5.2.3.

### 5.2.1 Non-parametric estimation by Block-SSBS

The original SSBS estimation is a diagonal method which may yield isolated spectral amplitudes and, thus, musical noise in speech enhancement. We can eliminate these isolated points by performing SSBS on blocks of time-frequency neighboring atoms. Such an approach is very similar to that proposed in [134] for denoising signals in the wavelet domain. However, the method we propose has some specific features.

First, it is carried out in the DCT domain for reasons evoked before. Second, speech or audio is not stationary but can be considered stationary on relatively small time-frequency zones. The same may hold for non-stationary noise as well. It follows that we must choose time-frequency zones in which speech and noise can reasonably be expected to be stationary. Such zones are unknown and highly dependent on the signal and noise of interest. The design of algorithms dedicated to the detection of such zones is beyond the scope of this thesis. In this work, we restrict attention to a regular splitting of the time-frequency domain in rectangular time-frequency boxes with the same size $(\Delta T, \Delta F)$, where $\Delta T$ is the number of time frames and $\Delta F$ is the number of frequency bins in each box. Values for $\Delta T$ and $\Delta F$ will hereafter be chosen so that audio signal and noise can acceptably be regarded as stationary in the resulting time-frequency boxes. If the audio and speech distribution in a given box is assumed to be unknown, the general methodology exposed in [134] can be adapted as follows for noisy speech and audio estimation in the DCT domain.

Since the signal distribution in a given box is unknown, the idea is to divide the box into non-overlapping rectangular blocks so that the signal can reasonably be considered to be deterministic

and unknown in each block. To reduce computational cost, we look for blocks with the same size inside a given box. The issue is then to find the optimal blocks size such that the overall estimation error in the box containing these blocks is minimal. On the one hand, when the box is filled with noise only, it makes seem to divide this box into small blocks. In this case, the optimal block size should be the minimum possible block size. On the other hand, when the box contains the signal of interest, it is expected that the deterministic assumption should lead the algorithm to find a relatively big optimal block size. Based on the aforementioned, the following estimation algorithm arises.

#### 5.2.1.1 Block-SSBS gain function

Consider a given box $\mathfrak{B}$ and a block $B$ within this box. As mentioned above, speech is assumed to be deterministic unknown in $B$. For noise estimation by various noise power spectrum estimators, noise is assumed to be centered and Gaussian distributed in the box under consideration, so that the noise variance is supposed to be the same in all blocks within this box. Let $\sigma_X^2(\mathfrak{B})$ stand for the noise power spectrum in $\mathfrak{B}$. Under these assumptions, in block $B$, the estimated *a posteriori* SNR $\widehat{\gamma}$ can be calculated by averaging the instantaneous noisy signal energies $Y^2[m,k]$ divided by the noise variance, so that:

$$\widehat{\gamma} = \overline{Y^2}/\sigma_X^2(\mathfrak{B}), \tag{5.11}$$

with

$$\overline{Y^2} = \frac{1}{|B|}\sum_{(m,k)\in B} Y^2[m,k], \tag{5.12}$$

where $|B|$ is the number of time-frequency points $(m,k)$ within $B$. Since we want to remove isolated time-frequency points, we proceed similarly to Equations. (5.9) and (5.10) for block thresholding, by choosing the SSBS gain function in block $B$ equal to:

$$G_{\tau,\lambda}^B(\widehat{\gamma}) = \frac{1}{1+e^{-\tau(\sqrt{\widehat{\gamma}}-\lambda)}}. \tag{5.13}$$

To implement the above SSBS gain function, we must choose the sizes of the boxes and blocks as well as parameters $\tau$ and $\lambda$.

#### 5.2.1.2 Size of the time-frequency boxes

With the notation introduced above, the larger $\Delta T$, the greater the time delay. Therefore, for real time processing applications, the length $\Delta T$ should be small enough. We have chosen $\Delta T = 8$ (*i.e* 128 ms in the our implementation) as a good trade-off between performance and time-delay. Furthermore, for taking into consideration that non-stationary noise impacts differently distinct frequency bands, we follow [10], which recommends to choose more than 6 bands, linearly spaced within the 8kHz bandwidth, to get good speech quality. Accordingly, and as a good trade-off between performance and computational load, we set $\Delta F = 16$, which corresponds to 8 bands linearly spaced.

#### 5.2.1.3 Time-frequency splitting by SURE

We now address the computation of the optimal block size within a given box $\mathfrak{B}$. The common size of the blocks is a pair henceforth denoted by $(L,W)$. The number of DCT coefficients pertaining to any block is thus $N = L \times W$. The computation of the optimal size $(L^*,W^*)$ for the blocks within a given box $\mathfrak{B}$ can be performed as in [134, 135], by resorting to the

SURE approach derived from Stein's Theorem [3]. However, in contrast to [134,135], the SURE approach is hereafter limited to the estimation of the optimal block size $(L^*, W^*)$ and will not be used to estimate $\lambda$ or $\tau$. Indeed, these two parameters can be evaluated via other means as we shall explain later.

For a given $\tau$ and $\lambda$, consider a box $\mathfrak{B}$. Split this box in $J$ non-overlapping rectangular blocks $B_1, \ldots, B_J$. The overall estimation risk for $\mathfrak{B}$ and its partition in $J$ boxes is thus:

$$\mathbf{R} = \sum_{j=1}^{J} \mathbf{R}_j, \tag{5.14}$$

where

$$\mathbf{R}_j = \sum_{(m,k)\in B_j} \mathbf{E}\left[|\widehat{S}[m,k] - S[m,k]|^2\right]$$

and

$$\widehat{S}[m,k] = G_{\tau,\lambda}^{B_j}(\widehat{\gamma}) Y[m,k]$$

for $(m,k) \in B_j$. Since the SSBS gain function is constant in each block and the blocks are constrained to have the same size, the overall risk depends on the block size $(L, W)$. The SURE Theorem now provides us with an unbiased estimate of $\mathbf{R}_j$. Therefore, we can calculate an unbiased estimate of the overall risk $\mathbf{R}$. It is then possible to look for the block size $(L^*, W^*)$ that minimizes this unbiased estimate of $\mathbf{R}$.

Specifically, we proceed as follows. Let $Y[m,k]$ with $(m,k) \in B_j$ be the $N$ available DCT values in block $B_j$. We can re-arrange these DCT values so as to form the $N$-dimensional random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$. Since the signal of interest is supposed to be deterministic unknown and noise to be Gaussian in $B_j$ with variance $\sigma_X^2$, we assume that

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{S}, \sigma_X^2 \mathbf{I}_N) \tag{5.15}$$

where $\mathbf{S}$ models the unknown clean signal in $B_j$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Now, define $\widehat{\mathbf{S}} : \mathbb{R}^N \to \mathbb{R}^N$ for any $\mathbf{y} \in \mathbb{R}^N$ by

$$\widehat{\mathbf{S}}(\mathbf{y}) = G(\mathbf{y})\mathbf{y}$$

and use Equations (5.11), (5.12) so that:

$$G(\mathbf{y}) = G_{\tau,\lambda}^{B_j}\left(\frac{\|\mathbf{y}\|_2^2}{N\sigma_X^2(\mathfrak{B})}\right)$$

where $\|\cdot\|_2$ stands for the usual Euclidean norm in $\mathbb{R}^N$. Readily, $\widehat{\mathbf{S}}$ is differentiable. Therefore, [136, Section 2] applies and the Stein's unbiased risk estimate of $\mathbf{R}_j$ is given by:

$$\widehat{\mathbf{R}_j}(\mathbf{y}) = -N\sigma_X^2 + \|\mathbf{y} - \widehat{\mathbf{S}}(\mathbf{y})\|_2^2 + 2\sigma^2 \sum_{n=1}^{N} \frac{\partial \widehat{S}_n}{y_n}(\mathbf{y}) \tag{5.16}$$

with $\widehat{\mathbf{S}} = (\widehat{S}_1, \ldots, \widehat{S}_N)$. Some easy algebra leads to (see Appendix C.1):

$$\widehat{\mathbf{R}_j}(\mathbf{y}) = N\sigma_X^2(\mathfrak{B})\left(2G_{\tau,\lambda}^{B_j}(\widehat{\gamma}) - 1\right) + \left(1 - G_{\tau,\lambda}^{B_j}(\widehat{\gamma})\right) \times \left(1 + \tau G_{\tau,\lambda}^{B_j}(\widehat{\gamma})/(N\sqrt{\widehat{\gamma}}) - G_{\tau,\lambda}^{B_j}(\widehat{\gamma})\right) \|\mathbf{y}\|_2^2, \tag{5.17}$$

We can then estimate $\mathbf{R}$ by:

$$\widehat{\mathbf{R}} = \sum_{j=1}^{J} \widehat{\mathbf{R}_\mathbf{j}} \tag{5.18}$$

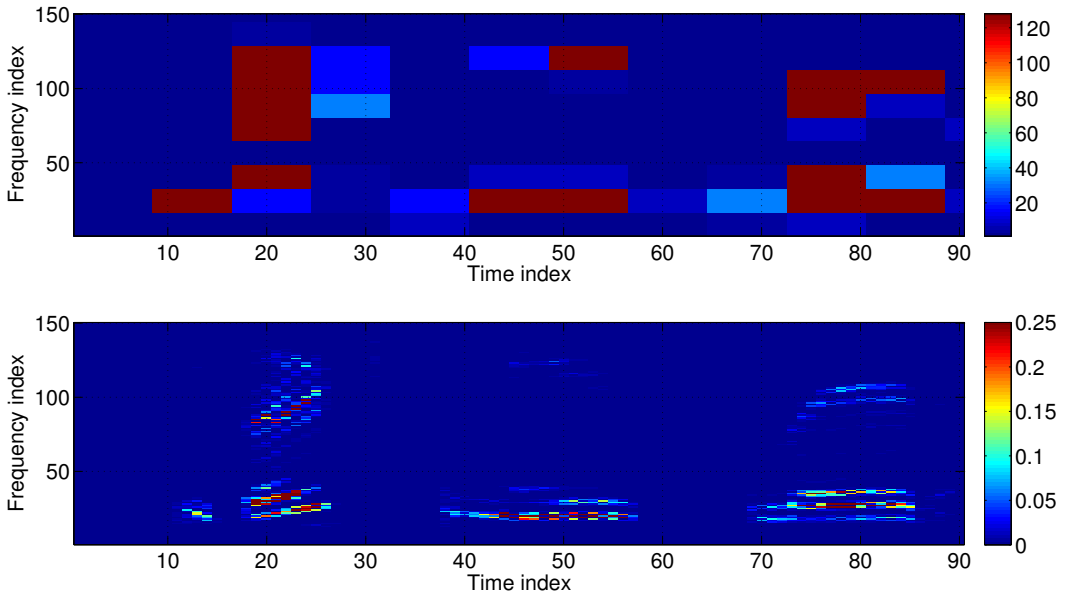



Figure 5.1 – A typical division of the time-frequency domain into boxes and blocks inside boxes shown in sub-figure above. This division is obtained by risk minimization for noisy white speech at SNR = 5dB. The time-frequency domain is first divided into non-overlapping rectangular boxes of size $2^3 \times 2^4$. Then, each box is split into blocks whose size is determined by minimizing the overall risk (5.18) via the SURE approach. We can see that this division matches rather well to the DCT spectrogram displayed by sub-figure below.

It then suffices to carry out an exhaustive search among all possible pairs $(L, W)$ so as to find the pair $(L^*, W^*)$ that minimizes $\widehat{\mathbf{R}}$. Note that the value of $\widehat{\mathbf{R}}_j$ does not only depend on $N$ but also on $L$ and $W$ through $\mathbf{y}$. With respect to the values $\Delta T$ and $\Delta F$ chosen above for the boxes $\mathfrak{B}$, it turns out that the set of all possible sizes $(L, W)$ contains 20 values only, which is easily tractable in practice. In addition, the noise variance $\sigma_X^2(\mathfrak{B})$ within a given box $\mathfrak{B}$ was estimated according to:

$$\sigma_X^2(\mathfrak{B}) = \frac{1}{|\mathfrak{B}|} \sum_{(m,k)\in\mathfrak{B}} \sigma_X^2[m,k], \tag{5.19}$$

where $|\mathfrak{B}|$ is the number of the time-frequency bin $[m, k]$ in $\mathfrak{B}$ and the values $\sigma^2[m, k]$ are the true spectrum if it is known, or estimated values of the spectrum otherwise.

Fig. 5.1 shows an example of box and block tiling obtained by minimization of the overall risk (5.18) on some noisy speech. In this figure, boxes have size $8 \times 16$ and the color of each box corresponds to the size determined by the SURE approach for the blocks within this box. For example, the rectangular box that spans from frames 41 to 48 and from frequency bins 16 to 32 is divided into blocks with size 128. Note that, as expected, the SURE approach yields a block size equal to the box size in time-frequency zones occupied by speech and noise. This is normal since, within a noisy box, speech is homogeneous. In contrast, in boxes where noise is only present, the SURE approach returns smaller block sizes because variations of speech inside these boxes require a finer analysis. This was expected as well.

#### 5.2.1.4 RDT-based selection of Block-SSBS parameters $\tau$ and $\lambda$

For speech enhancement applications, the two parameters $\tau$ and $\lambda$ in (5.8) are also key elements for controlling the performance of the proposed method and reaching the desired trade-off between signal distortion and noise reduction. As mentioned above, it is possible to estimate $\tau$ and $\lambda$ via the SURE approach. Such a possibility has not been tested in this work for the following reasons.

On the one hand, the SURE approach is particularly relevant to estimate local parameters, whereas the authors' feeling and experience with speech and images [137] suggests that $\tau$ can be adjusted as a global parameter. Indeed, $\tau$ is basically a slope which may vary from one signal to another but a global or average value for this parameter is not really detrimental. Basically, $\tau$ controls the level of attenuation applied by the SSBS gain function to the noisy signal. For a fixed $\lambda$ and when $\tau$ tends to infinity, the SSBS gain function behaves like hard thresholding gain function or binary masking. Binary masking or channel selection improves successfully speech intelligibility [1, pp.615]. Thus, shrinkage by the SSBS gain function can also be expected to bring some gain in speech intelligibility. Moreover, a large $\tau$ will affect speech quality. Some informal tests then led to choose

$$\tau \approx 4/\lambda, \tag{5.20}$$

as recommended in [2] for images, without resorting to any statistical approach.

As far as $\lambda$ is concerned, the question is a bit more intricate because $\lambda$ plays the role of a threshold that can be used to make a decision on speech presence or absence. This threshold may therefore vary significantly in the time-frequency domain with respect to the type of speech signal under observation. Thus the idea to estimate this threshold in each block, once $(L^*, W^*)$ has been calculated. Additionally, it is desirable to keep some control on the estimation performance, which is not actually feasible via the SURE approach. Thence, the interest of the non-parametric approach introduced below has the advantage of ensuring that the proposed choice for $\lambda$ is optimal in the particular sense of guaranteeing the false alarm probability of erroneously deciding that significant speech is present.

The method we propose below is based on the following rationale. Parameter $\lambda$ influences shrinkage performance by SSBS gain function because it affects the level of noise reduction applied to the noisy DCT coefficients. Although the SSBS gain function is smoother than hard thresholding, parameter $\lambda$ must however be carefully chosen to enhance speech quality. Indeed, suppressing too many speech components for reducing noise will necessary induce loss of speech quality. Otherwise said, when one aims at improving not only speech quality but also speech intelligibility, missing some important speech-carrying time-frequency channels may be more detrimental to speech enhancement than conserving more noise-only channels than strictly required. This favors the choice of small values for $\lambda$. On the other hand, the smaller $\lambda$, the smaller the signal distortion and musical noise, but the larger the background noise. Therefore, we cannot choose a too small value for $\lambda$. Hence, the necessity to achieve a trade-off between speech quality and denoising. A mean to achieve such a trade-off is to control the denoising by taking the outcome of some speech detector into account.

We follow a similar strategy by choosing $\lambda$ such that DCT coefficients with amplitudes above $\lambda$ with high probability pertain to relevant speech signal components, whereas DCT coefficients below $\lambda$ are more likely to be components of noise only or noisy speech coefficients that can be safely discarded. Since we accept that observations with amplitudes below $\lambda$ may contain information merely attenuated by the SSBS function, the choice of $\lambda$ is not derived hereafter from a detection problem as in [63, 137] for denoising images by wavelet shrinkage. Instead, we resort to the random distortion testing (RDT) approach [4].

Basically, with the notation and hypotheses of (5.15), the RDT approach amounts to testing whether $\|\mathbf{S}\|_2 \leqslant \delta$ or not when we observe $\mathbf{Y}$, where $\delta$ is a tolerance that is specified by the application. For better understanding the sequel, it must be noticed that this binary hypothesis test is invariant by orthogonal transform, in the sense that it remains identical if $\mathbf{Y}$ is transformed by any orthogonal transform of $\mathbb{R}^N$. This basically derives from the properties of the Gaussian distribution.

Let us decide that $\|\mathbf{S}\|_2 \leqslant \delta$ if $\|\mathbf{Y}\|_2 \leqslant \eta_\alpha(\delta)$ and that $\|\mathbf{S}\|_2 > \delta$ otherwise, where $\eta_\alpha(\delta)$ is the unique solution in $x$ to the equation $Q_{N/2}(\delta, x) = \alpha$ [1], where $Q_{N/2}(\cdot, \cdot)$ stands for the Generalized Marcum function [4] . According to [4, Proposition 2], the thresholding test satisfies several optimality properties with respect to the inherent invariant features of the problem. In particular, it is Uniformly Most Powerful Invariant (UMPI) with size $\alpha$ among all of the tests with level $\alpha$ that are invariant by orthogonal transforms. The reader is asked to refer to [4] for further details.

According to these properties, the threshold $\eta_\alpha(\delta)$ makes it possible to control the false alarm probability via $\alpha$ and guarantees optimal power or correct decision probability, without prior knowledge on the signal of interest, an appealing feature for speech enhancement. For homogeneity of the physical quantities in Equations (5.11) (5.12) and (5.13), we choose

$$\lambda = \eta_\alpha(\delta)/\sqrt{N}. \tag{5.21}$$

To clarify the use of RDT theory in speech denoising, Fig. 5.2 shows spectrograms when denoising is performed by SSBS on blocks and two different levels $\alpha$ are tested. The smaller the $\alpha$, the smaller the background noise. However, with $\alpha = 0.05$, some important frequency-time atoms are ignored (for instance, see the rectangle in Fig.5.2 (c)).

### 5.2.2 MMSE STSA in the DCT domain

Similarly to standard MMSE-based methods in the DFT domain [27], we compute the MMSE Bayesian estimator of the absolute value of the DCT clean signal coefficients. To this end, we need a model for the clean speech distribution. Motivated by the central limit theorem when the frame length is large enough, we assume that DCT coefficients of the clean signal have Gaussian prior density. Based on this assumption, the probability of each event $\phi_S = 1$ or $\phi_S = -1$ is equal to $1/2$. Thus, the probability density function of the amplitude of a given clean speech DCT coefficient $A_S$ has half-normal distribution:

$$f_{A_S}(a) = \frac{\sqrt{2}}{\sigma_S\sqrt{\pi}} \exp\left(-\frac{a^2}{2\sigma_S^2}\right) \mathbb{1}_{[0,\infty)}(a), \tag{5.22}$$

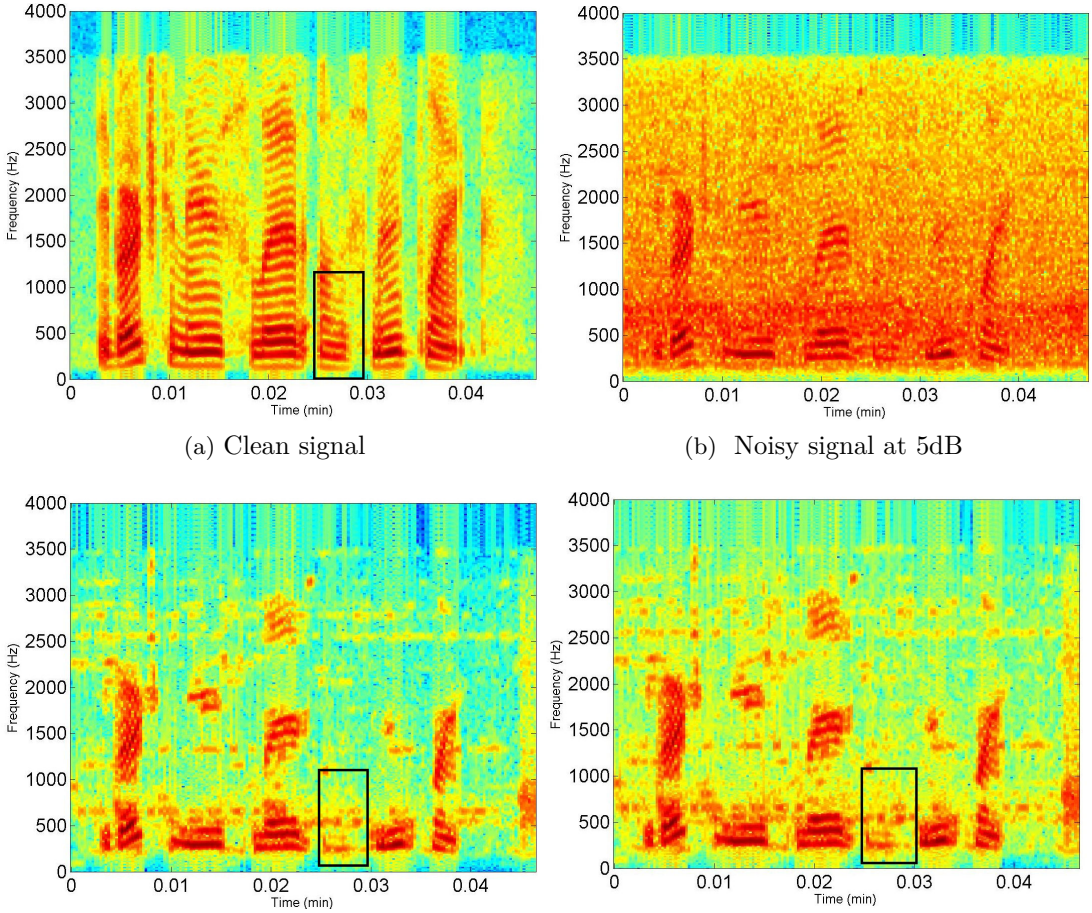
where $\mathbb{1}_{[0,\infty)}$ is the indicator function $\mathbb{1}_{[0,\infty)}(x) = 1$ if $x \geq 0$ and $\mathbb{1}_{[0,\infty)}(x) = 0$ otherwise. Moreover, noise is assumed to be Gaussian. Thus,

$$f_{Y|A_S}(y|a) = P(\phi_S = 1) f_{Y|A_S=a,\phi_S=1}(y) + P(\phi_S = -1) f_{Y|A_S=a,\phi_S=-1}(y) \tag{5.23}$$

so that $f_{Y|A_S}$ can be rewritten as:

$$f_{Y|A_S=a}(y) = \frac{1}{2\sigma_X\sqrt{2\pi}} \times \left(\exp\left(-\frac{(y-a)^2}{2\sigma_X^2}\right) + \exp\left(-\frac{(y+a)^2}{2\sigma_X^2}\right)\right). \tag{5.24}$$

[1] For any $x \in [0, \infty)$, $Q_{N/2}(\delta, x) = 1 - \mathbb{F}_{\chi^2_N(\delta^2)}(x^2)$ is the cumulative distribution function of the square root of non-central chi-square distribution with $N$ degrees of freedom and non-central parameter $\delta^2$.




(a) Clean signal

(b) Noisy signal at 5dB

(c) $\alpha = 0.01$

(d) $\alpha = 0.15$

Figure 5.2 – Spectrogram of clean speech (a), corresponding noisy car speech (b) and denoised speech by SSBS with two different levels: level = 0.01 (c) and level = 0.15 (d)

The Bayesian estimator of the speech short-time spectral amplitude (STSA) is a map $\psi$ of $\mathbb{R}$ into $[0, \infty)$ aimed at minimizing the mean-square error between the estimated and the true amplitude. It is known to be the conditional mean and is given for every $y \in \mathbb{R}$ by [124]:

$$\psi(y) = \frac{\int_0^\infty a f_{Y|A_S=a}(y) f_{A_S}(a) da}{\int_0^\infty f_{Y|A_S=a}(y) f_{A_S}(a) da}. \tag{5.25}$$

Given the DCT coefficient $Y$, the estimate $\widehat{A}_S$ of $A_S$ provided by this estimator is:

$$\widehat{A}_S = \psi(Y), \tag{5.26}$$

Injecting (5.22) and (5.24) into (5.25) yields:

$$\psi(y) = \frac{\int_0^\infty a \left[\exp\left(\frac{ay}{\sigma_X^2} - \frac{a^2}{2\sigma^2}\right) + \exp\left(-\frac{ya}{\sigma_X^2} - \frac{a^2}{2\sigma^2}\right)\right] da}{\int_0^\infty \left[\exp\left(\frac{ay}{\sigma_X^2} - \frac{a^2}{2\sigma^2}\right) + \exp\left(-\frac{ya}{\sigma_X^2} - \frac{a^2}{2\sigma^2}\right)\right] da}, \tag{5.27}$$

where

$$\sigma = \frac{\sigma_S \sigma_X}{\sqrt{\sigma_X^2 + \sigma_S^2}}. \tag{5.28}$$

As in [27], we can compute the gain function. By a direct computation from (5.27) (see Appendix C.2) or by using [138, Equations 3.462.1, 9.254.1, 9.254.2] successively, we get:

$$\psi(Y) = G(\xi, \gamma) A_Y, \tag{5.29}$$

where the gain function $G(\xi, \gamma)$ is given by:

$$G(\xi, \gamma) = \frac{\sqrt{\nu}}{\gamma} \frac{\sqrt{2} + \sqrt{\pi\nu}\,\mathrm{erf}(\sqrt{\nu/2})\exp(\nu/2)}{\sqrt{\pi}\exp(\nu/2)}, \tag{5.30}$$

where

$$\nu = \frac{\xi}{1+\xi}\gamma = \frac{\sigma^2 A_Y^2}{\sigma_X^4}. \tag{5.31}$$

and $\mathrm{erf}(\cdot)$ is the error function. This gain function depends on the *a priori* SNR $\xi$ and the *a posteriori* SNR $\gamma$. The *a posteriori* SNR is directly given by the observed amplitude $A_Y$. In contrast, the *a priori* SNR is unknown. This variable $\xi$ can be estimated via the decision directed approach [27]:

$$\xi[m,k] = \beta \frac{\widehat{A}_S^2[m-1,k]}{\sigma_X^2[m-1,k]} + (1-\beta)(\gamma[m,k]-1)_+, \tag{5.32}$$

where $0 < \beta < 1$ is the smoothing parameter and $\widehat{A}_S[m-1,k]$ is the estimated STSA at the previous frame.

Figure 5.3 (a) displays $G(\xi, \gamma)$ as a function of the *a posteriori* SNR $\gamma$ for fixed values of $\xi = 5, -5, -10$ dB. Alternatively, this gain function $G(\xi, \gamma)$ is plotted as a function of $\xi$ for fixed values of $\gamma = 5, -5, -10$ dB in Fig 5.3 (b). The gain function of the STSA MMSE in the DFT domain [27] is also a function of the same *a priori* and a posteriori SNRs. Therefore, for comparative purpose, Fig. 5.3 (a) and (b) also display the gain functions in the DFT domain with dashed lines in the same settings. In the two cases, the gain function of the STSA estimator in the DCT domain is shifted down by 2 dB with respect to the gain function of the STSA estimator in the DFT domain. This suggests that denoising in the DCT domain tends to reduce more the background noise.

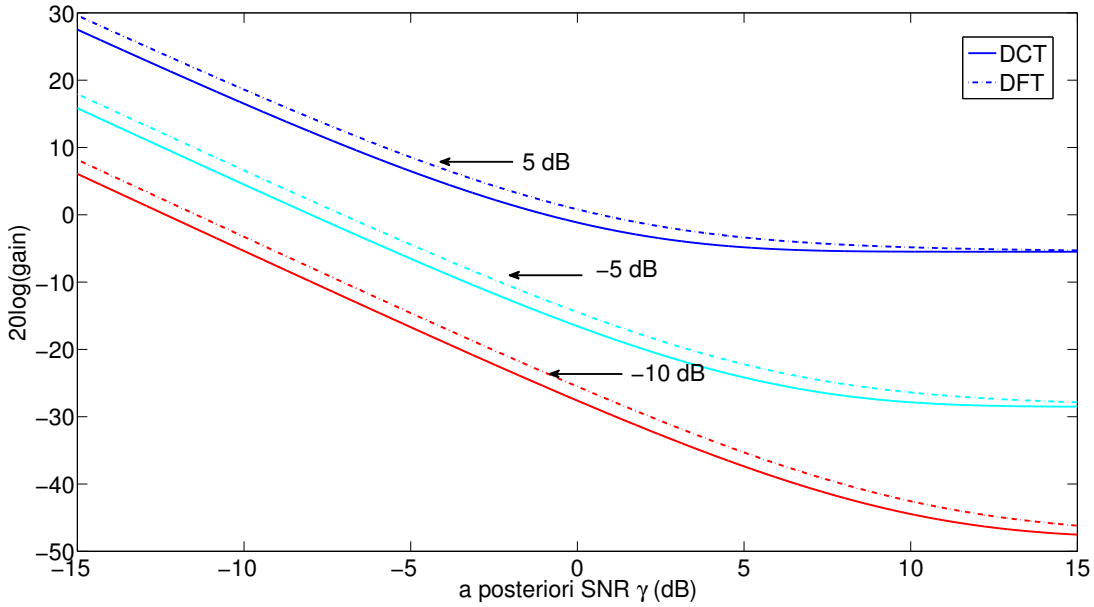
### 5.2.3 Combination method

After Block-SSBS, the transformed signal and noise are assumed to be Gaussian distributed. We then apply Bayesian STSA-MMSE. By doing so, prior knowledge on speech is incorporated to improve speech quality beyond speech intelligibility improvement achieved by Block-SSBS. Whence the following suggested combination of these parametric and non-parametric methods, which is summarized by Figure 5.4.

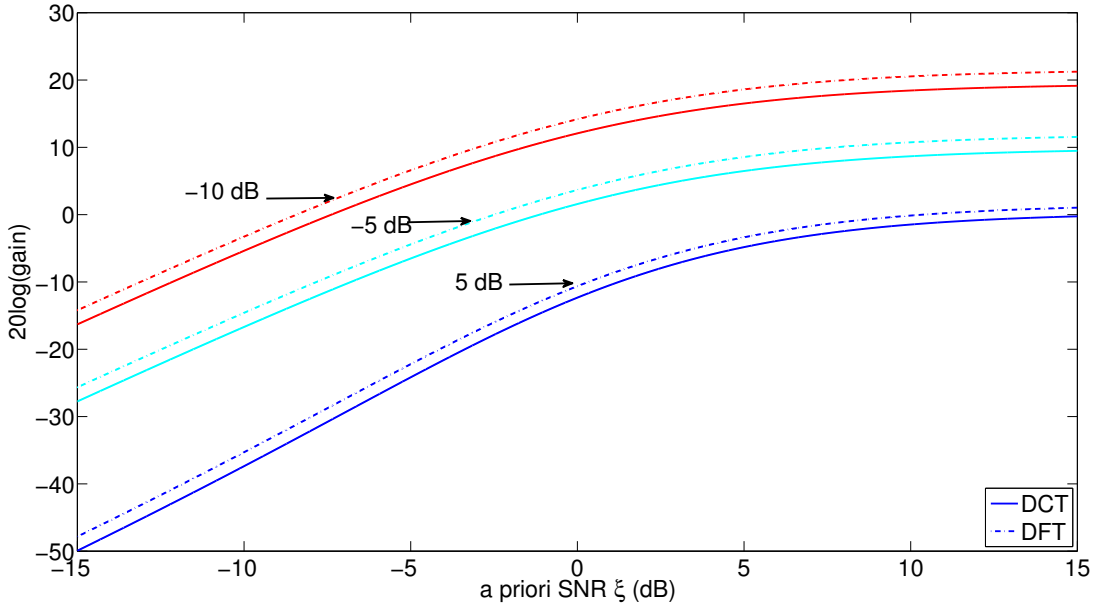
(i) **Signal decomposition**: The observed signal is segmented and transformed using DCT.

(ii) **Noise reduction**: The transformed coefficients are shrunk by the block SSBS gain function $G^B_{\tau,\lambda}(\widehat{\gamma})$ in each block $B$. Given a DCT coefficient $Y$ in this block, the estimate $\widehat{A}^*_S$ of the amplitude $A_S$ of the clean signal is calculated by:

$$\widehat{A}^*_S = G^B_{\tau,\lambda}(\widehat{\gamma}) A_Y \tag{5.33}$$




(a) Fixed $\xi$




(b) Fixed $\gamma$

Figure 5.3 – Gain functions of the STSA-MMSE estimators in the DCT and DFT domains as functions of $\xi$ and $\gamma$. In Fig. 5.3 (a) the gain functions vary with $\gamma$ at fixed values of $\xi$ whereas, in Fig. 5.3 (b), the gain functions vary with $\xi$ at fixed values of $\gamma$.

(iii) **Refined Estimation**: The Bayesian MMSE statistical estimator is applied to the coefficients shrunk by Block-SSBS so that the final estimate of the clean signal amplitude is:

$$\widehat{A}_S = G(\xi, |\widehat{A}_S^*|^2/\sigma_X^2)\widehat{A}_S^*, \tag{5.34}$$

where $G$ is the gain function of the STSA-MMSE Bayes estimator given by (5.30) and $\xi$ is calculated by the decision-directed approach (5.32).

(iv) **Signal reconstruction**: The enhanced signal is finally obtained from the estimated STSA $\widehat{A}_S$ and the noisy phase $\phi_Y$ by the overlap-add method [1].

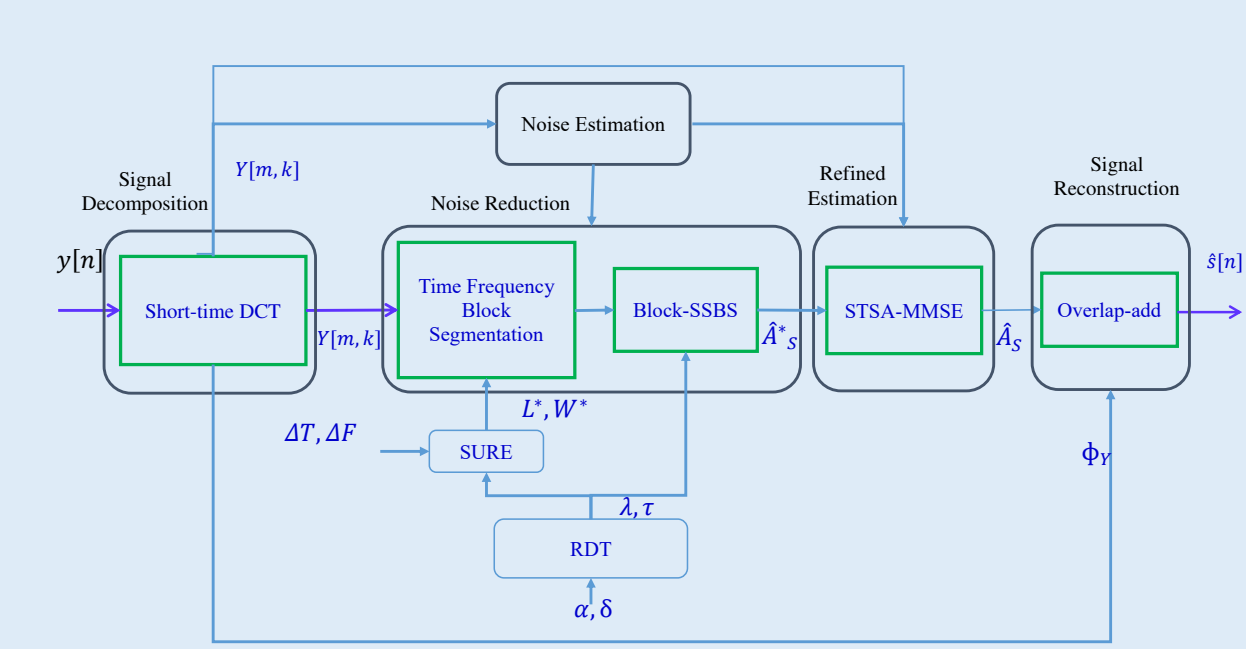



Figure 5.4 – Block overview of combination method where $y[n]$ is the input and $\Delta T$, $\Delta F$, $\delta$ and $\alpha$ are the parameters of the proposed combination method.

## 5.3 Experimental Results

### 5.3.1 Parameter adjustment

For Block-SSBS, the tolerance $\delta$ and the level $\alpha$ were chosen by maximizing the segmental SNR (SSNR) on sentences randomly chosen and corrupted by car noise at SNR level of 5 dB. These preliminary tests led us to choose $\alpha = 0.05$ and $\delta = 4$ dB for Block-SSBS.

### 5.3.2 Speech data set

The performance of all the methods were evaluated in two scenarios. In the first scenario, denoising is performed by using the reference noise power spectrum. This one is simply the theoretical power spectrum if noise is stationary. Otherwise, the reference noise power spectrum in a given bin $m$ is estimated as in [111] by:

$$\sigma_X^2[m,k] = \mu\sigma_X^2[m-1,k] + (1-\mu)A_X^2[m,k], \tag{5.35}$$

where $\mu = 0.9$. In the second scenario, for all the methods, the noise power spectrum was estimated using the E-DATE algorithm introduced in [89].

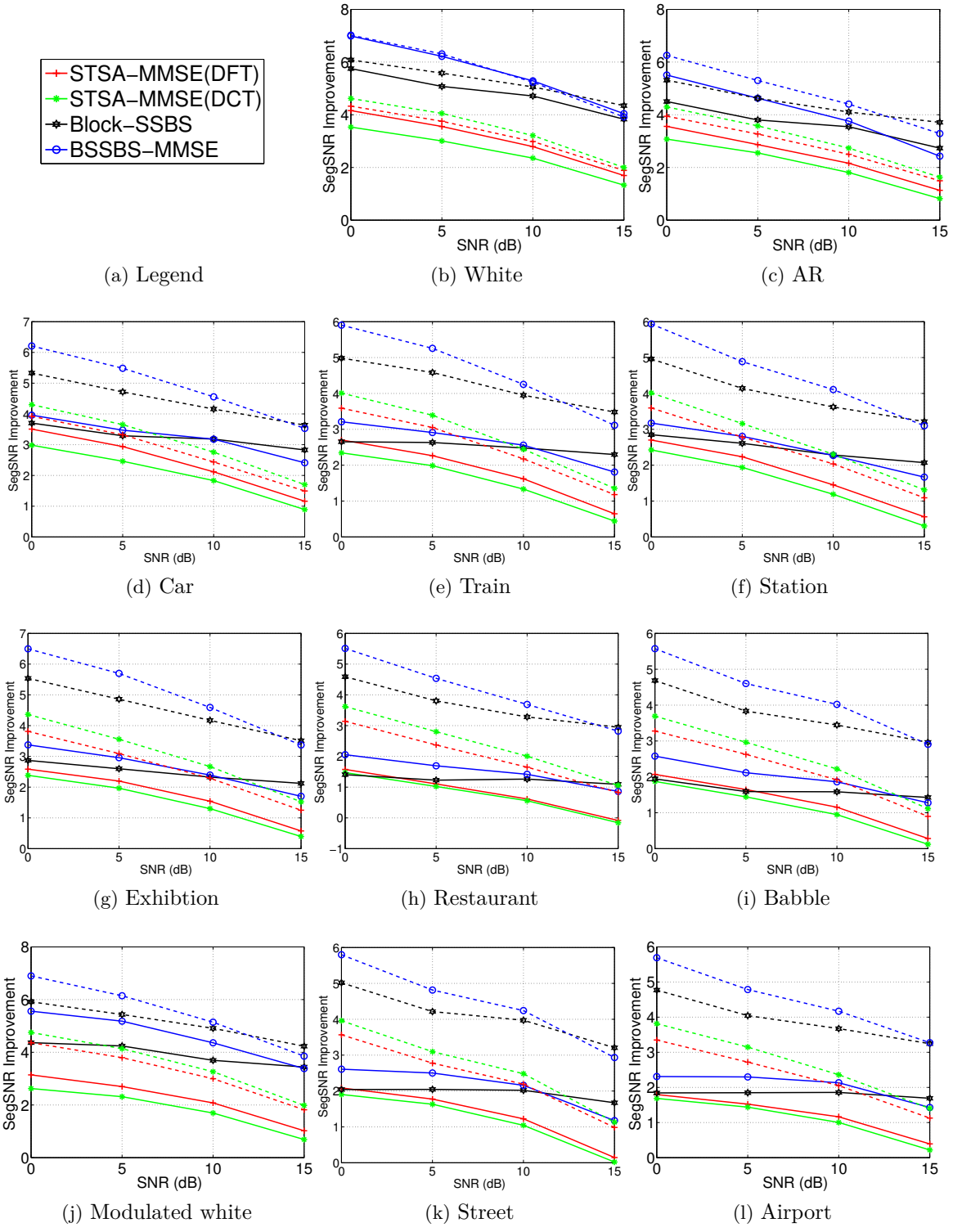
Experiments have been first conducted on the NOIZEUS database to evaluate the performance of the proposed methods in speech enhancement. The NOIZEUS database contains speech sentences degraded by noise environments from the AURORA database at various levels, namely 0, 5, 10 and 15 dB. The speech signals are sampled at 8 kHz. The noisy signals were segmented and windowed into 32-ms duration frames, and then transformed using STCT for STSA-MMSE(DCT), Block-SSBS, and BSSBS-MMSE and using STFT for STSA-MMSE(DFT) with 50% overlapped Hamming windows. STSA-MMSE(DFT) proposed in [27] is the referent method. This MMSE-based method is simple to implement and generally considered as a good reference method. As for STSA-MMSE(DCT) and Block-SSBS, this method is based on the MMSE criterion

#### 5.3.2.1 Speech objective Test

Speech quality and intelligibility were evaluated using objective quality as well as intelligibility criteria. Speech quality was assessed using the standard SSNR, SNRI, and the overall quality of speech criteria MARSovrl criterion. Speech intelligibility was first estimated by STOI criterion. A logistic function is applied to map the STOI measure to a meaningful intelligibility score [87].

The results are displayed in Figs. 5.5-5.8. Each figure has the same legend where STSA-MMSE(DFT), STSA-MMSE(DCT), Block-SSBS, and BSSBS-MMSE denote the STSA-MMSE in the DFT domain, the STSA-MMSE, the Block-SSBS and the combination of Block-SSBS and STSA-MMSE(DCT) methods in the DCT domain and designed by the red, green, black and blue lines with the circle, x-mark, plus and star makers, respectively as displayed in Fig. 5.5b. Moreover, the all measures obtained with the reference noise power spectrum and with B-E-DATE methods are drawn by dashed and solid lines, correspondingly. All algorithms have been benchmarked at four SNR levels and against various noise models, namely white Gaussian noise (White), 2nd-order auto-regressive (AR) noise, 4 usual types of quasi-stationary noise (car, train, station and street) and 4 kinds of non-stationary noise (airport, exhibition, restaurant and babble). AR noise was obtained by filtering white Gaussian noise by the discrete filter with transfer function $1/(1 + az^{-1})$ and $a = 0.5$. Fig. 5.5 shows the segmental SNR improvement obtained with the different denoising methods employing the reference noise power spectrum (dashed lines) and the noise power spectrum estimated by E-DATE (solid lines). We firstly consider the scenario where the reference noise power spectrum is used. The result for white and AR noises are given in Fig. 5.5b and 5.5c, respectively. The proposed BSSBS-MMSE method yields the highest segmental SNR improvement at three levels including 0, 5 and 10 dB, whereas the non-parametric Block-SSBS method achieves the best SSNR at 15dB. For non-stationary with slowly-varying noise spectrum like car, train, station and street noises, similar results are obtained: BSSBS-MMSE provides the largest SSNR improvement at the low and medium SNRs, while Block-SSBS performs better than BSSBS-MMSE at 15 dB. But the difference is small, as shown by Figs. 5.5d-5.5k. Figs. 5.5g- 5.5i present SSNR improvement for non-stationary noises. In this case, BSSBS-MMSE yields the best score at low and medium SNRs. At high SNR level, Block-SSBS and BSSBS-MMSE both lead to the same best measure. Remarkably, in comparison to STSA-MMSE in the DFT domain, the BSSBS-MMSE method has a gain of around $2.5 - 3$dB in this case.

The SSNR improvement, obtained in the more realistic case where the noise power spectrum is estimated by E-DATE, is also shown in Fig. 5.5 by solid lines. In this case, the BSSBS-MMSE method still yields the best score for all noise types from 0 dB to 10 dB, whereas Block-SSBS achieves the highest score at 15 dB. The gain now is about $0.5 - 1$ dB. Such results




(a) Legend (b) White (c) AR

(d) Car (e) Train (f) Station

(g) Exhibtion (h) Restaurant (i) Babble

(j) Modulated white (k) Street (l) Airport

Figure 5.5 – Speech quality evaluation after speech denoising: improvement of segmental SNR criterion. The result is displayed from stationary noise (White,AR) to quasi-stationary noise (train, car and station) and up to non-stationary noise (restaurant, exhibition, babble, street, modulated and airport). The legend is shown by Figure 5.5a.

basically relate to the sensitivity of STSA-MMSE and Block-SSBS in the DCT domain to noise estimation errors. In comparison with Fig. 5.5, STSA-MMSE(DCT), Block-SSBS and BSSBS-MMSE undergo performance loss by using E-DATE for noise power spectrum estimation. This loss is negligible for white and AR Gaussian noise and around 3 dB for other types of noise. Generally, although BSSBS-MMSE is sensitive to noise estimation errors, it keeps on yielding the best SSNR improvement. SNRI shown by Figure 5.6 confirms the performance of BSSBS-

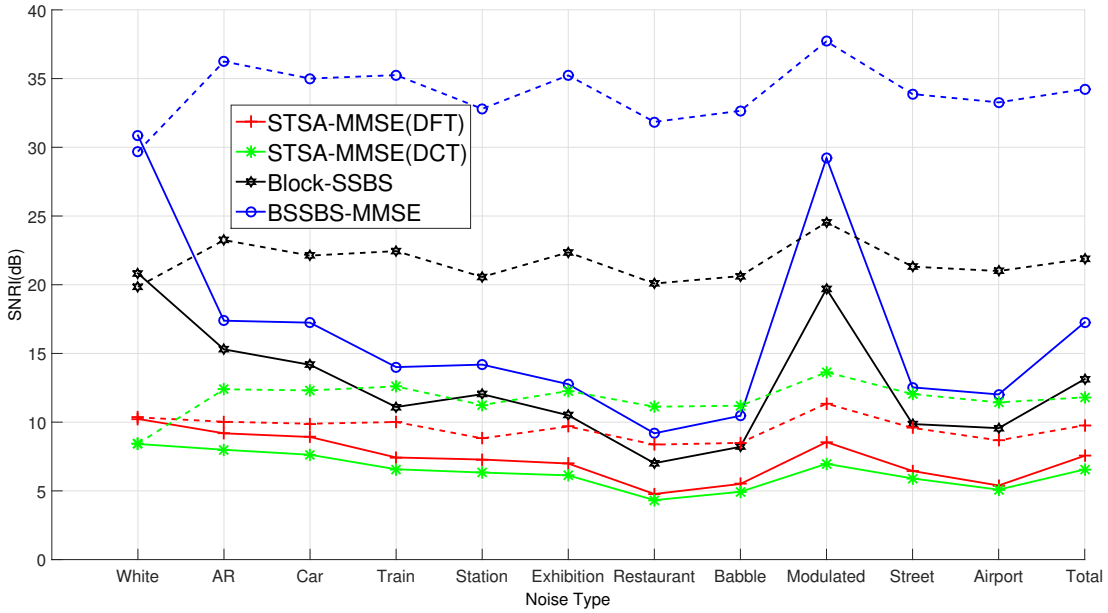
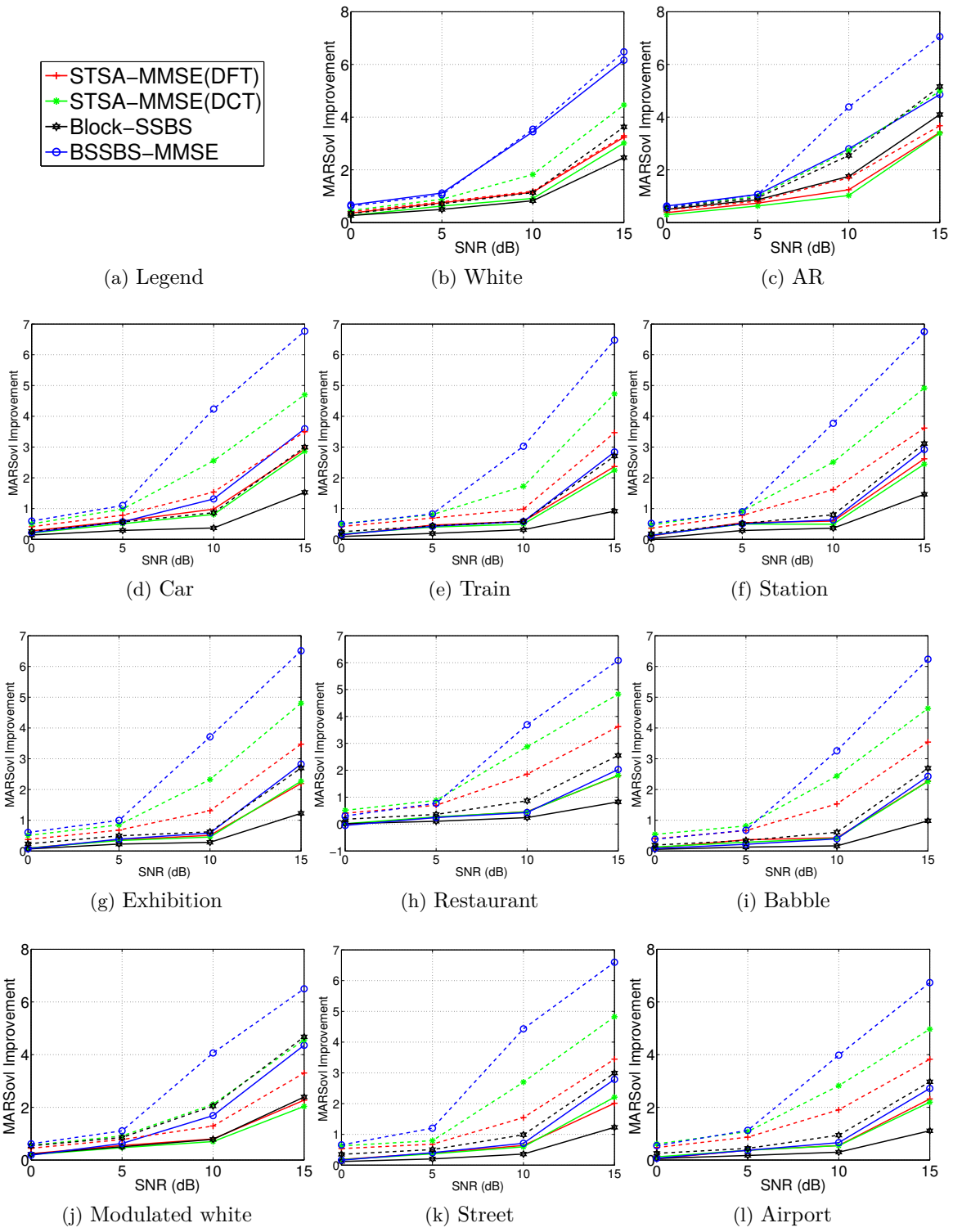



Figure 5.6 – SNRI with various noise types for all methods in two scenarios where the reference noise power spectrum is used or not. The legend is the same than in Fig. 5.5b.

MMSE in terms of the SNR where BSSBS-MMSE remains yield the best score in all situations. In the first scenario, compared to STSA-MMSE(DFT), the gain is at least 20 dB in the case of white noise and can be around 25 dB in the other kind of noise. The average gain indicated by label total reaches around 24.5 dB. Because BBSSB-MMSE is sensitive to noise estimation error, in the second scenario, the gain is around $10-20$ dB. Note that, like SSNR improvement score, STSA-MMSE(DCT) leads a better score in the first scenario and a lower score in the second scenario than STSA-MMSE(DFT).

In term of speech quality estimated by MARS overall, Fig. 5.7 (dashed lines) shows the improvement score when the reference noise power spectrum is used. With small *a priori* information about speech, the Block-SSBS method yields the lowest score in all situations. This remains true when E-DATE is employed to estimate the noise power spectrum (see the solid lines in Fig. 5.7). By taking into account the speech distribution at the refined estimation step, the good performance of the BSSBS-MMSE method is confirmed by MARS improvement measure obtained in the case of white Gaussian noise and AR noise (see Figs. 5.7b and 5.7c). For all types and levels of noises, BSSBS-MMSE provides the best MARS scores, except for babble and restaurant noises at low noise levels. However, in these cases, the MARS scores of the BSSBS-MMSE method is not significantly different from the best ones obtained by STSA-MMSE(DCT) (see Fig. 5.7d-5.7i).




(a) Legend (b) White (c) AR

(d) Car (e) Train (f) Station

(g) Exhibition (h) Restaurant (i) Babble

(j) Modulated white (k) Street (l) Airport

Figure 5.7 – Speech quality evaluation after speech denoising: improvement of MARSovrl composite criterion. The legend is shown in Figure 5.7a.

When combining denoising with E-DATE noise estimation, the MARS overall improvement is presented in Fig. 5.7 by solid lines. It turns out that the speech quality obtained by STSA-MMSE(DFT) is not really affected by errors in the noise spectrum estimation (compare the dashed lines to solid line in Fig. 5.7). In contrast, for non-stationary noise, the sensitivity to noise estimation errors mentioned above for methods carried out in the DCT domain is greater. Thereby, BSSBS-MMSE, STSA-MMSE(DCT) and STSA-MMSE(DFT) yield very similar results for this type of noise, especially at low and medium SNR levels (see Figs. 5.7e-5.7i). For stationary noise, Figs. 5.7b and 5.7c show that the BSSBS-MMSE method remains better than the other methods, without real performance loss due to noise spectrum estimation by E-DATE.

In terms of speech intelligibility, the intelligibility score (IS) obtained by mapping the STOI measure is shown in Figs. 5.8. At high SNR, the scores obtained by all the methods are not significantly different. At low SNR and in presence of AR and white Gaussian noise, Block-SSBS and BSSSBS-MMSE behave similarly in the two scenarios (with and without reference noise power spectrum). For non-stationary noises and when using the reference noise spectrum, BSSBS-MMSE yields the highest scores. In comparison with the worst results, the gain is around $10-15\%$. When noise spectrum is estimated by E-date, the best performance is attained by Block-SSBS. In comparison to the STSA-MMSE(DFT) method, BSSBS-MMSE provides often better score in the case of non-stationary noise with a gain $5-15\%$, whereas Block-SSBS method leads to the improvement $10-20\%$ (see Figs. 5.8h, 5.8l and 5.8i).

**Summarily, the BSSBS-MMSE method achieves a better trade-off between speech quality and intelligibility than the other methods.**

#### 5.3.2.2 Subjective Test

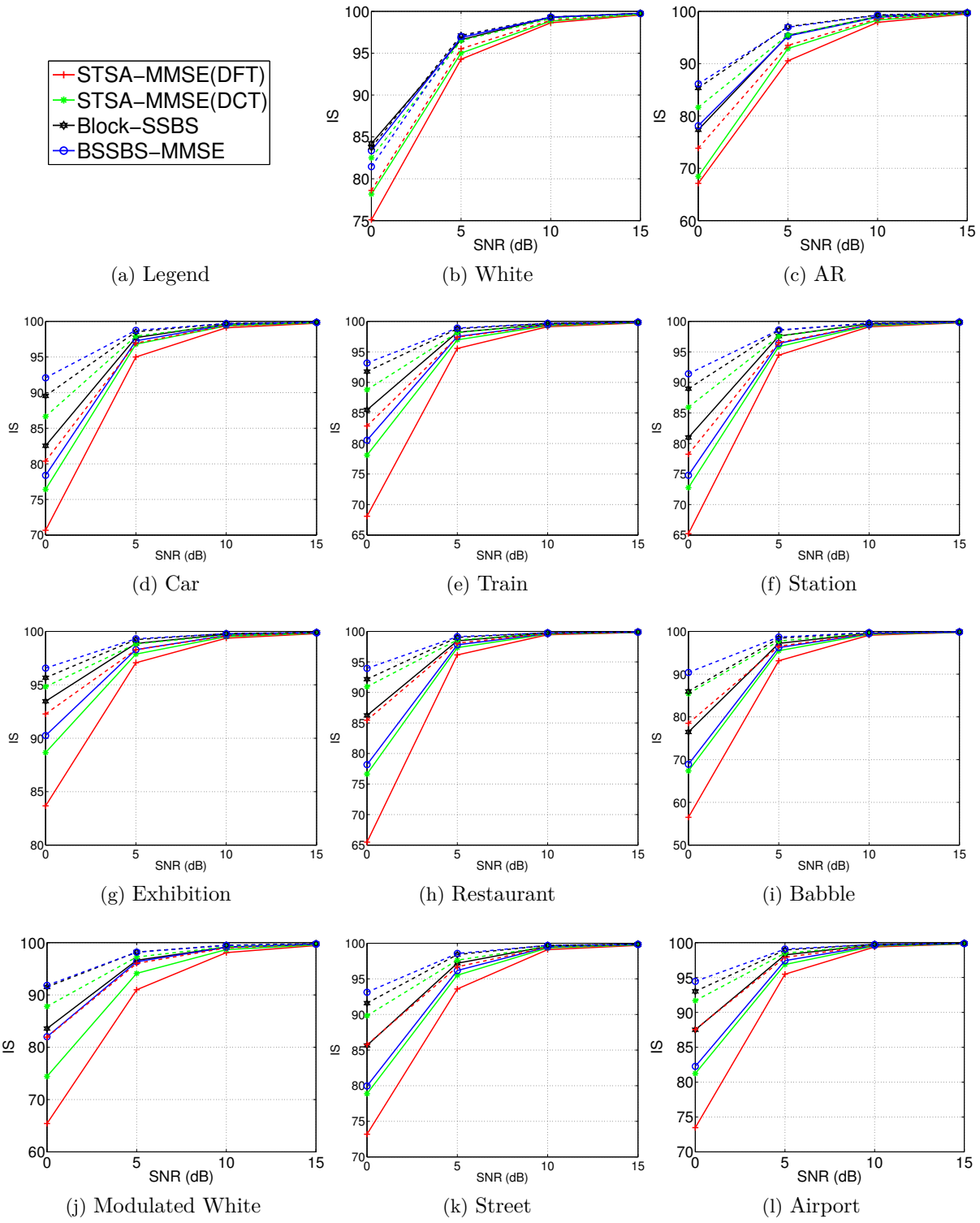
We have also conducted a subjective test for comparing two methods namely BSSBS-MMSE and STSA-MMSE(DFT). This test was performed with ten listeners using the mean opinion score (MOS) method recommended by the IEEE Subcommittee [88]. Two SNR noise levels namely 5dB and 10 dB and three different kinds of noise from stationary white noise to speech-like non-stationary babble noise up to fast-changing non-stationary street noise are used.

The results obtained with different noise types and SNR values are compiled in Table 5.1. For each scenario, noise types and considered noise level, the best results are shown in boldface. In

Table 5.1 – MOS obtained with BSSBS-MMSE and STSA-MMSE(DFT) in the two scenarios

| | | Reference noise | | E-DATE | |
|---|---|---|---|---|---|
| | | STSA-MMSE | SBSSB-MMSE | STSA-MMSE | SBSSB-MMSE |
| White | 5 dB | 2.35 | **2.4** | 2.15 | **2.31** |
| | 10 dB | 2.80 | **3.00** | **2.78** | **2.78** |
| Street | 5 dB | 2.1 | **2.68** | 1.73 | **2.35** |
| | 10 dB | 3.08 | **3.2** | 2.73 | **2.85** |
| Babble | 5 dB | 2.53 | **2.88** | **2.05** | 1.75 |
| | 10 dB | 3.23 | **3.25** | 2.28 | **2.98** |

the first scenario, BSSBS-MMSE yields a better score than STSA-MMSE(DFT) in all situations. In the second scenario, at 5 dB SNR level, for babble noise, MMSE-STSA leads to better scores, whereas, for the other case, BSSBS-MMSE provides similar or better results than the STSA-MMSE(DFT).




(a) Legend (b) White (c) AR

(d) Car (e) Train (f) Station

(g) Exhibition (h) Restaurant (i) Babble

(j) Modulated White (k) Street (l) Airport

Figure 5.8 – Speech intelligibility evaluation after speech denoising: Intelligibility score by mapping STOI criterion.

### 5.3.3 Music data set

For demonstrating the robustness and universality of our algorithms, we also assessed our proposed method on music audio signal recorded from the solo wind instrument (oboe). Experiments have been realized with 6 types of noise (White, car, train, babble, street and airport noise) and at 4 levels (0, 5, 10 and 15 dB). The music signal is sampled at $11kHz$ and transformed into time-frequency domain by using 50ms frames with 50% overlapped Hamming windows. The reference noise power spectrum as given in section above is also used.

#### 5.3.3.1 Objective criterion

In order to evaluate the algorithm performance, we used the SSNR criterion for the objective test. The results are shown in the Figure 5.9. For white noise displayed by Figure 5.9a, at low SNR levels, Block-SSBS and BSSBS-MMSE yield similar and better measures than STSA-MMSE(DFT) and STSA-MMSE(DCT) with gains of 3 dB. At high SNR levels, Block-SSBS leads to the best score with a gain of 4.5 dB in comparison to STSA-MMSE(DFT). The same behavior for car, train and airport noise is seen in Figures 5.9b, 5.9c and 5.9e. For street and babble noise, in all situations, Block-SSBS achieves the best score (see Figures 5.9d and 5.9f), albeit with lower gain compared to BSSBS-MMSE. Note that as for speech data set, when using the reference noise, STSA-MMSE(DCT) outperforms slightly the STSA-MMSE(DCT).

**In conclusion, Block-SSBS gives the best score in terms of SSNR for most situations, with a gain in range $0.5 - 6$ dB in comparison to the reference method. Although BSSBS-MMSE provides a lower score than Block-SSBS, which is a non-parametric method, there is not much difference.**

#### 5.3.3.2 Subjective criterion

For further analyzing the robustness of the semi-parametric approach, we have also performed subjective test on the same music database with 15 raters some of them being music players. We consider only the white noise (5 and 10 dB) where its power spectral can be theoretically determined at two SNR levels. Moreover, this subjective test has been conducted for comparing BSSBS-MMSE to STSA-MMSE(DFT).

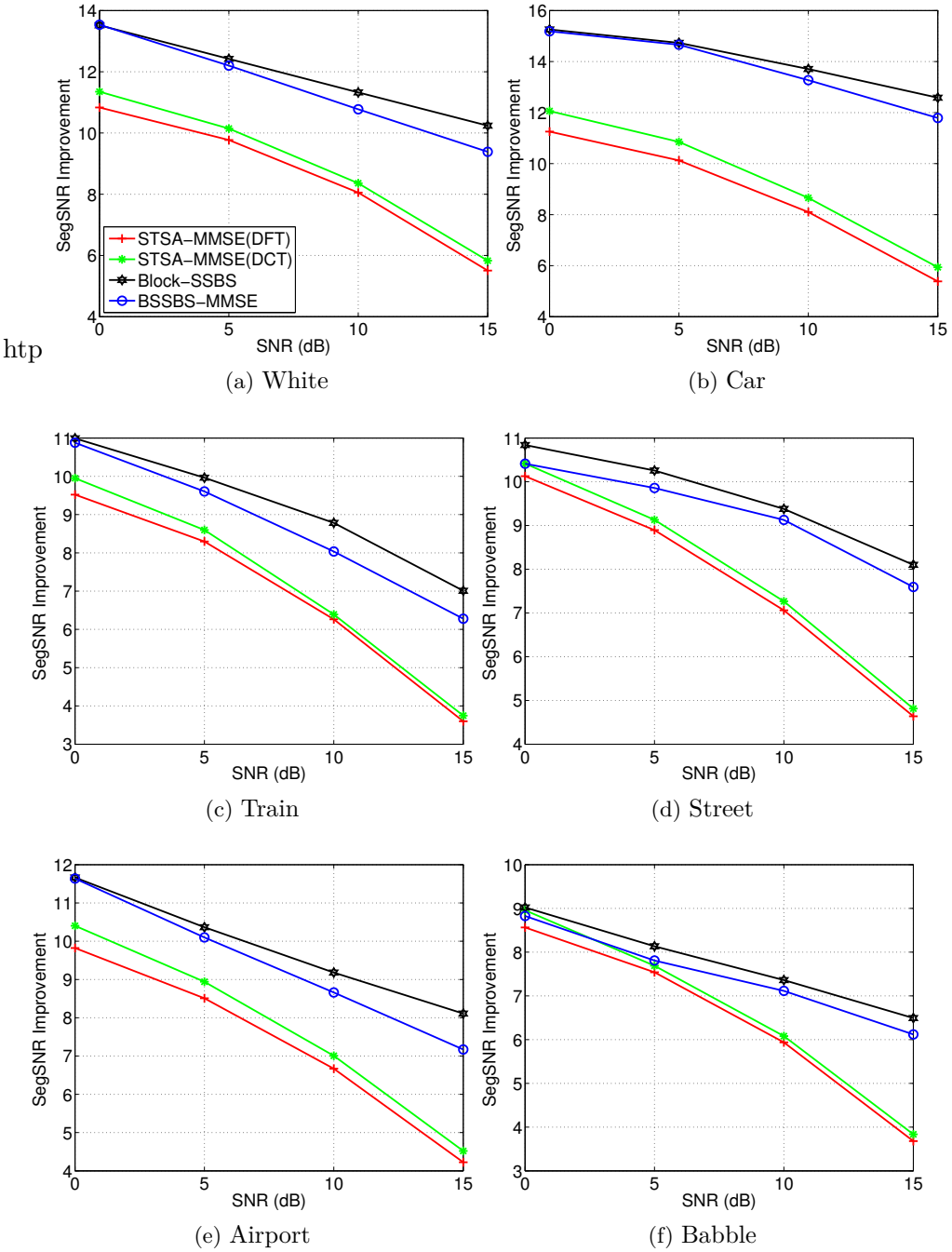
The MOS score obtained is given by Table 5.2. In each element, the first number is the average score and the second couple numbers is the 95% confidence interval (CI). In two SNR levels, SSBS-MMSE provides a higher average score and also the CI than STSA-MMSE(DFT). This result confirms the relevance of BSSBS-MMSE for audio signal denoising.

Table 5.2 – MOS for music signal obtained with BSSBS-MMSE and STSA-MMSE(DFT)

| | STSA-MMSE(DFT)**(CI)** | SBSSB-MMSE **(CI)** |
|---|---|---|
| 5 dB | 2.78 (**[2.65 2.91]**) | 3.82 (**[3.67 3.97]**) |
| 10 dB | 3.21 (**[3.09 3.34]**) | 4.02 (**[3.89 4.16]**) |

## 5.4 Conclusion

In this chapter, we have introduced several speech denoising methods in the DCT domain, which makes it possible to get rid of the phase estimation problem. These methods are Block-SSBS, STSA-MMSE(DCT) and BSSBS-MMSE. Block-SSBS is non-parametric and can be seen as a smooth shrinkage of DCT coefficients. Its parameters are optimized by SURE and RDT




htp

(a) White (b) Car

(c) Train (d) Street

(e) Airport (f) Babble

Figure 5.9 – The SSNR improve for audio signal with the reference noise for 6 kinds of noise from stationary noise (white) to slow-changing non-stationary noise (car and train noise) and up to speech-like and fast-changing non-stationary noise (street, airport and babble noise).

approaches, which are also non-parametric methods for statistical inference. STSA-MMSE(DCT) is a Bayesian estimator. BSSBS-MMSE combines Block-SSBS and STSA-MMSE(DCT) so as to benefit from the advantages of each method. Namely, Block-SSBS achieves good performance in terms of speech intelligibility by background noise reduction; STSA-MMSE improves speech quality by enhancing speech contained in small coefficients returned by the shrinkage.

The performance evaluation was conducted on the NOIZEUS database, with and without noise power spectrum reference. Various types of stationary and non-stationary noises were considered. When the noise spectrum is unknown, it is estimated by an up-to-date method. In addition, objective and subjective tests were used to assess the speech estimators, in comparison to a reference approach. Subjective tests involved a statistically significant number of raters. The experimental results show that BSSBS-MMSE performs better than the other methods in most situations. These experiments also confirm the relevance of working in the DCT domain.

The STSA-MMSE is devised under the Gaussian assumption for the Block-SSBS outcome. Asymptotic statistics could perhaps help justifying this assumption. However, the task seems rather difficult and, in any case, the experimental results provide evidence that such a Gaussian assumption leads to an STSA-MMSE good enough to retrieve relevant speech contents in small DCT coefficients.

All these results have proved the relevance of the proposed methods for speech signals. Noting however that the theoretical framework is based on very general assumptions, it can be wondered whether the proposed methods could be used to denoise other kinds of signals. As a proof of concept, these methods were applied here to denoise music signals. Subjective tests involving several raters confirm these very good and promising results. They demonstrate the robustness of the approaches independently of the nature of the signal of interest.

# Part IV

# Conclusion

Chapter 6

# Conclusions and Perspectives

*Good, better, best. Never let it rest. 'Til your good is better and your better is best.*

St. Jerome

## 6.1 Conclusion

The objective of this thesis work was to propose a complete speech enhancement system with innovative techniques in signal processing for applications such as assisted listening for hearing aids, cochlear implants and voice communication applications with lack of resources. In such areas of applications, the complete speech enhancement system should not only further enhance speech quality but also speech intelligibility. Moreover, this system is expected to have low computational cost, low power usage and operate without help of any database. In order to overcome these constraints, this research intended to investigate how far we can get in speech denoising by using unsupervised statistical methods only, without resorting to psycho-acoustics or machine learning (supervised) based approach. In this respect and taking into account the large amount of results provided by the literature on the topic, this research involved both parametric and non-parametric statistics for audio denoising, when the signal of interest is degraded by uncorrelated and independent additive noise.

In the first part, noise power spectrum estimation, the main block of most single micro speech enhancement system, has been considered. We have proposed a novel method for noise power spectrum estimation, called Extended-DATE (E-DATE). This method extends the d-dimensional amplitude trimmed estimator (DATE), originally introduced for additive white Gaussian noise power spectrum estimation to the more challenging scenario of non-stationary noise. The key idea is that, in each frequency bin and within a sufficiently short time period, the noise instantaneous power spectrum can be considered as approximately constant and estimated as the variance of a complex Gaussian noise process possibly observed in the presence of the signal of interest. The proposed method relies on the fact that the Short-Time Fourier Transform of noisy speech signals is sparse in the sense that transformed speech signals can be represented by a relatively small number of coefficients with large amplitudes in the time-frequency domain. The E-DATE estimator is robust in that it does not require prior information about the signal probability distribution except for the weak-sparseness property. In comparison to other state-of-the-art methods, the E-DATE is found to require the smallest number of parameters (only two). Two practical implementations of E-DATE algorithm, namely the B-E-DATE and SW-E-DATE, achieve good performance with and without noise reduction. In general, the E-DATE estimator yields the most accurate noise power spectrum estimate for speech enhancement in presence of various noise types and levels. This estimator has also shown its relevance for both speech quality and intelligibility improvement when incorporated into a complete system based STSA, a standard noise reduction algorithm. Although B-E-DATE is the straightforward block-based implementation of the E-DATE, but it entails an estimation delay. This can however be circumvented by resorting to the sliding window implementation SW-E-DATE.

After noise power spectrum estimation by E-DATE, Part III focused on noise reduction techniques. We have considered two different approaches for recovering the signal of interest: a parametric one and a non-parametric one. In the two approaches, we exploited a joint detection and estimation strategy to further remove or reduce background noise, without increasing the signal distortion. This strategy was motivated by the fact that for the signal of interest in noise has a sparse representation that can often be found on an appropriate orthogonal basis. Thus, the signal of interest can be reasonably assumed to be not always present in the time-frequency domain. These two joint strategies have been applied to speech enhancement in a parametric and a semi-parametric approaches presented in Chapters 4 and 5, respectively. In Chapter 4, the approach is purely parametric, whereas Chapter 5 focuses on non-parametric approaches as well as their combination with parametric statistics.

More specifically, in Chapter 4, we proposed some novel methods for estimating the STSA and LSA of speech. These methods are based on the combination of parametric detection and estimation theories. The main idea is to take into consideration speech presence and absence in each time-frequency bin for improving performance of minimum mean square error based estimators. Optimal detectors have been derived where they enable to figure out the absence or presence of speech in each time-frequency bin based on the estimators. Conversely, the estimators take into account the feedback from these detectors to improve them. Two signal models including strict and uncertain presence/absence of speech have been considered. Depending on the signal model, the STSA was either forced to zero or replaced by a small spectral floor for reducing musical noise when speech absence has been detected. These methods have been assessed in two scenarios, that is, with and without reference noise power spectrum. The objective tests confirmed the relevance of these approaches both in terms of speech quality and intelligibility.

Joint detection and estimation can be viewed as a form of Smoothed Sigmoid-Based Shrinkage (SSBS). For further improvement of performance and robustness in audio denoising, a semi-parametric approach was proposed in Chapter 5. As is well known, the *narrowband* Fourier transform has good frequency resolution. Thus, most speech enhancement algorithms use it to transform the observed signal into the time-frequency domain. However, the Fourier coefficients are complex so that these algorithms require phase spectrum estimation. To bypass this issue, Chapter 5 presents a novel estimator for estimating the amplitude of the speech coefficients in the cosine time-frequency domain after discrete cosine transform (DCT). This estimator aims at minimizing the mean square error of the absolute value of speech DCT coefficients. In order to take advantage of both parametric and non-parametric approaches, Chapter 5 studies also the combination of block shrinkage and Bayesian statistical estimation. In such a combination, the absolute value of the clean coefficient is firstly estimated by block smoothed sigmoid-based shrinkage (Block-SSBS). The block size required by Block-SSBS is obtained by statistical optimization via application of the SURE theorem. This step enables us to improve speech intelligibility as achieved by smoothed binary masking. In addition, it makes it possible to deal with various types of signals. Secondly, for refining the estimation, an optimal statistical estimator is added to handle musical noise. For evaluating the performance of the proposed method, objective as well as informal subjective test were used. The experiments described in this chapter yield promising results, both in terms of speech quality and intelligibility.

In summary, we have proposed several speech enhancement algorithms that are all based on a joint detection/estimation strategy. These enable us to improve quality and intelligibility, for speech as well as audio signals, in comparison to standard estimators. Figure 6.1 displays all methods in the plan of quality and intelligibility. Parametric methods (STSA-MMSE) provide a higher score in terms of quality and a lower score of intelligibility than non-parametric approach (Block-SSBS). Semi-parametric approach allow us to take advantage of two methods. It is worth noticing that, both parametric and semi-parametric approaches were exploited and that each of them has been shown to have its own relevance. Therefore, depending on the considered application, a suitable estimator should be chosen. The parametric estimators proposed above are more efficient to reduce musical noise in speech enhancement, whereas non-parametric estimators have been shown to be more relevant to denoise other types of audio signal, like music for instance.

## 6.2 Perspectives

Although the present work focused on noise reduction in speech enhancement systems using the DFT, it must be emphasized that the E-DATE estimator introduced in Chapter 3 is neither

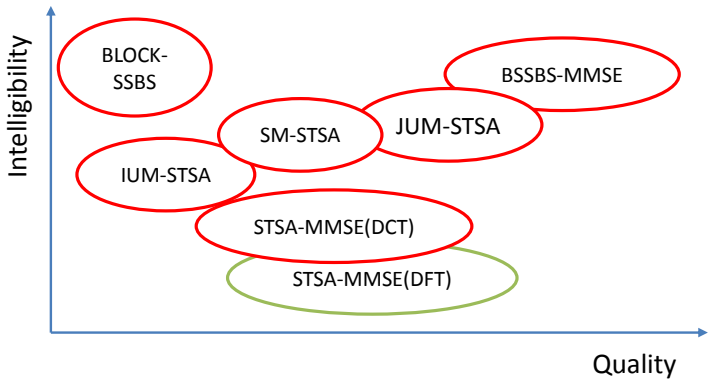



Figure 6.1 – A general view of all noise reduction methods based on STSA-MMSE considered in this thesis.

restricted to the DFT domain nor to speech signals. Therefore, it could find other applications in any scenario where noisy signals have a weakly-sparse representation. For instance, we have successfully considered the use of the E-DATE in the DCT in Chapter 5. For many signals of interest, not limited to speech, such a weakly-sparse representation can be provided by an appropriate wavelet transform. In this respect, the application of the E-DATE algorithm to audio source separation could be considered in continuation of [104]. The E-DATE estimator fundamentally relies on the DATE estimator which, as emphasized in [65], can be regarded as an outlier detector. Consequently the E-DATE can also be used as an outlier detector in each frequency bin. This opens interesting perspectives in voice activity detection based on frequency analysis as well as in the detection and estimation of chirp signals in various types of noise.

In Chapter 4, to take into account the presence or absence of speech, a novel estimator has been proposed relying on joint detection and estimation. This estimator is based on STSA and LSA where a Gaussian distribution of DFT coefficients is assumed. But other distributions for the DFT coefficients could be investigated. In addition, in this chapter, several strategies that combine detection and estimation for improving the performance of Bayesian estimators in speech enhancement have been proposed. The efficiency of all these approaches is highly dependent on the quality of the detector. In addition, all detectors are based on the Gaussian assumption for speech signals. Because this assumption may not be satisfied, other types of speech detectors in each time-frequency bin could be considered. A promising approach in this respect is proposed in [4]. This detector is based on the RDT algorithm and can provide good performance without knowledge about the distribution of signal of interest as already discussed in Chapter 5. Therefore, the semi-parametric estimators could be paired with the RDT detector in each time-frequency bin.

Chapter 5 investigated denoising methods using the DCT. Since it does not make any assumption on the signal of interest, Block-SSBS can be applied to other applications like image denoising. In this chapter, we also derived an STSA-MMSE in the DCT domain by making a Gaussian assumption on the DCT coefficients. It is thus natural to wonder again whether other distributions could be more relevant for modeling DCT coefficients. In addition, it has been observed that although DCT has a real and more compact representation than DFT, Block-SSBS,

STSA-MMSE in the DCT domain and BSSBS-MMSE are more sensitive to noise estimation errors than STSA-MMSE in the DFT domain. This point requires further investigation.

As a final note, it is worth noticing that all the speech enhancement methods exposed in this thesis were proposed in the context of a single microphone and were based on statistical approaches only. As such, a few promising perspectives arise as a natural generalization of our results. First and as discussed in the introduction, multi-microphone speech enhancement systems can immediately apply and benefit from the proposed methods at the output of the beam former. Second, the performance of our speech enhancement algorithms may be further improved by incorporating perceptual information, in continuation of researches such as [139,140]. Finally, although we have restricted attention to unsupervised approaches, the proposed methods can be used to post-process supervised approaches like the Wiener filter or the MMSE estimator.

# Appendices

Appendix A

# Lemma of the integral optimization problem

This appendix recalls a lemma that allows us to derive the Neyman-Pearson test and the combination detection and estimation presented in Chapter 4 proposed in [141].

**Lemma 1** *If $f : \mathbb{R}^n \to \mathbb{R}$ is a function defined on $\mathbb{R}^2$, thus:*

$$[f < 0] = \underset{B \in \mathcal{B}(\mathbb{R}^n)}{\operatorname{argmin}} \int_B f(x)dx, \tag{A.1}$$

*where $\mathcal{B}(\mathbb{R}^n)$ is the Borel algebra and $[f < 0] = \{x \in \mathbb{R}^n : f(x) < 0\}$*

*Proof:* For any function $f$ we can have: $f(x) = f\mathbb{1}_{[f \geq 0]}(x) + f\mathbb{1}_{[f<0]}(x)$. Therefore, $\forall B \in \mathcal{B}(\mathbb{R}^n)$:

$$\begin{aligned} I = \int_B f(x)dx &= \int_B f\mathbb{1}_{[f \geq 0]}(x)dx + \int_B f\mathbb{1}_{[f<0]}(x)dx \\ &= \int_{B \bigcap [f \geq 0]} f(x)dx + \int_{B \bigcap [f<0]} f(x)dx \\ &= \int_{B \bigcap [f \geq 0]} |f(x)|dx - \int_{B \bigcap [f<0]} |f(x)|dx. \end{aligned} \tag{A.2}$$

In fact, $\int_{B \bigcap [f \geq 0]} |f(x)|dx \geq 0$ so that

$$I \geq -\int_{B \bigcap [f<0]} |f(x)|dx.$$

Because $B \bigcap [f < 0] \subset [f < 0]$,

$$I \geq -\int_{[f<0]} |f(x)|dx.$$

Moreover, $-\int_{[f<0]} |f(x)|dx = \int_{[f<0]} f(x)dx$, $\forall B \in \mathcal{B}(\mathbb{R}^n)$, we have

$$\int_B f(x)dx \geq \int_{[f<0]} f(x)dx. \tag{A.3}$$

To reach the equality, we need only set $B = [f < 0]$. ■

Appendix B

# Detection threshold under joint detection and estimation



This appendix provides the computation of detection threshold in Chapter 4 based on Gaussian assumption.

## B.1 Strict model

In the strict presence/absence model, the threshold $\tau$ can be chosen by fixing the false alarm probability to a specified value. Thanks to the Gaussian assumption, the pdf of $A_Y$ under $H_0$ is Rayleigh [1, p.212] and given by:

$$f_{A_Y}(a; H_0) = \frac{2a}{\sigma_X^2} \exp\left(-\frac{a^2}{\sigma_X^2}\right). \tag{B.1}$$

Consider the decision

$$|Y| \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} \tau, \tag{B.2}$$

The false alarm probability is:

$$\mathbb{P}_{H_0}(\mathbf{D}=1) = \int_\tau^\infty \frac{2r}{\sigma_X^2} \exp\left(-\frac{r^2}{\sigma_X^2}\right) dr = \exp\left(-\frac{\tau^2}{\sigma_X^2}\right). \tag{B.3}$$

Therefore the threshold to guarantee a false alarm probability equal to $\alpha$ is

$$\tau^{\mathrm{SM}}(\alpha) = \sigma_X \sqrt{-\log(\alpha)} \tag{B.4}$$

Therefore, the detection thresholds in the strict speech presence/absence model are determined as follows:

$$\tau^{\mathrm{SM}}_{\mathrm{STSA}}(\alpha) = \mathcal{D}^{\mathrm{SM}}_{\mathrm{STSA}}(\tau^{\mathrm{SM}}(\alpha)), \tag{B.5}$$

$$\tau^{\mathrm{SSM}}_{\mathrm{LSA}}(\alpha) = \mathcal{D}^{\mathrm{SSM}}_{\mathrm{LSA}}(\tau^{\mathrm{SM}}(\alpha)). \tag{B.6}$$

## B.2 Uncertain model

Generally, in the uncertain model, the detection threshold $\tau$ is obtained by imposing $\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \alpha$. In addition, in the uncertain model, the distribution of $A_Y$ under hypothesis $H_0$ is given by

$$f_{A_Y}(a; H_0) = \frac{2a}{\sigma_X^2(1+\beta)} \exp\left(-\frac{a^2}{\sigma_X^2(1+\beta)}\right). \tag{B.7}$$

However, for the sake of simplicity, we propose the different choices of the threshold for the corresponding situations.

### B.2.1 Independent estimators

In this situation, because $\mathbb{D}_1(y;\psi_1) = \mathbb{D}_0(y;\psi_0)$, the choice of threshold $\tau$ does not affect $\mathcal{A}$. Thus, focusing on the estimation problem and with regard to Equations (4.84) and (4.100), it seems coherent to choose $\tau_{\text{STSA}}^{\text{IUM}}$ in the form of a likelihood ratio. In this respect, we propose to choose:

$$\tau_{\text{STSA}}^{\text{IUM}} = \tau_{\text{LSA}}^{\text{IUM}} = \Lambda(\xi, \gamma_0), \tag{B.8}$$

where $\gamma_0$ is the smallest *a posteriori* SNR above which we have a false alarm. Since the probability density function under $H_0$ is given by Equation (B.7), Equation (B.2) can be rewritten as

$$\gamma \underset{\mathbf{D}=0}{\overset{\mathbf{D}=1}{\gtrless}} -(1+\beta)\log(\alpha), \tag{B.9}$$

where $\alpha$ is the false probability. It follows that $\gamma_0 = -(1+\beta)\log(\alpha)$.

### B.2.2 Joint estimator

#### B.2.2.1 STSA-based estimator

In this section, the Bayesian risk under $H_0$ is given by:

$$\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \int r_{10}(y;\psi_1)\delta(y) f_Y(y; H_0)dy + \int r_{00}(y;\psi_0)(1-\delta(y)) f_Y(y; H_0)dy, \tag{B.10}$$

where

$$\begin{aligned} r_{10}(y;\psi_1) &= \int c_{10}(\psi_1, a_0) f_{A_0|Y=y}(a_0)da_0 = \int a_0^2 f_{A_0|Y=y}(a_0)da_0, \\ r_{00}(y;\psi_0) &= \int c_{00}(\psi_0, a_0) f_{A_0|Y=y}(a_0)da_0 = r_{10}(y;\psi_1) - \left(\psi_{\text{STSA}}^{\text{JUM(0)}}(y)\right)^2. \end{aligned}$$

According to [1, Equation (7.94)], we have:

$$r_{10}(y;\psi_1) = \mathbf{E}_0[A_0^2|Y=y] = \frac{\beta}{1+\beta}\left(\frac{1+\nu_\beta}{\gamma}\right)|y|^2 = G_0|y|^2, \tag{B.11}$$

where $\nu_\beta = \gamma\beta/(1+\beta)$ and $G_0 = \dfrac{\beta}{1+\beta}\left(\dfrac{1+\nu_\beta}{\gamma}\right)$. Moreover (see Equation (4.88):

$$\psi_{\text{STSA}}^{\text{JUM(0)}}(y) = G_{\text{STSA}}(\beta,\gamma)|y|. \tag{B.12}$$

The Bayesian risk under $H_0$ is now

$$\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \int_{\mathcal{A}} G_0|y|^2 f_Y(y; H_0)dy + \int_{\mathcal{A}^c} (G_0 - G^2_{\mathrm{STSA}}(\beta,\gamma))|y|^2 f_Y(y; H_0)dy \tag{B.13}$$

where $\mathcal{A}$ is the critical region given by Equation (4.72). Theoretically, $G_0$ and $G_{\mathrm{STSA}}(\beta,\gamma)$ are dependent on $y$. However, $0 \leq \beta \ll 1$ so that $G_0$ and $G_{\mathrm{STSA}}(\beta,\gamma)$ are smaller than 1. We can reasonably assume that $G_0$ and $G_{\mathrm{STSA}}(\beta,\gamma)$ are independent on $y$. By taking into account that $\beta$ is less than $\xi$, it follows from Equations (4.93) and (4.94) that the decision amounts to comparing the absolute value of the observation to a threshold $\tau^*$. Therefore, the Bayesian risk is approximated by:

$$\begin{aligned}\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) =& G_0 \int_{\tau_*}^{\infty} \frac{2r^3}{\sigma_X^2(1+\beta)} \exp\left(-\frac{r^2}{\sigma_X^2(1+\beta)}\right) dr \\ &+ (G_0 - G^2_{\mathrm{STSA}}(\beta,\gamma)) \int_0^{\tau_*} \frac{2r^3}{\sigma_X^2(1+\beta)} \exp\left(-\frac{r^2}{\sigma_X^2(1+\beta)}\right) dr,\end{aligned} \tag{B.14}$$

After a change of variable and an integration by parts, some routine algebra leads to:

$$\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \sigma_X^2(1+\beta)(G_0 - G^2_{\mathrm{STSA}}(\beta,\gamma)) + \left(G^2_{\mathrm{STSA}}(\beta,\gamma)\tau_*^2 + \sigma_X^2(1+\beta)\right) \exp\left(-\frac{\tau_*^2}{\sigma_X^2(1+\beta)}\right).$$

By solving equality $\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \alpha$, we can find $\tau_*(\alpha)$. Therefore the detection threshold $\tau^{\mathrm{JUM}}_{\mathrm{STSA}}(\alpha)$ is given by:

$$\tau^{\mathrm{JUM}}_{\mathrm{STSA}}(\alpha) = \mathcal{D}^{\mathrm{JUM}}_{\mathrm{STSA}}(\tau_*(\alpha)) \tag{B.15}$$

#### B.2.2.2 LSA-based estimator

With the same methodology as above, we have first:

$$\begin{aligned} r_{10}(y;\psi_1) &= \int (\log(a_0+1))^2 f_{A_0|Y=y}(a_0)da_0, \\ r_{00}(y;\psi_0) &= \int c_{00}(\psi_0, a_0) f_{A_0|Y=y}(a_0)da_0 = r_{10}(y;\psi_1) - \left(t^{\mathrm{JUM}(0)}_{\mathrm{LSA}}(y)\right)^2, \end{aligned}$$

where $t^{\mathrm{JUM}(0)}_{\mathrm{STSA}}(y) = \log[G_{\mathrm{LSA}}(\beta,\gamma)|y| + 1]$. The $r_{10}(y;\psi_1)$ is hardly tractable in theory. In order to evaluate $\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D})$, we can reasonably assume that $A_0 = \beta_0 A_Y$ where $0 \leq \beta_0 \ll 1$. Moreover, $\log(x+1) \approx x$ with $x \ll 1$. Therefore, according to Equation (B.11):

$$\begin{aligned} r_{10}(y;\psi_1) &= G_0|y|^2 \\ r_{00}(y;\psi_0) &= \left(G_0 - G^2_{\mathrm{LSA}}(\beta,\gamma)\right)|y|^2, \end{aligned}$$

In the similar way as section above,

$$\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \sigma_X^2(1+\beta)(G_0 - G^2_{\mathrm{LSA}}(\beta,\gamma)) + \left(G^2_{\mathrm{LSA}}(\beta,\gamma)\tau_*^2 + \sigma_X^2(1+\beta)\right) \exp\left(-\frac{\tau_*^2}{\sigma_X^2(1+\beta)}\right).$$

where $\tau_*(\alpha)$ is found by solving $\mathbf{R}_0(\widehat{A}_1, \widehat{A}_0, \mathbf{D}) = \alpha$. Thus, the detection threshold is given by:

$$\tau^{\mathrm{JUM}}_{\mathrm{LSA}}(\alpha) = \mathcal{D}^{\mathrm{JUM}}_{\mathrm{LSA}}(\tau_*(\alpha)). \tag{B.16}$$

Appendix C

# Semi-parametric approach



## C.1 The unbiased estimate risk of block for Block-SSBS

In the block $B_j$, the estimate risk $\widehat{R}_j$ in Equation 5.17 is calculated by SURE theorem given by Equation 5.16. We derive now this result. We consider first that the estimated risk is the sum of the three part following:

$$\widehat{R}_{j1} = -N\sigma_X^2(\mathfrak{B}), \tag{C.1}$$

$$\widehat{R}_{j2} = \|\mathbf{y} - \widehat{\mathbf{S}}(\mathbf{y})\|_2^2, \tag{C.2}$$

$$\widehat{R}_{j3} = 2\sigma_X^2(\mathfrak{B}) \sum_{n=1}^{N} \frac{\partial \widehat{S}_n}{y_n}(\mathbf{y}). \tag{C.3}$$

The second part $\widehat{R}_{j2}$ is the difference between the estimated $\widehat{\mathbf{S}}$ and the observation $\mathbf{Y}$. Readily, $G(\mathbf{y})$ is a scalar gain so that this difference can be written as:

$$\widehat{R}_{j2} = (1 - G(\mathbf{y}))^2 \|\mathbf{y}\|_2^2. \tag{C.4}$$

In the fact that $\widehat{S}_n(\mathbf{y}) = G(\mathbf{y})y_n$. Therefore the partial derivative $\dfrac{\partial \widehat{S}_n}{y_n}(\mathbf{y})$ can be calculated as:

$$\frac{\partial \widehat{S}_n}{y_n}(\mathbf{y}) = \frac{\partial G(\mathbf{y})}{y_n} y_n + G(\mathbf{y}). \tag{C.5}$$

Moreover, the gain function in the $B_j$ block can be explicitly expressed as a function of $y_n$ whereas other are considered as the parameter. Thus, using the chain rule for computing the derivative, we have

$$\frac{\partial G(\mathbf{y})}{y_n} = \tau \frac{e^{-\tau(\sqrt{\widehat{\gamma}}-\lambda)}}{\left(1 + e^{-\tau(\sqrt{\widehat{\gamma}}-\lambda)}\right)^2} \frac{1}{2\sqrt{\widehat{\gamma}}} \frac{2y_n}{N\sigma_X^2(\mathfrak{B})} = \tau G(\mathbf{y})\left(1 - G(\mathbf{y})\right) \frac{y_n}{\sqrt{\widehat{\gamma}} N\sigma_X^2(\mathfrak{B})}, \tag{C.6}$$

where the estimated *a posteriori* SNR $\widehat{\gamma} = \frac{\|\mathbf{y}\|_2^2}{N\sigma_X^2(\mathfrak{B})}$. The third part of the estimated block risk is leaded to:

$$\widehat{R}_{j3} = 2N\sigma_X^2(\mathfrak{B})G(\mathbf{y}) + 2\tau G(\mathbf{y})\left(1 - G(\mathbf{y})\right) \frac{\|\mathbf{y}\|_2^2}{N\sqrt{\widehat{\gamma}}}. \tag{C.7}$$

Taking the sum of the three part $\widehat{R}_{j1}$, $\widehat{R}_{j2}$ and $\widehat{R}_{j1}$, we obtain the estimated risk of the block $B_j$ as Equation (5.17).

## C.2 The MMSE gain function in the DCT domain

The gain function $G(\xi, \gamma)$ for STSA-MMSE in the DCT domain is derived from Equation (5.27) by two ways and shown by Equation (5.30). This appendix section provides the computational detail of the direct method for calculating this gain function. Without the loss of generality, we suppose that $y \geq 0$ so that $y = |y|$. Let us denote,

$$I_{1\pm} = \int_0^\infty a \exp\left(-\frac{a^2}{2\sigma^2} \pm \frac{ay}{\sigma_X^2}\right) da, \tag{C.8}$$

$$I_{2\pm} = \int_0^\infty \exp\left(-\frac{a^2}{2\sigma^2} \pm \frac{ay}{\sigma_X^2}\right) da, \tag{C.9}$$

the nominator $I_1$ and the denominator $I_2$ of Equation (5.27) are respectively given by:

$$I_1 = I_{1+} + I_{1-}, \tag{C.10}$$

$$I_2 = I_{2+} + I_{2-}. \tag{C.11}$$

Moreover, we have

$$I_{1\pm} = \int_0^\infty (-\sigma^2)\left(\frac{-a}{\sigma^2} \pm \frac{y}{\sigma_X^2}\right) \exp\left(-\frac{a^2}{2\sigma^2} \pm \frac{ay}{\sigma_X^2}\right) da \pm \left(\frac{y\sigma^2}{\sigma_X^2}\right) I_{2\pm}. \tag{C.12}$$

Therefore, the nominator can be written as $I_1 = I_{1+} + I_{1-} = 2\sigma^2 + \left(\frac{y\sigma^2}{\sigma_X^2}\right)(I_{2+} - I_{2-})$. In addition,

$$I_{2\pm} = \int_0^\infty \exp\left[-\left(\frac{a}{\sqrt{2\sigma^2}} \mp \frac{y\sqrt{\sigma^2}}{\sqrt{2}\sigma_X^2}\right)^2\right] \exp\left(\frac{\sigma^2 y^2}{2\sigma_X^4}\right) da. \tag{C.13}$$

Thus, the denominator and the nominator can be evaluated as:

$$I_2 = I_{2+} + I_{2-} = 2\sqrt{2\sigma^2} \exp\left(\frac{\sigma^2 y^2}{2\sigma_X^4}\right) \int_0^\infty \exp(-t^2) dt, \tag{C.14}$$

$$I_1 = 2\sigma^2 + \left(\frac{y\sigma^2}{\sigma_X^2}\right)(I_{2+} - I_{2-}) = 2\sigma^2 + \frac{y\sigma^2}{\sigma_X^2} 2\sqrt{2\sigma^2} \exp\left(\frac{\sigma^2 y^2}{2\sigma_X^4}\right) \int_{-y^*}^{+y^*} \exp(-t^2) dt, \tag{C.15}$$

Considering the zero-mean Gaussian distribution with its variance equal to 1/2, its probability density function is $\phi(t) = \frac{e^{-t^2}}{\sqrt{\pi}}$. It makes possible that $\int_0^\infty \exp(-t^2)dt = \sqrt{\pi}/2$ and $\int_{-y^*}^{+y^*} \exp(-t^2)dt = \sqrt{\pi}/2 \operatorname{erf}(y^*)$. Replacing these results to Equations (C.14) and (C.15), we can compute $\psi(y)$ and $G(\xi, \gamma) = \psi(y)/|y|$, Equation (5.30) is proved.

Appendix D

# Author Publications

**(J-5)** V.K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, "MMSE-based Speech Enhancement: From Parametric to Non-Parametric Approaches", in preparation.

**(J-4)** V.K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, "Speech Enhancement with Spectral Amplitude and Log-Spectral amplitude Estimators Based on Joint Detection and Estimation", *IEEE/ACM Trans. Audio, Speech, Language Processing*, in preparation.

**(J-3)** V.K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, "Semi-Parametric Block Shrinkage and Application to Musical Audio Denoising", *IEEE Signal Processing Letters*, submitted.

**(J-2)** V.K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, "Robust Speech Denoising by Non-diagonal Smoothed Sigmoid-Based Shrinkage in the DCT Domain", *IEEE/ACM Trans. Audio, Speech, Language Processing*, submitted.

**(J-1)** V.K. Mai, D. Pastor, A. Aissa-El-Bey, R. Le-Bidan, "Robust Estimation of Non-Stationary Noise Power Spectrum for Speech Enhancement.", *IEEE/ACM Trans. Audio, Speech, Language Processing*, n. 23(4), pp. 670-682, 2015.

# Bibliography


[1] P. C. Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[2] A. M. Atto, D. Pastor, and G. Mercier. Smooth sigmoid wavelet shrinkage for non-parametric estimation. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, pages 3265–3268, 2008.

[3] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[4] D. Pastor and Q. T. Nguyen. Random distortion testing and optimality of thresholding tests. *IEEE Trans. Signal Process.*, 61(16):4161–4171, 2013.

[5] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen. On optimal multichannel mean-squared error estimators for speech enhancement. *IEEE Signal Process. Lett.*, 16(10):885–888, 2009.

[6] R. C. Hendriks, T. Gerkmann, and J. Jensen. DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. *Synthesis Lectures on Speech and Audio Processing*, 9(1):1–80, 2013.

[7] R. Le Bouquin-Jeannès, P. Scalart, G. Faucon, and C. Beaugeant. Combined noise and echo reduction in hands-free systems: A survey. *IEEE Trans., Speech, Audio, Process.*, 9(8):808–820, 2001.

[8] M. R. Weiss, E. Aschkenasy, and T. W. Parsons. Study and development of the INTEL technique for improving speech intelligibility. Technical report, DTIC Document, 1975.

[9] M Berouti, R Schwartz, and John Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE Int. Conf. Acoustics, Speech, Signal Process.(ICASSP)*, volume 4, pages 208–211. IEEE, 1979.

[10] S. Kamath and P. C. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, volume 4, pages IV–4164. IEEE, 2002.

[11] M. T. Sadiq, N. Shabbir, and W. J. Kulesza. Spectral subtraction for speech enhancement in modulation domain. *IJCSI Int. Journal, Computer, Science*, 10(2), 2013.

[12] N. Upadhyay and A. Karmakar. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 54:574–584, 2015.

[13] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, 1949.

[14] P. Scalart and J. VIEIRA FILH0. Speech enhancement based on a priori signal to noise estimation. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process. (ICASSP)*, volume 2, pages 629–632. IEEE, 1996.

[15] J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction wiener filter. *IEEE Trans. audio, speech, and lang. process.*, 14(4):1218–1234, 2006.

[16] B. Xia and C. Bao. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, 60:13–29, 2014.

[17] N. Upadhyay and R. Jaiswal. Single channel speech enhancement: Using wiener filtering with recursive noise estimation. *Procedia Computer Science*, 84:22–30, 2016.

[18] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Commun.*, 10(1):45–57, 1991.

[19] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech, Audio, Process.*, 3(4):251–266, 1995.

[20] J. Huang and Y. Zhao. A DCT-based fast signal subspace technique for robust speech recognition. *IEEE Trans. Speech, Audio, Process.*, 8(6):747–751, 2000.

[21] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech, Audio, Process.*, 11(4):334–341, 2003.

[22] J. Sun, C.and Xie and Y. Leng. A signal subspace speech enhancement approach based on joint low-rank and sparse matrix decomposition. *Archives of Acoustics*, 41(2):245–254, 2016.

[23] U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and D. Wang. Speech intelligibility of ideal binary masked mixtures. In *18th Europ. Signal Process. Conf.*, pages 1909–1913. IEEE, 2010.

[24] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters. The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses. *IEEE Trans. Audio, Speech, and Lang. Process.*, 21(1):63–72, 2013.

[25] R. Koning, N. Madhu, and J. Wouters. Ideal time–frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners. *IEEE Trans. Biomed. Eng.*, 62(1):331–341, 2015.

[26] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech, Signal Process.*, 28(2):137–145, 1980.

[27] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech, Signal Process.*, 32(6):1109–1121, 1984.

[28] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal, Process.*, ASSP-33(2):443–445, Apr. 1985.

[29] T. Lotter and P. Vary. Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP J. applied, signal, process.*, 2005:1110–1126, 2005.

[30] P. J. Wolfe and S. J. Godsill. Simple alternatives to the ephraim and malah suppression rule for speech enhancement. In *2001. Proc. IEEE Signal Process. Workshop, Stat. Signal Process.,*, pages 496–499. IEEE, 2001.

[31] C. Breithaupt and R. Martin. MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.(ICASSP)*, volume 1, pages I–896. IEEE, 2003.

[32] I. Cohen. Speech enhancement using super-gaussian speech models and noncausal a priori snr estimation. *Speech commun.*, 47(3):336–350, 2005.

[33] B. Chen and P. C. Loizou. A laplacian-based MMSE estimator for speech enhancement. *Speech commun.*, 49(2):134–143, 2007.

[34] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(6):1741–1752, 2007.

[35] P. C. Loizou. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. *IEEE Trans. Speech, Audio, Process.*, 13(5):857–869, 2005.

[36] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech, Audio, Process.*, 13(5):845–856, 2005.

[37] I. Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Process. Lett.*, 9(4):113–116, 2002.

[38] Y. Soon, S. N. Koh, and C. K. Yeo. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. *Signal Process.*, 75(2):151–159, 1999.

[39] A. Abramson and I. Cohen. Simultaneous detection and estimation approach for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(8):2348–2359, 2007.

[40] T. Sreenivas and P. Kirnapure. Codebook constrained wiener filtering for speech enhancement. *IEEE Trans. Speech, Audio, Process.*, 4(5):383–389, 1996.

[41] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(1):163–176, 2006.

[42] S. Srinivasan, J. Samuelsson, and W. B. Kleijn. Codebook-based bayesian speech enhancement for nonstationary environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(2):441–452, 2007.

[43] Q. He, C. Bao, and F. Bao. Multiplicative update of ar gains in codebook-driven speech enhancement. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, pages 5230–5234. IEEE, 2016.

[44] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Trans. Signal, Process.*, 40(4):725–735, 1992.

[45] D. Y. Zhao and W. B. Kleijn. HMM-based gain modeling for enhancement of speech in noise. *IEEE Trans. Audio, Speech, and Lang. Process.*, 15(3):882–892, 2007.

[46] H. Veisi and H. Sameti. Speech enhancement using hidden markov models in mel-frequency domain. *Speech Communication*, 55(2):205–220, 2013.

[47] N. Mohammadiha, R. Martin, and A. Leijon. Spectral domain speech enhancement using HMM state-dependent super-gaussian priors. *IEEE Signal, Process. Lett.*, 20(3):253–256, 2013.

[48] N. Mohammadiha and A. Leijon. Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(5):998–1011, 2013.

[49] J. Xu, Y.and Du, L. R. Dai, and C. H. Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.*, 21(1):65–68, 2014.

[50] J. Xu, Y.and Du, L. R. Dai, and C. H. Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio, Speech,Lang. Process.*, 23(1):7–19, 2015.

[51] T. T. Vu, B. Bigot, and E. S. Chng. Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, pages 499–503. IEEE, 2016.

[52] C. D. Sigg, T. Dikk, and J. M. Buhmann. Speech enhancement with sparse coding in learned dictionaries. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, pages 4758–4761. IEEE, 2010.

[53] C. D. Sigg, T. Dikk, and J. M. Buhmann. Speech enhancement using generative dictionary learning. *IEEE Trans. Audio, Speech, and Lang. Process.*, 20(6):1698–1712, 2012.

[54] Y. Zhou, H. Zhao, and T. Shang, L.and Liu. Immune K-SVD algorithm for dictionary learning in speech denoising. *Neurocomputing*, 137:223–233, 2014.

[55] S. Mavaddaty, S. M. Ahadi, and S. Seyedin. Modified coherence-based dictionary learning method for speech enhancement. *IET Signal Process.*, 9(7):537–545, 2015.

[56] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process. (ICASSP)*, pages 4029–4032, 2008.

[57] N. Mohammadiha, T. Gerkmann, and A. Leijon. A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization. In *IEEE Workshop Appli. Signal Process. Audio, Acoust. (WASPAA)*, pages 45–48. IEEE, 2011.

[58] N. Mohammadiha, W. B. Kleijn, and A. Leijon. Gamma hidden markov model as a probabilistic nonnegative matrix factorization. In *21st European Signal, Process. Conf. (EUSIPCO)*, pages 1–5. IEEE, 2013.

[59] K. Kwon, J. W. Shin, and N. S. Kim. Nmf-based speech enhancement using bases update. *IEEE Signal, Process. Lett.*, 22(4):450–454, 2015.

[60] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. signal, process.*, 54(11):4311–4322, 2006.

[61] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.

[62] D. L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory*, 41(3):613–627, 1995.

[63] A. M. Atto, D. Pastor, and G. Mercier. Detection threshold for non-parametric estimation. *Signal, Image and Video processing*, 2(3):207–223, 2008.

[64] A. M Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80, 2012.

[65] D. Pastor and F. Socheleau. Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences. *IEEE Trans. Signal, Process.*, 60(4):1545–1555, Apr. 2012.

[66] B. P. Rao. *Nonparametric functional estimation*. Academic press, 2014.

[67] G. V. Moustakides. Optimum joint detection and estimation. In *Proc. IEEE Int. Inf. Theory*, pages 2984–2988, 2011.

[68] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang. Joint detection and estimation: Optimum tests and applications. *IEEE Trans. Inf. Theory*, 58(7):4215–4229, 2012.

[69] P. Stoica and R. L. Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, 1997.

[70] R. Martin. An efficient algorithm to estimate the instantaneous snr of speech signals. In *Proceeding of the Eurospeech*, volume 93, pages 1093–1096, 1993.

[71] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. European, Signal, Process. Conf. (EUSIPCO)*, pages 1182–1185, 1994.

[72] M. Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Trans., Acoust., Speech, Signal Process.*, 28(1):55–69, 1980.

[73] R. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Trans., Acoust., Speech, Signal Process.*, 28(1):99–102, 1980.

[74] S. Nawab, T. Quatieri, and J. Lim. Signal reconstruction from short-time fourier transform magnitude. *IEEE Trans., Acoust. , Speech, Signal Process.*, 31(4):986–998, 1983.

[75] D. Griffin and J. Lim. Signal reconstruction from short-time fourier transform magnitude. *IEEE Trans., Acoust. , Speech, Signal Process.*, 32(2):236–243, 1984.

[76] G. T. Beauregard, X. Zhu, and L. Wyse. An efficient algorithm for real-time spectrogram inversion. In *Int. conf. digital audio eff. (DAFx)*, pages 116–118, 2005.

[77] N. Sturmel and L. Daudet. Signal reconstruction from stft magnitude: a state of the art. In *Int. conf. digital audio eff. (DAFx)*, pages 375–386, 2011.

[78] J. H. L. Hansen and B. L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Int. Conf. Spoken Language (ICSLP)*, volume 7, pages 2819–2822, 1998.

[79] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective measures of speech quality*. Prentice Hall, 1988.

[80] IUT. ITU recommendation, G. 160. *Voice Enhancement Devices for Mobile Networks*, 2005.

[81] A. W. Rix, John G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, volume 2, pages 749–752. IEEE, 2001.

[82] Y. Hu and P. C. Loizou. Evaluation of objective measures for speech enhancement. In *Proc. Interspeech*, pages 1447–1450, 2006.

[83] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(1):229–238, 2008.

[84] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac. On the evaluation of the conversational speech quality in telecommunications. *EURASIP Journal, Advances, Signal Process.*, 2008:93, 2008.

[85] ANSI. S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19:90–119, 1997.

[86] Ray L Goldsworthy and Julie E Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Amer.*, 116(6):3679–3689, 2004.

[87] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, 2011.

[88] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock. Ieee recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.*, 17(3):225–246, 1969.

[89] V. K. Mai, D. Pastor, A. Aïssa-El-Bey, and R. Le-Bidan. Robust estimation of non-stationary noise power spectrum for speech enhancement. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23(4):670–682, 2015.

[90] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process. (ICASSP)*, volume 1, pages 153–156, Detroit, Michigan, USA, May 1995.

[91] B. Ahmed and W. H. Holmes. A voice activity detector using the chi-square test. In *IEEE Int. Conf. Acoust., Speech, Signal Process.*, volume 1, pages I–625, Montreal, Quebec, Canada, 2004.

[92] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech, Audio, Process.*, 11(5):466–475, Sep. 2003.

[93] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech, Audio Process.*, 9(5):504–512, Jul. 2001.

[94] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal, Process. Lett.*, 9(1):12–15, Jan. 2002.

[95] S. Rangachari and P. C. Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech communications*, 48(2):220–231, Feb. 2006.

[96] R. Yu. A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process. (ICASSP)*, pages 4421–4424, Taipei, Taiwan, Apr. 2009.

[97] T. Gerkmann and R. C. Hendriks. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1383–1393, May 2012.

[98] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani. Noise power spectral density tracking: A maximum likelihood perspective. *IEEE Signal, Process. Lett.*, 19(8):495–498, Aug. 2012.

[99] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and wiener filtering. In *Proc. IEEE Inter. Conf. Acoust., Speech, Signal Process. (ICASSP)*, volume 3, pages 1875–1878 vol.3, 2000.

[100] P.L. Davies and U. Gather. The identification of multiple outliers (with discussion). *J. Amer. Statist. Assoc.*, (423):782 – 801, 1993.

[101] N. N. Lebedev. *Special Functions and their Applications*. Prentice-Hall, Englewood Cliffs, 1965.

[102] D. Pastor. A theoretical result for processing signals that have unknown distributions and priors in white gaussian noise. *Computational Statistics & Data Analysis, CSDA*, 52(6):3167 – 3186, 2008.

[103] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[104] S. M. Aziz Sbai, A. Aïssa-El-Bey, and D. Pastor. Contribution of statistical tests to sparseness-based blind source separation. *EURASIP journal, applied, signal, process.*, Jul. 2012.

[105] S. M. Berman. *Sojourns and extremes of stochastic processes*. Wadsworth, Reading, MA, January 1992.

[106] S. Mallat. *A wavelet tour of signal processing, second edition*. Academic Press, 1999.

[107] R. J. Serfling. *Approximations theorems of mathematical statistics*. Wiley, 1980.

[108] A. M. Atto, D. Pastor, and G. Mercier. Detection thresholds for non-parametric estimation. *Signal, Image, Video process.*, 2(3):207–223, February 2008.

[109] O. Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.*, 52(7):1830–1847, July 2004.

[110] D. Pastor and A. M. Atto. *Wavelet shrinkage: from sparsity and robust testing to smooth adaptation; In Fractals and Related Fields, Eds: J. Barral & S. Seuret.* Birkhaüser, 2010.

[111] R. C. Hendriks, J. Jensen, and R. Heusdens. Noise tracking using DFT domain subspace decompositions. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(3):541–553, Mar. 2008.

[112] C. H. You, S. N. Koh, and S. Rahardja. $\beta$-order MMSE spectral amplitude estimation for speech enhancement. *IEEE Trans. Speech, Audio, Process.*, 13(4):475–486, 2005.

[113] E. Plourde and B. Champagne. Auditory-based spectral amplitude estimators for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(8):1614–1623, 2008.

[114] B. J. Borgström and A. Alwan. Log-spectral amplitude estimation with generalized gamma distributions for speech enhancement. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, pages 4756–4759. IEEE, 2011.

[115] Ioannis Andrianakis and Paul R White. MMSE speech spectral amplitude estimators with chi and gamma speech priors. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, volume 3, pages 1068–1071. IEEE, 2006.

[116] D. Malah, R. V. Cox, and A. J. Accardi. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, volume 2, pages 789–792. IEEE, 1999.

[117] N. S. Kim and J. H. Chang. Spectral enhancement based on global soft decision. *IEEE Signal Process. Lett.*, 7(5):108–110, 2000.

[118] T. Gerkmann, C. Breithaupt, and R. Martin. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(5):910–919, 2008.

[119] Y. Hu and P. C. Loizou. Subjective comparison of speech enhancement algorithms. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, volume 1, pages I–1523–I–156. IEEE, 2006.

[120] P. C. Loizou and G. Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):47–56, 2011.

[121] J. Jensen and R. C. Hendriks. Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(1):92–102, 2012.

[122] Y. Hu and P. C. Loizou. Techniques for estimating the ideal binary mask. In *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, pages 154–157, 2008.

[123] H. V. Poor. *An introduction to signal detection and estimation.* Springer Science & Business Media, 2013.

[124] S. M Kay. Fundamentals of statistical signal processing, volume i: estimation theory. 1993.

[125] A. Papoulis and U. Pillai, S. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[126] A. Aziz-Sbaï, S. M.and Aïssa-El-Bey and D. Pastor. Contribution of statistical tests to sparseness-based blind source separation. *EURASIP J. Adv. Signal, Process.*, 2012(1):1–15, 2012.

[127] G.V. Moustakides. Optimum joint detection and estimation. Technical Report SSP-2010-1, Department of Electrical and Computer Engineering University of Patras, GREEC, 2010.

[128] Y. Soon, S. N. Koh, and C. K. Yeo. Noisy speech enhancement using discrete cosine transform. *Speech communication*, 24(3):249–257, 1998.

[129] Saeed Gazor and Wei Zhang. Speech enhancement employing laplacian-gaussian mixture. *IEEE Trans. Speech, Audio, Process.*, 13(5):896–904, 2005.

[130] K. R. Rao and P. Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[131] N. Ahmed, T. Natarajan, and K. R Rao. Discrete cosine transfom. *IEEE trans., Comput.*, (1):90–93, 1974.

[132] M. Elad. *Sparse and redundant representations*. Springer, 2010.

[133] R. Tavares and R. Coelho. Speech enhancement with nonstationary acoustic noise detection in time domain. *IEEE Signal Process. Lett.*, 23(1):6–10, 2016.

[134] T. T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Annals of statistics*, pages 898–924, 1999.

[135] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Trans. Signal Process.*, 56(5):1830–1839, 2008.

[136] R. Tibshirani and L. Wasserman. Stein's unbiased risk estimate. *Course notes from "Statistical Machine Learning, Spring 2015"*, pages 1–12, 2015.

[137] D. Pastor and A. M. Atto. *Wavelet shrinkage: from sparsity and robust testing to smooth adaptation; In Fractals and Related Fields, Eds: J. Barral & S. Seuret*. Birkhaüser, 2010.

[138] A. Jeffrey and D. Zwillinger. *Table of integrals, series, and products*. Academic Press, 2007.

[139] A. Amehraye. Débruitage perceptuel de la parole. *Ecole Nationale Supérieure des Télécommunications de Bretagne, Thèse de Doctorat*, 2009.

[140] A. Amehraye, D. Pastor, A. Tamtaoui, and D. Aboutajdine. From maskee to audible noise in perceptual speech enhancement. *Int., Journal, Signal, Process.,*, 5(2):93–96, 2009.

[141] D. Pastor. *Lecture notes on decesion theory*. Ecole Nationale Supérieure des Télécommunications de Bretagne, Poly, 2016.

## Résumé

Cette thèse traite d'un des problèmes les plus stimulants dans le traitement de la parole concernant la prothèse auditive, où seulement un capteur est disponible avec de faibles coûts de calcul, de faible utilisation d'énergie et l'absence de bases de données. Basée sur les récents résultats dans les deux estimations statistiques paramétriques et non-paramétriques, ainsi que la représentation parcimonieuse. Cette étude propose quelques techniques non seulement pour améliorer la qualité et l'intelligibilité de la parole, mais aussi pour s'attaquer au débruitage du signal audio en général.

La thèse est divisée en deux parties ; Dans la première partie, on aborde le problème d'estimation de la densité spectrale de puissance du bruit, particulièrement pour le bruit non-stationnaire. Ce problème est une des parties principales du traitement de la parole du mono-capteur. La méthode proposée prend en compte le modèle parcimonieux de la parole dans le domaine transféré. Lorsque la densité spectrale de puissance du bruit est estimée, une approche sémantique est exploitée pour tenir compte de la présence ou de l'absence de la parole dans la deuxième partie.

En combinant l'estimation Bayésienne et la détection Neyman-Pearson, quelques estimateurs paramétriques sont développés et testés dans le domaine Fourier. Pour approfondir la performance et la robustesse de débruitage du signal audio, une approche semi-paramétrique est considérée. La conjointe détection et estimation peut être interprétée par Smoothed Sigmoid-Based Shrinkage (SSBS). Ainsi, la méthode Bloc-SSBS est proposée afin de prendre en compte les atomes voisinages dans le domaine temporel-fréquentiel. De plus, pour améliorer fructueusement la qualité de la parole et du signal audio, un estimateur Bayésien est aussi dérivé et combiné avec la méthode Bloc-SSBS. L'efficacité et la pertinence de la stratégie dans le domaine transformée cosinus pour les débruitages de la parole et de l'audio sont confirmées par les résultats expérimentaux.



**Mots clés:** Enrichissement de la parole et de l'audio, Débruitage statistique, Représentation parcimonieuse, Estimation paramétrique, Combinaison de détection et estimation, Seuillage parcimonieux, Estimation non-paramétrique



## Abstract

This PhD thesis deals with one of the most challenging problem in speech enhancement for assisted listening where only one micro is available with the low computational cost, the low power usage and the lack out of the database. Based on the novel and recent results both in non-parametric and parametric statistical estimation and sparse representation, this thesis work proposes several techniques for not only improving speech quality and intelligibility and but also tackling the denoising problem of the other audio signal.

In the first major part, our work addresses the problem of the noise power spectrum estimation, especially for non-stationary noise, that is the key part in the single channel speech enhancement. The proposed approach takes into account the weak-sparseness model of speech in the transformed model. Once the noise power spectrum has been estimated, a semantic road is exploited to take into consideration the presence or absence of speech in the second major part.

By applying the joint of the Bayesian estimator and the Neyman-Pearson detection, some parametric estimators were developed and tested in the discrete Fourier transform domain. For further improve performance and robustness in audio denoising, a semi-parametric approach is considered. The joint detection and estimation can be interpreted by Smoothed Sigmoid-Based Shrinkage (SSBS). Thus, Block-SSBS is proposed to take into additionally account the neighborhood bins in the time-frequency domain. Moreover, in order to enhance fruitfully speech and audio, a Bayesian estimator is also derived and combined with Block-SSBS. The effectiveness and relevance of this strategy in the discrete Cosine transform for both speech and audio denoising are confirmed by experimental results.



**Keywords:** Speech and audio enhancement, Noise reduction, Spare representation, Parametric estimator, Joint detection and estimation, Sparse thresholding, Non-parametric estimator.



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

**N° d'ordre : 2017IMTA0008**

IMT Atlantique Bretagne-Pays de la Loire - www.imt-atlantique.fr

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 03
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler - La Chantrerie
CS 20722
44307 Nantes Cedex 3
T +33 (0)2 51 85 81 00
F +33 (0)2 51 85 81 99

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
T +33 (0)2 99 12 70 00
F +33 (0)2 99 12 70 08