Studies in Signal Processing for Robust Speech Recognition in Noisy and Reverberant Environments

Kenko Ota

N° d'ordre : 71

ÉCOLE CENTRALE DE LILLE
UNIVERSITÉ DOSHISHA

THÈSE

présentée en vue d'obtenir le grade de

DOCTEUR

Spécialité : Automatique et Informatique Industrielle

par

Kenko OTA

Traitement du signal pour la reconnaissance de la parole robuste dans des environnements bruités et réverbérants

Doctorat délivré conjointment par l'École Centrale de Lille
et l'Université Doshisha

Soutenue le 19 janvier 2008 devant le jury constitué de :

Mr Noureddine Ellouze, Professeur, ENI Tunis, *Rapporteur*
Mr Hugo Leonardo Rufiner, Professeur, SINC, *Rapporteur*
Mr Yoichi Yamashita, Professeur, Université Ritsumeikan, *Rapporteur*
Mr Mehdi Nouri-Shirazi, Professeur, Institut des Technologies d'Osaka, *Président*
Mr Emmanuel Duflos, Professeur, EC Lille, *Directeur de thèse*
Mr Philippe Vanheeghe, Professeur, EC Lille, *Directeur de thèse*
Mr Masuzo Yanagida, Professeur, Université Doshisha, *Directeur de thèse*

Résumé

Les méthodes de reconnaissance de la parole sont utilisables d'un point de vue pratique a condition d'utiliser un microphone proche de la bouche et que l'environnement ne soit pas perturbé. Dans ce cas, un « casque » est souvent utilisé pour maintenir un petit microphone à proximité de la bouche. Si ce type de dispositif n'est pas utilisé, il est nécessaire de faire attention à la distance entre le microphone et la bouche du locuteur pour éviter de tenir compte du souffle dû à la respiration. Il est donc intéressant de développer des méthodes qui permettront au locuteur de ne pas devoir faire attention à la position de son microphone lors d'un discours. De plus, si le microphone est placé loin du locuteur, le taux de reconnaissance de la parole décroît rapidement en présence de bruits ou de phénomènes de réflexion. Si la reconnaissance est utilisée pour commander un appareil, cet appareil peut également émettre des sons. Ceux-ci seront alors considérés comme des bruits qui contribueront à la diminution du taux de reconnaissance. Par conséquent, il est également nécessaire de développer des méthodes de réduction de bruit et de dé-réverbération pour concevoir des systèmes de reconnaissance de la parole efficaces.

C'est l'objet de cette thèse qui est constituée de six chapitres. Ceux-ci décrivent quatre méthodes de traitement des signaux pour la reconnaissance de la parole robuste dans un environnement réel.

Le chapitre 1 aborde la nécessité de la réduction de bruit. Il présente une analyse bibliographique sur les travaux en reconnaissance de la parole et les caractéristiques des méthodes à développer pour obtenir une reconnaissance robuste.

Le chapitre 2 propose une technique pour réduire les bruits constitués par les sons émis par un appareil à commander. Pour réduire ces bruits, une méthode classique de soustraction spectrale peut être utilisée. Cependant, la soustraction spectrale ne donne pas de bons résultats si la fonction de transfert entre les haut-parleurs de l'appareil et le microphone varie dans le temps. Un mécanisme d'adaptation est alors proposé pour faire face à la variation temporelle de la fonction de transfert. Bien que l'annulation adaptative de bruit (ANC) soit une méthode connue pour résoudre le problème précédent, l'excès de soustraction peut provoquer une distorsion du signal de parole estimé. En effet, l'ANC coupe une partie du signal de parole dans le cas où il y a une forte corrélation entre le signal produit par l'appareil à commander et le bruit. Pour résoudre ce problème, le système proposé dans cette thèse, utilise la structure harmonique des segments voisés que la méthode ANC conventionnelle ne prend pas en compte.

Dans une première étape, la méthode proposée est l'extraction de la fréquence fondamentale du signal reçu. Puis la fréquence fondamentale et le spectre du signal sont utilisés pour classifier le signal reçu en 3 types de segments :

- Segments voisés,

- Segments non voisés,
- Segments sans parole.

Dans le cas de segments voisés, la structure harmonique de la parole cible est extraite en utilisant la fréquence fondamentale. L'estimation de la structure harmonique est soustraite du signal reçu puis le bruit présent dans le signal reçu par le microphone est estimé. L'estimation du bruit est utilisée pour déterminer la fonction de transfert qui permet l'analyse du cadre courant avec le bruit connu qui est directement déterminé à partir de l'appareil à commander. Dans le cas de segments non-voisés, le signal cible n'a pas de structure harmonique, mais les caractéristiques de la fonction de transfert acoustique entre l'appareil à contrôler et le microphone, tel entre un poste de télévision et un microphone, ne varie pratiquement pas entre 2 trames d'analyse consécutives. Les caractéristiques de la fonction de transfert de la trame courante sont considérées comme étant les mêmes que celles obtenues pour l'analyse de la trame précédente. Pour les segments sans parole, la méthode ANC conventionnelle est utilisée.

Des résultats expérimentaux sont présentés dans ce chapitre. Ils montrent que la méthode proposée donne de meilleurs résultats pour l'estimation du spectre et pour la reconnaissance de la parole que la méthode de soustraction spectrale classique. Ce résultat est vrai y compris dans des environnements acoustiques sévères caractérisés par des rapports signal à bruit inférieurs à 9dB.

Le chapitre 3 propose une nouvelle technique de dé-réverbération en considérant les caractéristiques fréquentielles des surfaces réfléchissantes. Les excès de soustraction se produisent lorsque les caractéristiques fréquentielles des surfaces réfléchissantes sont plates. En effet, dans ce cas l'estimation du temps de réverbération donne un résultat plus long que ce qu'il devrait être sur plusieurs bandes de fréquences. Pour résoudre ce problème, il est nécessaire d'estimer le temps de réverbération en tenant compte des caractéristiques fréquentielles de la réflexion. La méthode proposée dans cette thèse est une technique de dé-réverbération aveugle simple canal qui est basée sur l'utilisation de la fonction d'auto-corrélation de séquences temporelles des composantes fréquentielles.

Les étapes de cette technique sont les suivantes. La fonction d'auto-corrélation pour chaque bande de fréquence est calculée en utilisant une série temporelle de fréquences du spectre. Le temps de réverbération pour chaque bande de fréquence est déterminé comme un retard temporel au-delà duquel la fonction d'auto-corrélation devient suffisamment petite. La fonction d'auto-corrélation jusqu'au temps de réverbération estimé est vue comme un affaiblissement caractéristique de la réflexion. La dé-réverbération est réalisée en utilisant l'estimation du temps de réverbération et le taux d'affaiblissement dans le domaine fréquentiel. Il est supposé dans ces travaux que le signal reçu est constitué de l'onde directe et de la somme des composantes réfléchies. La technique proposée procède par soustraction des composantes réfléchies une par une.

En s'appuyant sur les performances obtenues sur des données simulées et réelles, la méthode proposée s'avère capable d'estimer le temps de réverbération en fonction de la fréquence et donne ainsi de meilleurs résultats que les méthodes conventionnelles de dé-réverbération. Cette méthode a été appliquée à la commande vocale d'un téléviseur (voir Appendice C.1 de la thèse). Les résultats de l'évaluation employant ce système prouve que le taux de reconnaissance de la parole est amélioré dans les cas où le SNR = 0∼12dB. La reconnaissance de la parole est également correctement réalisée même dans le cas où la distance entre le microphone et un utilisateur serait de 50∼100cm.

Il est nécessaire de développer des méthodes qui peuvent estimer un signal de bruit pour améliorer les performances de la technique décrite au chapitre 2. Les procédures itératives ont besoin de temps de calcul importants pour réduire le bruit. Ainsi, le système ne peut pas réaliser la reconnaissance de la parole en ligne, à moins que la charge de traitement soit réduite. Hors, si le nombre maximum d'itérations est limité, la reconnaissance de la parole ne peut pas atteindre les meilleures performances. Les chapitres 4 et 5 proposent des techniques de réduction de bruit permettant de prendre en compte ces compromis.

Le chapitre 4 propose une technique permettant de solutionner le problème de permutation qui apparaît lorsque l'on utilise l'Analyse en Composante Indépendante (ICA) dans le domaine fréquentiel appliqué à la Séparation de Source Aveugle (BSS). La méthode BSS tente de séparer les signaux mélangés en signaux source sans information *a priori*. La méthode ICA dans le domaine fréquentiel donne des performances intéressantes même dans le cas de signaux mélangés résultant d'une convolution ou dans des cas de réverbération. Cependant, il existe une difficulté appelée « problème de permutation » lors de l'utilisation de la méthode ICA dans le domaine fréquentiel. L'affectation d'une source identifiée à chaque bande de fréquence est nécessaire après la séparation, mais il est difficile de disposer de moyens efficaces pour cela. Di Persia propose une méthode ICA sans permutation (PF - ICA) pour séparer les signaux provenant d'une convolution en signaux sources. Cependant cette technique à un défaut, celui de devoir supposer une directivité commune a toutes les bandes de fréquence. Il permet néanmoins d'éviter le problème de permutation en rassemblant toutes les bandes de fréquence en un seul vecteur. Une méthode ICA multi bandes (MB - ICA) est proposée dans cette thèse comme une version modifiée de PF - ICA. Elle effectue la séparation après avoir assemblée un nombre défini de bandes de fréquence adjacentes. MB - ICA peut faire face aux caractéristiques fréquentielles locales et apporter une solution stable au problème de permutation. Cependant, MB - ICA a un inconvénient : le nombre de bandes de fréquence qui doivent être assemblées pour obtenir le meilleur résultat possible dépend de l'environnement.

Le chapitre 5 propose une technique pour l'estimation du spectre de la parole en

utilisant un filtre particulaire et un microphone unique. Il existe dans la littérature une méthode utilisant un filtre particulaire pour estimer à la fois le bruit et le signal de parole propre modélisé par un modèle à mélange de gaussiennes (GMM). Cependant, un point essentiel pour mettre en œuvre cette méthode est la construction d'un GMM approprié. Un très grand nombre de données de parole est nécessaire pour la construction de ce modèle. Le mélange de processus de Dirichlet (DPM) est un modèle permettant le mélange d'une infinité de distributions de probabilité et il est possible de décider du nombre de distributions gaussiennes (ou autres densités d'ailleurs) nécessaires. La méthode proposée dans cette thèse permet l'estimation du bruit et du spectre de la parole sans avoir besoin de construire de GMM. Au lieu d'utiliser un GMM, le modèle du spectre de la parole est basé sur un DPM. Un processus de Dirichlet (DP) est une distribution de probabilité non paramétrique sur un espace constitué de toutes les distributions possibles. Le DP est utilisé comme une distribution *a priori* du DPM qui nous permet alors de mélanger un nombre infini de distributions de probabilité. En utilisant un modèle basé sur l'utilisation de DPM pour l'estimation du spectre de la parole, nous pouvons donc envisager de développer une méthode pour une estimation adaptative du spectre. Lors de sa mise en œuvre nous avons pu constater que nous obtenons une meilleure estimation du bruit et un meilleur taux de reconnaissance de la parole que les méthodes conventionnelles utilisant un GMM.

Le chapitre 6 conclue cette thèse et présente les perspectives de ces travaux.

Abstract

Speech recognition technology reaches almost a practical level if we use a close contact microphone in quiet environments. However, speech input using a close contact microphone has a disadvantage, that is to say a headset must be put on in order to keep a small microphone just in front of the mouth. If we do not use a close contact microphone, we have to speak to a microphone paying attention to the distance to the microphone in order to avoid breath puff to the microphone. So, it is desirable to develop a system which does not force users to pay attention to microphone position during utterance. If microphones are located at a distant position from a speaker, the speech recognition rate decreases drastically by noise and reflected waves. Furthermore, sometimes sounds emitted by the target system itself turn into noise which also decreases the speech recognition rate. Hence, it is necessary to develop noise reduction and dereverberation techniques for building practical voice-controlled systems.

This dissertation consists of the following six chapters which describe four signal processing techniques to achieve noise robust speech recognition in real environments.

Chapter 1 explains the necessity of noise reduction in speech recognition, and describes the past speech recognition researches and desirable signal processing techniques for realizing the robust speech recognition.

Chapter 2 proposes a technique for reducing obstructive sounds emitted by the target apparatus to be controlled. To reduce these sounds, the Spectral Subtraction (SS) is a popular and possible means. However, the SS cannot give a satisfactory performance in case temporal changes occur in the transfer functions between the loudspeakers of the apparatus and the microphone. An adaptive scheme is introduced to cope with time varying situations. Although the Adaptive Noise Canceller (ANC) is available to solve this problem, the excess subtraction might cause distortion in the estimated speech signal since the ANC sometimes cuts a part of the speech signal in case there is high correlation between the target signal and the noise signal. To solve this problem, the proposed system uses the harmonic structure of the voiced segments which the conventional ANCs do not directly take into account.

First, the proposed method extracts the fundamental frequency of the received signal. Then, using the fundamental frequency and the frequency spectrum, the system classifies the received signal into three segment types: (1) voiced segments; (2) unvoiced segments; and (3) non-speech segments. In case of voiced segments, the harmonic structure of the target speech is extracted based on the fundamental frequency. The estimated harmonic structure is subtracted from the received signal and then the noise signal in the mixed signal received by the microphone is estimated. The estimated noise signal is used for calculating the transfer characteristics for the

current analysis frame, together with the known noise which can be directly obtained from the target apparatus. In case of unvoiced segments, however, the target signal does not have harmonic structure, but the acoustic transfer characteristics from the target system, such as a TV set, to the microphone does not differ so much between every two consecutive frames. The transfer characteristics in the current analysis frame are assumed to remain the same as those obtained in the previous analysis frame. For non-speech segments, the proposed method is designed to work as same as the conventional ANCs.

Experimental results show that the proposed method realizes better frequency spectra estimation and speech recognition rate than the conventional SS in acoustically severe environments characterized by Signal to Noise Ratio (SNR) less than 9dB.

Chapter 3 proposes a new dereverberation technique by considering the frequency characteristics on reflective surfaces. Over-subtraction occurs in the conventional dereverberation in case flat frequency characteristics are assumed at reflection on surfaces because the reverberation time is estimated longer than actual values at several frequency bins. To overcome this problem, it is required to estimate the actual reverberation time by assuming the frequency characteristics of reflection. Proposed is a single channel blind dereverberation technique which was the auto-correlation functions of the time sequences of the frequency components.

The procedure of this technique is described as follows: the auto-correlation function for each frequency bin is calculated using a time series of frequency spectra. The reverberation time for each frequency bin is determined as the time delay beyond which the auto-correlation function gets enough small. The auto-correlation function up to the estimated reverberation time is regarded as the decay characteristics of the reflection. Dereverberation is performed using the estimated reverberation time and the decay rate in the frequency domain. It is assumed that the received signal consists of the direct wave and the sum of reflection components. The proposed technique subtracts the reflection components one by one.

From the performance on simulated and actual data, the proposed method proves to be able to estimate the reverberation time as a function of frequency, and thus yield better results than the conventional dereverberation methods. The proposed method is applied to a voice-controlled TV system described in Appendix C.1. Evaluation results using this system shows that the speech recognition rate is improved in the cases where SNR= 0$\sim$12dB and the speech recognition was successful even in case the distance between the microphone and a user is 50$\sim$100cm.

It is required to develop a method that can estimate a noise signal for improving the performance of the technique described in Chapter 2. Iterative procedures require long processing time to reduce the noise. So, the system cannot achieve on-line sequential speech recognition unless the processing load is reduced. However, if the maximum iteration count is limited to be small, speech recognition cannot reach the

best performance. From the point of this trade-off, chapters 4 and 5 propose noise reduction techniques expected in the future.

Chapter 4 proposes a technique to escape from the permutation problem which appears in the frequency-domain Independent Component Analysis (ICA) applied to the Blind Source Separation (BSS). BSS tries to separate mixed signals into source signals without knowing any *a priori* information. The frequency-domain ICA is employed and gives fair performance even for convolutively mixed signals or reverberant cases. In the frequency-domain ICA, convolutional mixture is converted into simple additive mixture in each frequency bin by converting the observed signals into the frequency domain. However, there exists a tough issue called "permutation problem" in frequency-domain ICA. Consistent assignment of source identification for each frequency bin is required after the separation for each frequency bin, but we do not have any reliable means for that. Di Persia proposed the permutation-free ICA (PF-ICA) for separating convolutively mixed signals into source signals escaping from the permutation problem. This technique, however, has a defect that it assumes a single directivity common to all frequency bins, though it can avoid the permutation problem connecting all the frequency bins into one long vector. A Multi-bin ICA (MB-ICA) is proposed here as a revised version of PF-ICA. It performs separation after connecting a definite number of adjacent frequency bins. By connecting adjacent frequency bins, MB-ICA can cope with local frequency characteristics and stably solves the permutation problem. However, MB-ICA has a drawback that the number of frequency bins to be connected for giving the best result depends on the environment.

Chapter 5 proposes a technique to estimate speech spectrum using a particle filter with a single microphone. There is a method for estimating both noise signal employing particle filter and clean speech signal employing a Gaussian Mixture Model (GMM). However, an essential point of this technique is to construct an accurate GMM in advance and huge number of speech data are required to construct the model. Dirichlet Process Mixture is a model to mix infinite probability distributions and it can decide the number of required Gaussian distributions. Proposed is a technique to estimate noise and speech spectra without building the GMM. Instead of the GMM, the speech spectrum is modeled based on a Dirichlet Process Mixture (DPM). The Dirichlet Process (DP) is a non-parametric probability distribution over a space consisting of all possible distributions. The DP is used as the prior distribution of the DPM. As the DP is a generative model for infinite distributions, DPM allows us to mix the infinite probability distributions. Using a model based on DPM in the estimation process of the speech spectrum, it is expected to develop a method to estimate the spectrum adaptively. In evaluation using speech recognition, the proposed method realizes better noise estimation and achieves better speech recognition rate than the conventional method using GMM.

Chapter 6 concludes this dissertation and describes future works.

概要

音声認識技術の向上により，比較的雑音の少ない環境において接話マイクロホンによる音声認識は，現在，ほぼ実用的なレベルに達している．しかし，接話マイクロホンを用いる場合は，ヘッドセットマイクなどを常に身に着けていなければならず，非常にわずらわしい．そのため，ユーザがマイクロホンに気を配ることなくシステムを利用できるようにする必要がある．つまり，音声認識システムの実用化には，マイクロホンから離れた位置からのユーザ発話でも正しく認識できる必要がある．しかし，話者から離れたところにマイクロホンを設置すると，認識対象音声のパワーが小さくなり，周囲の雑音や壁や天井からの反射音の影響により音声認識率は大幅に低下する．また，操作する対象の機器 (例えば TV) 自体が音を発する場合はその影響も受ける．それゆえ，実用的なシステムを構築するためには，雑音除去技術や残響除去技術の導入が不可欠である．

本論文では，実環境において頑健に動作する音声認識のための信号処理技術について述べている．本論文は以下の６章から構成されている．

第１章は本論文の導入であり，これまでの音声認識研究および，音声認識率改善のための信号処理技術について述べている．

第２章は，音声により操作する対象機器自体が発する音を除去する技術について述べている．TV のように，操作対象自体が音を発する機器を音声によって操作する場合，機器自体の発する音が雑音となり音声認識率が低下する．このような問題に対しては，操作対象自体が発する音は既知という条件の下で，適応ノイズキャンセラを用いることができるが，話者の発話中においても TV-マイクロホン間の伝達特性が人の動きや室温，湿度の変化などにより変動する．また，TV 音の場合，壁などからの反射音とシステム利用者の音声は相関が高いため，従来の適応ノイズキャンセラでは十分な雑音除去性能が得られないという事情もある．そこで，従来の適応ノイズキャンセラでは用いられていない音声の調波構造を考慮した雑音除去法を開発することによりこの問題を解決する手法を提案している．

提案法では，入力信号をその振幅により認識対象音声が含まれる区間であるか，そうでないかを判定する．そして認識対象音声が含まれる区間はさらにスペクトルの特徴から有声音であるか，無声音であるかを判定する．有声音の区間に関しては基本周波数を基に調波構造を推定し，推定した調波構造を入力信号から減算することによって，雑音信号を推定する．そして，この雑音の推定値と既知としている操作対象自体が発する音を用いて伝達特性の学習が行われる．無声音の区間に関しては，調波構造を用いることができないため，前フレームで推定した伝達特性を流用し，雑音信号の推定を行う．認識対象音声が含まれない区間に関しては従来の適応ノイズキャンセラと同じ動作をする．

提案法は，SN 比が 0〜9dB の環境において，従来の単純なスペクトルサブトラクション法や Wiener フィルタと音声認識率を比較しても十分よいもしくは同程度の性能を持つことを示している．ただし，SN 比が 12dB 以上の良い音響環境での音声認識率は，改善の余地が無く，事実改善が認められない．

第３章は，反射の周波数特性を考慮した反射音除去技術について述べている．第２章の雑音除去法では，雑音源とマイクロホン間の伝達特性を考慮しているが，話者とマイクロホン間の伝達特性を考慮していない．従来の反射音除去処理では，壁や天井での反射の周波数特性が平坦であることを仮定した複数のマイクロホンによる手法が用いられていた．しかし，これらの手法では，残響時間が真の値よりも長く推定されることにより過剰な減算が起こる可能性があった．そこで，この問題を解決するために，ここでは単一マイクロホンによる各周波数ビンでの自己相関関数を基に，残響時間を周波数の関数として推定する手法を提案している．これにより壁や天井での反射の周波数特性を考慮した反射音除去処理が可能になり，過剰な減算を抑制することができる．シミュレーションおよび実際に収録した音声を用いて評価を行ったところ，過剰な減算が抑制され，SN 比が改善することが明らかになっている．また付録 B に示す音声操作による TV システムを用いて，音声認識率の評価を行ったところ，SN 比が 0〜12dB の範囲で認識率の改善が得られ，マイクロホンから 50〜100cm 離れた位置からの音声認識が可能であることが明らかになった．

第２章で述べた雑音除去技術を実現するためには，雑音のみの信号が得られなければならない．また，逐次的な手法で精度良く雑音を除去するためには，伝達特性の学習のために多くの時間を要する．そのため，システムは計算量を減らさない限り連続的に音声認識をすることが困難である．しかし，逐次計算の繰り返し回数を制限するなどで計算回数を減らすと音声認識率が低下する．以上のことより，第４，５章では今後音声認識に期待される雑音除去技術を提案している．

第４章は，周波数領域での独立成分分析 (ICA) によるブランインド音源分離 (BSS) で問題となるパーミュテーション問題を抑制する ICA の提案を行っている．BSS は音源に関する事前情報が無い状況で，複数の音源からの音を波形レベルで分離する技術である．BSS の手法の一つとして，ICA に基づく手法が近年盛んに研究されており，瞬時 (畳み込みでない) 混合に対しては十分な分離性能が得られている．また，畳み込み混合に対処する方法としては周波数領域 ICA がある．これは信号を短時間フーリエ変換により周波数領域に変換することによって，周波数ビン毎の瞬時混合として ICA を適用するものである．しかし，BSS に周波数領域 ICA を適用した場合，各チャネルの周波数ビンの入れ違いを正しく並び替える (パーミュテーション問題を解決する) 必要がある．しかし，いまだパーミュテーション問題の完全な解決には至っていない．Di Persia らは，これまでに全周波数ビンを連結して１本のベクトルと考えて分離を行うパーミュテーション・フリー ICA（PF-ICA）を提案している．この手法では，分離された信号の周波数ビン間の入れ替わりが生じない利点があるが，分離行列の周波数特性を考慮することができないという欠点があった．そこで，筆者は，対象とする周波数ビンの前後数本の周波数ビンを連結することによって，当該周波数ビンに関する分離行列を安定して求めるマルチビン ICA（MB-ICA）を提案し，分離行列の周波数特性を考慮した分離ができるようにした．

第５章は，粒子フィルタを用いて単一マイクロホンで音声のスペクトルを推定する手法について述べている．システムの実用化を考えると，少数のマイクロホンでシ

ステムを実現することが望ましい．それゆえ，単一マイクロホンで雑音除去，および残響除去を行う技術の研究が要請される．単一マイクロホンによって，粒子フィルタを用いた雑音のスペクトルの推定と，ガウス混合モデル (GMM) を用いた音声のスペクトルの推定法が開発されている．粒子フィルタは近年，計算機性能の向上により普及したものであり，従来のカルマンフィルタが線形・ガウス型の状態空間にしか適用できなかったものを，非線形・非ガウス型の状態空間にも適用できるように拡張したフィルタである．しかし，GMM を用いた手法は音声信号の追従性能が GMM の学習精度に依存する．一方，音声のスペクトルのモデル化に GMM ではなくディリクレ混合過程 (DPM) に基づくモデルを用いると，混合分布数をデータから自動的に得ることができるため，予めモデルを学習する必要がなくなり，より柔軟なモデル化が可能になる．DPM とは，ディリクレ過程により生成された確率分布の混合モデルであり，ディリクレ過程は混合する各要素の確率分布の確率分布を表現するものであり，無限個の確率分布を生成するためのモデルである．また，状態空間モデルに話者とマイクロホン間の伝達特性の影響を導入し，雑音除去とともに残響抑圧も行う手法も提案している．実環境音声・音響データベースに収録されている３種類の雑音データ，および４種類の SN 比（0, 3, 6, 9dB）を用いて，DPM に基づくモデルを用いた雑音除去法の評価を行っている．評価の結果， DPM に基づくモデルを用いた提案手法が GMM による手法よりも音声認識率を改善することを確認している．

第６章は，本論文のまとめおよび今後の課題について述べている．

Acknowlegdments

This dissertation is the outcome of a joint Ph.D program between the graduate school of Engineering, Doshisha University and École Centrale de Lille, France. I went to France to obtain a doctoral degree during my Ph.D student period. I was a member of *Laboratoire d'Automatique Génie Informatique et Signal* (LAGIS) in *École Centrale de Lille* for 10 months, from 1st September 2006 to 30th June 2007.

I greatly thank my principal Ph.D adviser, Professor Masuzo Yanagida, for everything he has done. I cannot accomplish the work reported in this dissertation without him. I am very glad to study under him.

I also thank my second Ph.D advisers, Professor Philippe VanHeeghe and Professor Emmanuel Duflos, for their patience and kindly teaching. I was not able to well speak French even English, nevertheless they took long time to discuss with me. I could not achieve my overseas study without them.

I thank Professor Yoichi Yamashita, Professor Mehdi Nouri-Shirazi and Professor Seiichi Yamamoto for their useful comments. I also thank the reporters, professor Hugo Leonardo Rufiner and Professor Noureddine Ellouze.

Moreover, I am also deeply grateful to Professor Sylvienne Wignacourt, Ms. Monique Bukowski, Ms. Marie-Françoise Toricot, Ms. Christine Yvoz and Ms. Michiko Jono for their kindly support and accommodation. I was able to concentrate entirely on my work owing to them.

I would like to express my gratitude to the members at Intelligent Mechanism Laboratory for their assistance for my work, especially Toshiyuki, Akihiro, Hidehiko, Tadashige, Masanori, Tsubasa, Kouhei and Ayako. I am thankful to the members at *Laboratoire d'Automatique Génie Informatique et Signal* (LAGIS), especially François, Corentin and Thomas.

I would like to also express my gratitude to all of my friends in France for their kindness. I thank to my parents in France, Françoise and Gérard, for their warmhearted accommodation.

Contents

List of Figures

List of Tables

List of Publications

Journal Papers

[1]Kenko Ota, Emmanuel Duflos, Philippe VanHeeghe and Masuzo Yanagida, "Bayesian Inference for Speech Density Estimation by the Dirichlet Process Mixture", Journal of Studies in Informatics and Control, vol. 16, No. 3, pp. 227–244, 2007.
[2]Kenko Ota, Tadashige Noguchi, Kohei Yasui, Leandro Di Persia and Masuzo Yanagida,"Frequency Domain ICA Connecting of Adjacent Frequency Bins", IEICE transaction, D Vol.J91-D, No.01, pp. 130-135, 2008.

International Conferences with Review

[3]Kenko Ohta and Masuzo Yanagida, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-wise Time Sequences of Frequency Components", IWAENC, No. 49, Paris, Sep., 2006.
[4]Kenko Ohta, Keiji Yasuda, Genichiro Kikui and Masuzo Yanagida, "Quantitative Evaluation of Effects of Speech Recognition Errors on Speech Translation Quality," EUROSPEECH, 4CP1c-6, Lisboa, Sep., 2005.
[5]Kenko Ohta, Leandro Di Persia and Masuzo Yanagida, "Removing Reflected Waves Using Temporal and Spectral Subtraction Without a priori Knowledge," IEEE NSIP, 19PM1D-02, Sapporo, May, 2005.

International Conferences without Review

[6]Kenko Ota, Tatsuya Yamazaki and Masuzo Yanagida, "Adaptive Filtering Using Harmonic Structure of Voiced Speech for Reducing Non-stationary Known Noise," ASA/ASJ Joint Meeting, 4aSP9, Honolulu, Nov., 2006.
[7]Kenko Ota and Masuzo Yanagida, "Blind Dereverberation Based on Auto-Correlation Functions of Frame-wise Time sequences of Frequency Components", ASA/ASJ

Joint Meeting, 1pSP7, Honolulu, Dec., 2006.
[8]Tadashige Noguchi, Kenko Ohta, Leandro Di Persia and Masuzo Yanagida, "Frequency Domain Independent Component Analysis by overlap piecewise integration of separation processing," ASA/ASJ Joint Meeting, 1pSP12, Honolulu, Nov., 2006.
[9]Masanori Enoki, Tadashige Noguchi, Tsubasa Arai, Ayako Miyake, Kohei Yasui, Kenko Ota, Masuzo Yanagida and Masaki Ida, "Speech Interface for Operating Information Technological Home Appliances," ASA/ASJ Joint Meeting, 1pSC19, Honolulu, Nov., 2006.
[10]Kenko Ohta and Masuzo Yanagida, "Blind Dereverberation Using Temporal and Spectral Subtraction," Forum Acousticum, 388-0, Budapest, Sep., 2005.

Domestic Meetings

[11]Yoshihiro Morimoto, Tadashige Noguchi, Leandro Di Persia, Kenko Ota and Masuzo Yanagida, "Effects of Wiener Filter on Blind Source Separation by post-processing on sequentially connected bins", 2007 Kansai-section Joint Convention of Institutes of Electrical Engineering, Kobe Univ., Nov., 2007
[12]Tadashige Noguchi, Kenko Ota, Leandro Di Persia and Masuzo Yanagida, "Effects of Wiener Filter on Blind Source Separation by processing on sequentially connected bins", Technical Meeting on Electrical Acoustic, NAIST, Sep., 2007
[13]Masanori Enoki, Kenko Ota, Masaki Ida, Hiroshi Nagaoka, Shigeki Matsuda, Satoshi Nakamura, Yutaka Kida and Masuzo Yanagida, "Introducing Speech Interface to IT Home Appliances in the Next Generation," 2007 IEICE General Conference, D-14-6, Meijyo Univ., Mar., 2007.
[14]Kohei Yasui, Tadashige Noguchi, Kenko Ota, Leandro Di Persia and Masuzo Yanagida, "Source Separation by Frequency-Domain ICA on Locally Connected Frequency Bins," Symposium on Signal Processing, Vol.2006, No.136, Nagoya Univ., Dec., 2006.
[15]Ayako Miyake, Tsubasa Arai, Masanori Enoki, Kenko Ota, Masaki Ida and Masuzo Yanagida, "A Study on a TV Program Retrieval for a Voice-Controlled TV system," ASJ Kansai brunch Meeting, No. 4, Kyoto Campus Plaza, Dec., 2006.
[16]Kenko Ota, Masanori Enoki and Masuzo Yanagida, "Robust Speech Recognition in a Known-Noise Case," 2006 Kansai-section Joint Convention of Institutes of Electrical Engineering, G16-3, Osaka Institute of Technology, Nov., 2006.
[17]Kohei Yasui, Tadashige Noguchi, Kenko Ota, Leandro Di Persia and Masuzo Yanagida, "Permutation-Free ICA Considering the Frequency Characteristics," 2006 Kansai-section Joint Convention of Institutes of Electrical Engineering, G16-13, Osaka Institute of Technology, Nov., 2006.
[18]Tsubasa Arai, Ayako Miyake, Masanori Enoki, Kenko Ota and Masuzo Yanagida, "Introduction of TV Program Retrieval into Voice-Controlled TV System," 2006 Kansai-section Joint Convention of Institutes of Electrical Engineering, G12-23, Osaka Institute of Technology, Nov., 2006.

[19]Kenko Ota, Masanori Enoki, Tatsuya Yamazaki and Masuzo Yanagida, "Suggestion of a Non-Stational-Noise Reduction Method in Case of Known-Noise," Symposium on Signal Processing, C6-4, Kyoto Univ. Assembly Hall, Nov., 2006.
[20]Kenko Ota, Tadashige Noguchi, Leandro Di Persia and Masuzo Yanagida, "Frequency Domain ICA Using a Connection of Adjacent Frequency Bins," Symposium on Signal Processing, C9-3, Kyoto Univ. Assembly Hall, Nov., 2006.
[21]Kenko Ota, Kohei Yasui and Masuzo Yanagida, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-wise Time Sequences of Frequency Components," ASJ, 1-1-19, Kanazawa Univ., Sep., 2006.
[22]Tadashige Noguchi, Kenko Ota, Leandro Di Persia and Masuzo Yanagida, "Frequency Domain ICA Connecting Adjacent Frequency Bins," ASJ, 1-1-12, Kanazawa Univ., Sep., 2006.
[23]Kenko Ota, Masanori Enoki and Masuzo Yanagida, "Investigation of Speech Recognition Rate and Processing Speed of a TV System introduced processing of Reduction of Known Noise," ASJ, 3-6-5, Kanazawa Univ., Sep., 2006.
[24]Kenko Ota, Tatsuya Yamazaki and Masuzo Yanagida, "A Study on Effectiveness of a Known-Noise Reduction Method," Technical Meeting on Speech, SP2006-2, Doshisha Univ., May, 2006.
[25]Leandro Di Persia, Tadashige Noguchi, Kenko Ota and Masuzo Yanagida, "Performance of Permutation-Free ICA," Technical Meeting on Speech, SP2006-1, Doshisha Univ., May, 2006.
[26]Leandro Di Persia, Kenko Ota and Masuzo Yanagida, "A Method for Solving the Permutation Problem in ICA," Technical Meeting on Speech and Hearing, SP2005-193, Jyochi Univ., Mar., 2006.
[27]Kenko Ota, Tatsuya Yamazaki and Masuzo Yanagida, "Comparison of the Effect of Known-Noise Reduction Method in Different Environments," ASJ, 2-11-1, Nihon Univ., Mar., 2006.
[28]Toshiyuki Sakai, Kenko Ota, Tatsuya Yamazaki and Masuzo Yanagida, "Speech Recognition introducing word weights controlled by the state of information appliances," SIG-SLUD, SIG-SLUD-A503-6, The National Institute for Japanese Language, Mar., 2006.
[29]Akihiro Handa, Leandro Di Persia, Kenko Ota and Masuzo Yanagida, "Separation of mixed speech signals of short duration using Wiener filter as postprocessing for Frequency-Domain ICA," Technical Meeting on Speech, Vol.2006, No.12, Atami, Feb., 2006.
[30]Kenko Ota and Masuzo Yanagida, "Adaptive Known-Noise Reduction Using Harmonic Structure of Voiced Speech," ASJ Kansai brunch Meeting, No. 3, Kyoto Campus Plaza, Dec., 2005.
[31]Masanori Enoki, Kenko Ota and Masuzo Yanagida, "Target Speech Enhancement by a delay-and-sum array avoiding interporation," 2005 Kansai-section Joint Convention of Institutes of Electrical Engineering, G16-13, Kyoto Univ., Nov., 2005.
[32]Tadashige Noguchi, Toshiyuki Sakai, Kenko Ota and Masuzo Yanagida, "Effects of Noise and Reverberation on Speech Recognition Rate," 2005 Kansai-section Joint

Convention of Institutes of Electrical Engineering, G16-15, Kyoto Univ., Nov., 2005.
[33]Kenko Ota and Masuzo Yanagida, "Adaptive Noise Reduction in the Known Noise Case," ASJ, 3-7-1, Tohoku Univ., Sep., 2005.
[34]Kenko Ota and Masuzo Yanagida, "Noise Reduction Based on the Harmonic Structure of the Speech Signal in the Known Noise Case," FIT2005, G-015, Chuo Univ., Sep., 2005.
[35]Kenko Ohta, Keiji Yasuda, Genichiro Kikui and Masuzo Yanagida, "Quantitative Evaluation of Effects of Speech Recognition Errors on Speech Translation Quality", NLP2005, A3-4, Kagawa Univ., Mar., 2005.
[36]Kenko Ota and Masuzo Yanagida, "Removing Reflected Waves Using Delayed Wave Subtraction Combined with Spectral Subtraction", 6-th IPSJ SIG-SLP, Vol.2004, No.131, ATR Dec. 2004.
[37]Kenko Ota and Masuzo Yanagida, "Removing Reflected Waves and Reverberant Component Using Waveform Subtraction Combined with Spectral Subtraction", 2004 Kansai-section Joint Convention of Institutes of Electrical Engineering, G16-11, Doshisha Univ., Nov. 2004.
[38]Kenko Ota and Masuzo Yanagida, "Removing reflected waves using delay time detected on auto-correlation functions", ASJ, 2-4-20, Ryukyu Univ., Sep. 2004.
[39]Kenko Ohta and Masuzo Yanagida, "Removing Reflected Waves Using Delay Time Detected by Majority Decision on Auto-correlation Functions", FIT2004, G-010, Doshisha Univ., Sep. 2004.
[40]Kenko Ohta and Masuzo Yanagida, "Reverberation Removal for Improving Speech Recognition Rate in a Real Environment", 66-th JPSJ, 4L-6, Keio Univ., Mar. 2004.

Chapter 1

Introduction

Speech recognition technologies have achieved rapid advancement. The reason for it is that the probability model, called Hidden Markov Model (HMM), was introduced and large-scale corpora have been developed. HMM, which is now widely used in speech recognition, was proposed in late 70s expressing the temporal and spatial fluctuation of speech. The Defense Advanced Research Projects Agency (DARPA) project began to use the HMM for speech recognition and to develop a common large-scale corpus. Owing to promotion of this project, speech recognition technologies have achieved advancement. Before the project started, speech recognition technologies could only cope with the word recognition for a specific speaker within a closed task. Now, they can cope with more complicated tasks. As the results of the research, current speech recognition technologies reach almost a practical level in case of being used in quiet environments with a close contact microphone.

However, to use speech input with a close contact microphone forces us to inconveniently wear a headset that mounts a small microphone kept a fixed distance to the mouth. So, it is required to develop a system which does not require users to pay attention to the microphone in practical use. If microphones are allocated in a position far from the speaker, the speech recognition rate decreases drastically affected by noise and reflected waves. In such environments, the reason why the speech recognition performance decreases is that the acoustical environment differs from one in which the acoustic model was built. Moreover, precise detection of speech segments also proves to be important for speech recognition in practical applications.

If microphones are placed not near the speaker, speech recognition is severely affected by noises and room acoustics. Moreover, we have to take the human factors in their response into account for building a speech dialogue system.

Various techniques have been proposed to cope with noises and reverberation.

There can be the following three types of processing schemes for improving the situation: i) extraction of feature parameters representing perceptual characteristics, ii) removing noise and normalizing channel distortion and iii) adaptation to environments.

i) extraction of feature parameters representing perceptual characteristics

Techniques for feature extraction for speech recognition are studied to realize robust extraction of speech features to the environmental changes. Recently mel frequency cepstral coefficients are widely used as feature parameters for speech recognition, as they are robust to the environmental changes. Moreover, the Perceptual Linear Prediction (PLP) [1], RASTA [2] and dynamic cepstrum [3] draw attentions as these parameters take the auditory characteristics into account.

ii) removing additive noise and normalizing channel distortion

This processing can be classified into the single microphone processing and multiple microphone processing.

In single microphone cases, Spectral Subtraction (SS) [4], missing feature theory [5] and Cepstral Mean Subtraction (CMS) [6] are generally used. SS detects noisy segments and subtracts estimated power spectrum from the received signal estimating the noise power spectrum. In low signal-to-noise ratio (SNR) cases, however, noise removal is difficult. Moreover, in case the noise signal is non-stationary, it becomes much more difficult to estimate the noise spectrum precisely. The missing feature theory is a method to recover information buried in noise, while the CMS can normalize channel distortion. In the cepstrum domain, convolution is represented by linear expression, so the effects of room acoustics are reduced by subtracting noise cepstrum from the cepstrum of the observed signal.

In multiple microphone cases, we can use microphone array, multi-channel SS, blind source separation and so forth. Two-channel SS estimates the frequency spectrum of the target source by subtracting noise spectrum estimated from two-channel mixed signals. BSS (Blind Source Separation) is to separate mixed signals into each source signal without knowing any *a priori* information. Microphone array processing is a method for suppressing noise signals controlling the directivity of signal enhancement with a multi-channel microphone system. To suppress noise signals using a microphone array, the directions of the target and noise sources are detected and directivity is formed by delay-and-sum among microphones in the frequency domain. Cross-power Spectrum Phase Analysis (CSP) [7] is often used for estimation of source direction. There are two types of microphone array, additive array and subtractive array, for forming directivity. An additive array forms directivity for target source direction and a subtractive array forms null directivity for noise source.

iii) adaptation to environments

Adaptation to environments means automatic modification of the acoustic model for speech data obtained in arbitrary acoustic environments in which speech recognition is performed. Adaptation techniques are classified into one of the following three types: a) mapping onto a feature vector space, b) adaptation of the acoustic model parameters and c) acoustic model composition and/or decomposition. The first one, e.g. SPLICE [8], maps features of observed signal onto a feature vector space for signals affected by noises and reverberation. The second one leads the Maximal Likelihood Linear Regression (MLLR) and the Maximal *a posteriori* Estimation (MAP). The last one, e.g. [9], [10], [11], [12], is a method which composes an adapted acoustic model from noise HMMs and speech HMMs.

The objective of the research described in this dissertation is to develop signal processing techniques for realizing a robust speech recognition. The dissertation focused on the removing noise and reverberation for improving speech recognition. Proposed in this dissertation are known noise reduction, a dereverberation technique using a single microphone, BSS using Independent Component Analysis (BSS) and single channel noise reduction based on a Dirichlet Process Mixture (DPM) model.

We want to develop user interface that is easy to use for everyone. Speech recognition draws our attention as interface because speech is the most important and convenient communication media for almost all people.

In case of using distant microphones, speech recognition rate decreases due to noise, reverberation and, in some cases, the sound emitted by the target apparatus to be controlled, such as a TV set. SS can be used to reduce the sounds emitted by the target apparatus itself. However, SS cannot give satisfactory noise removal performance under situations where the transfer functions between the loudspeakers of the apparatus and the microphone is time variant.

Although Adaptive Noise Canceller (ANC) is available in time varying cases, over-subtraction might cause distortion in the estimated speech signal if there is high correlation between the target signal and the noise signal. Introduced in chapter 2 is an ANC using the harmonic structure of voiced speech segments which the conventional ANCs have not directly taken into account. Sounds from the target apparatus to be controlled can be removed as they are known to the system and the transfer characteristics from the apparatus to the microphone can be calculated by the proposed method in chapter 2. The sounds from the target apparatus to the microphone can be calculated as convolution of the source sound and the transfer characteristics of the path. While the reflected speech uttered by the speaker should be removed with some means of dereverberation for speech recognition. So, we introduced dereverberation techniques shown in Chapter 3.

In Chapter 3, proposed is a technique in frequency domain using a single microphone. Over-subtraction occurs in conventional dereverberation in case flat frequency characteristics are assumed on reflective surfaces because the reverberation time is estimated longer than actual values at several frequency bins. It is required

to estimate the actual reverberation time assuming the frequency characteristics of reflection. Proposed is a single channel blind dereverberation technique using auto-correlation functions on the time sequences of frequency components.

The procedure of this technique is described as follows: The auto-correlation function for each frequency bin is calculated using a time series of frequency spectra. The reverberation time for each frequency bin is determined as the time delay beyond which auto-correlation function gets enough small. The auto-correlation function up to the estimated reverberation time is regarded as the decay characteristics of reflection. Dereverberation is performed using the estimated reverberation time and the decay rate in the frequency domain. It is assumed that the received signal consists of the direct wave and the sum of the reflection components. The proposed technique subtracts the reflection components one by one.

It is required to develop a method that can estimate a noise signal for improving the performance of the method described in Chapter 2. It is hard for the system to achieve on-line continuous/sequential speech recognition unless the processing load is reduced. However, iterative procedures require long processing time to reduce the noise. If we limit the maximum iteration count to be small, speech recognition cannot reach the best performance. From the point of this trade-off, a system should introduce noise reduction techniques other than iterative methods.

BSS is a good means for noise removal in case we do not have any *a priori* information about source signals. BSS by independent component analysis (ICA) is employed in Chapter 4. The frequency-domain ICA is employed for separating convolutively mixed signals in reverberant cases. In the frequency-domain ICA, convolutional mixture is converted into simple additive mixture in each frequency bin, by taking Fourier transform. However, there exists a difficulty called the "permutation problem" in the frequency-domain ICA. The permutation problem requires us correct assignment of source identification for every frequency bins after separation for each frequency bin. There are many techniques proposed to cope with the permutation problem, but the problem has not been completely solved yet. Di Persia proposed a permutation-free ICA (PF-ICA) for separating convolutively mixed signals into source signals without permutation errors. This technique has an advantage that it can avoid the permutation problem, but has a defect that it assumes a single directivity common to all frequency bins. Multi-bin ICA (MB-ICA) is proposed here as a revised version of PF-ICA. It performs separation after connecting a definite number of adjacent frequency bins.

It is desirable for a more practical system to develop a method requiring only a single microphone. Although ICA is a powerful means, it requires more than two microphones.

In a single microphone case, it is necessary to estimate a noise signal precisely. Currently, a particle filtering is widely used in a speech recognition field due to advancement of computer abilities. Speech and noise spectrum estimation technique

with a single microphone is proposed in Chapter 5. There is a method for estimating noise signal employing a particle filter and for estimating speech spectrum employing a Gaussian Mixture Model (GMM). However, an essential point of this technique is to construct an accurate GMM in advance and huge number of speech data are required to construct the accurate GMM. Dirichlet Process Mixture is a model to mix infinite probability distributions and it can decide the number of required Gaussian distributions. Proposed is a technique to estimate the noise and speech spectrum without building the GMM. Instead of the GMM, the speech spectrum is modeled using a model based on Dirichlet Process Mixture (DPM). The Dirichlet Process (DP) is a non-parametric probability distribution over a space of all possible distributions. The DP is used as the prior distribution of the DPM. The DP is a generative model for infinite distributions. So, DPM allows us to mix the infinite probability distributions. By using a model based on the DPM in the estimation process of the speech spectrum, we estimate this spectrum adaptively.

Chapter 2

Known noise reduction using harmonic structure

2.1 Introduction

Our research objective is to realize an effective TV control system using speech recognition which can operate in noisy environments. Even if we want to control the TV set by voice, we do not want to use a close contact microphone, but rather prefer using microphones placed somewhere from the speaker, not knowing nor being aware of source location. In case of using distant microphones, however, the speech recognition rate decreases making due to the sound that the TV set itself is emitting.

To reduce the effect of the TV sound, the spectral subtraction (SS) is a possible means [4]. However, we cannot expect satisfactory performance from SS because of temporal changes of the transfer functions between the loudspeakers of the TV set and the microphone. So, we have to take an adaptive scheme to cope with the time varying situation [20]. If there is high correlation between a target signal and a noise signal, excess subtraction might cause distortion on the estimated speech signal since the adaptive filter sometimes functions to cut a part of the speech signal [21]. To avoid this situation, the filter coefficients are updated only for non-speech segments. The transfer function of the path from the source to the microphone might be easily affected by disturbances in real environments [22]. Thus, in case fluctuation of the transfer function occurs during speech segments, noise removal performance will be degraded, resulting in serious decrease in speech recognition rate. So, adaptation is required for modifying the filter coefficients even for speech segments.

We focused attention on the harmonic structure of voiced speech segments though the conventional ANCs have not directly taken it into account [23], [24]. The technical objective of these research is how to accelerate the convergence speed of calculat-

ing filter coefficients. So, introducing the harmonic structure and the fundamental frequency into the noise reduction process, we can expect more satisfactory results in noise reduction.

2.2 The method

In case we want to change TV channels, the TV set should be already ON and may be emitting sounds from its loudspeakers. The microphone in the room inevitably receives the sound from the TV set itself together with command speech which the system has to recognize. In this case, however, the sound from the loudspeaker of the TV set is completely known though the sound may be considerably modified on the way to the microphone. If we know the transfer characteristics of the path from the TV set to the microphone including reflections on the walls, ceiling, floor and furniture, we will be able to estimate the sounds from the TV set arriving at the microphone. To estimate the transfer characteristics, we employ an ANC technique using the harmonic structure of the command speech.

First, the system extracts the fundamental frequency of the command speech which the system has to recognize. Secondly, the spectrum of the command sound is estimated based on a harmonic structure model using the fundamental frequency. We call the mechanism of these two steps the "Harmonic Structure Estimator". According to this processing, we can extract the command speech from the received signal leaving the TV sound as extraction residuals. From the results, the transfer characteristics from the TV set to the microphone can be iteratively estimated using the TV sound even under existence of command speech. In the following section, the algorithm of the conventional ANC is explained and, then, the above-mentioned idea is formulated.

2.2.1 Conventional ANC

Figure 2.1 shows a scheme for conventional ANC using a digital filter of finite impulse response (FIR). To reduce the noise $\boldsymbol{n}'(k)$ in the received signal $\boldsymbol{x}(k)$, we need to

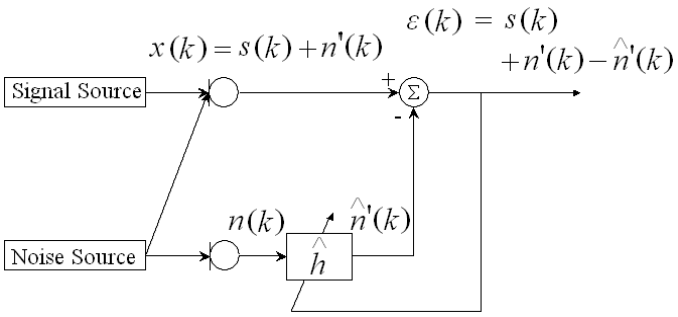

Figure 2.1: A scheme of conventional ANC

estimate the noise. The received noise is formulated as convolution of known noise signal $\boldsymbol{n}(k)$ and the time varying impulse response of the path from the source to the microphone approximated by an FIR filter

$$\hat{\boldsymbol{h}}(k) = [\hat{h}_0(k), \hat{h}_1(k), \cdots, \hat{h}_{N-1}(k)]^T$$

of length N, whose time varying coefficients are estimated adaptively. The noise signal received by the microphone is formulated as follows:

$$\hat{n'}(t_k) = \hat{\boldsymbol{h}}(k)^T \boldsymbol{n}(k) \tag{2.1}$$

where

$$\boldsymbol{n}(k) = [n(t_k), n(t_k - 1), \cdots, n(t_k - N + 1)]^T$$

and $\boldsymbol{n}(k)$ is a part of the noise signal or a k-th frame. Subtracting $\hat{n'}(t_k)$ from the received signal $x(t_k)$ yields the error signal

$$\epsilon(t_k) = x(t_k) - \hat{n'}(t_k) = s(t_k) + \{n'(t_k) - \hat{n'}(t_k)\}$$

To estimate the filter coefficients, it is necessary to minimize

$$\begin{aligned} \mathrm{E}[\epsilon^2(t_k)] = \mathrm{E}[s^2(t_k)] &+ \mathrm{E}[\{n'(t_k) - \hat{n'}(t_k)\}^2] \\ &+ 2\mathrm{E}[s(t_k)\{n'(t_k) - \hat{n'}(t_k)\}] \end{aligned}$$

Then, to minimize $\mathrm{E}[\epsilon^2(t_k)]$, $\mathrm{E}[\epsilon^2(t_k)]$ is differentiated by h_i as follows:

$$\begin{aligned} \frac{\partial \mathrm{E}[\epsilon^2(t_k)]}{\partial h_i} &= -2n(t_k - i)\{n'(t_k) - \hat{n}'(t_k)\} + 2s(t_k)n(t_k - i) \\ &= -2n(t_k - i)\{s(t_k) + n'(t_k) - \hat{n}'(t_k)\} \\ &= -2n(t_k - i)\epsilon(t_k) \end{aligned}$$

When the noise is stationary, the optimum filter $\hat{\boldsymbol{h}}_{opt}$ can be realized as the Wiener solution. TV sound, however, cannot be assumed to be stationary. So, the stochastic gradient, least-mean-square (LMS) or normalized LMS algorithm are widely used for updating the filter coefficients in case the noise is non-stationary. For example, the LMS algorithm updates the filter coefficients as follows:

$$\hat{\boldsymbol{h}}(k) = \hat{\boldsymbol{h}}(k-1) + \mu \boldsymbol{n}(k)\epsilon(t_k) \tag{2.2}$$

$$= \hat{\boldsymbol{h}}(k-1) + \mu \boldsymbol{n}(k)\{s(t_k) + n'(t_k) - \hat{n'}(t_k)\} \tag{2.3}$$

where μ is a step size controller.

2.2.2 ANC using harmonic structure of voiced speech

Figure 2.2 shows the proposed scheme for ANC using harmonic structure of voiced speech. The received noise can be alternatively thought to be obtained by subtract-

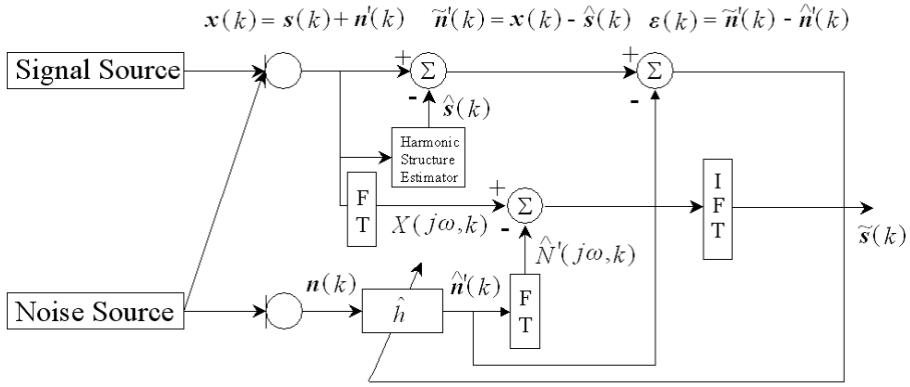

Figure 2.2: The proposed scheme for ANC

ing the estimated speech signal $\hat{s}(t_k)$ from the received signal $x(t_k)$. We have:

$$\tilde{n'}(t_k) = x(t_k) - \hat{s}(t_k)$$

Here, we design the filter $\hat{\boldsymbol{h}}(k)$ so as to minimize the energy of the error of the difference between $\hat{n'}$ and $\tilde{n'}$. Subtracting $\hat{n'}(t_k)$ from $\tilde{n'}(t_k)$ yields the error signal,

$$\begin{aligned}\epsilon(t_k) &= \tilde{n'}(t_k) - \hat{n'}(t_k) \\ &= s(t_k) + n'(t_k) - \hat{s}(t_k) - \hat{n'}(t_k) \\ &= \{s(t_k) - \hat{s}(t_k)\} + \{n'(t_k) - \hat{n'}(t_k)\}\end{aligned}$$

So, mean square error is obtained as,

$$\begin{aligned}\mathrm{E}[\epsilon^2(t_k)] &= \mathrm{E}[\{s(t_k) - \hat{s}(t_k)\}^2] + \mathrm{E}[\{n'(t_k) - \hat{n'}(t_k)\}^2] \\ &\quad +2\mathrm{E}[\{s(t_k) - \hat{s}(t_k)\}\{n'(t_k) - \hat{n'}(t_k)\}]\end{aligned}$$

Then, to minimize $\mathrm{E}[\epsilon^2(t_k)]$, $\mathrm{E}[\epsilon^2(t_k)]$ is differentiated by h_i as follows:

$$\begin{aligned}\frac{\partial \mathrm{E}[\epsilon^2(t_k)]}{\partial h_i} &= -2n(t_k - i)\{n'(t_k) - \hat{n}'(t_k)\} \\ &\quad +2\{s(t_k) - \hat{s(t)}\}n(t_k - i) \\ &= -2n(t_k - i)\{s(t_k) + n'(t_k) - \hat{s(t_k)} - \hat{n}'(t_k)\} \\ &= -2n(t_k - i)\epsilon(t_k)\end{aligned}$$

LMS algorithm updates the filter coefficients as:

$$\hat{\boldsymbol{h}}(k) = \hat{\boldsymbol{h}}(k-1) + \mu \boldsymbol{n}(k)\epsilon(t_k) \tag{2.4}$$

$$= \hat{\boldsymbol{h}}(k-1) + \mu \boldsymbol{n}(k)\{s(t_k) + n'(t_k) - \hat{s}(t_k) - \hat{n'}(t_k)\} \tag{2.5}$$

Here, we explain how to estimate $\hat{s}(t_k)$. First, the system extracts the fundamental frequency in the received signal. Then, using the fundamental frequency and

the frequency spectrum, the system classifies the received signal into three segment types: (1) voiced segments; (2) unvoiced segments; and (3) non-speech segments.

In case of voiced segments, $\hat{S}(j\omega, k)$ shows harmonic structure whose frequency components at integer multiples of the fundamental frequency. We can put:

$$\hat{S}(j\omega, k) = \sum_{p=1}^{P} a_p \delta(\omega - 2\pi p f_{0k}) \tag{2.6}$$

where a_p denotes the amplitude of p-th harmonic component, P denotes the number of harmonic components and f_{0k} denotes the fundamental frequency of kth frame. Assuming that the maximal amplitude of the signal expressed in Eq. (2.6) does not exceed that of $\hat{S}(j\omega, k)$ at the same analysis frame. It leads to the following definition of the amplitude a_p[1],

$$a_p = \frac{\max\limits_{0 \le \nu < N-1} x(t_k + \nu)}{P} \tag{2.7}$$

Equation (2.7) declares that a_p is set to be the same value for all the harmonics in a frame.

In case of unvoiced segments, $\hat{S}(j\omega, k)$ cannot be assumed to have harmonic structure. Here, we assume that the transfer characteristics of consecutive two frames do not differ so much. So, we assume that the following frequency spectrum $\hat{N}'(j\omega, k)$ can be described using the transfer characteristics of the previous analysis frame.

$$\hat{N}'(j\omega, k) = \hat{H}(j\omega, k-1) N(j\omega, k)$$

where $N(j\omega, k)$ is the frequency spectrum of $\boldsymbol{n}(k)$. Hence, $\hat{S}(j\omega, k)$ is expressed as follows:

$$\hat{S}(j\omega, k) = X(j\omega, k) - \hat{N}'(j\omega, k)$$

where $X(j\omega, k)$ is the frequency spectrum of $\boldsymbol{x}(k)$.

In case of non-speech segments, we put:

$$\hat{S}(j\omega, k) = 0$$

In any segment, $\hat{\boldsymbol{s}}(k)$ is obtained as Fourier Inverse Transform of $\hat{S}(j\omega, k)$. Here, we focus on Eqs.(2.3) and (2.5). If $\hat{s}(t_k)$ in Eq. (2.5) is replaced with $\hat{s}(t_k) = 0$, Eqs.(2.3) and (2.5) are completely identical. It means that the proposed method works same as conventional ANC for non-speech frames, and achieves better speech recognition rate than that of conventional ANC assuming the harmonic structure for voiced segments. The proposed method needs to classify the received signal into three segment types. The problem, however, is not so critical.

[1]We consider some definitions of the amplitude, for example a linear prediction analysis. However, this simple definition provided the best result

Finally, the signal for recognition is obtained according to the following procedure. First, the frequency spectra $X(j\omega, k)$ and $\hat{N}'(j\omega, k)$ are obtained as Fourier Transforms of the received signal $\boldsymbol{x}(k)$ and $\hat{\boldsymbol{n}}'(k)$, respectively. Then, subtracting $X(j\omega, k)$ from $\hat{N}'(j\omega, k)$ in the frequency domain yields $\tilde{S}(j\omega, k)$, from which the estimated source signal $\tilde{\boldsymbol{s}}(k)$ is obtained as Fourier Inverse Transform of $\tilde{S}(j\omega, k)$. $\tilde{\boldsymbol{s}}(k)$ is used for speech recognition.

2.3 Performance evaluation in a short reverberation time environment

Preliminary experiments were carried out in a sound proof cabin. The assumed situation is that the distance between a microphone and a loudspeaker, pretending a user, is 50cm and the distance between a microphone and a TV set is 100cm. Recording was made using the following speech and noise sources mixing them up to obtain signal to noise ratios (SNR) 0, 6, 12 and 18 dB.

- Speech source
  - Number of speech data: 50 utterances by four subjects (three males and one female) each
  - Contents of the utterance: TV controlling commands (i.e. "terebi keshite (TV off)", "onryo agete (Volume up)", "terebi Asahi (channel name)", "nyu:su (news)", etc.)
  - Recording procedure: Recording was made with a close contact microphone in a sound proof cabin, where subjects were asked to repeat until the utterance was correctly recognized
- Noise source
  - Noise type: Mixture of speech on music

Compared are three speech recognition schemes: (1) without any preprocessing in case of 0dB in SNR; (2) conventional Spectral Subtraction; and (3) the proposed method. Speech recognition rate is employed to evaluate the performance. Results of the preliminary experiment are shown in Fig. 2.3 where significant improvement by the proposed method is recognized. Figure 2.4 compares spectrograms of the original speech signal, received signal, estimated signals obtained by SS, by conventional ANC and by the proposed method. Spectral overreduction of harmonic components observed in SS (see Fig. 2.4 (d)) may lower speech recognition rate. Moreover, SS does not achieve enough noise reduction performance in the low frequency region. So, the proposed method shows the best performance among the three speech recognition schemes in the sound proof cabin. However, performance

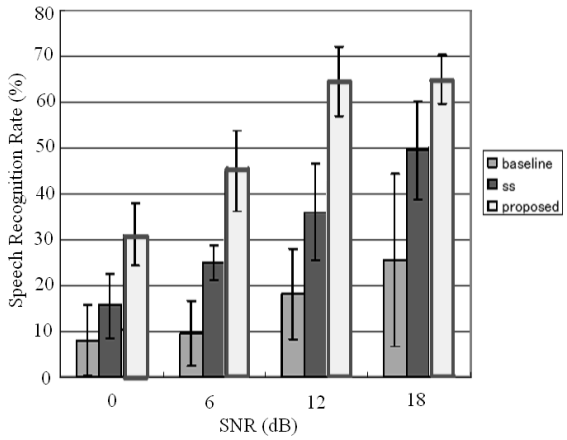

Figure 2.3: Speech recognition rate obtained by the preliminary experiment

in practical situations should be investigated to confirm the effect of the proposed method for practical use.

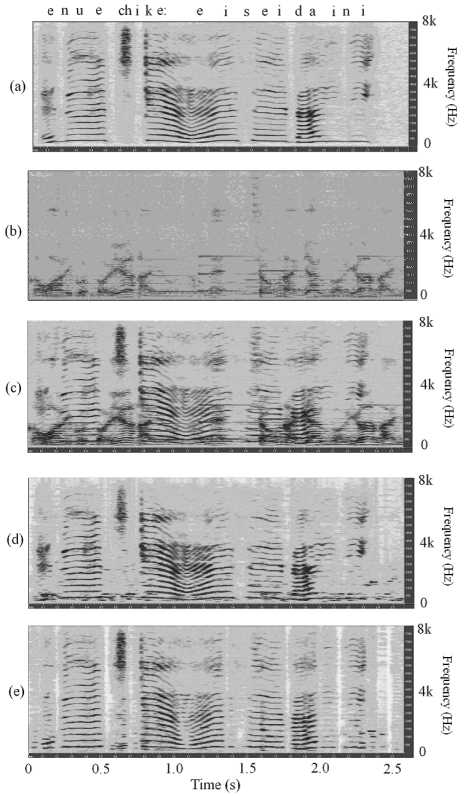

Figure 2.4: Comparison based on the spectrograms (a: the source signal, b: noise signal, c: received signal, d: conventional SS, e: proposed method)

2.4 Performance evaluation in a long reverberation time environment

Recording condition

Effect of the proposed method on speech recognition rate was evaluated in Living Room Simulator (LS) and Ubiquitous Home (UH). Figures 2.5 and 2.6 show the positions of the loudspeakers and microphone in each room. This situation assumes that a system user utters TV control commands to the microphone placed on a table. Temperature and humidity in these rooms are not controled so there is a possibility that they randomly fluctuate.

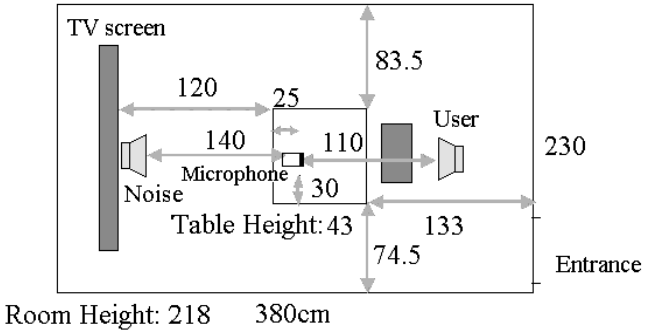

Figure 2.5: Allocation of loudspeakers and microphone in the Living Room Simulator

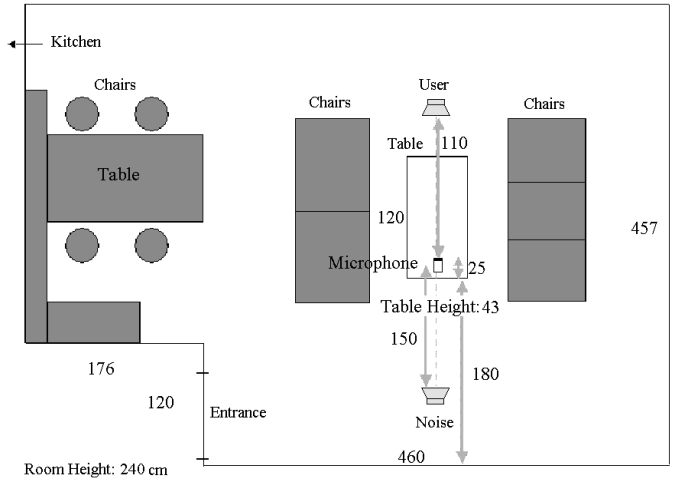

Figure 2.6: Allocation of loudspeakers and microphone in the Ubiquitous Home

The reverberation times of the rooms were measured using a time-stretched pulse as four times of the time duration required to decay 15dB in power instead of

measuring the time duration required to decay 60dB, setting the loudspeakers and microphones as depicted in Figs. 2.5 and 2.6. The reverberation time of LS is 313ms and that of UH is 353ms. The reason why the measured reverberation time of UH is longer than that of LS is thought to lie in the material difference of the surface of the tables. Recording was made using the following speech data and noise sources mixed up to obtain signal to noise ratios (SNR) 0, 3, 6, 9, 12, 15, 18, 21 and 24dB. Speech data employed for evaluations are 100 utterances by 15 subjects (10 males and 5 female) each, and other specifications are the same as those in the preliminary experiment.

Table 2.1: Relationship between SNR and the volume of TV sound

	distance between the loudspeaker and a microphone	
volume	50cm	100cm
fairly large	5.6dB	1.4dB
a little bit large	7.2dB	3.0dB
normal	9.4dB	5.2dB

The relationship between SNR and the volume of TV sound was investigated setting user's speech amplitude at ordinary level. Table 2.1 shows that SNR gets worse as TV volume increases. Based on table 2.1, we have decided to focus our research interest on the improvement of speech recognition rate under the situation where SNR falls between 0dB to 9dB.

Experimental Set-up

A speech recognition decoder "Julian" is employed for the experiment [25]. Table 2.2 shows conditions of acoustic analysis for speech recognition. The vocabulary size of the dictionary is 1342 and the number of grammatical rule is 14 (see Appendix A).

Experimental Results

We compare the speech recognition rate of the following two recognition schemes: (1) without any preprocessing, (2) a conventional SS, and (3) the proposed method. Figures 2.7 and 2.8 show the speech recognition rates for the data recorded in LS and UH, respectively. The abscissa discriminates SNR, and the ordinate represents speech recognition rate. We do not employ any adaptation techniques to environment, so the speech recognition rate in case ∞dB remains about 80%.

Speech recognition rates of the two environments show improvement in low SNR cases by the proposed method. Degradation of speech recognition rates, however, is observed in high SNR cases for both the environments. We will discuss the point in the next section.

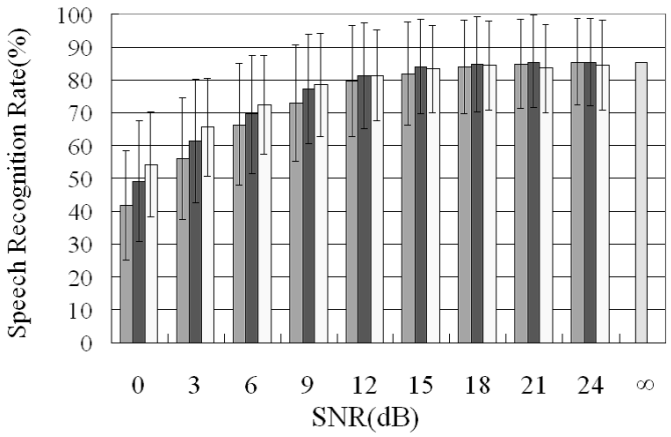

Figure 2.7: Speech recognition rate in LS (left: without noise reduction, center: conventional SS, right: proposed method)

Table 2.2: Conditions of acoustic analysis

Demension of analysis	25
Feature parameter	mfcc
Window	Hamming
Frame size	25 ms
Frame shift	10 ms
Sampling rate	16 ksamples/sec
Quantization bit	16 bits

2.5 Robustness of the proposed method

In this section, the robustness in detection errors of speech segments and in estimation errors of fundamental frequency are discussed. The proposed method performs noise reduction by detecting speech segments of the target speech estimating the harmonic structure of voiced speech based on the fundamental frequency. So, performance of noise reduction is degraded due to the detection errors of speech segments or the errors in estimating fundamental frequency. In general, it is difficult to extract the fundamental frequency of noisy speech. Hence, we discuss how much accuracy of the speech interval detection and the fundamental frequency estimation does it require to work correctly.

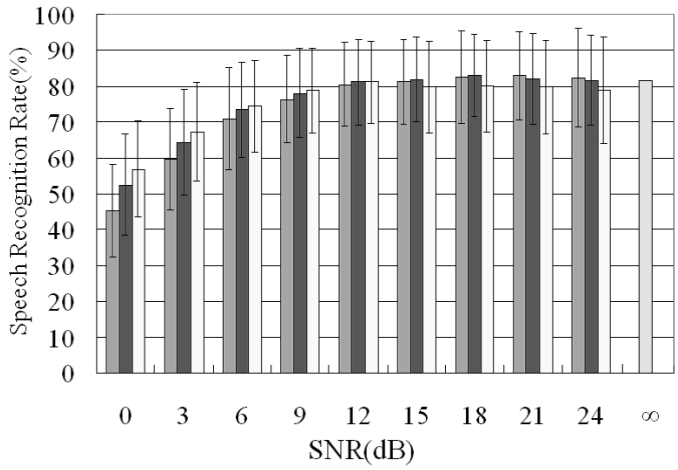

Figure 2.8: Speech recognition rate in UH (left: without noise reduction, center: conventional SS, right: proposed method)

Errors in detecting speech segments

Detection errors are classified into (i) discriminating errors between speech and non-speech, (ii) detection errors of unvoiced segments and (iii) detection errors of voiced segments. The error rate of each error is shown in Fig. 2.9. From Fig. 2.9, it is

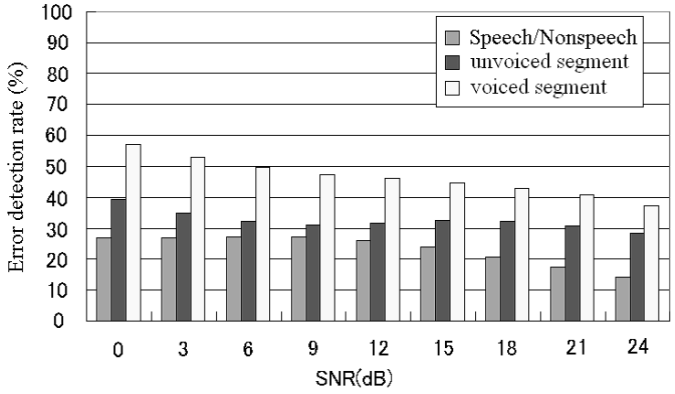

Figure 2.9: Comparion on detection error rate

clear that the error rate of detecting voiced segment is higher than the others. So, this result shows that it is necessary to detect voiced segments accurately.

Necessity of modeling of unvoiced segments

In general, it can be said that unvoiced segments are difficult to detect. To model the speech signal more simply, we confirm the speech recognition rate without detecting

the unvoiced segments. That is to say, the system classifies the received signal into two segment types, voiced segment and non-speech segment. This result is shown in Fig. 2.10. From Fig. 2.10, the difference between the speech recognition rate in

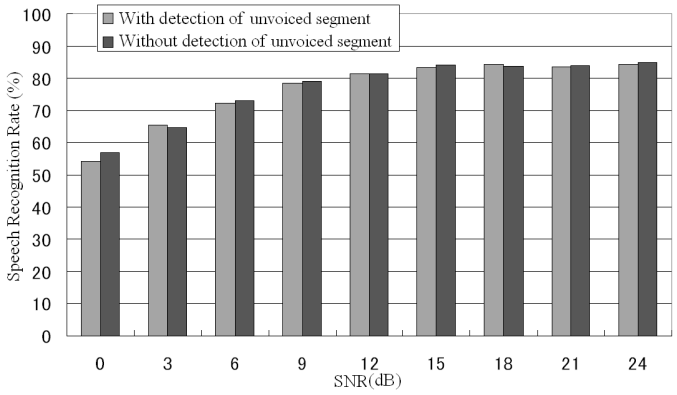

Figure 2.10: Comparison of the speech recognition rates between with and without unvoiced segment detection

case with detecting unvoiced segments and without that, is a little bit and there is no significant difference. So, it is not necessary to detect the unvoiced segments.

Robustness in estimating the fundamental frequency

Firstly, fundamental frequency is estimated using clean speech. This estimation result is regarded as the criterion for the comparison. Then we compare the fundamental frequency estimated from the received signal with its criterion and obtain the accordance rate of fundamental frequency shown in Fig. 2.11.

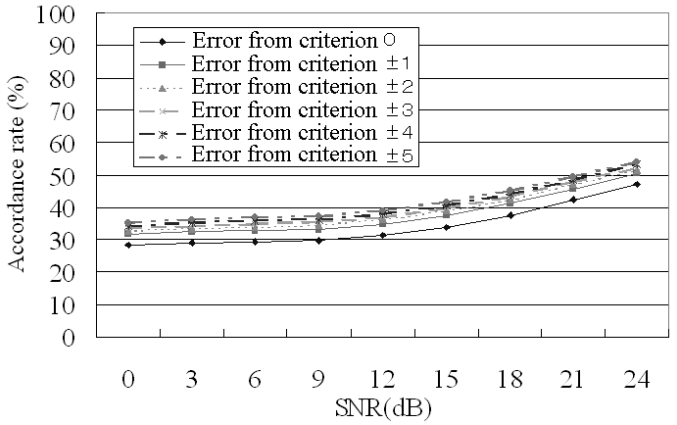

Figure 2.11: Accordance rate of fundamental frequency on various SNRs

Figure 2.11 shows the accordance rate in case where we accepted the difference between the criterion and the fundamental frequency using the proposed method within ±5 samples, one sample equals 15.625Hz. From Fig. 2.11, it is clear that the accordance rate is about 30% in case where the SNR is between 0∼12dB, nevertheless the speech recognition rate was improved by the proposed method. The reason why we got this result is that the proposed method can work as the conventional ANC in non-speech segments. That is, about the frames in which the fundamental frequency is not estimated correctly, considering them as the non-speech segments, we can expect the same performance with the conventional ANC at the worst. Whereas, in case where the SNR is between 15∼24dB, speech recognition rate cannot be improved by the proposed method though the accordance rate is about 50%.

Dependency of speech recognition rate on the accuracy of fundamental frequency estimation

To investigate the reason that speech recognition rate is not improved in the high SNR, the dependent degree of speech recognition rate on the accuracy of the fundamental frequency estimation is checked. 9 males and 4 females are used for this evaluation. Figure 2.12 shows the speech recognition rate which is obtained employ-

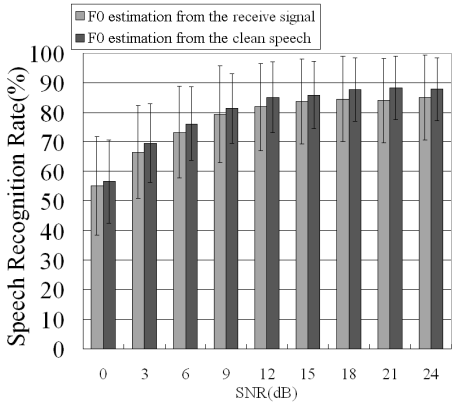

Figure 2.12: Speech recognition rate of the proposed method using fundamental frequency estimated from both the original clean speech and the received signal

ing the proposed method using the fundamental frequency estimated from both the clean speech and the received signal. From Fig. 2.12, it is clear that the speech recognition rate is improved significantly by the proposed method in case where the SNR is 12, 18, 21, 24dB. That is, if we want to get better result in the high SNR region, it is necessary to estimate fundamental frequency more correctly than the above-mentioned evaluation.

2.6 Comparison with the Wiener filter

The proposed method performs an adaptive processing by the LMS algorithm. So, we compare the proposed method with a Wiener filter.

Wiener Filter

We assume the following signal model.

$$x(t) = s(t) + n(t) \tag{2.8}$$

where $x(t)$ is an observed signal, $s(t)$ is a target signal and $n(t)$ is a noise signal. Wiener filter gives the optimal estimation of the target signal. Wiener filter can be obtained by minimizing the mean square of the difference between the input and the filter output. The error spectrum of the target signal is defined as follows:

$$\begin{aligned} E_s(f) &= U(f)S(f) - S(f) \\ &= (U(f) - 1)S(f) \end{aligned}$$

where $U(f)$ is the Wiener filter's system function and $S(f)$ is the spectrum of the target signal. The error spectrum of the noise signal is defined as follows:

$$E_n(f) = U(f)N(f)$$

where $N(f)$ is the spectrum of the noise signal. From the Parseval's theorem, the mean square error between the input and the filter output is proportional to the integration of the power spectrum of the total error over all frequencies.

$$\begin{aligned} \int_{-\infty}^{\infty} e^2(t)dt &\propto \int_{-\infty}^{\infty} |E_s(f) + E_n(f)|^2 df \\ &= \int_{-\infty}^{\infty} |(U(f) - 1)S(f)|^2 + |U(f)N(f)|^2 df \end{aligned} \tag{2.9}$$

where we assume no-correlation between the target signal and the noise signal. Wiener filter $U(f)$ is obtained by minimizing the error power by nullifying differential of Eq. (2.9) with $U(f)$. As a result, we have

$$U(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2} \tag{2.10}$$

Comparison result

Figure 2.13 shows speech recognition rates of the proposed method and a Wiener filter.

From Fig. 2.13, you can see that the proposed method is more effective than the Wiener filter in the low SNR region, while a Wiener filter is better than the proposed

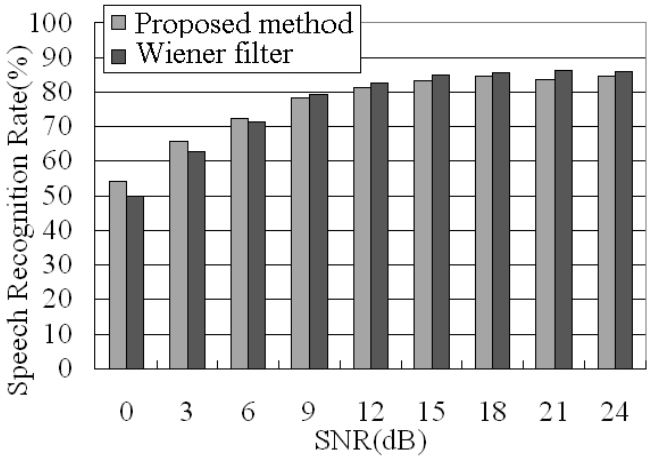

Figure 2.13: Comparison between the proposed method and Wiener filter

method in the high SNR region, the difference is about 2%. From this result, using the harmonic structure of the voiced speech, we can get the improvement of speech recognition rate. These evaluations are performed offline. On the other hand, if we apply the proposed method to the real environments, the SNR is estimated in the noise segments and then according to the SNR, we should switch the methods.

2.7 Discussions and conclusions

Looking at the speech recognition rates for the speech data without any preprocessing, recognition rate for SNR 12dB in UH, for example, shows about the same value as that of the noise free case. So, speech recognition rate for SNR better than 12dB decreases. Moreover, we can observe that adding noise sometimes yields improvement in speech recognition of noise free speech. The tendency becomes clear in high SNR cases in particular. We guess a possible reason that noise might hide reverberation [26]. Confirmation of this is left as future investigation.

Looking the recognition rates by the proposed method, the rate for SNR less than 12dB in UH are higher than the rate without preprocessing.

We proposed a new algorithm for adaptive filtering using harmonic structure of voiced segments for reducing non-stationary known noise. The preliminary experimental results show that the proposed method realizes better frequency spectra estimation and speech recognition rate than the conventional SS. Moreover, effectiveness of the proposed method is confirmed in different practical environments, UH and LS. To get better improvement, we have to apply a dereverberation method to the resultant signal of the proposed method [27].

Chapter 3

Speech dereverberation in the frequency domain

3.1 Introduction

Performance of automatic speech recognition has reached a practical level in case a close contact microphone is used in quiet environments. In practical situations, however, speech recognition rate decreases drastically due to environmental noises and reflected waves. There can be two approaches for improving the speech recognition rate in actual environments as long as employing the same recognition engine. One is signal manipulation on input signals and the other is introducing adaptation techniques in the speech recognition process. The proposed method is an approach classified to the former.

Inverse filtering of source-microphone transfer functions is widely employed for suppressing the effects of reflected waves [28], but this method cannot be adopted to cases where the transfer functions from the source to microphones cannot be obtained nor to time variant cases. Several methods of spectral subtraction have been proposed to cope with these cases [29] [30] [31]. Yanagida *et al.* formulated least-squares source sound separation introducing generalized convolutional inverse matrices [32]. However, most of them require measurement of transfer functions among sources and microphones.

Unoki *et al.* proposed a Modulation Transfer Function (MTF) based method, which does not require measuring transfer functions [33], but a source signal and transfer characteristics are modeled by MTF for recovering the power envelope of the source wave from the received reverberant signal. Although the method requires a procedure for estimating the reverberation time and path amplitude, no definite method for determining them has been developed yet.

Takiguchi *et al.* proposed an adaptive method which does not require transfer functions [34]. The method, however, cannot yield sufficient improvement in case the reverberation time is long [18].

Nakatani *et al.* proposed a method exploiting the harmonic features of voiced speech [35]. Their method constitutes an inverse filter based on a large number of reverberant speech data. The method, however, takes a lot of time to design an accurate inverse filter. Hence, it is difficult to put their method into practical uses.

There are some methods based on linear prediction analysis [36] [37]. These methods, however, require obtaining linear prediction coefficients of speech in advance.

The proposed methods, however, do not require transfer functions, inverse filters nor linear prediction coefficients. We developed two different methods for dereverberation. One is a method in the time- and frequency-domains. The other is a method in the frequency-domain.

3.2 A conventional method in the time domain

In this chapter, we assume an environment where one signal source exists. A signal $x_i(t)$ received by microphone $\#i$ consists of several waves from the source including a direct wave and reflected waves and $x_i(t)$ is represented by convolution of source signal $s(t)$ and an impulse response of a set of paths from the source to microphone $\#i$. The signal $x_i(t)$ received by microphone $\#i$ can be expressed by the following equation.

$$x_i(t) = s(t) * h_i(t)$$

where * denotes convolution, $h_i(t) = \sum_{j=0}^{J} h_{ij}(t)$, where $h_{ij}(t)$ represents the impulse response of the j-th path from the source to microphone $\#i$ including the direct path ($j = 0$).

The reflected signal along the j-th path is assumed to have amplitude of a constant decay rate α_{ij} ($0 < \alpha_{ij} \leq 1$) and a certain amount of delay time l_{ij}, that is, the reflected signal received by microphone $\#i$ from the source via the j-th path is expressed as $\alpha_{ij} s(t - l_{ij})$ assuming that all the paths are flat in frequency characteristics. Here, dominant component of each impulse response is assumed to be single reflection, and waves of multiple reflections are supposed to decay much more compared with single reflection waves. That is, the effect of reflection waves is expected to be nullified or almost reduced by removing single reflection waves. Based on this assumption, the following equation is expected to estimate the direct wave $\alpha_{i0} s(t - l_{i0})$

$$\alpha_{i0} s(t - l_{i0}) \simeq x_i(t) - \sum_{j=1}^{J} \alpha_{ij} s(t - l_{ij}) \tag{3.1}$$

where J denotes the effective number of reflected waves that reach microphone $\#i$.

We use the received signal with delay l_{ij}, as an aproximation of $s(t - l_{ij})$, thus

$$\alpha_{i0} s(t - l_{i0}) \simeq x_i(t) - \sum_{j=1}^{J} \alpha_{ij} x(t - l_{ij})$$

In discrete form we have

$$\alpha_{i0} s(k - l_{i0}) \simeq x_i(k) - \sum_{j=1}^{J} \alpha_{ij} x(k - l_{ij}) \tag{3.2}$$

where k denotes the k-th sample point and j denotes the path ID for the microphone $\#i$. In fact, however, l_{ij} should be written as l_{ij}/T, where T is a sampling interval, we might employ l_{ij} unless we misunderstand. Here, we simply use l_{ij} to represent time delay counted by sampling interval. The frequency characteristics of reflective surfaces are assumed to be flat [30].

Equation (3.2) is rewritten into an iterative form. The basic idea of the proposed method is to remove reflected signals one by one, so at the starting point of this algorithm we set an initial value as follows:

$$x_i^{(0)}(k) = x_i(k)$$

Let $x_i^{(j)}(k)$ denote the received signal with the 1st through the j-th reflected signals removed. Then, $x_i^{(j)}(k)$ is expressed as follows:

$$x_i^{(j)}(k) = x_i(k) - \sum_{l=1}^{j} \alpha_{il} x(k - l_{il}) \tag{3.3}$$

From Eq. (3.3), we get a recursive form as follows:

$$x_i^{(j)}(k) = x_i^{(j-1)}(k) - \alpha_{ij} x_i^{(j-1)}(k - l_{ij}) \qquad j = 1 \ldots J \tag{3.4}$$

$$\hat{s}_i(k) \simeq x_i^{(J)}(k)$$

where $\hat{s}_i(k)$ is the estimated source signal, $\alpha_{i0} = 1$ and $l_{i0} = 0$.

3.2.1 Estimation of the delay time

As the auto correlation function (ACF) of the source signal itself is not flat even if it is not affected by reflection, ACF of the received signal, consisting of the direct wave and reflected waves, would show local peaks attributed to either reflection or local peaks of the ACF of the source signal itself. That is, even if the ACF of a received signal shows a local peak at a particular time lag, the time lag cannot be attributed to reflection. A power-normalized ACF can be used in place of ACF to avoid the above-mentioned problem.

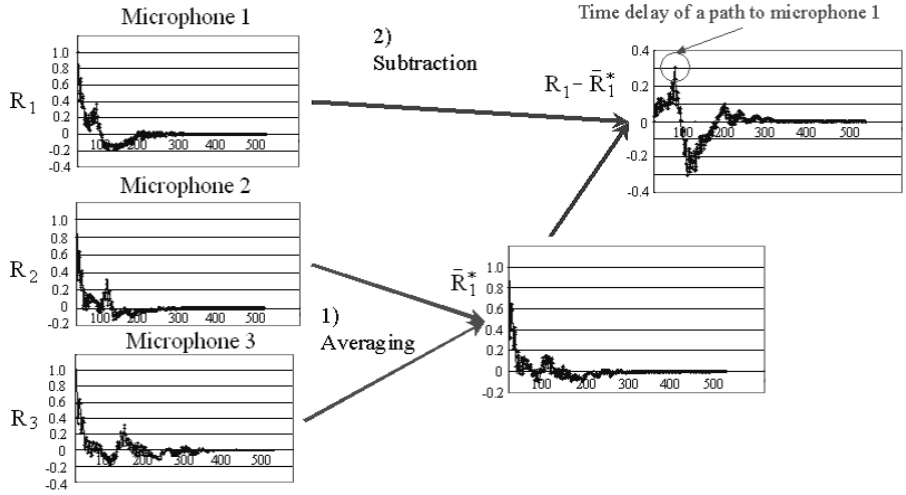

Figure 3.1: Estimation of time delays on ACF

The proposed method solves the problem using a certain number of microphones. Figure 3.1 shows how to detect time delays for a three microphones case.

Assume that we want to remove a principal reflection from the signal $x_1(k)$ received by microphone #1. First, the average ACF $\bar{R}_1^*(\tau)$ is calculated by averaging ACFs of the signals received by microphones other than #1, i.e. microphones #2 and #3 for this case. The ACF of the source signal $R_s(\tau)$ is approximated by $\bar{R}_1^*(\tau)$. Furthermore, $\bar{R}_1^*(\tau)$ is used later as the reference for estimating the path amplitude α_{1j} of the reflected wave along the j-th path. Then, the difference between the ACF $R_1(\tau)$ of the signal $x_1(k)$ and $\bar{R}_1^*(\tau)$ is calculated to extract the time delay that gives the maximum value on $R_1(\tau) - \bar{R}_1^*(\tau)$. The amount of delay due to reflection would be dependent on the relative position of microphones and walls. So, the time lag at which $R_1(\tau) - \bar{R}_1^*(\tau)$ gets large is thought to be the time difference between the j-th reflection path and the direct path. The delay time l_{1j} of j-th path to microphone #1 is estimated by detecting the positive maximal value of the difference $R_1(\tau) - \bar{R}_1^*(\tau)$.

3.2.2 Estimation of the decay rate

The proposed method requires estimation of two parameters. One is the delay time l_{ij} and the other is the decay rate α_{ij}, where i and j denote the microphone ID and the path ID, respectively. Explained in this section is how to estimate the decay rate α_{ij}. The time delay l_{ij} obtained as described in the previous section and the average ACF are used to estimate the decay rate. Here, $\bar{R}_i^*(\tau)$ is assumed to approximate $R_s(\tau)$, so it is used as the reference in the recursive procedure, which will be explained soon, to estimate the decay rate by minimizing the difference between $R_i(\tau)$ and $\bar{R}_i^*(\tau)$ at the time delay l_{ij}. Figure 3.2 shows the algorithm for estimating path amplitude.

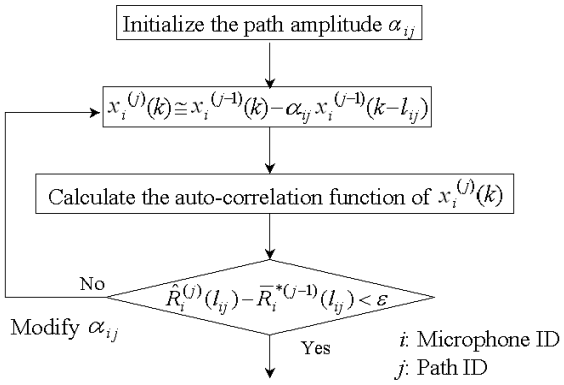

Figure 3.2: An algorithm to estimate the decay rate α_{ij} of the path $\#i$

First, the initial value for the decay rate α_{ij} is set to be null. Then, the ACF of the estimated signal, $\hat{R}_i^{(j)}(\tau)$, is calculated based on $x_i^{(j)}(k)$ expressed by Eq. (3.4).

Next, $\Delta R_i(l_{ij}) = \hat{R}_i^{(j)}(l_{ij}) - \bar{R}_i^{*(j-1)}(l_{ij})$ is calculated, where $\hat{R}_i^{(j)}(l_{ij})$ and $\bar{R}_i^{*(j-1)}(l_{ij})$ are the values of $\hat{R}_i^{(j)}(\tau)$ and $\bar{R}_i^{*(j-1)}(\tau)$, respectively, at the delay time l_{ij}. If the difference $\Delta R_i(l_{ij})$ is less than 10^{-6}, then take the current α_{ij} to be the decay rate for the j-th path, then the process goes to the next step. Otherwise, the decay rate is increased to be checked further. Figure 3.3 explains the process of reducing the difference on ACFs at l_{ij}. Convergence of this estimation algorithm is shown in Appendix B.

3.2.3 Segmentation of the received signal

In our previous method [38], presumed reflection waves are subtracted one by one from the received wave in the time domain. The method, however, has difficulties of over-subtraction in fricative and nasal segments because of poor power compared with other phoneme segments. Over reduction sometimes occurs by waveform subtraction in case the received wave $x_i(k)$ is used as the source wave $s(k)$ in Eq. (3.1) even if both the delay time and decay rate are properly estimated. So, our previous method could not make speech recognition rate satisfactory for speech signals picked up in reverberant environments.

The proposed method is designed to detect fricative- or nasal-like segments in input signals and leave them as they are to avoid over-subtraction. Fricatives and nasals are detectable as they show power concentration in high and low frequency regions, respectively. Temporal waveform subtraction is employed only for the speech segment except fricative-like or nasal-like segments, and conventional spectral subtraction is exclusively employed for non-speech segments. To subtract reflected waves properly, the input signal is segmented into speech or non-speech segments.

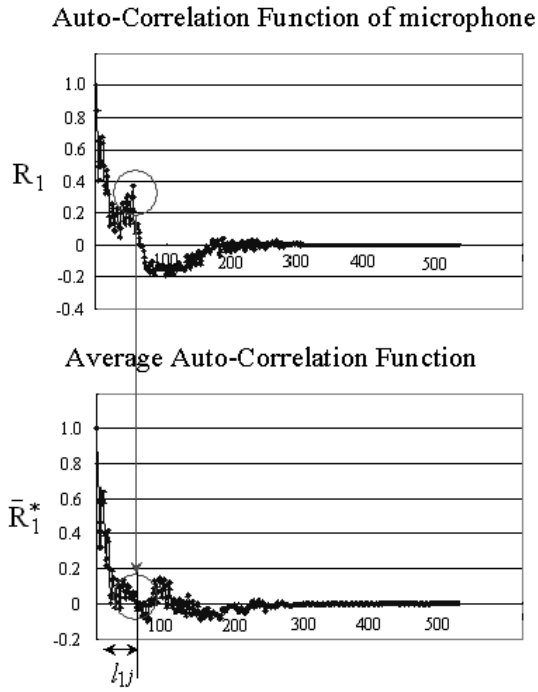

Figure 3.3: Explanatory sketch of the processing on ACFs

Explained below is how to partition a received signal into speech and non-speech segments, and then, further partitioning of the speech segment into fricative-like or nasal-like segments and the rest. First, a received signal is partitioned into a type of segments (class "S") whose short-term powers are larger than a threshold employed to detect utterance initials, and the other type of segments (class "O") whose short-term powers are smaller than that. If the power spectrum of a segment of class "O" shows the maximum value at frequency beyond 4kHz (assuming that the sampling rate is 16ksamples/sec), the segment is classified into class "F", a fricative-like segment. Then, a segment having its spectral peak in the low frequency region below 1kHz with fundamental frequency below 400Hz is regarded as a segment of class "N", a nasal-like segment, because nasals are periodic and have power concentration in low frequency regions. The rest of the input signal is regarded as non-speech segment, or class "NS". Following the partitioning procedure described above, each segment in a received signal is classified into one of classes S, O, F, N or NS as shown in Fig. 3.4.

3.2.4 The processing scheme

Figure 3.5 shows the processing scheme of the proposed method. Firstly, utterance initials of speech signals are detected by a double threshold method. Then, $R_i(\tau)$, the ACF of the signal $x_i(t)$ received by microphone $\#i$ is calculated. Next, l_{ij}, the

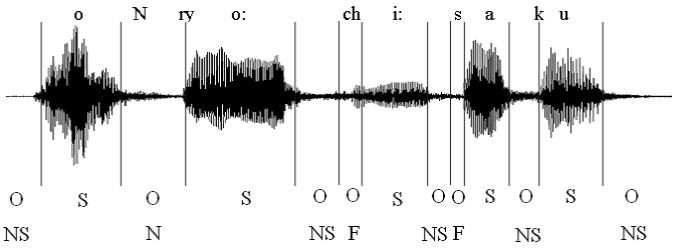

Figure 3.4: Segmentations of a received signal and their classification into classes S, O and F, N, NS

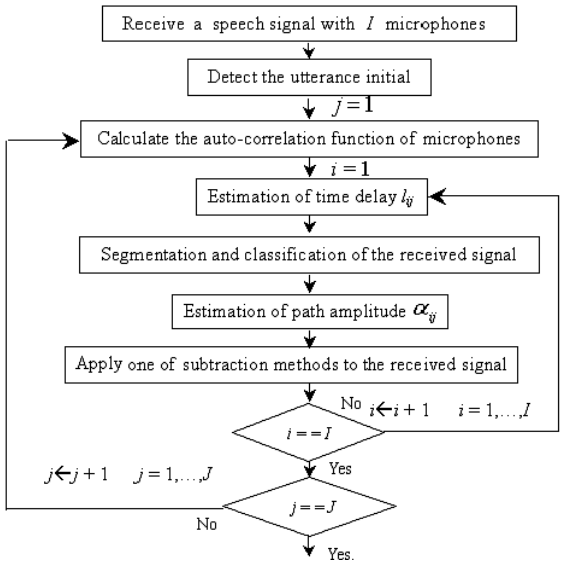

Figure 3.5: A processing scheme of removing reflected waves. (I : the number of microphones, J : the number of reflected waves to be removed)

delay time of the j-th path is estimated as the time lag τ that gives the maximum value for $\Delta R_i(\tau) = \hat{R}_i^{(j)}(\tau) - \bar{R}_i^{*(j-1)}(\tau)$, where $\bar{R}_i^*(\tau)$ denotes an approximation of the ACF of the source signal and $\hat{R}_i^{(j)}(\tau)$ denotes the ACF of the estimated signal. Then, the received signal is partitioned into segments of either large amplitude segment "S" or small amplitude segment "O", and then, "O" is further classified into fricative-like segment "F", nasal-like segment "N" or non-speech segment "NS". Finally, the decay rate α_{ij} is obtained as the value that minimizes the difference $\Delta R_i(l_{ij})$. Then the presumed reflection wave is subtracted according to Eq. (3.4). Proceed by increasing $i \leftarrow i+1$ until i reaches I, then $j \leftarrow j+1$ until j reaches J.

3.2.5 Evaluations

Experimental set-up

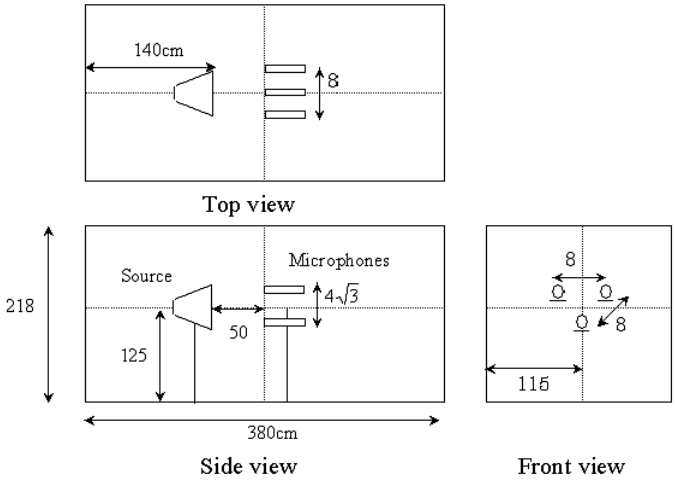

Figure 3.6: The recording configuration in the experimental living room

To investigate the performance of the proposed method, an experimental "living room" having reverberation time 440ms is used as a reverberant room. One loudspeaker and three microphones are used in this evaluation. Figure 3.6 shows the configuration of the loudspeaker and three microphones in the "living room". The parameter J is set for four according to a preliminary experiment using small data set.

Speech data and conditions for speech recognition

Speech data played back for the evaluations are those recorded with a close contact microphone in a sound proof cabin. The speech samples are 250 Japanese words uttered by one female and four males. Contents of the speech data are commands in Japanese for controlling TV sets. For example, "terebi oN (TV on)", "chaN-neru ichi (channel one)", "nyuusu (news)" and so on. The vocabulary size is 99 and the number of grammar rules is 13. Speech sounds are reproduced by the loudspeaker and are sampled at 16ksamples/sec with 16bit accuracy. "Julian" is employed as the speech recognition decoder [25].

Results

The frequency spectrum of the received signal $x_i(k)$ and that of the estimated signal $\hat{s}_i(k)$ are compared to evaluate the improvements in the sound quality and recognition rate.

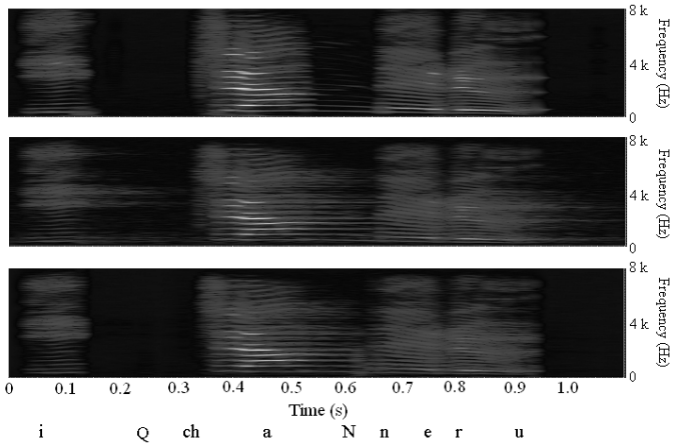

Figure 3.7: Comparison of spectrograms. (Top: source signal, center: signal received in the "living room", bottom: estimated signal)

Figure 3.7 shows spectrograms of the source, the received and estimated signals. As shown in Fig 3.7, reflected waves are effectively removed around 0.2s and 1.0s and fricative- and nasal-like sounds well remain not over-subtracted around 3.5s and 0.6s. As the result of applying the proposed method to reverberant signals, the source signal is approximately recovered from the reverberant signals. We can also recognize that reverberation in non-speech segments is sufficiently removed. Listening to the reflection-removed signals, slight improvement in sound quality is perceived.

Table 3.1: Comparison of speech recognition rates (%) recalculated employing the majority decision among three microphone ($* < 0.05$, $** < 0.01$)

speker	without any processings	proposed
M1	84	90
M2	67	88**
M3	82	82
M4	90	96
F1	76	86
Ave.	80	89*

Table 3.1 shows the speech recognition rate obtained by introducing the majority decision among the speech recognition results of the three microphones. It is improved by 9% from 80% to 89% by applying the proposed method to reverberant signals. Significant difference is recognized between speech recognition rates of received signals and recovered signals only for speaker M2 among 5 speakers accord-

ing to a sign test at 1% hazard rate. Moreover, significant difference is recognized between average speech recognition rates according to a t-test at 5% hazard rate.

3.2.6 Discussions and conclusions

The result of t-test using all 250 data between average speech recognition rates of received raw signals and reflection removed signals shows significant difference at 5% hazard rate. So, it can be barely said that the proposed method well removes reflected waves and improves the speech recognition rate.

The current evaluation was carried out on the condition that a speaker, a loudspeaker here, does not move. So, it is problematic whether a system can manage moving speakers or not. The proposed method, however, estimates a reflection wave one by one. So, it is expected to be robust with respect to the movement of speakers.

To improve the recognition rate further, it would be necessary to refine the method. For example, in the proposed method, we do not take the frequency characteristics of reflective surfaces into consideration, and furthermore, spectral subtraction is not applied to fricative-like or nasal-like segments. There exist further reflection waves that the present system cannot remove as the system deals with only single reflection.

Proposed is a method to remove reflected waves from a signal received in a reverberant room. The proposed method has solved some of the problems that our previous method had. Problems improved or almost solved by the proposed method are: unreliability in delay time estimation, approximation errors in estimated source waves, and over-subtraction for fricative- and nasal-like segments. Concretely speaking, delay time estimation is improved by majority decision on ACFs, over-subtraction is suppressed for consonants such as fricative- and nasal-like segments by classifying each speech segment into subcategories. Applying the proposed method, the recognition rate is improved from 80% to 89%.

3.3 A new dereverberation method in the frequency domain

The proposed method introduced in the previous section cannot cope with the frequency characteristics. So, in this section, we introduce a method which can estimate reverberation time and decay rate at each frequency bin.

3.3.1 Basis of dereverberation

A signal $x(t)$ received by a microphone generally consists of several waves from a source including the direct wave and reflected waves. The signal $x(t)$ from a source is represented by convolution of source signal $s(t)$ and impulse response $h(t)$ consisting

of those for possible paths from the source to the microphone. The received signal $x(t)$ is expressed as:

$$x(t) = s(t) * h(t) \tag{3.5}$$

where * denotes convolution. Taking short term Fourier transform, Eq. (3.5) can be rewritten as follows:

$$X(\omega_n, k) = S(\omega_n, k)H(\omega_n, k) \tag{3.6}$$

where ω_n denotes the n-th frequency bin and k denotes the frame ID. Here, $H(\omega_n, k)$ can be divided into the direct path component $D(\omega_n, k)$ and the total sum of reflection components $R(\omega_n, k)$. So, Eq. (3.6) is rewritten as follows:

$$X(\omega_n, k) = S(\omega_n, k)\{D(\omega_n, k) + R(\omega_n, k)\} \tag{3.7}$$

The frequency spectrum $X(\omega_n, k)$ at frequency ω_n in frame k of the received signal can be approximated by the convolution of the frequency spectrum $S(\omega_n, k)$ with $\alpha_{\omega_n}(k)$, the impulse response of the integrated propagation paths for frequency ω_n.

$$X(\omega_n, k) \simeq \sum_{l=0}^{L_n} \alpha_{\omega_n}(l)S(\omega_n, k-l) \tag{3.8}$$

where, L_n represents the time delay of n-th frequency bin and $\alpha_{\omega_n}(l)$ denotes the decay by distance and reflection characteristics at frequency ω_n. Then, the component for $l = 0$ in the summation for $X(\omega_n, k)$ can be regarded as the direct path component $S(\omega_n)D(\omega_n)$ and the other components in $X(\omega_n, k)$ can be regarded as the total sum of reflection components $S(\omega_n)R(\omega_n)$. So, Eq. (3.8) can be rewritten as follows:

$$X(\omega_n, k) \simeq \alpha_{\omega_n}(0)S(\omega_n, k) + \sum_{l=1}^{L_n} \alpha_{\omega_n}(l)S(\omega_n, k-l) \tag{3.9}$$

Here, if we estimate the direct path component, Eq. (3.9) is rewritten as follows:

$$\hat{S}(\omega_n, k) \simeq X(\omega_n, k) - \sum_{l=1}^{L_n} \alpha_{\omega_n}(l)S(\omega_n, k-l) \tag{3.10}$$

where, $\hat{S}(\omega_n, k) = \alpha_{\omega_n}(0)S(\omega_n, k)$ denotes the direct path component. As the result of this processing, the frequency spectrum of source signal is estimated.

3.3.2 Estimation of the reverberation time and decay rate

The auto-correlation function is employed for estimating the reverberation time and decay rate for each frequency bin. The power spectrum of a received signal can be

approximated by Eq. (3.8). So, the auto-correlation function of the received signal is expressed as:

$$\begin{aligned}\phi_{XX}(\omega_n,\kappa) &= \sum_{m=0}^{M}\log(X(\omega_n,m)+1)\log(X(\omega_n,m+\kappa)+1)\\ &\simeq \sum_{m=0}^{M}X(\omega_n,m)X(\omega_n,m+\kappa)\\ &= \sum_{m=0}^{M}\left\{\left(\sum_{l_1=0}^{L_n}\alpha_{\omega_n}(l_1)S(\omega_n,m-l_1)\right)\left(\sum_{l_2=0}^{L_n}\alpha_{\omega_n}(l_2)S(\omega_n,\kappa+m-l_2)\right)\right\}.\end{aligned}$$

By separating the production of the inner $\sum$ into two cases, one for $l_1 = l_2 - \kappa$ and the other for $l_1 \neq l_2 - \kappa$, the equation described above is rewritten as follows:

$$\begin{aligned}&\phi_{XX}(\omega_n,\kappa)\\ &\simeq \sum_{m=0}^{M}\left(\sum_{l=0}^{L_n-\kappa}\alpha_{\omega_n}(l)\alpha_{\omega_n}(l+\kappa)S^2(\omega_n,m-l)\right)\\ &+\sum_{m=0}^{M}\left\{\left(\sum_{\substack{l_1=0\\ l_1\neq l_2-\kappa}}^{L_n}\alpha_{\omega_n}(l_1)S(\omega_n,m-l_1)\right)\left(\sum_{\substack{l_2=0\\ l_2\neq l_1+\kappa}}^{L_n}\alpha_{\omega_n}(l_2)S(\omega_n,\kappa+m-l_2)\right)\right\}\end{aligned}$$

Here, if each component of $S(\cdot, m-l)$ is mutually independent or $S(\cdot,m-l_1)S(\cdot,\kappa+m-l_2)$ is enough smaller than $S^2(\cdot,m-l)$, $\phi_{XX}(\omega_n,\kappa)$ can be approximated as:

$$\phi_{XX}(\omega_n,\kappa)\simeq\sum_{m=0}^{M}\left\{\left(\sum_{l=0}^{L_n-\kappa}\alpha_{\omega_n}(l)\alpha_{\omega_n}(l+\kappa)S^2(\omega_n,m-l)\right)\right\} \tag{3.11}$$

where we cannot obtain $S(\omega_n,k)$, so we perform the above procedure as follows:

$$\begin{aligned}\phi_{XX}(\omega_n,\kappa) &= \xi\phi_{XX}(\omega_n,\kappa)\\ &\simeq \sum_{m=0}^{M}\left\{\left(\sum_{l=0}^{L_n-\kappa}\alpha_{\omega_n}(l)\alpha_{\omega_n}(l+\kappa)S^2(\omega_n,m-l)\right)\right\}.\end{aligned}$$

By normalizing Eq. (3.11) we get

$$\begin{aligned}\Phi_{XX}(\omega_n,\kappa) &= \frac{\sum_{m=0}^{M}\left\{\left(\sum_{l=0}^{L_n-\kappa}\alpha_{\omega_n}(l)\alpha_{\omega_n}(l+\kappa)S^2(\omega_n,m-l)\right)\right\}}{\sum_{m=0}^{M}\left\{\left(\sum_{l=0}^{L_{max}}\alpha_{\omega_{max}}(l)\alpha_{\omega_{max}}(l)S^2(\omega_{max},m-l)\right)\right\}}\\ &= f_n(\omega_n,\kappa)\end{aligned}$$

where $\omega_{max} = \max_{\omega_n} \phi_{XX}(\omega_n, 0)$, $f_n(\omega_n, \kappa)$ denotes the frequency characteristics of the auto-correlation function $\Phi_{XX}(\omega_n, \kappa)$ for delay κ. However, $\Phi_{XX}(\omega_n, \kappa)$ is nearly zero for κ larger than L_n the reverberation time for the n-th frequency bin. The effective reverberation time L_n for each frequency bin is assumed as the point beyond which $\Phi_{XX}(\omega_n, \kappa)$ is smaller than ϵ, where ϵ is an enough small value. Figure 3.8 shows an example of auto-correlation function.

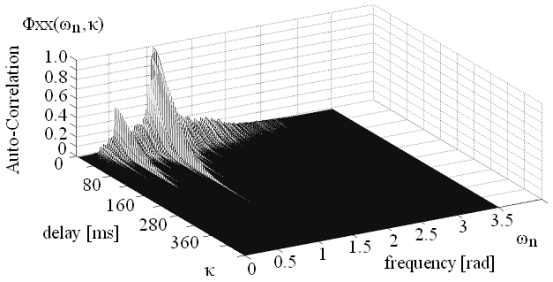

Figure 3.8: An example of auto-correlation function

Moreover, $\Phi_{XX}(\omega_n, \kappa)$ is regarded as the decay rate for delay κ for frequency ω_n.

$$\alpha_{\omega_n}(\kappa) = \Phi_{XX}(\omega_n, \kappa) \qquad \kappa = 1 \cdots L_n$$

Reflected waves have a high correlation with the direct signal. The auto-correlation of the received signal at each time lag represents the degree of similarity between reflected waves and the direct signal. So, the auto-correlation function up to the estimated time delay can be regarded as the decay characteristics of reflection.

3.3.3 Subtracting reflected waves on the power spectrum

There should be clean frames having no effects of reflection at the beginning of utterance before the first reflection reaches the microphone. Dereverberation can be achieved by Eq. (3.10). A processing scheme for Eq. (3.10) is depicted in Fig. 3.9, where the abscissa corresponds to the time axis plotted frame by frame and the ordinate symbolically represents amplitude at frequency ω_n. Assume that the total tick width over the time sequences of the amplitude component represents effective reverberation time L_n. The gray bar in the upper half of this figure represents the direct component and black bars represent the reflected wave components. The lower half of this figure shows the time sequences of the amplitude component after subtracting the reflected components of $S(\omega_n, 0)$.

3.3.4 Performance evaluation

Two types of evaluation are carried out to confirm the validity and effectiveness of the proposed method. One is evaluation using simulated data and the other is evaluation using actual data.

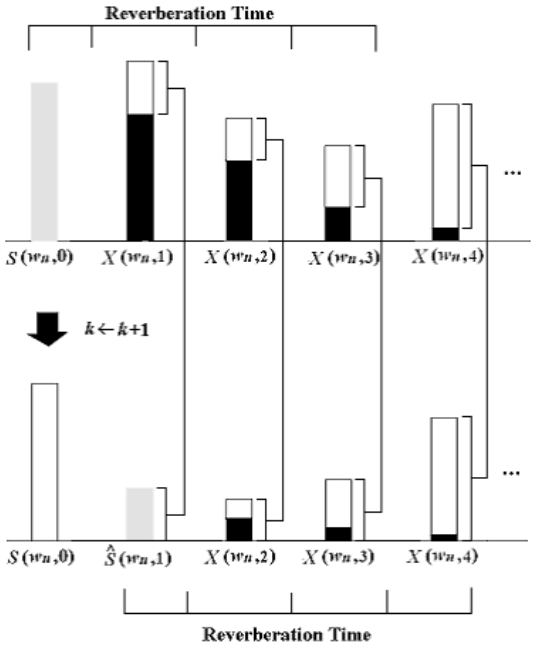

Figure 3.9: Spectral subtraction on the time sequence of a frequency bin corresponding to a certain frequency ω_n.

Evaluation using simulated data

First, an evaluation in which the reverberation time can be estimated correctly by the proposed method, is performed to confirm the validity of the method.

Frequency characteristics shown in Fig. 3.10 are convolved with a swept sinusoidal signal for simulation. Figure 3.11 shows the spectrograms before and after performing the proposed method.

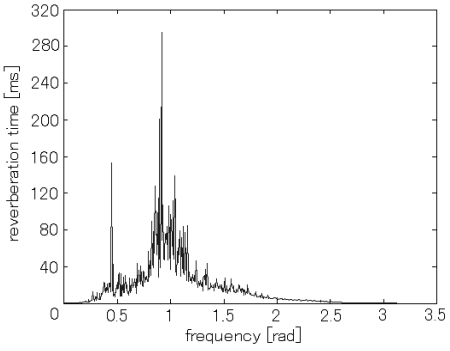

Figure 3.10: The frequency characteristics assumed for integrated paths from a source to a microphone for generating a received signal

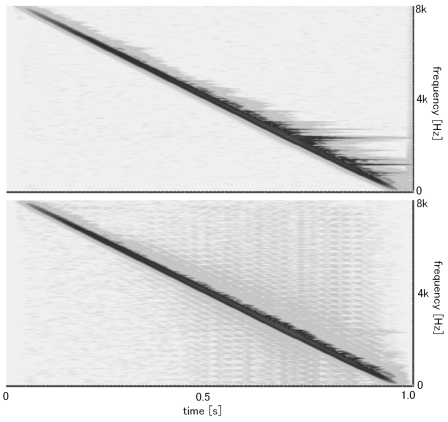

Figure 3.11: Spectrograms comparing on a swept sinusoidal signal (Top: simulated reverberation characteristics, bottom: results of dereverberation)

From Fig. 3.11, the proposed method can reduce reverberation mainly for reflected waves in the low frequency region. Figure 3.12 shows the estimated reverberation time by the proposed method.

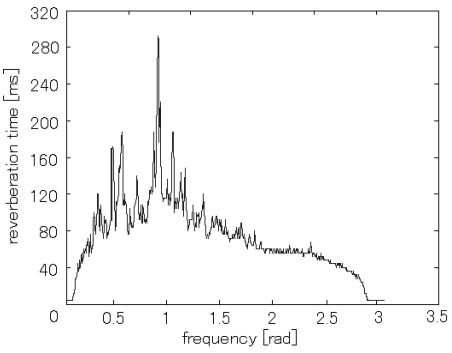

Figure 3.12: The estimated reverberation time for the simulation data

Comparing of Figs. 3.10 and 3.12, the proposed method seems to be able to estimate the longest reverberation time. On the other hand, the proposed method tends to estimate too long reverberation time for the frequencies at which the reverberation time is short. This problem is caused by the correlation of the signal itself.

Secondly performed is simulation to compare the proposed method with the method that assumes flat frequency characteristics at reflection on walls. Figure 3.13 shows the spectral comparison of the simulated data. Comparing (c) and (d) in Fig. 3.13, you can see that the proposed method well suppresses over-subtraction frequently observed in the conventional spectral subtraction. As an index for evalu-

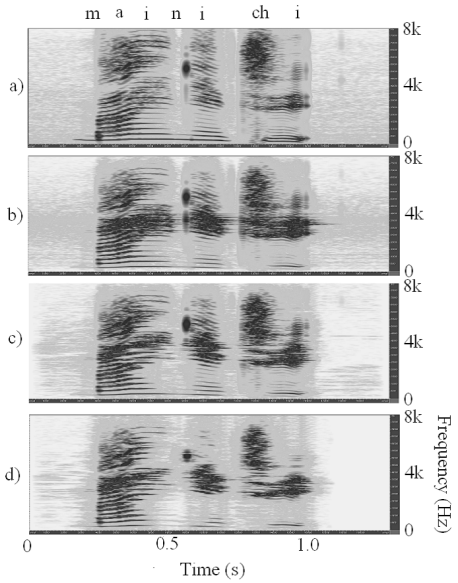

Figure 3.13: Performance comparison on spectrograms for the simulated data. (a: Original speech, b: reverberant speech, c: dereverberated speech by the proposed method, d: dereverberated speech by a conventional method assuming flat frequency characteristics)

ating dereverberation performance, employed is segmental SNR defined as follows:

$$SNR(k) = \frac{\sum_{n=0}^{N-1} |S(\omega_n, k)|^2}{\sum_{n=0}^{N-1} \left| S(\omega_n, k) - \hat{S}(\omega_n, k) \right|^2} \tag{3.12}$$

where, N denotes the number of frequency bins on the output of FFT. The average segmental SNR of speech signal before processing is -2.4 dB. The average segmental SNR of the speech signal processed by the proposed method is 0.7 dB, while that by the conventional method assuming flat frequency characteristics is 0.5 dB. Figure 3.14 shows the estimated frequency characteristics of the reverberation time by the two methods. Here, short term Fourier transform is performed under the condition of 64 ms frame length and 4 ms shifting interval. From Fig. 3.14, the estimated longest reverberation time is about 800 ms around $3 \sim 3.5$ kHz. The frequency having the longest reverberation time in Fig. 3.14 corresponds to the frequency having long lasting tails in Fig. 3.13 b). If we take the dotted line given by the conventional method assuming flat frequency characteristics as the reverberation time, it is clear that over-subtraction occurs because the estimated reverberation time is too long.

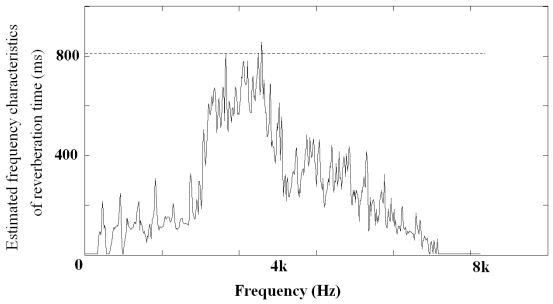

Figure 3.14: The stimated frequency characteristics of the reverberation time. (The solid line represents the results obtained by the proposed method and the horizontal dotted line is that obtained by the conventional spectral subtraction assuming flat frequency characteristics.)

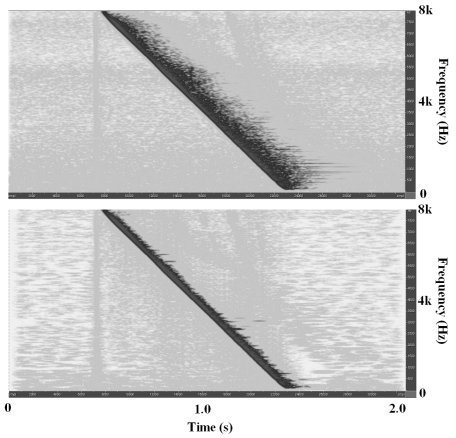

Figure 3.15: The pectrogram of the swept sinusoidal signal. (Top: Received, Bottom: Processed by the proposed method)

Evaluation using actual data

Evaluation data are recorded in a small cabin (230 cm × 380 cm, H :218 cm). An evaluation result for a swept sinusoidal signal is shown in Fig. 3.15, which shows the spectrograms of the swept sinusoidal signal before and after processing. Figure 3.16 shows the frequency characteristics of the reverberation time estimated by the proposed method.

From this figure, the longest reverberation time is estimated as about 740 ms. Comparing Fig. 3.15 with 3.16, the contour of estimated frequency characteristics of the reverberation time is similar to the outline of the spectrogram of the swept sinusoidal signal. So, we may be able to suppose that the frequency characteristics are successfully estimated in case of actual data. To confirm the performance of

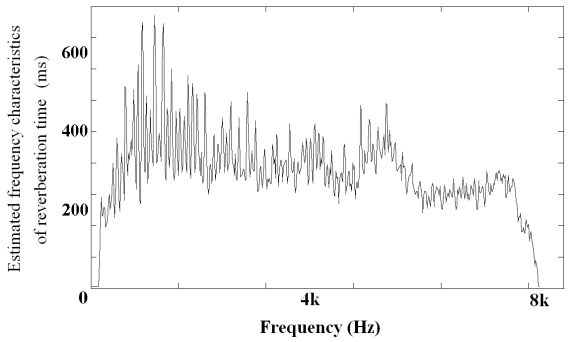

Figure 3.16: The estimated frequency characteristics of the reverberation time.

the proposed method, the reverberation curve is depicted in Fig. 3.17, from which reverberation time appears to be 390 ms.

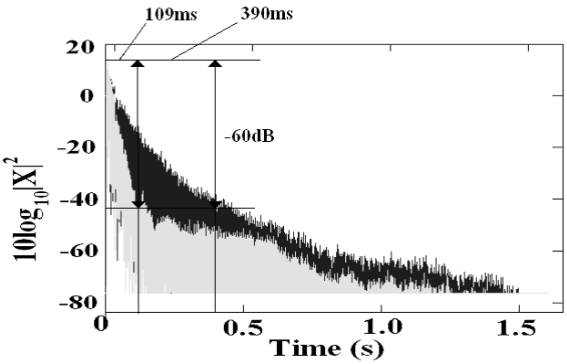

Figure 3.17: Reverberation curves. (Black: Received, Gray: Processed by the proposed method)

On the other hand, the reverberation time is shortened to be 109 ms using the proposed method. Figure 3.17 shows that the proposed method drastically reduces the initial reflections.

Next, Fig. 3.18 shows the processing result for the actual speech data.

The average SNR of these signals are -1.0 dB and -0.1 dB respectively. Figure 3.19 shows the estimation result of the reverberation time for the speech signal.

3.3.5 Discussions and conclusions

The proposed dereverberation method is adopted each frequency bin. The processing expressed in Eq. (3.10) can be handled in a group within narrow frequency ranges as the frequency characteristics are supposed to be similar within a narrow frequency range. So, to get better dereverberation performance, we propose a scheme in which

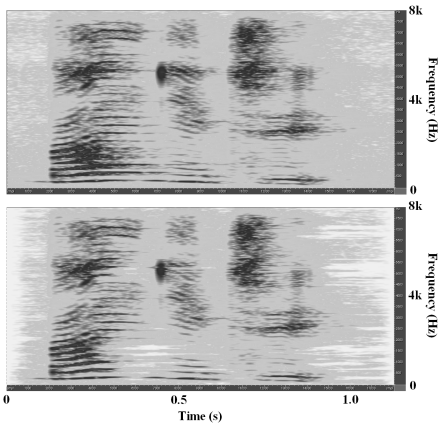

Figure 3.18: The spectrogram of an actual speech signal. (Top: received, Bottom: processed by the proposed method)

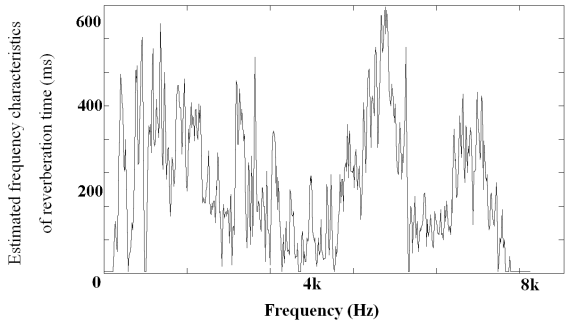

Figure 3.19: The estimated frequency characteristics of the reverberation time for the speech signal.

a certain number of adjacent frequency bins are collectively processed in succession shifting along the frequency axis. That is to say, we introduce smoothing into the proposed method. Figure 3.20 shows average segmental SNRs obtained for various numbers for dividing the total frequency range.

From this figure, the best performance seems to be obtained by the proposed method at 4 to 64 separations in case of 1024-point FFT.

This technique is evaluated with a voice-controlled TV system shown in Appendix C.1. Data for this evaluation were recorded in a sound proof cabin usign a close contact microhone. Four males and three females uttered 100 short phrases of TV control commands. These speech data were played back in the Living Room Simulator whose reverberation time is 313ms. Recorded speech and noise are added on a computer making signal to noise ratio 0dB, 6dB, 12dB and 18dB. Figure 3.21 shows a speech recognition rate.

From Fig. 3.21, the proposed method can achieve improvement of the speech

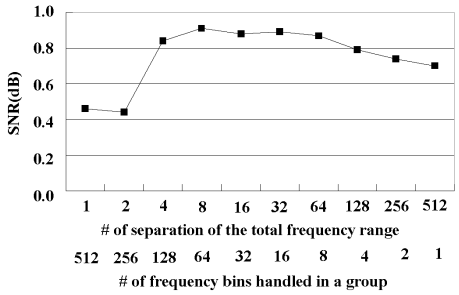

Figure 3.20: Average segmental SNRs obtained for several separation numbers of the total frequency range.

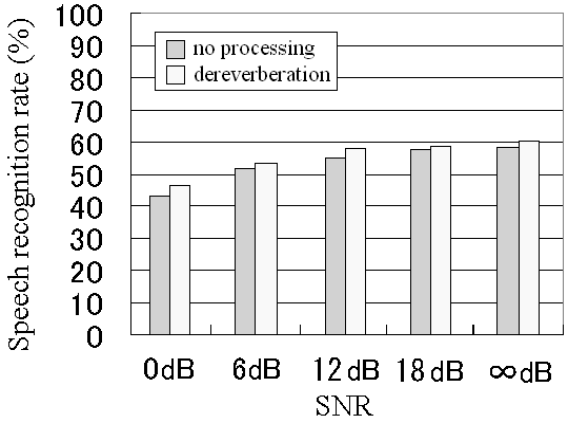

Figure 3.21: Comparison on speech recognition rate

recognition rate in case of all SNRs. In case of ∞dB, the speech recognition rate remains about 60% because any adaptation techniques are not employed and speech data which is failed to detect a speech segment are counted as failures of speech recognition.

This chapter proposed a single channel blind dereverberation method based on auto-correlation functions of time sequences of frequency components on running power spectra. The proposed method estimates the reverberation time and the frequency characteristics of decay. So, the proposed method achieves dereverberation without any preparatory measurement or *a priori* information. From the performance evaluation on simulated data, the proposed method shows ability to estimate the reverberation time almost correctly. Moreover, the proposed method yields better results than the conventional spectral subtraction methods. From the performance evaluation on actual data, the proposed method also shows better reduction in the reverberation time. Introducing smoothing techniques is expected to lead to further better results.

Chapter 4

Solving the permutation problem in the frequency-domain ICA for BSS

4.1 Introduction

BSS (Blind Source Separation) is one of techniques for source separation. BSS is to separate mixed signals into each source signal without using any *a priori* information. Up to now, time domain ICA (Independent Component Analysis) [43] [44] is the principal means for realizing BSS.

In early 80s, ICA was first used in the context of neural network modeling. ICA is a statistical and a computational technique for revealing hidden factors underlying random variables, measurements and signals. The model of ICA is represented as the linear or nonlinear mixture of hidden factors, which is assumed to be non-Gaussian variables and mutually independent. Moreover, the mixing process is also assumed to be unknown. ICA can be seen as an extension to principal component analysis and factor analysis. ICA is, however, a much more effective technique capable of finding hidden factors or signal sources even in the case where classical methods fail completely. So, ICA can be applied to various types of application fields, including digital images and document databases, as well as economic indicators and psychometric measurements. In many cases, the measurements are given as a set of parallel signals or time series. The typical example is realizing the cocktail party effect. In mid-90s, some research groups showed the effectiveness of ICA on realizing of the cocktail party effect.

Time domain ICA is successful in case mixture is not convolutional. Though separation by time domain ICA is hard in reverberant environments, frequency-domain ICA is employed and gives fair performance even in reverberant cases. In frequency-domain ICA, convolutional mixture is converted into simple additive mixture in each

frequency bin by taking Fourier transform.

However, in frequency-domain ICA, there exists a tough issue called "permutation problem". The permutation problem requires us correct assignment of source identification for every frequency bins after separation for each frequency bin. There are many techniques proposed for coping with the permutation problem. Kurita *et al.* [52] proposed to use directivity characteristics of each frequency bin. Ikeda *et al.* [53] proposed to use the correlation between frequency bins. Sawada *et al.* [54] proposed combining above-mentioned two techniques. Moreover, Sawada *et al.* [55] proposed an improved technique using harmonic structure of voiced signal. Mukai *et al.* [56] proposed a technique that estimates a speaker position using a near-field model. However, the permutation problem has not been completely solved yet.

Di Persia *et al.* [57] [58] proposed permutation-free ICA (PF-ICA) for separating convolutively mixed signals into source signals without permutation errors. This technique has an advantage that it can avoid the permutation problem, but has a defect that it assumes a single directivity common to all frequency bins.

In this chapter, multi-bin ICA (MB-ICA) is proposed as a revised version of PF-ICA. It performs separation after connecting a definite number of adjacent frequency bins. By connecting adjacent frequency bins, MB-ICA can cope with local frequency characteristics and stably solves the permutation problem. That means MB-ICA can manage versatile directivity of each frequency bin, so we can expect to obtain more accurate separation result.

A method to avoid the permutation problem was proposed by Kim *et al.* [59]. This method extended an univariate cost function to a multivariate cost function to avoid the permutation problem. A method to perform the ICA using the simultaneous processing of adjacent frequency bins was proposed by Robledo-Arnuncio *et al.* [60]. The target of this method is to avoid degradation of learning performance of separation matrices in case of short phrases. In this method, a method by Sawada *et al.* [55] is used to cope with the permutation problem.

4.2 Basis of ICA in additive cases

We suppose a non-reverberant room in which P loudspeakers emit signals $s_p(t)$ received by M microphones. The problem does not lost generality even if loudspeakers are located at an equal distance from microphones. The received signals $x_m(t)$ are represented with the following linear combination.

$$x_m(t) = \sum_{p=1}^{P} h_{mp} s_p(t) \quad m = 1, \cdots, M,$$

where h_{mp} is unknown constant coefficient that corresponds to the decay rate of the path from source P to microphone m. The source signal $s_p(t)$ is also unknown. In this situation, we want to estimate the source signals from the received signals. This problem is called BSS.

If we can obtain a separation matrix W consisting of coefficients w_{pm}, the source signals can be obtained as follows:

$$y_p(t) = \sum_{m=1}^{M} w_{pm} x_m(t) \quad p = 1, \cdots, P$$

The separation matrix W could be found as the (pseudo-)inverse of the matrix H that consists of unknown mixing coefficients h_{mp}.

The problem is how to estimate the coefficients w_{pm} without knowing H. One solution for this problem is to introduce statistical independence of source signals. This method is called Independent Component Analysis (ICA). To estimate the separation matrix W by maximizing the independence of $y_p(t)$ that leads to estimation of $s_p(t)$. However, a further requirement is necessary to realize estimation of W. It is the requirement that the independent components must have non-Gaussian distributions. If the source signals are Gaussian, the source and the mixed distributions would be identical. So, it is impossible to infer the separation matrix from mixed signals in that case.

Independence is a stronger property than uncorrelatedness. It is obvious that Principal Component Analysis and Factor Analysis cannot separate a mixed signal into independent signals. Uncorrelatedness, signifying zero covariance, is not enough for realizing ICA but higher-order statistics are demanded.

Independence implies nonlinear uncorrelatedness, which means that any separated signals are uncorrelated and the separated signals obtained as nonlinear transforms are also uncorrelated. The means to choose nonlinear functions are the maximum likelihood method [61] [62] [63] [64] [65] and mutual information [66] [67].

Independent components maximize non-Gaussanity. Non-Gaussianity can be measured by kurtosis, which is a higher-order cumulant. We can perform ICA by maximizing kurtosis. This is based on the central limit theorem. The distribution summing up some non-Gaussian distributions comes close to the Gaussian distribution. If an independent component is found, kurtosis gets maximum.

There are two ambiguities in the ICA model. One is amplitude ambiguity and the other is order ambiguity. These ambiguities are induced by the fact that both the source signals and the separation matrix are unknown.

4.3 Conventional frequency-domain ICA

Frequency-domain ICA performs separation,

$$\boldsymbol{Y}(f_n, k) = \boldsymbol{W}(f_n)\boldsymbol{X}(f_n, k)$$

where f_n denotes n-th frequency bin and k represents the frame-ID, using a separation matrix $\boldsymbol{W}(f_n)$ learned from $M \times I$ observation matrix consisting of time series of n-th frequency bin

$$\boldsymbol{X}(f_n, k) = [X_1(f_n, k), X_2(f_n, k), \cdots, X_M(f_n, k)]^T$$

obtained on k-th frame signal $\boldsymbol{x}_m(k)$ received by M microphones by applying short term Fourier Transform. Although there are several methods to obtain a separation matrix $\boldsymbol{W}(f_n)$, JADE [68] is employed here. A $P \times 1$ vector of P source signals at k-th frame is written as

$$\boldsymbol{S}(f_n, k) = [S_1(f_n, k), S_2(f_n, k), \cdots, S_P(f_n, k)]^T.$$

The mixing process of the signals is represented as follows:

$$\boldsymbol{X}(f_n, k) = \boldsymbol{H}(f_n)\boldsymbol{S}(f_n, k)$$

where $\boldsymbol{H}(f_n)$ denotes the frequency characteristics of the mixing system. A vector of P separated source signals at k-th frame is obtained as

$$\boldsymbol{Y}(f_n, k) = [Y_1(f_n, k), Y_2(f_n, k), \cdots, Y_P(f_n, k)]^T.$$

ICA performs separation using independency among the separated source signals as a criterion. So, there is a possibility that the independency can be maximized even if rows of the separation matrix are permuted. That is to say, there is a possibility that a frequency bin at frequency f_n of a source signal $\boldsymbol{S}_i$ substitutes other source signal $\boldsymbol{S}_j$ at the same frequency. Conventional frequency-domain ICA needs to absolutory cope with the permutation problem, but the problem has not been completely solved yet.

4.4 Conventional methods to cope with the permutation problem

There is a method to treat the permutation problem by calculating directivity for each frequency bin and estimating the direction of noise sources [52]. The optimal null, however, is not always formed for all frequency bins in the low frequency region in particular [54]. On the other hand, Ikeda *et al.* proposed a method using the correlation among frequency bins [53]. It is reasonable that frequency components forming spectral envelopes have large correlation among the adjacent frequency bins belonging to the same source. So, one decision strategy is to maximize the summation of the correlation between the two adjacent frequency bins. This strategy, however, cannot solve the permutation problem stably because once it fails at one frequency bin, it continues failing to decide the permutation for the subsequent frequency bins.

There can be a method which starts to decide the permutation from a frequency bin for which separation performance is better than for other frequency bins. Normally, spectral components have large correlation among the adjacent frequency bins belonging to the same source, but this method assumes the large correlation even among distant frequency bins, so this method is not adequate.

To solve the problem mentioned above, Sawada *et al.* proposed a method based on the estimation of the directions of sources and the correlation among frequency bins [54]. In this method, firstly, directivity is calculated for each frequency bin. Based on this result, the method decides the permutation of definite number of frequency bins for which the permutation is decided with high reliability. Next, this method decides the permutation for undetermined frequency bins by maximizing the summation of the correlation between the neighboring frequency bins for which permutation is already decided. This method cannot decide the permutation stably in case reverberation time is long. The reason for the instability in long reverberation cases is the difficulty in estimating the direction of noise sources in the low frequency region.

To solve this problem, Sawada *et al.* proposed a method based on the harmonic structure of voiced speech [55]. Speech signals have high correlation at the integer multiples of the fundamental frequency. Using this feature, it can be expected to be able to decide the permutation more accurately. These methods, however, still have possibility to fail.

4.5 Permutation free ICA

In order to avoid the permutation problem, a smart method treating all the frequency bins at one time is proposed by the authors [57] [58]. In this method, temporal sequences of all frequency bins are connected together to form one long vector for each received signal before applying ICA to a set of vectors at one time. Components of a frequency bin are represented as the sequence of the spectral values along the time axis at frequency f_n on the series of short-time frequency spectrum of the signal received by microphone m as follows:

$$\boldsymbol{X}_m(f_n) = [X_m(f_n, 1), X_m(f_n, 2), \cdots, X_m(f_n, I)].$$

$\boldsymbol{X}_m(f_n)$ is a $1 \times I$ vector and $X_m(f_n, k)$ is a scalar. Then, the connected form of this sequence is expressed as a long column vector written as

$$\boldsymbol{X}_m = [\boldsymbol{X}_m(f_1), \boldsymbol{X}_m(f_2), \cdots, \boldsymbol{X}_m(f_N)]^T.$$

where N is the number of frequency bins which is equal to (frame size)/2 + 1, I is the total number of frames and $\boldsymbol{X}_m$ is a $1 \times IN$ vector. Figure 4.1 shows structure of a long vector $\boldsymbol{X}_m$ in PF-ICA.

Separation is executed on the set of M vectors. So, the separation process for PF-ICA is represented as follows:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}$$

where $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M]^T$ and the matrix sizes of $\boldsymbol{Y}$, $\boldsymbol{W}$ and $\boldsymbol{X}$ are $P \times IN$, $P \times M$ and $M \times IN$.

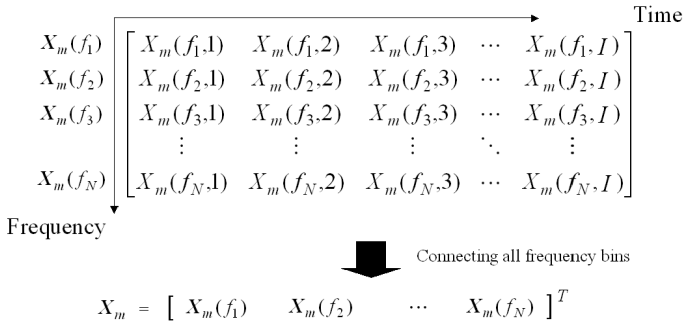

Figure 4.1: Structure of a long vector for microphone m in PF-ICA

To obtain a separation matrix $\boldsymbol{W}$ using JADE, firstly, an $M \times M$ covariance matrix $\boldsymbol{R}_x$ of $\boldsymbol{X}$ is calculated, then a $P \times M$ whitening matrix $\boldsymbol{B}$ is calculated using $\boldsymbol{R}_x$. Secondly, a $P^2 \times P^2$ kurtosis matrix $\boldsymbol{K}$ of the whitened sequence $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{X}$ is calculated, then P groups of the eigenvalues and eigenvectors of the kurtosis matrix $\boldsymbol{K}$, $\{\lambda_r, \boldsymbol{D}_r | 1 \leq r \leq P\}$, is selected from the eigenvalues and eigenvectors of the kurtosis matrix $\boldsymbol{K}$ arranging in descending order, where λ_r is an eigenvalue and $\boldsymbol{D}_r$ is $P \times P$ matrix which consists of P eigenvectors. Then, an independence criterion is defined as the following fourth order cross-cumulant:

$$c(\boldsymbol{V}_{opt}) = \sum_{i,j,l=1\cdots P} |\mathrm{Cum}(\boldsymbol{A}_i, \boldsymbol{A}_i^*, \boldsymbol{A}_j, \boldsymbol{A}_l^*)|^2$$

where $\boldsymbol{A}$ is obtained separating the whitened sequence $\boldsymbol{Z}$ by the $P \times P$ matrix $\boldsymbol{V}_{opt}$. An optimum $\boldsymbol{V}_{opt}$ is estimated by minimizing $c(\boldsymbol{V}_{opt})$, where $(\cdot)^*$ denotes the conjugation and the suffix for $\boldsymbol{A}$ corresponds to the source signal ID. In the processing scheme of JADE, minimization of $c(\boldsymbol{V}_{opt})$ is not realized directly. Instead of that, JADE estimates the optimum $\boldsymbol{V}_{opt}$ by joint diagonalization of a set of P matrices $\mathcal{N}$. Concretely, a set of P matrices $\mathcal{N}$ is defined representing as $\mathcal{N} = \{N_r = \lambda_r D_r | 1 \leq r \leq P\}$, where N_r is a $P \times P$ matrix, then the optimum $\boldsymbol{V}_{opt}$ is estimated minimizing the following criterion by Givens transformation.

$$c(\boldsymbol{V}_{opt}, \mathcal{N}) = \sum_{r=1\cdots P} |\mathrm{diag}(\, V_{opt}^H N_r \, V_{opt})|^2$$

$|\mathrm{diag}(\cdot)|$ represents the norm of diagonal components of matrix $(\cdot)$ and $(\cdot)^H$ denotes complex conjugate transposition. Finally, the separation matrix is obtained as $\boldsymbol{W} = \boldsymbol{V}_{opt}\boldsymbol{B}$. After separation, a time signal of a frequency bin is extracted from the corresponding part on the long separated vector.

PF-ICA does not need to sort frequency bins after separation because PF-ICA executes separation on a set of all frequency bins. PF-ICA, however, cannot solve the amplitude indeterminacy. So, to cope with this problem, "Minimal distortion

principle [69]" is employed. Moreover, PF-ICA cannot cope with frequency characteristics of separation directivity because PF-ICA estimates the only one $P \times M$ separation matrix. That is, PF-ICA does not give the optimal separation results.

4.6 Multi-bin ICA

As explained above, PF-ICA uses a single separation matrix common to all frequency bins, so it cannot express the different directivity for each frequency. In order to take directivity for each frequency into consideration, an idea is introduced into the PF-ICA dividing the total frequency range into bands of considerable width overlapping boundary regions. Vectors for separation are built placing time series of the frequency bin in concern in the center of the corresponding vectors by placing a marginal bins before and after the central bin. However, in case frequency bins before or after the frequency bin in concern are not sufficiently available, processing is done using only available bins. That is, the vector separated by Multi-bin ICA is defined by connecting vectors $\boldsymbol{X}_m(f_n)$. We have

$$
\begin{aligned}
& \cdots \\
\boldsymbol{X}_m(\boldsymbol{f}_{a+1}) &= [\boldsymbol{X}_m(f_1), \cdots, \\
&\quad \boldsymbol{X}_m(f_{a+1}), \cdots, \boldsymbol{X}_m(f_{2a+1})]^T \\
\boldsymbol{X}_m(\boldsymbol{f}_{a+2}) &= [\boldsymbol{X}_m(f_2), \cdots, \\
&\quad \boldsymbol{X}_m(f_{a+2}), \cdots, \boldsymbol{X}_m(f_{2a+2})]^T \\
& \cdots
\end{aligned}
$$

where $\boldsymbol{X}_m(\boldsymbol{f}_{a+1})$ is the vector having the center bin whose frequency ID is $a+1$. Separation is executed using a set of vectors having the same frequency components at the center.

$$\boldsymbol{Y}(\boldsymbol{f}_n) = \boldsymbol{W}(\boldsymbol{f}_n)\boldsymbol{X}(\boldsymbol{f}_n)$$

where $\boldsymbol{X}(\boldsymbol{f}_n) = [\boldsymbol{X}_1(\boldsymbol{f}_n), \boldsymbol{X}_2(\boldsymbol{f}_n), \cdots, \boldsymbol{X}_M(\boldsymbol{f}_n)]^T$ and the sizes of $\boldsymbol{Y}(\boldsymbol{f}_n)$, $\boldsymbol{W}(\boldsymbol{f}_n)$ and $\boldsymbol{X}(\boldsymbol{f}_n)$ are $P \times I(2a+1)$, $P \times M$ and $M \times I(2a+1)$, respectively, placing a marginal bins before and after the central bin.

After separation, the central frequency bins of the vectors in concern are extracted for further processing. Figure 4.2 shows how to build the target vectors. The lower half of Fig. 4.2 shows vectors for which the frequency in concern is f_1, f_2 and f_N, respectively.

4.7 Evaluation

Evaluation Setup

Evaluation is carried out in an experimental cabin that represents both laterally balanced and unbalanced situations. Loudspeakers and microphones are located as

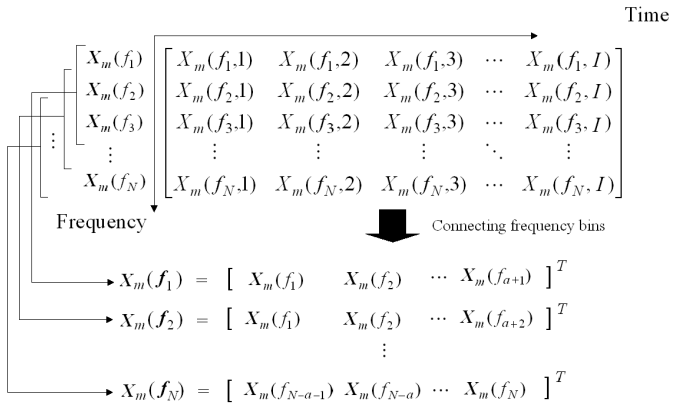

Figure 4.2: Structure of vectors in MB-ICA

shown in Figs. 4.3 and 4.4.

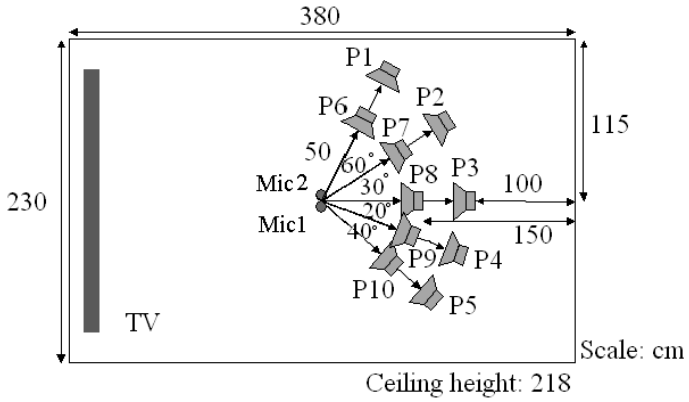

Figure 4.3: Positioning of microphones and loudspeakers in an experimental cabin (Environment 1)

In the environment shown in Fig. 4.3, the frequency characteristics of directionality are almost the same for the both sides of the pair microphones. On the other hand, in another environment Fig. 4.4, the frequency characteristics of directionality are made different by putting reflective and absorbing objects. Interval of two microphones is fixed to be 5cm for both the situations.

In environment 2, the frequency characteristics for each environment was measured to confirm widely different frequency characteristics of a mixing system according to directions. Environment 1 was measured using a loudspeaker located at position P3. Environment 2 was measured using a loudspeaker located at perpendicular position of 100cm from the microphones. If the frequency characteristics of a mixing system are greatly different according to directions, there is a large difference between the frequency characteristics of two microphones. Figure 4.5 shows

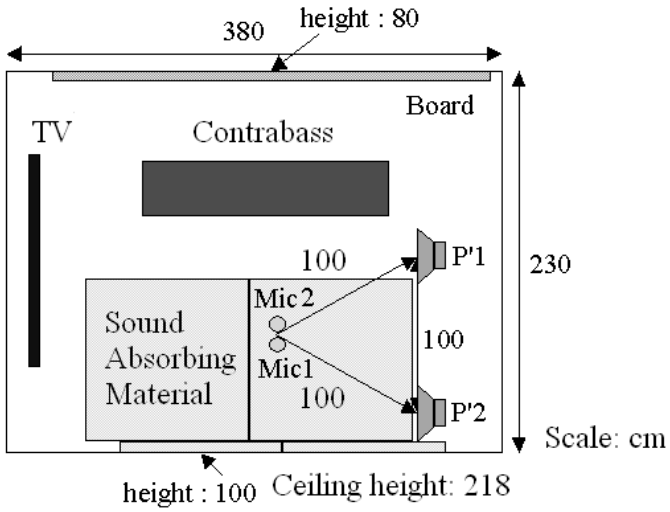

Figure 4.4: Positioning of microphones, loudspeakers and some objects in an experimental cabin (Environment 2)

the difference between the frequency characteristics of two microphones measured in environments 1 and 2. The difference between the frequency characteristics of two microphones is calculated by the following equation.

$$\Delta\hat{H}(f_n) = 10\log(\frac{|\hat{H}_1(f_n)|}{|\hat{H}_2(f_n)|}) \tag{4.1}$$

$\hat{H}_1(f_n)$ denotes the frequency characteristics of the mixing system measured by microphone 1 and $\hat{H}_2(f_n)$ denotes that measured by microphone 2.

From Fig. 4.5, it is clear that the frequency characteristics of the mixing system measured in environment 2 has a large difference according to directions. The phase characteristics have the frequency dependency according to the difference of arrival time between microphones. So, difference dependent on the direction in phase characteristics measured in environment 1 are not small.

Source signals for evaluations are recorded in a sound proof cabin. Source data are 10 short sentences uttered by one male and two females. Then, these source data are emitted from the loudspeakers in the environments depicted in Figs. 4.3 and 4.4, and 400 mixed data were made for each combination of loudspeakers (P1-2, P1-3, P1-5, P1-8, P2-9, P3-10, P4-6, P5-7, P7-10, P8-10) in Fig. 4.3 and a combination of loudspeakers (P'1-2) in Fig. 4.4. That is, in case of Fig. 4.3, the total number of the evaluation data is 4000 mixed short sentences.

Table 4.1 shows the condition of acoustical analysis.

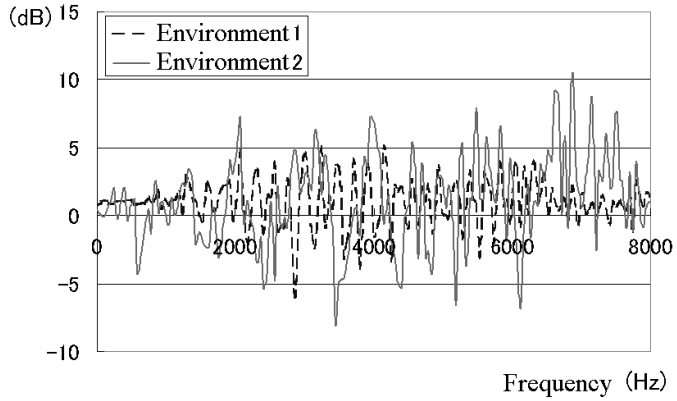

Figure 4.5: Frequency characteristics difference between signals received by two microphones

Table 4.1: Condition of acoustical analysis

Sampling rate	16000 samples/sec
Frame size	1024
Frame shift	512
Window	Hamming

Results

Compared are three processing schemes: a conventional method using JADE employing correlation for resolving the permutation problem, PF-ICA and MB-ICA. Parameter a for MB-ICA is set as 1, 2, 4, 8, 16, 32, 64, 128.

Signal to Noise Ratio (SNR), defined as follows, is employed for the evaluation index of separation performance.

$$SNR_i(k) = \frac{\sum_{n=0}^{N-1} |S_i(f_n, k)|^2}{\sum_{n=0}^{N-1} \left| S_i(f_n, k) - \hat{Y}_i(f_n, k) \right|^2} \tag{4.2}$$

where i denotes the microphone ID, $S_i(f_n, k)$ and $\hat{Y}_i(f_n, k)$ represent the spectrum of a source signal and the spectrum of a separated signal, respectively.

Figures 4.6 and 4.7 show the separation results in the environment shown in Figs. 4.3 and 4.4, respectively. The abscissas in Figs. 4.6 and 4.7 distingwish the conventional ICA, PF-ICA and MB-ICAs of $a = 1,\ 2,\ 4,\ 8,\ 16,\ 32,\ 64,\ 128$.

The ordinates in Figs. 4.6 and 4.7 show the average SNR for all combinations of loudspeakers.

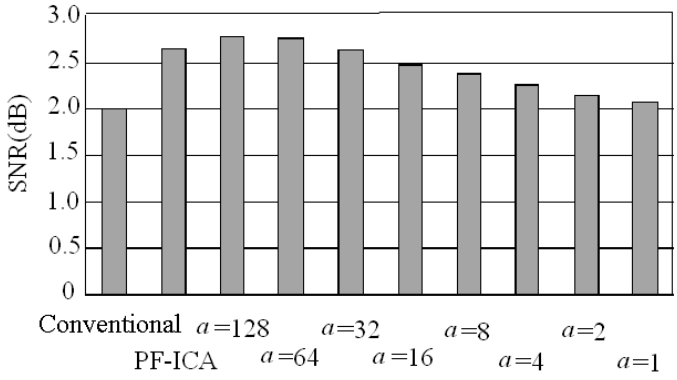

Figure 4.6: Segmental SNR for several versions of ICA in case of the environment shown in Fig. 4.3

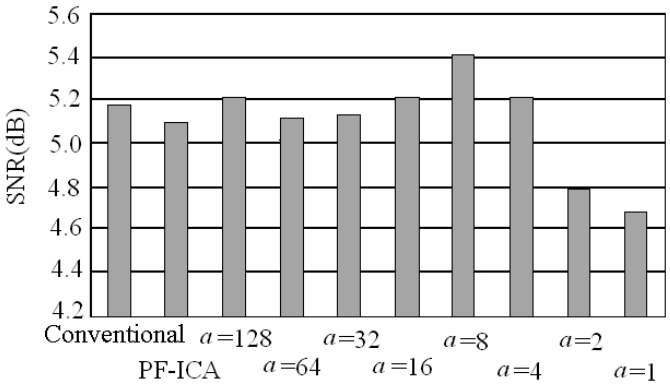

Figure 4.7: Segmental SNR for several versions of ICA in case of the environment shown in Fig. 4.4

From Fig. 4.6, or the environment in which frequency characteristics of directionality are almost the same, MB-ICA and PF-ICA yield almost the same separation performance in case of $a = 128$. However, the separation performance gets nearer to that obtained by the conventional method as parameter a for MB-ICA becomes the smaller, while in environment having innegligible frequency characteristics, MB-ICA yields the best separation results in case of $a = 8$.

4.8 Discussions and conclusions

Generally, if the permutation problem is solved completely, the conventional ICA would yield the best results. In fact, however, it would be impossible to solve the permutation problem completely, so we got the results shown above. In case frequency characteristics of directionality are uniform as the environment shown in Fig. 4.3, PF-ICA is expected to give the best separation results, but such a situation is rare in real environments. In Fig. 4.7, for environment shown in Fig. 4.4, which is closer to actual environments, separation performance of PF-ICA is inferior to the conventional ICA. Hence, in the environment of Fig. 4.4, frequency characteristics have directionality and PF-ICA, having fixed directionality, cannot improve the separation performance. On the other hand, it can be stated from Fig. 4.7 that MB-ICA can yield the best separation performance in case of $a = 8$. In the scheme of MB-ICA, vectors for separation are built placing the frequency bin in concern in the center of the corresponding vectors accompanying a marginal bins before and after the central bin. So, there are common frequency bins between the adjacent vectors and the separation matrix for adjacent frequencies have similar characteristics. As the result, it can be stated that MB-ICA can take the frequency characteristics of directionality into consideration suppressing the permutation errors.

A type of frequency-domain ICA is realized by connecting adjacent frequency bins to avoid the permutation problem. PF-ICA and MB-ICA of large a are proved to yield better results than the conventional ICA in case frequency characteristics of directionality do not drastically change, and MB-ICA of $a = 8$ gives the best results in case frequency characteristics of directionality are different as shown in Fig. 4.4. However, as you can see in Fig. 4.7, the separation performance by MB-ICA depends on parameter a. Parameter a for obtaining the best performance is different between Figs. 4.6 and 4.7. Hence, it can be said that the optimal value for a depends on the directionality of frequency characteristics of the environment. The decision of a is left for feature works.

Chapter 5

Estimation of speeh spectrum based on the Dirichlet process mixture model

5.1 Introduction

Currently, noise robustness is one of the most important problems for developing the effective speech recognition systems in real environments. Several techniques using array microphone are proposed, e.g. delay-and-sum array [14], Griffith-Jim array [73] etc. in order to improve speech recognition rate in real environments. Moreover, as a different approach, Independent Component Analysis [66] attracted the interest in order to solve the Blind Source Separation problem.

On the other hand, S.F. Boll proposed Spectral Subtraction [4] as a technique with a single microphone. In general, the techniques with a single microphone demand the accurate noise estimation. It is not difficult to accurately estimate the noise sequence of stationary noise, e.g. white noise. However, many non-stationary noises, e.g. TV set sound or human voice etc., exist in real environments. So, it is difficult to improve speech recognition rate using simple Spectral Subtraction.

To solve such problems, several estimation methods of non-stationary noises based on a sequential EM (Expectation Maximization) algorithm are reported [45] [46] [47] [48] that can effectively estimate noises. However, their computation costs are expensive because frame by frame iterative estimation is required for the convergence of noise parameters. Owing to the advancement of computer performance, particle filter-based sequential estimation methods [49] [50] have attracted attention and been applied to various research fields. The particle filtering is a Bayesian estimation method, whose main estimation framework is based on a sequential Monte

Carlo method. Thus, the computational costs of the particle filter are cheaper than the sequential EM algorithm because iterative estimation is not always required. Within the field of speech recognition, Fujimoto *et al.* proposed a noise estimation technique based on a particle filtering [51]. This technique consists of the following two parts: one is a noise estimation based on particle filtering and the other is a minimum mean square error (MMSE) based estimation with a Gaussian Mixture Model (GMM) of the speech. An essential point of this technique is to develop an accurate GMM beforehand. To develop the accurate GMM it is necessary to use huge number of speech data.

This chapter proposes a technique for the estimation of noise and speech spectrum without developing the GMM. Instead of the GMM, the speech spectrum is modeled using a DPM [74]. The Dirichlet Process (DP) [75] is a non-parametric probability distribution over the space of all possible distributions. The DP is used as the prior of the DPM. The DP can be considered as the probability distribution for the probability distribution of mixture components. The DP is a generative model for infinite distribution. So, DPM allows us to mix the infinite probability distribution. By using DPM in the estimation process of the speech spectrum, it is expected to estimate the spectrum more flexibly.

There are several researches on the nonparametric density estimation using DPM [76], [77]. Caron *et al.* [78] applied the DPM to the density estimation in the context of dynamic models. Caron *et al.* can achieve the improvement of the performance of standard algorithms when the noise pdfs are unknown. Hence, in case where the speech spectra are unknown, we also expect to get better result than the standard algorithms.

5.2 Theoretical concepts of Bayesian algorithms

The objective of the dynamic state estimation by the Bayesian approach is to construct the *posterior* probability density function (pdf) $p(n_k|x_{1:k})$ based on the observed sequence $x_{1:k} = \{x_1, x_2, \cdots, x_k\}$, where x_k stands for the measurement vector at time k and n_k stands for the state vector at time k. To define the problem of linear/nonlinear filtering, the state evolves according to the following model:

$$n_k = f_{k-1}(n_{k-1}, w_{k-1}) \tag{5.1}$$

where f_{k-1} is a known, linear/nonlinear function of the state n_{k-1} and of the process noise w_{k-1}. The measurement is related to the state via the measurement model:

$$x_k = g_k(n_k, v_k) \tag{5.2}$$

where g_k is a known, linear/nonlinear function and v_k is measurement noise. The pdf $p(n_k|x_{1:k})$ is obtained recursively via Eqs. (5.1) and (5.2) from the pdf $p(n_{k-1}|x_{1:k-1})$ in the following two stages: prediction and update [49], [79].

We suppose that the pdf $p(n_{k-1}|x_{1:k-1})$ is available. Firstly, at the prediction stage, the prediction density $p(n_k|x_{1:k-1})$ of the state at time k can be obtained via the following Chapman-Kolmogorov equation:

$$p(n_k|x_{1:k-1}) = \int p(n_k|n_{k-1})p(n_{k-1}|x_{1:k-1})dn_{k-1}$$

where the pdf $p(n_k|n_{k-1})$ is defined by the Eq. (5.1). Secondly, at the update stage, when the measurement x_k is observed, the updated pdf can be obtained from the prediction pdf via the following Bayesian rule:

$$\begin{aligned} p(n_k|x_{1:k}) &= p(n_k|x_k, x_{1:k-1}) \\ &= \frac{p(x_k|n_k, x_{1:k-1})p(n_k|x_{1:k-1})}{p(x_k|x_{1:k-1})} \\ &= \frac{p(x_k|n_k)p(n_k|x_{1:k-1})}{p(x_k|x_{1:k-1})} \end{aligned} \tag{5.3}$$

where the normalizing constant

$$p(x_k|x_{1:k-1}) = \int p(x_k|n_k)p(n_k|x_{1:k-1})dn_k \tag{5.4}$$

depends on the likelihood function $p(x_k|n_k)$ defined by the Eq. (5.2). In general, the pdfs given by Eqs. (5.3) and (5.4) cannot be determined analytically. In case where the functions f_{k-1} and g_k are linear and the pdf $p(n_k|x_{1:k})$ is Gaussian, an optimal algorithm, Kalman filter, can be formulated. In the other cases, we have to use approximations or suboptimal Bayesian algorithms, Extended Kalman filter, Particle filter. Brief descriptions of these algorithms are presented in the following sections.

Kalman filter

The Kalman filter [49], [79] is assumed that the *posterior* pdf at every time is Gaussian and the functions f_{k-1} and g_k are linear. That is, Eqs. (5.1) and (5.2) can be rewritten as:

$$\begin{aligned} n_k &= F_{k-1}n_{k-1} + w_{k-1} \\ x_k &= G_k n_k + v_k \end{aligned}$$

where F_{k-1} and G_k are the matrices defining the linear functions, w_{k-1} and v_k are mutually independent zero-mean White Gaussian whose covariances are Q_{k-1} and R_k respectively. The Kalman algorithm, derived by Eqs. (5.3) and (5.4), can be considered as the following recursive relationships:

$$\begin{aligned} p(n_{k-1}|x_{1:k-1}) &= \mathcal{N}(n_{k-1}; \hat{n}_{k-1|k-1}, P_{k-1|k-1}) \\ p(n_k|x_{1:k-1}) &= \mathcal{N}(n_k; \hat{n}_{k|k-1}, P_{k|k-1}) \\ p(n_k|x_{1:k}) &= \mathcal{N}(n_k; \hat{n}_{k|k}, P_{k|k}) \end{aligned}$$

where $\mathcal{N}(n; m, P)$ is a Gaussian density with argument n, the state, mean m and covariance P. The appropriate mean and covariance of the Kalman filter are computed as follows:

$$\begin{aligned} \hat{n}_{k|k-1} &= F_{k-1}\hat{n}_{k-1|k-1} \\ P_{k|k-1} &= Q_{k-1} + F_{k-1}P_{k-1|k-1}F_{k-1}^T \\ \hat{n}_{k|k} &= \hat{n}_{k|k-1} + K_k(x_k - G_k\hat{n}_{k|k-1}) \\ P_{k|k} &= P_{k|k-1} - K_k A_k K_k^T \end{aligned}$$

where

$$\mathrm{A}_k = G_k P_{k|k-1} G_k^T + R_k$$

is the covariance matrix of $x_k - G_k\hat{n}_{k|k-1}$, and

$$K_k = P_{k|k-1}G_k^T S_k^{-1}$$

is the Kalman gain.

Extended Kalman filter

In the real situations, the optimal filter (i.e. Kalman filter) is hard to use because of the nonlinearity of the target state. Instead, we have to use approximations or suboptimal Bayesian algorithms. In this section, we introduce the Extended Kalman Filter (EKF) [49], [79].

The EKF can be applied for nonlinear function f_{k-1} and g_k with additive noise. So, Eqs. (5.1) and (5.2) can be rewritten as follows:

$$n_k = f_{k-1}(n_{k-1}) + w_{k-1} \tag{5.5}$$
$$x_k = g_k(n_k) + v_k \tag{5.6}$$

Then, the nonlinear functions in Eqs. (5.5) and (5.6) are approximated by the first term in their Taylor series expansion. The mean and covariance of the EKF are computed as follows:

$$\begin{aligned} \hat{n}_{k|k-1} &= f_{k-1}(\hat{n}_{k-1|k-1}) \\ P_{k|k-1} &= Q_{k-1} + \hat{F}_{k-1}P_{k-1|k-1}\hat{F}_{k-1}^T \\ \hat{n}_{k|k} &= \hat{n}_{k|k-1} + K_k(x_k - g_k(\hat{n}_{k|k-1})) \\ P_{k|k} &= P_{k|k-1} - K_k A_k K_k^T \end{aligned}$$

where

$$\begin{aligned} A_k &= \hat{G}_k P_{k|k-1}\hat{G}_k^T + R_k \\ K_k &= P_{k|k-1}\hat{G}_k^T S_k^{-1} \end{aligned}$$

$\hat{F}_{k-1}$ and $\hat{G}_k$ are the local linearization of functions f_{k-1} and g_k respectively.

$$\begin{aligned}\hat{F}_{k-1} &= \left.\frac{\partial f_{k-1}}{\partial n_{k-1}}\right|_{n_{k-1}=\hat{n}_{k-1|k-1}} \\ \hat{G}_k &= \left.\frac{\partial g_k}{\partial n_{k-1}}\right|_{n_{k-1}=\hat{n}_{k|k-1}}\end{aligned}$$

Unscented Kalman filter

Unscented Kalman Filter (UKF) proposed by Julier [80] performs approximation of *a posteriori* density by a Gaussian density. Unlike EKF, which approximates nonlinear functions f_{k-1} and g_k with linear functions, UKF approximates a probability density with a set of weighted sample points chosen by a deterministic method. These points are transformed by the nonlinear functions f_{k-1} and g_k for obtaining an updated probability density. This approximation is called as unscented transform.

Unscented transform

Unscented transform is a method permitting to calculate statistics of a random variable which suffer nonlinear transformation [81] [82]. We consider a following nonlinear system

$$y = f(x)$$

where x is a random variable with a mean μ_x and a covariance P_{xx} and y is a random variable of statistics to be determined. Weighted sample points $(\mathcal{X}_i, W_i)$ are deterministically chosen so that they completely describe the true mean μ_x and covariance P_{xx}. The nonlinear function f is applied to each sample point to obtain a set of transformed points with the mean μ_y and covariance P_{yy}.

The probability density of a random variable x is approximated by $2n + 1$ weighted sample points shown as

$$\begin{aligned}\mathcal{X}_0 &= \mu_x & W_0 &= \frac{\kappa}{n+\kappa} \\ \mathcal{X}_i &= \mu_x + \left(\sqrt{(n+\kappa)P_{xx}}\right)_i & W_i &= \frac{\kappa}{2(n+\kappa)} \\ \mathcal{X}_{i+n} &= \mu_x - \left(\sqrt{(n+\kappa)P_{xx}}\right)_i & W_{i+n} &= \frac{\kappa}{2(n+\kappa)}\end{aligned}$$

where $\kappa \in \mathbb{R}$, $\left(\sqrt{(n+\kappa)P_{xx}}\right)_i$ is the i-th row of the matrix square root $(n+\kappa)P_{xx}$ and W_i is the weight of i-th sample point. Transformation procedure is shown as follows:

1. Each sample point $\mathcal{X}_i$ is transformed by the nonlinear function f to obtain a set of transformed sample points

$$\mathcal{Y}_i = f(\mathcal{X}_i).$$

2. The mean μ_y is given by the mean of transformed sample points

$$\mu_y = \sum_{i=0}^{2n} W_i \mathcal{Y}_i.$$

3. The covariance matrix P_{yy} is given by

$$P_{yy} = \sum_{i=0}^{2n} W_i (\mathcal{Y}_i - \mu_y)(\mathcal{Y}_i - \mu_y)^T$$

Particle filter

Particle filter [49], [79] [83] is also a suboptimal filter. The particle filter can be applied to nonlinear and nonGaussian problems. In this section, a particle filtering based on the sequential importance sampling is introduced. The fundamental idea of the particle filter is that the posterior density $p(n_{0:k}|x_{1:k})$ are approximated by the particles generated from the importance density.

$$p(n_{0:k}|x_{1:k}) \simeq \sum_{j=1}^{J} \omega_k^{(j)} \delta(n_{0:k} - n_{0:k}^{(j)})$$

where j is the particle ID, J is the total number of the particles, $\omega_k^{(j)}$ is the particle weight, the particles consist of $\omega_k^{(j)}$ and $n_{0:k}^{(j)}$ and $\delta(\cdot)$ is a delta function. If the samples $n_{0:k}^{(j)}$ are drawn from an importance density $q(n_{0:k}|x_{1:k})$, then the weigth $\omega_k^{(j)}$ is represented as follows:

$$\omega_k^{(j)} \propto \frac{p(n_{0:k}^{(j)}|x_{1:k}^{(j)})}{q(n_{0:k}^{(j)}|x_{1:k}^{(j)})} \tag{5.7}$$

$\propto$ represents that the left term is proportional to the right term. $p(n_{0:k}|x_{1:k})$ is written by the following recursive formula using the Bayesian rule.

$$\begin{aligned}
& p(n_{0:k}|x_{1:k}) \\
= & \frac{p(x_k|n_{0:k}, x_{1:k-1}) p(n_{0:k}|x_{1:k-1})}{p(x_k|x_{1:k-1})} \\
= & \frac{p(x_k|n_{0:k}, x_{1:k-1}) p(n_k|n_{0:k-1}, x_{1:k-1}) p(n_{0:k-1}|x_{1:k-1})}{p(x_k|x_{1:k-1})} \\
= & \frac{p(x_k|n_k) p(n_k|n_{k-1})}{p(x_k|x_{1:k-1})} p(n_{0:k-1}|x_{1:k-1}) \\
\propto & \; p(x_k|n_k) p(n_k|n_{k-1}) p(n_{0:k-1}|x_{1:k-1})
\end{aligned} \tag{5.8}$$

If $q(n_{0:k}|x_{1:k})$ can be expressed by the following recursive formula

$$q(n_{0:k}|x_{1:k}) = q(n_k|n_{0:k-1}, x_{1:k}) q(n_{0:k-1}|x_{1:k-1}) \tag{5.9}$$

then sample weight $\omega_k^{(j)}$ can be represented as the following recursive formula by substituting Eqs. (5.8) and (5.9) into Eq. (5.7)

$$\omega_k^{(j)} \propto \omega_{k-1}^{(j)} \frac{p(x_k|n_k^{(j)})p(n_k^{(j)}|n_{k-1}^{(j)})}{q(n_k^{(j)}|n_{0:k-1}^{(j)}, x_{1:k})}$$

5.3 Problem statements

Now, we want to realize the speech recognition with a single microphone in noisy and reverberant environments. In this problem, the accuracy of noise estimation is one of the most important things. Therefore, we want to estimate not only a speech spectrum but also a noise spectrum. Fujimoto *et al.* dealt with the noise estimation problem using a particle filtering and a speech spectrum estimation by a GMM. However, by using DPM in the estimation process of the cspeech spectrum, it is expected to estimate the speech spectrum more flexibly.

5.4 Conventional method using GMM

In this section, we ignore the effect of the reflected waves. In the frequency domain, we have the following relationship between speech S and noise signal N:

$$X = S + N$$

where X is an observed signal. Speech recognition is generally performed in the log spectral domain. So, if we define $X = \exp(x)$, $S = \exp(s)$ and $N = \exp(n)$, we can get

$$\begin{aligned} \exp(x) &= \exp(s) + \exp(n) \\ \log(\exp(x)) &= \log(\exp(s) + \exp(n)) \\ x &= s + \log(1 + \exp(n - s)) \end{aligned}$$

where, x, s and n denote X, S and N in the log spectral domain respectively. The above model has been proposed by Segura *et al.* in [84]. So, it is necessary to consider the nonlinear relationship between the original speech spectrum and the noise spectrum. In this conventional method, a particle filter base noise estimation is used.

Dynamic model for a conventional method

Conventional method based on the utilization of GMM proposed by Fujimoto *et al.*. They employed the observed signal model proposed by Segura *et al.* for each

particle as follows [84]:

$$\begin{aligned} x_k &= s_{k,r_k} \\ &+ \log(I + \exp(n_k - s_{k,r_k})) + v_k \qquad (5.10) \\ &= g(s_{k,r_k}, n_k) + v_k \\ n_k &= n_{k-1} + w_{k-1} \qquad (5.11) \\ v_k &\sim \mathcal{N}(0, \Sigma_{s,r_k}), w_k \sim \mathcal{N}(0, \Sigma_w) \end{aligned}$$

where, k is a frame ID. A frame is a time interval for performing a short-term Fourier transform. s_{k,r_k} is modeled by a GMM representing as $S = \sum_r P_{s,r}\mathcal{N}(\mu_{s,r}, \Sigma_{s,r})$ and s_{k,r_k} is generated as follows:

$$r_k \sim P_s$$

where r_k is randomly chosen according to the mixture weight vector P_s for each Gaussian distribution and then

$$s_{k,r_k} \sim \mathcal{N}(\mu_{s,r_k}, \Sigma_{s,r_k})$$

where, μ_{s,r_k} and Σ_{s,r_k} denote the mean vector and diagonal covariance matrix of the r_k-th Gaussian mixture component.

Conventional algorithm

The noise samples and noise covariance, call the parameters in the following, are estimated by a particle filter. When we use the word "the noise sample", it is from the application point of view not particle filtering point of view. This particle filter consists of an EKF for parameter updating, a sample weight computation [49], residual resampling and a Markov Chain Monte Carlo with Metropolis-Hastings sampling [85] for random variable drawing. The speech spectrum s is estimated by MMSE estimation. The initial noise sample is drawn as

$$\begin{aligned} n_0^{(j)} &\sim \mathcal{N}(\mu_n, \Sigma_n) \\ \Sigma_{n_0}^{(j)} &= \Sigma_n \end{aligned}$$

where, μ_n and Σ_n denote the mean vector and diagonal covariance matrix of initial noise spectrum distribution respectively. μ_n and Σ_n are estimated by the first 5 frames of the observed signal with no speech spectrum in the observed signal.

The tracking performances of noise sequences depends on the accuracy of GMM. In order to develop an accurate GMM, it takes very long time and needs huge volume of data. It will be a problem for applying to various applications.

Polyak averaging and a switching dynamical system

In order to improve the performance of noise estimation, Fujimoto *et al.* employ a Polyak averaging and a switching dynamical system [86]. In real situations, noise

spectrum is not always random, so it is necessary to accurately model the noise spectrum. The Polyak averaging is expressed as follows:

$$\begin{aligned} n_k^{(j)} &= (1-\alpha_p)n_{k-1}^{(j)} + \alpha_p \hat{n}_{k-1} \\ &+ \alpha_p \beta_p (\mu_{n,t}^{(j)} - n_{k-1}^{(j)}) + w_{k-1}^{(j)} \\ \hat{n}_{k-1} &= \Sigma_{j=1}^{J} \omega_k^{(j)} n_{k-1}^{(j)} \\ \mu_{n,k}^{(j)} &= \frac{1}{T_p} \Sigma_{s=k-T_p+1}^{k} n_{s-1}^{(j)} \end{aligned}$$

In real situations, the aspect of noise fluctuation is also time variant. So, parameters for the Polyak averaging, α_p, β_p and T_p need to change according to time. To realize this mechanism, a switching dynamical system is introduced leading to a Jump Markov System. This switching dynamical system has several dynamical systems with different parameter settings, and switches suitable parameters for the next frame according to the index of the current model $m_k^{(j)}$. The target model at the next time instance is randomly selected according to the transition probability from the current model $m_k^{(j)}$ to the target model $m_{k+1}^{(j)}$. The transition probability is defined as follows:

$$p_{m_k, m_{k+1}}^{(j)} = \gamma^{|m_{k+1}^{(j)} - m_k^{(j)}|}$$

where the range of γ is $0 \leq \gamma \leq 1$. Also, the transition probability $p_{m_k, m_{k+1}}^{(j)}$ is normalized as $\Sigma_{m_{k+1}} p_{m_k, m_{k+1}}^{(j)} = 1$.

Parameter updating by Extended Kalman Filter (EKF)

To update the noise parameters an EKF is applied. This EKF is derived from the Eqs. (5.10) and (5.11).

$$\begin{aligned} n_{k|k-1}^{(j)} &= (1-\alpha_p)n_{k-1}^{(j)} + \alpha_p \hat{n}_{k-1} \\ &+ \alpha_p \beta_p (\mu_{n,k}^{(j)} - n_{k-1}^{(j)}) + w_{k-1}^{(j)} \\ \hat{n}_{k-1} &= \Sigma_{j=1}^{J} \omega_k^{(j)} n_{k-1}^{(j)} \\ \mu_{n,k}^{(j)} &= \frac{1}{T_p} \Sigma_{s=k-T_p+1}^{k} n_{s-1}^{(j)} \end{aligned} \tag{5.12}$$

$$\begin{aligned} \Sigma_{n_{k|k-1}}^{(j)} &= F_{k-1}^{(j)} \Sigma_{n_{k-1}}^{(j)} F_{k-1}^{(j)T} + \Sigma_w \\ K_k^{(j)} &= \Sigma_{n_{k|k-1}}^{(j)} G_k^{(j)T} [G_k^{(j)} \Sigma_{n_{k|k-1}}^{(j)} G_k^{(j)T} + \Sigma_{s, r_k^{(j)}}]^{-1} \\ F_{k-1}^{(j)} &= \alpha_p (-1 + \omega_{k-1}^{(j)} + \frac{\beta_p}{T_p}) \\ G_k^{(j)} &= \frac{\partial}{\partial n_{k|k-1}^{(j)}} \left\{ g(s_{k, r_k^{(j)}}^{(j)}, n_{k|k-1}^{(j)}) \right\} \end{aligned} \tag{5.13}$$

$$\hat{n}_k^{(j)} = n_{k|k-1}^{(j)} + K_k^{(j)} \left(x_k - g(s_{k,r_k^{(j)}}^{(j)}, n_{k|k-1}^{(j)}) \right) \tag{5.14}$$

$$\Sigma_{n_k}^{(j)} = \Sigma_{n_{k|k-1}}^{(j)} - K_k^{(j)} G_k^{(j)} \Sigma_{n_{k|k-1}}^{(j)} \tag{5.15}$$

Equations (5.12) and (5.13) are equations for the prediction and a Polyak averaging [86] is employed. $K_k^{(j)}$ is the Kalman gain. $G_k^{(j)}$ is the linearization function. Equations (5.14) and (5.15) are equations for the update.

Sequential importance sampling for particle filtering

In the particle filtering algorithm, *a posteriori* pdf $p(n_{0:k}|x_{0:k})$ is approximated by Monte Carlo sampling as follows:

$$\begin{aligned} p(n_{0:k}|x_{0:k}) &\simeq \frac{1}{J} \Sigma_{j=1}^{J} \delta(n_{0:k} - n_{0:k}^{(j)}) \\ &\simeq \Sigma_{j=1}^{J} \omega_k^{(j)} p(n_{0:k}^{(j)}|x_{0:k}) \end{aligned}$$

In the sequential importance sampling, sample weight $\omega_k^{(j)}$ can be represented as the following recursive formula

$$\omega_k^{(j)} \propto \omega_{k-1}^{(j)} \frac{p(n_k^{(j)}|n_{k-1}^{(j)}) p(x_k|n_k^{(j)})}{q(n_k^{(j)}|n_{0:k-1}^{(j)}, x_{0:k})} \tag{5.16}$$

where $q(\cdot|\cdot)$ is an *importance density*.

If it is assumed that the pdf $p(n_k^{(j)}|n_{k-1}^{(j)})$ is equal to $q(n_k^{(j)}|n_{0:k-1}^{(j)}, x_{0:k})$, then the expression of Eq. (5.16) can be rewritten

$$\omega_k^{(j)} \propto \omega_{k-1}^{(j)} p(x_k|n_k^{(j)})$$

where

$$p(x_k|n_k^{(j)}) = \mathcal{N}(x_k; g(s_{k,r_k^{(j)}}^{(j)}, n_k^{(j)}), \Sigma_{s,r_k^{(j)}}).$$

That is to say, Fujimoto *et al.* employed bootstrap filter.

Residual resampling

After calculating sample weights, some of the samples become insignificant. These samples will degenerate the estimation. So, residual resampling step [49] is introduced after the weight calculation. In the residual resampling step, the samples are generated by a resampling with replacement which is proportional to their weights. This method can avoid degeneracy problem by discarding samples with insignificant weights, and to maintain a constant number of samples.

The residual resampling can reduce the effects of degeneracy. However, it causes the other problem which is the particles having high weights are selected many times. As a result, this leads to a loss of diversity among the particles.

Markov chain Monte Carlo step

After the residual resampling step, there is a possibility that most of particles have a same value. To avoid the loss of diversity among the particles, Fujimoto *et al.* introduced a Metropolis-Hasting (MH) sampling [85] in each sample. To simplify the calculation, Fujimoto *et al.* assume that the importance distribution is symmetric. So, the acceptance probability is given by

$$\nu = \min\left\{1, \frac{\omega_k^{*(j)}}{\omega_k^{(j)}}\right\}$$

where $\omega_k^{*(j)}$ denotes the sample weight computed by the MH sampling. The state transition by MH sampling is derived as:

$$\Phi_k^{(j)} = \begin{cases} \Phi_k^{*(j)} & \text{if } u \leq \nu \\ \Phi_k^{(j)} & \text{otherwise} \end{cases}$$

where $\Phi_k^{(j)} = (\omega_k^{(j)}, \hat{n}_k^{(j)}, \Sigma_{n_k}^{(j)})$, $\Phi_k^{*(j)}$ is samples drawn by the MH sampling step, or the outputs from EKF, and u is drawn from the uniform distribution $0 \leq u \leq 1$.

5.5 Estimation of speech spectrum from mixed sound based on DPM

We propose the modeling of the speech spectrum using DPM instead of GMM. By introducing DPM, we expect more flexible estimation of speech spectrum. Because DPM allows us to mix infinite probability distribution. Moreover, DPM can adapt automatically the number of Gaussian distributions needed. If we want to mix other distributions than Gaussian, it is also possible.

Density estimation of the speech spectrum

We consider the speech spectrum distributed according to an unknown probability density F

$$s_k \sim F(s_k) \tag{5.17}$$

We want to estimate this probability density F based on the known samples s_k in the baysian framework. We are interested in the probability density class representing on the following mixture model

$$F(s) = \int_\Theta f(s|\theta) d\mathbb{G}(\theta) \tag{5.18}$$

where $\theta \in \Theta$ is a latent variable, $f(\cdot|\theta)$ is the known mixture density and $\mathbb{G}$ is the mixture distribution. The mixture distribution $\mathbb{G}$ is supposed to be unknown and

to distribute according to $P(\mathbb{G})$. $\mathbb{G}$ is called Random Probability Measure (RPM). Equations 5.17 and 5.18 are reformulated as a following hierarchical form:

$$\begin{aligned} \mathbb{G} &\sim P(\mathbb{G}) \\ \theta_k &\sim \mathbb{G} \\ s_k &\sim f(\cdot|\theta_k) \end{aligned}$$

where $\mathbb{G}$ is a Random Probability Measure (RPM), $P(\cdot)$ is *a priori* distribution, θ_k is called the latent variable, $f(\cdot|\theta_k)$ is a mixed probability density function and s_k is a speech spectrum. In this model, the problem is how to define *a priori* distribution.

In the parametric model framework, the random distribution $\mathbb{G}$ is considered to be characterized by a parameter $\phi \in \Phi$ with unknown finite dimension. The random density F belongs to a space of functions $\mathcal{F}$ of finite dimension. The *a priori* distribution is defined on ϕ and the hierarchical model is reformulated as follows:

$$\begin{aligned} \phi &\sim p(\phi) \\ \theta_k|\phi &\sim \mathbb{G}(\cdot|\phi) \\ s_k|\theta_k &\sim f(\cdot|\theta_k) \end{aligned}$$

In some cases, however, to suppose the probability density taking the certain parametric form limits the inference realized by such models. We are interested in non parametric models which define *a priori* distribution on the large space. The non parametric models are defined as the parametric models with infinite parameters. In the non parametric models, the random distribution $\mathbb{G}$ belongs to a space of functions $\mathcal{F}$ of infinite dimension. In the Bayesian framework, it is supposed that the RPM $\mathbb{G}$ is distributed according to a certain *a priori* distribution, that is, a distribution on a set of probability distributions. We employ here the RPM following a Dirichlet Process (DP) prior.

Dirichlet Processes

Ferguson *et al.* [75] defined two properties for the adequate *a priori* distribution for $P(\cdot)$.

1. The support of the prior distribution should be large.
2. Posterior distribution given a sample of observation from the true probability distribution should be manageable analytically.[1]

In [75], the authors introduced the DP as a probability measure on the space of probability measures, which satisfies the above properties. A probability distribution $\mathbb{G}$ is drawn from $DP(\mathbb{G}_0, \alpha)$ where a probability measure $\mathbb{G}_0$ is defined on a

[1] This property lost the importance by the development of Monte Carlo Method.

measurable space $(\Omega, \mathcal{A})$, α is a positive real number called scale factor. The Dirichlet distribution is the unique distribution over the space of all possible distributions on $\mathcal{A}$ and satisfies the following relation

$$(\mathbb{G}(A_1), \cdots, \mathbb{G}(A_k)) \sim \mathcal{D}(\mathbb{G}_0(A_1), \cdots, \mathbb{G}_0(A_k), \alpha)$$

where $\mathcal{D}$ is a Dirichlet distribution and $A_i \in \mathcal{A}$ [79].

The DP is an extension of the Dirichlet Distribution to the continuous space. Some properties of DP (conjugation and formation by the Polya urn model) are similar to those of Dirichlet Distribution. The detail of the Dirichlet Distribution is denoted in Appendix.

Many probability distributions can be obtained using urn models. The urn model that corresponds to the Dirichlet distribution is the Polya urn model [87]. Polya urn model is defined as follows: Consider a bag with α balls. Initially the number of balls of color j is m_j. We draw balls at random from the bag and at each step we replace the ball that we drew by two same color balls. Then, the probability of the obtaining a ball of color j at the ith step $P(X_i = j)$ is represented as follows:

$$P(X_i = j | X_{1:i-1}) = \frac{m_j + \sum_{k=1}^{i} \delta(X_k = j)}{\alpha + i}.$$

A method for obtaining the Dirichlet process is to consider the limit of the number of colors in the Polya urn model. Moreover, Blackwell *et al.* [87] showed that the predictive distribution is given by the Polya urn model as follows

$$\theta_{k+1} | \theta_k \sim \frac{\alpha}{\alpha + k} \mathbb{G}_0 + \frac{1}{\alpha + k} \sum_{j=1}^{k} \delta(\theta - \theta_j).$$

Dirichlet Process Mixture

It is now possible to reformulate the density estimation problem using the following hierarchical model known as DPM [78]:

$$\begin{aligned} \mathbb{G} &\sim DP(\mathbb{G}_0, \alpha) \\ \theta_k &\sim \mathbb{G} \\ s_k &\sim f(\cdot | \theta_k) \end{aligned}$$

where the RPM $\mathbb{G}$ is the mixture distribution distributed according to $DP(\mathbb{G}_0, \alpha)$. The latent variables θ_k are distributed according to $\mathbb{G}$. $f(\cdot|\theta_k)$ is a mixed probability density function. The following flexible model is adopted for the unknown distribution F

$$F(s) = \int_{\Theta} f(s|\theta) d\mathbb{G}(\theta)$$

with $\theta \in \Theta$.

Estimation of speech spectrum based on the Dirichlet process mixture

In the Bayesian framework, our problem of estimating both a noise spectrum and a speech spectrum, is equivalent to the determination of the probability $p(n_{0:k}, s_{1:k}|x_{1:k})$. A speech spectrum s_k is supposed to be distributed according to a DPM of base mixed distribution $\mathcal{N}(\mu_k, \Sigma_k)$ and scale parameter α [78]. Instead of developing an accurate GMM, we introduce the estimation of speech spectrum with the DPM model which will adapt automatically the number of Gaussian laws to use for the modeling of the speech spectrum.

The problem is now to determine the probability $p(n_{0:k}, \theta_{1:k} \mid x_{1:k})$, which can be written as follows:

$$p(n_{0:k}, \theta_{1:k} \mid x_{1:k}) = p(n_{0:k} \mid \theta_{1:k}, x_{1:k}) p(\theta_{1:k} \mid x_{1:k})$$

where, θ_k consists of the mean vector μ_k and covariance matrix Σ_k of speech spectrum and is drawn from the following Dirichlet process.

$$\begin{aligned} \mathbb{G} &\sim DP(\mathbb{G}_0, \alpha) \\ \theta_k &\sim \mathbb{G} \end{aligned}$$

Then a speech spectrum is drawn from

$$s_k \sim f(\cdot|\theta_k)$$

A probability measure $\mathbb{G}_0$ denotes, a Normal-inverse Wishart base distribution which is usually used when θ_k are a mean μ_k and a covariance Σ_k of Gaussian distribution:

$$\mathbb{G}_0 = \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$$

with $\mu_0, \kappa_0, \nu_0, \Lambda_0$ the hyperparameters of the Normal-inverse Wishart. Sample from the Normal-inverse Wishart distribution is represented as follows:

$$\begin{aligned} \mu|\Sigma &\sim \mathcal{N}(\mu_0, \frac{\Sigma}{\kappa_0}) \\ \Sigma^{-1} &\sim W(\nu_0, \Lambda_0^{-1}) \end{aligned}$$

where $\mathcal{N}$ is a Gaussian distribution and W is the Wishart distribution. The parameters ν_0 and Λ_0 are the degree of freedom and the scale parameter of Wishart distribution respectively. μ_0 is the mean vector and κ_0 is also a scale parameter.

As $p(n_{0:k} \mid \theta_{1:k}, x_{1:k})$ can be computed using the EKF defined by Fujimoto *et al.* [51], we only need to estimate the probability $p(\theta_{1:k} \mid x_{1:k})$ using a particle method. At the k-th frame, it follows that $p(n_k, \theta_{1:k} \mid x_{1:k})$ is approximated through a set of J particles by the following empirical distribution

$$P_N(n_k, \theta_{1:k} \mid x_{1:k}) = \sum_{j=1}^{J} \tilde{\omega}_k^{(j)} p(n_k \mid \theta_{1:k}^{(j)}, x_{1:k})$$

with

$$p(n_k \mid \theta_{1:k}^{(j)}, x_{1:k}) \simeq \mathcal{N}(\hat{n}_{k|k}(\theta_{1:k}^{(j)}), \Sigma_{n_{k|k}}^{(j)}(\theta_{1:k}^{(j)})).$$

The parameters $\hat{n}_{k|k}(\theta_{1:k}^{(j)})$ and $\Sigma_{n_{k|k}}^{(j)}(\theta_{1:k}^{(j)})$ are computed recursively for each particle j using the EKF. On the other hand, the posterior $p(\theta_{1:k}^{(j)} \mid x_{1:k})$ is proportional to $p(\theta_{1:k-1}^{(j)} \mid x_{1:k-1})$ as follows:

$$\begin{aligned} & p(\theta_{1:k}^{(j)} \mid x_{1:k}) \\ \propto \; & p(\theta_{1:k-1}^{(j)} \mid x_{1:k-1}) p(x_k \mid \theta_{1:k}^{(j)}, x_{1:k-1}) p(\theta_k^{(j)} \mid \theta_{1:k-1}^{(j)}) \end{aligned}$$

where

$$\begin{aligned} p(x_k \mid \theta_{1:k}^{(j)}, x_{1:k-1}) &= p(x_k \mid \theta_k^{(j)}, \theta_{1:k-1}^{(j)}, x_{1:k-1}) \\ &= \mathcal{N}(\hat{x}_k(\theta_{1:k}^{(j)}), \hat{\Sigma}_x^{(j)}(\theta_{1:k}^{(j)})) \end{aligned}$$

and

$$\begin{aligned} \hat{x}_k(\theta_{1:k}^{(j)}) &= s_k^{(j)} + \log(I + \exp(n_k^{(j)} - s_k^{(j)})) \\ \hat{\Sigma}_x(\theta_{1:k}^{(j)}) &= G_k^{(j)} \Sigma_{n_k}^{(j)} G_k^{(j)T} + \Sigma_{s,k} \\ G_k^{(j)} &= \frac{\partial}{\partial n_k^{(j)}} \left\{ s_k^{(j)} + \log(I + \exp(n_k^{(j)} - s_k^{(j)})) \right\} \\ s_k^{(j)} &\sim \mathcal{N}(\mu_k^{(j)}, \Sigma_k^{(j)}) \end{aligned}$$

Finally, sample weights are calculated using the following estimates

$$\tilde{\omega}_k^{(j)} \propto \omega_{k-1}^{(j)} \mathcal{N}(\hat{x}_k(\theta_{1:k}^{(j)}), \hat{\Sigma}_x(\theta_{1:k}^{(j)}))$$

because we chose the following importance distribution

$$q(\theta_k | \theta_{1:k-1}^{(j)}, x_{1:k}) = p(\theta_k | \theta_{1:k-1}^{(j)})$$

where $p(\theta_k^{(j)} \mid \theta_{k-1}^{(j)})$ is determined using the polya urn representation [78].

Introduction of reverberation and reflection into the proposed model

In this section, we introduce reverberation and reflected waves into the proposed model. In real situations, speech signals are affected by reverberation and reflected waves. Also, speech signals decays when microphones are located far from the speakers. Let h denotes the transfer characteristics in the log spectral domain and we assume a classical convolution in the time domain. We can get the following equation as an observation equation.

$$\begin{aligned} x_k &= s_k + h_k + \log(I + \exp(n_k - s_k - h_k)) + v_k \\ &= g(x_k, n_k, h_k) + v_k \end{aligned}$$

A transition equation for $h_k^{(j)}$ is defined as follows:

$$\begin{aligned} h_k^{(j)} &= h_{k-1}^{(j)} + u_{k-1}^{(j)} \\ u_{k-1}^{(j)} &\sim \mathcal{N}(0, \Sigma_u) \end{aligned}$$

EKF is modified as follows:

$$\begin{aligned} n_{k|k-1}^{(j)} &= (1-\alpha_p) n_{k-1}^{(j)} + \alpha_p \hat{n}_{k-1} \\ &+ \alpha_p \beta_p (\mu_{n,k}^{(j)} - n_{k-1}^{(j)}) + w_{k-1}^{(j)} \\ \Sigma_{n_{k|k-1}}^{(j)} &= F_{k-1}^{(j)} \Sigma_{n_{k-1}}^{(j)} F_{k-1}^{(j)T} + \Sigma_w \\ h_{k|k-1}^{(j)} &= h_{k-1}^{(j)} + u_{k-1}^{(j)} \\ \Sigma_{h_{k|k-1}}^{(j)} &= \Sigma_{h_{k-1}}^{(j)} + \Sigma_u \\ K_{\eta,k}^{(j)} &= \Sigma_{\eta_{k|k-1}}^{(j)} G_{\eta,k}^{(j)T} S_k^{(j)\ -1} \\ S_k^{(j)} &= G_{n,k}^{(j)} \Sigma_{n_{k|k-1}}^{(j)} G_{n,k}^{(j)T} + G_{h,k}^{(j)} \Sigma_{h_{k|k-1}}^{(j)} G_{h,k}^{(j)T} + \Sigma_k \\ F_{k-1}^{(j)} &= \alpha_p (-1 + \omega_{k-1}^{(j)} + \frac{\beta_p}{T_p}) \\ G_{\eta,k}^{(j)} &= \frac{\partial}{\partial \eta_{k|k-1}^{(j)}} \left\{ g(\Psi^{(j)}) \right\} \\ \hat{\eta}_k^{(j)} &= \eta_{k|k-1}^{(j)} + K_{\eta,k}^{(j)} \left(x_k - g(\Psi^{(j)}) \right) \\ \Sigma_{\eta_k}^{(j)} &= \Sigma_{\eta_{k|k-1}}^{(j)} - K_{\eta,k}^{(j)} G_{\eta,k}^{(j)} \Sigma_{\eta_{k|k-1}}^{(j)} \end{aligned}$$

where $\Psi^{(j)} = \{s_k^{(j)}, n_{k|k-1}^{(j)}, h_{k|k-1}^{(j)}\}$ and $\eta = [n|h]$. In order to estimate n_k, h_k and s_k, it is necessary to determine the probability $p(n_{0:k}, h_{0:k}, \theta_{1:k} \mid x_{1:k})$ which can be decomposed as follows:

$$\begin{aligned} & p(n_{0:k}, h_{0:k}, \theta_{1:k} \mid x_{1:k}) \\ = \ & p(n_{0:k} \mid \theta_{1:k}, x_{1:k}) p(h_{0:k} \mid \theta_{1:k}, x_{1:k}) p(\theta_{1:k} \mid x_{1:k}) \end{aligned}$$

$p(n_{0:k} \mid \theta_{1:k}, x_{1:k})$ and $p(h_{0:k} \mid \theta_{1:k}, x_{1:k})$ are calculated by the EKF respectively and $p(\theta_{1:k} \mid x_{1:k})$ is calculated by the particle filtering which is shown in 5.5.

Detection of speech/non-speech segment

In the high SNR region, there was the possibility that the noise estimation performance by the proposed method degrade [88]. We introduce detection of speech/non-speech frame into the proposed method. Detection is performed based on the dis-

tance defined as follows:

$$\begin{aligned} d_{s_k} &= (x_k - (\hat{s}_k + \log(1 + \exp(\hat{n}_k - \hat{s}_k))))^2 \\ d_{n_k} &= (x_k - \hat{n}_k)^2 \\ \Delta d_k &= d_{s_k} - d_{n_k} \end{aligned}$$

where $\hat{s}_k$ and $\hat{n}_k$ are the estimated speech spectrum and noise spectrum, respectively. If Δd_k is larger than a threshold obtained from the average of Δd_k over first 5 frames, the current frame is considered as the speech frame and modified as follows:

$$\begin{aligned} \hat{s}_k &= \hat{s}_k + \xi\sqrt{d_{s_k}} \\ \hat{n}_k &= \hat{n}_k - \xi\sqrt{d_{s_k}} \end{aligned}$$

In the reverse case, the signs of above equations are inverted. The proposed method can finally be represented as the following algorithm.

5.6 Simulation

Simulation Setup

We compare three processing schemes: first one is a method proposed by Fujimoto *et al.* [86] where Vector Taylor Series method and MMSE are not employed (conventional)[2], second one is the proposed method without considering transfer characteristics and third one is the proposed method with considering transfer characteristics. [3] Three types of data set are made for evaluations. First one is clean speeches recorded in a sound proof cabin, second one is noisy speeches which are artificially generated by adding three types of noises and third one is noisy reverberant speeches which are artificially generated by convolving transfer characteristics with the noisy speeches. Noise data are taken from "Sound Scene Database in Real Acoustical Environment" [89]. We employ white noise, particle noise and shaver noise. Then, these noises are artificially added to clean speeches with SNRs from 0 to 9dB. Transfer characteristics are simulated using the image method [90]. Reverberation time of the simulated data is about 500ms. 100 utterances uttered by four males and two females are used for this evaluation. The contents of the utterances are TV controlling commands, e.g. "volume up", "turn off" and so on. The total number of evaluation data for each SNR is 3,600 short phrases.

GMM with 256 mixture distributions is trained using 500 utterances uttered by 3 males and 2 females.

[2]These processings require large processing costs, so we estimted the clean speech as $\hat{s}_t = \Sigma_{j=1}^{J} \omega_t^{(j)} s_t^{(j)}$ after the particle filtering step.

[3]We compared the proposed method with the Spectral Subtraction method. We cannot obtain the improvement of the speech recognition rate.

Initialization
$j = 1, \cdots, J$
$n_0^{(j)} \sim \mathcal{N}(\mu_n, \Sigma_n), \quad h_0^{(j)} = 0, \quad \omega_k^{(j)} = \frac{1}{J}$
end
$k = 1, \cdots, T$
calculate μ_0, Λ_0
$j = 1, \cdots, J$
if $k == 1$
$\theta_k^{(j)} \sim \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$
else
$\theta_k^{(j)} \sim p(\theta_k^{(j)} | \theta_{k-1}^{(j)})$
end
$s_k^{(j)} \sim \mathcal{N}(\mu_k^{(j)}, \Sigma_k^{(j)}) \quad \theta_k^{(j)} = \{\mu_k^{(j)}, \Sigma_k^{(j)}\}$
switching dynamical system [86]
EKF
$[\hat{x}_k(\theta_{1:k}^{(j)}), \hat{\Sigma}_x(\theta_{1:k}^{(j)}), n_k^{(j)}, \Sigma_{n_k}^{(j)}, h_k^{(j)}, \Sigma_{h_k}^{(j)}]$
$= EKF(n_{k-1}^{(j)}, \Sigma_{n_{k-1}}^{(j)}, \theta_k^{(j)}, h_{k-1}^{(j)}, \Sigma_{h_{k-1}}^{(j)}, x_k)$
calculate sample weights
$\tilde{\omega}_k^{(j)} \propto \omega_{k-1}^{(j)} \mathcal{N}(\hat{x}_k(\theta_{1:k}^{(j)}), \hat{\Sigma}_x(\theta_{1:k}^{(j)}))$
end
$\Sigma_{j=1}^J \tilde{\omega}_k^{(j)} = 1$
Compute $N_{eff} = \left\{ \sum_{j=1}^J \left(\tilde{\omega}_k^{(j)} \right)^2 \right\}^{-1}$
if $N_{eff} \leq \eta$, resample the particles and $\omega_k^{(j)} = \frac{1}{J}$
$\hat{n}_k = \Sigma_{j=1}^J \omega_k^{(j)} n_k^{(j)}, \quad \hat{s}_k = \Sigma_{j=1}^J \omega_k^{(j)} s_k^{(j)}$
Detection of speech/non-speech frame
end

An acoustic model for speech recognition is developed using the Acoustical Society of Japan (ASJ) continuous speech corpus [91]. The training data are about 30,000 sentences uttered by 150 males and 150 females. The feature parameters for the acoustic model is composed of 39 Mel Frequency Cepstral Coefficients (MFCCs) [92] with 13 MFCCs (with zero-th MFCC) and their first and second order derivatives. At the feature extraction stage, Cepstral Mean Subtraction (CMS) [6] is applied to each sentence.

The parameter α for DPM is different according to the length of utterance. Because, as the result of a preliminary experiment, it is clear that short phrases can be recognized even if α is a small number, while in order to recognize the long phrases, it is necessary that α is a large number.

We have no *a priori* information on the speech signal distribution. The value

of the hyperparameters being not known a priori, a simple estimation process is introduced. This estimation bases on the difference between the received signal and the received signal estimated using the estimated speech spectrum at $k-1$ and the estimated noise spectrum at k. That is to say, at the time k, the speech spectrum is estimated roughly as follows:

$$\begin{aligned} \tilde{s}_k^{(j)} &= s_{k-1}^{(j)} + \Delta s^{(j)} + z_k^{(j)} \\ \Delta s_k^{(j)} &= x_k - \hat{x}_k^{(j)}(\bar{s}_{k-1}^{(j)}, \bar{n}_k^{(j)}) \\ &= x_k - (\bar{s}_{k-1}^{(j)} + \log(1 + \exp(\bar{n}_k^{(j)} - \bar{s}_{k-1}^{(j)}))) \end{aligned} \tag{5.19}$$

where $\bar{n}_k^{(j)}$ is obtained from the Polyak averaging [51], $\bar{s}_t^{(j)}$ is obtained from the average over the 5 past frames, $z_{k-1}^{(j)} \sim \mathcal{N}(0, \Sigma_z)$ and $\hat{x}_k^{(j)}(\cdot)$ is an estimated observation signal given $(\cdot)$. $\Delta s^{(j)}$ is determined from the past errors and the effect of the past error decays according to the exponential function. Then, the mean vector and covariance matrix of these particles are calculated and we regard these values as μ_0 and Λ_0 of hyperparameters.

$$\begin{aligned} \mu_0 &= \frac{1}{J}\Sigma_{j=1}^{J}\tilde{s}_k^{(j)} \\ \Lambda_0 &= \sqrt{\frac{1}{J}\Sigma_{j=1}^{J}(\tilde{s}_k^{(j)} - \mu_k)^2} \end{aligned}$$

Then $\kappa_0 = 1$ and $\nu_0 = 500$ for our case.

Parameters for the particle filtering is as follows: w_k is set to $\Sigma_w = 0.1$, u_k is set to $\Sigma_u = 0.0001$ and z_k is set to $\Sigma_z = 1$. The number of particles is 100. Parameters for the Polyak averaging and feedback have four states respectively, e.g. $\alpha_p = \{0.05, 0.1, 0.15, 0.2\}$, $\beta_p = \{0.5, 1.0.1.5, 2.0\}$ and $T_p = \{5, 10, 15, 20\}$. Moreover, A parameter for the switching dynamical system is $\gamma = 0.5$ [86]. μ_N and Σ_N are calculated from the first 5 obserbed samples.

Results

Firstly, the noise and speech spectrum estimation results are shown. Figure 5.1 shows one example of the noise and speech spectrum estimation results by the proposed method. The abscissa is the number of frame and the ordinate is the average energy of filter bank output in the log spectral domain. It is clear that the proposed method can estimate the noise spectrum in case SNR is 3dB.

Figure 5.2 shows one example of the difference between the true noise spectrum and the estimated noise spectrum. You can see that the conventional method fails to estimate a noise spectrum at the start of phrase. On the other hand, the proposed method works more badly than the conventional method at the end of phrase. The reason is that the conventional method using GMM cannot estimate the sudden change of the noise spectrum. However, the proposed method using the DPM permits a flexible noise spectrum estimation.

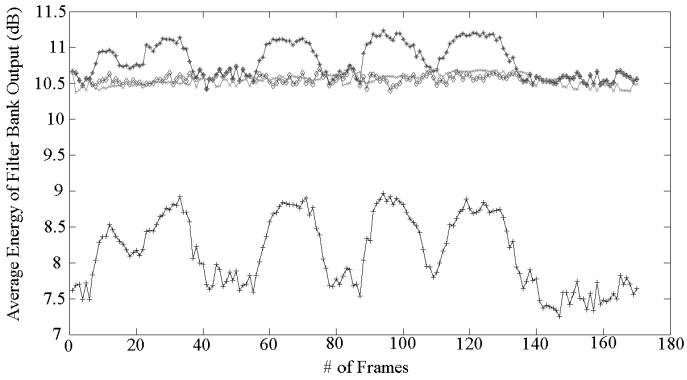

Figure 5.1: Estimation result of the proposed method (in case where the noise signal is a white noise and SNR is 3dB) $*$: received (observed) signal, $\diamondsuit$: true noise signal, $\times$: estimated noise signal, $+$: estimated clean signal

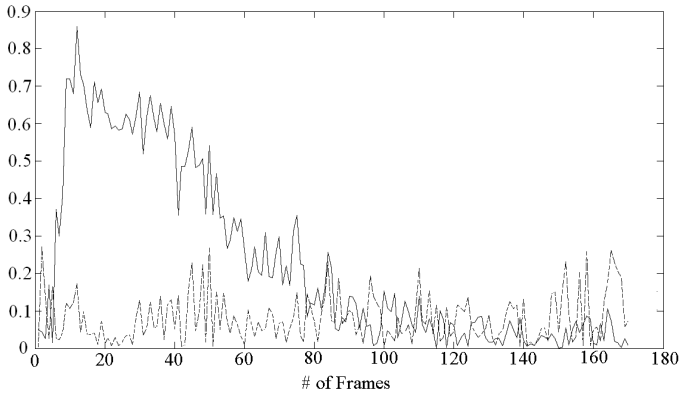

Figure 5.2: Difference average filter bank output between an estimated noise spectrum and a true noise spectrum in case where the noise signal is a white noise and SNR is 3dB (solid line: the method using GMM, dotted line: the proposed method)

Secondly, the speech recognition rates are compared. Evaluations are performed using speech recognition decoder "Julian" [25]. Clean speeches are recorded in a sound proof cabin using a close contact microphone. Table 5.1 shows the condition of acoustical analysis.

Figures 5.3 and 5.4 show speech recognition rates in case we did not considered the effect of the reverberation and the reflected waves and we considered respectively. In these tables, the speech recognition rate for three types of noise data (white noise, shaver noise, particle noise) are shown. Moreover, for each noise data, there are the speech recognition rates of three processing schemes (no processing, proposed method using EKF, conventional method using GMM).

From these figures, it can be seen that the speech recognition rates are improved

Table 5.1: Condition of acoustical analysis

Sampling rate	16000 samples/sec
Frame size	512
Window size	400
Frame shift	160
Feature parameter	39 dimensional mfcc (mfcc+C0+Δmfcc+ΔC0 +ΔΔmfcc+ΔΔC0)
Cepstum coefficient	24 dimension

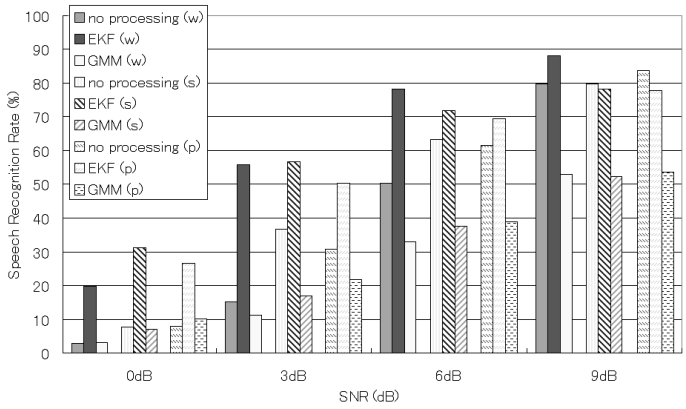

Figure 5.3: Speech recognition rate for three recognition schemes and three noises without reverberation (w: white noise, s: shaver noise, p: particle noise)

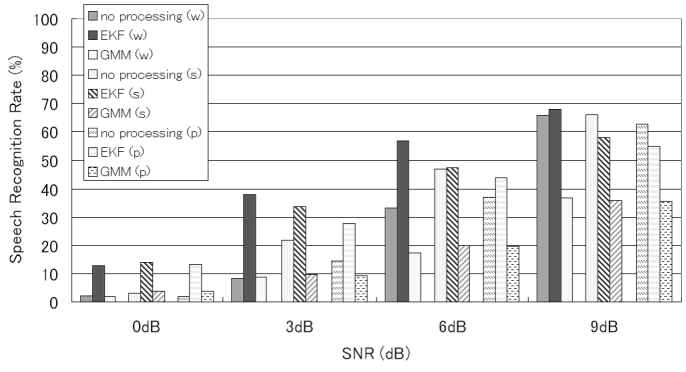

Figure 5.4: Speech recognition rate for three recognition schemes and three noises with reverberation (w: white noise, s: shaver noise, p: particle noise)

using the proposed method in case where the SNRs are 0, 3 and 6dB. On the other hand, in case the SNR is 9dB, the speech recognition rates are degraded except for

the case of white noise. The reason why this degradation of speech recognition rate is that the noise spectrum estimation performance degrades as shown in Fig. 5.5. In

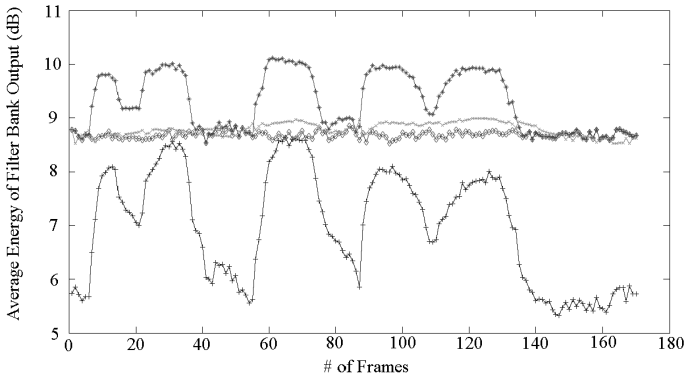

Figure 5.5: Estimation result of the proposed method (in case where the noise signal is a white noise and SNR is 9dB) $*$: received (observed) signal, $\diamondsuit$: true noise signal, $\times$: estimated noise signal, $+$: estimated clean signal

case SNR is 9dB, the noise signal power is small and the fluctuation of noise signal is also small, while the fluctuation of clean speech is large. The EKF fails to estimate a noise signal. As the result, the estimated noise spectrum becomes larger than the true noise spectrum, on the other hand estimated speech spectrum becomes smaller than the true speech spectrum. The speech recognition rate by the conventional method is lower than even that with no processing. The reason is that the time allocated to the GMM learning is not enough long. [4]

In order to improve the speech recognition rate in 9dB SNR case, we employ the UKF. It can be expected to estimate more accurately noise sequence applying the UKF. As a result, we can expect to improve speech recognition rate. Figure 5.6 shows speech recognition rate for three processing schemes (no processing, proposed method using the EKF and UKF).

In the evaluation using speech recognition, our proposed method can improve the speech recognition rate in the SNRs 0dB, 3dB, 6dB and 9dB except for the particle noise. In Fig. 5.7, we show noise spectrum estimation result using the UKF.

From this figure, the proposed method using the UKF gives better estimation result than that using the EKF.

[4]Although this is one reason, we obtained better speech recognition rate by employing the conventional method with VTS and MMSE on the limited data set. The required processing time became 10 times more than that of the convential method without VTS and MMSE.

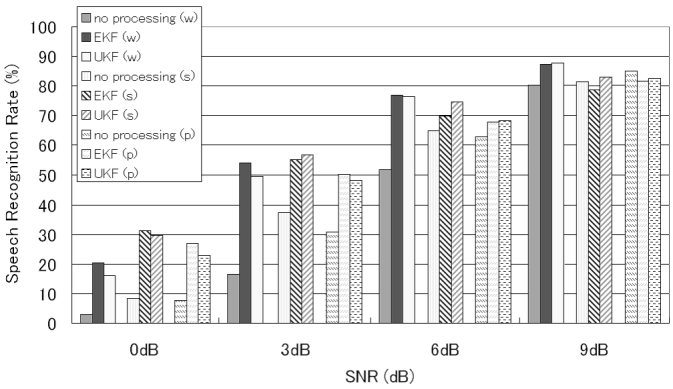

Figure 5.6: Speech recognition rate for three recognition schemes and three noises without reverberation (w: white noise, s: shaver noise, p: particle noise)

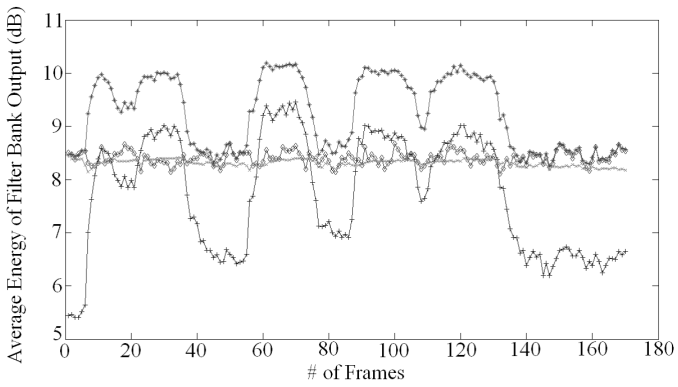

Figure 5.7: Estimation result of the proposed method (in case where a noise signal is a particle noise and SNR is 9dB) $*$: received (observed) signal, $\diamondsuit$: true noise signal, $\times$: estimated noise signal, $+$: estimated clean signal

5.7 Discussions and conclusions

In this chapter, we proposed a method for estimating both the speech spectrum using DPM and noise spectrum using particle filtering. Our proposed method realizes better noise estimation accuracy than the method using the GMM. In the evaluation using speech recognition, our proposed method using the EKF can improve the speech recognition rate in the SNRs 0dB, 3dB, 6dB except for White noise. On the other hand, in case of high SNR, estimation performance degrades. However, we can obtain better speech recognition rate using the UKF than that using the EKF. In the evaluation using speech recognition, our proposed method using UKF can improve the speech recognition rate in the SNRs 0dB, 3dB, 6dB, 9dB except for the particle noise.

Chapter 6

Conclusion

In this dissertation, we proposed four signal processing techniques for realizing a robust speech recognition.

Chapter 2 solved the problem, when we cannot estimate the transfer function of the path between a noise source and a microphone during a double talk, by using the harmonic structure of voiced speech. We proposed an adaptive filtering algorithm using the harmonic structure of voiced segments to reduce non-stationary known noise. The preliminary experiment showed that the proposed method improves both frequency spectra and speech recognition rates in a sound proof cabin more than the SS. Moreover, effectiveness of the proposed method was also confirmed in different practical environments which reverberation time was long, UH and LS.

Chapter 3 solved the dereverberation problem proposing two dereverberation methods in the time- and frequency-domains. One is a method to remove reflected waves from the received signal in the time and frequency domains. The proposed method solved some of the problems that our previous method had. Problems improved or almost solved by the proposed method are: (1) the unreliability in the delay time estimation, (2) the approximation errors in estimated source waves, and (3) the over-subtraction for fricative- and nasal-like segments. Concretely speaking, the delay time estimation was improved by the majority decision on ACFs, over-subtraction is suppressed for consonants such as fricative- and nasal-like segments by classifying each speech segment into subcategories. Applying the proposed method, the recognition rate was improved from 80% to 89% for speech data each for four males and one female.

The other method proposed in chapter 3 was a single channel blind dereverberation method based on auto-correlation functions of time sequences of frequency

components on running power spectra. The proposed method estimated the reverberation time and the frequency characteristics of decay. So, the proposed method achieved dereverberation without any preparatory measurement or *a priori* information. From the performance evaluation on simulated and actual data, the proposed method showed ability to estimate the reverberation time almost correctly. Moreover, the proposed method yielded better reduction results than conventional spectral subtraction methods. Introducing smoothing techniques was expected to lead to further better results.

Chapter 4 solved the permutation problem in the frequency-domain ICA (Independent Component Analysis). Proposed was a type of frequency-domain ICA realized by connecting adjacent frequency bins for separation processing to avoid the permutation problem. The separation performance by MB-ICA (Multi Bin ICA) depended on parameter a. In case where frequency characteristics of directionality do not drastically change, PF-ICA (Permutation Free ICA) and MB-ICA of a large value for a yields better result than the conventional ICA, while in case frequency characteristics of directionality are diverse, MB-ICA of $a = 8$ gave the best results. The optimal value for parameter a to obtain the best performance depended on acoustic environments.

Chapter 5 solved the speech spectrum estimation problem using the Bayesian inference. We proposed a method for estimating both the speech spectrum using DPM (Dirichlet Process Mixture) and noise spectrum using the particle filtering. Our method realized better noise estimation than the method using GMM. In evaluation using speech recognition, our method using EKF (Extended Kalman Filter) can improve the speech recognition rate in case of SNRs 0dB, 3dB, 6dB but not for White noise case. On the other hand, in case of high SNR, the estimation performance degraded. However, we can obtain better speech recognition rate using UKF (Unscented Kalman Filter) than using EKF. In evaluation using speech recognition, our method using UKF improved the recognition rate in cases of SNRs 0dB, 3dB, 6dB, 9dB but not for particle noise.

From evaluation results in Chapters 2, 3, 4 and 5, we saw that these methods gave significant improvement in SNR and speech recognition rate. So, the following issues are left for the future works in order to apply these methods to practical use.

Future works

About the reduction of known noise, we need to employ an accurate estimation technique of the fundamental frequency. As seen in sections 2.5, the cepstrum method gives unreliable estimation of fundamental frequency in noisy environments.

About the dereverberation method, we should apply the proposed dereverberation methods to ICA as a post processing. Currently, ICA achieves high separation

performance. A remaining problem is dereverberation or deconvolution as the post-processing of ICA.

About Independent Component Analysis, we should employ the natural gradient learning for estimating the separation matrix. The parameter space of ICA is not always Euclidean but has a Riemannian metric structure [93]. In such a case, the steepest direction of the cost function is given by the natural gradient. In case the utterance duration is enough long to learn a separation matrix, the natural gradient learning gives better separation results than the other learning rules.

About the Bayesian inference, we should introduce a more precise jump Markov system than it is. We have already introduced a switching dynamical system which is same as jump Markov system. However, the state transition probability is defined based only on the current state. In fact, we should take histories of the state transition process into consideration.

Bibliography

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", Journal of Acoustical Society of America, Vol. 87, pp. 1738–1752, 1990.

[2] H. Hermansky, N. Morgan, "RASTA processing of speech", IEEE trans. Speech Audio Processing 2, pp. 578–589, 1994.

[3] K. Aikawa, H. Singer, H. Kawahara, Y. Tohkura, "Cepstrum representation of speech motivated by time-frequency masking", Journal of Acoutical Sciety America, Vol. 100, pp. 603–614, 1996.

[4] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction,"*IEEE Trans. ASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[5] R.P.Lippman and B.A. Carlson, "Using missing theory to actively select features for robust speech recognition with interruptions," Proc. Eurospeech'97, pp.37-40.

[6] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am. 55 (6), pp. 1304-1312, 1974.

[7] D. Giuliani, M. Omologo, P. Svaizer, "Talker localization and speech recognition using a microphone array and a crosspower spectrum phase analysis", Proc. Int. Conf. Spoken Language Processing, ICSLP94, S22-1, pp. 1243–1246, 1994.

[8] L. Den, A. Acero, L. Jiang, J. Droppo, X. Huang, "High performance robust speech recognition using stereo training data", Proc. Int. Conf. Acoust. Speech Signal Processing, ICASSP01, 2001.

[9] A. P. Varga, R. Moore, "Hidden Markov Model decomposition of speech and noise", Proc. Int. Conf. Acoust. Speech Signal Processing, pp. 845–848, ICASSP90, 1990.

[10] M. J. F. Gales, S. J. Young, "Robust continuous speech recognition using paralel model combination", IEEE Trans. Speech and Audio Processing 4, pp. 352–359, 1996.

[11] F. Martin, K. Shikano, Y. Minami, Y. Okabe, "Recognition of noisy speech by composition of hidden Markov models", IEICE technical report, SP92-96, 1992.

[12] T. Takaguchi, S. Nakamura, K. Shikano, "HMM-Separation based speech recognition for a distant moving speaker", IEEE Trans. Speech and Audio Processing Vol.9, No. 2, pp. 127–140, 2001.

[13] Nakamura, S., "Towards robust speech recognition in real acoustic environments", Technical report of IEICE, SP2002-12, pp.31-36, 2002.

[14] J. L. Flanagan, J. D. Johnston, R. Zahn and G.W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", JASA, Vol. 78, pp. 1508-1518, Nov. 1985.

[15] K. OTA and M. YANAGIDA, "Noise Reduction Based on the Harmonic Structure of the Speech Signal in the Known Noise Case", FIT2005, G-015, pp. 295-296, 2005.

[16] K. OTA and M. YANAGIDA, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-wise Time Sequences of Frequency Components", IWAENC, No. 49, Paris, Sep., 2006.

[17] T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano, "Model adaptation based on HMM decomposition for reverberant speech recognition", Proc. ICASSP-97, vol. 2,pp. 827-830,Apr.1997.

[18] A. Baba et al., 2002. Speech recognition by reverberation adapted acoustic model. Proc. of ASJ, 1-9-14,pp. 27-28.

[19] H. Yamamoto, T. Nishimoto, S. Sagayama, "Reverberant Speech Recognition Using Frame-Synchronous Model Composition", IEICE technical report, NLC2003-71, pp.127-132, Dec., 2003.

[20] J. Benesty and Y.Huang, Adaptive Signal Processing, Springer, pp. 137–144, 2003.

[21] William A. Harrison, Jae S. Lim, Elliot Singer, "A new application of adaptive noise cancellation" IEEE Trans. ASSP, vol. 34, no. 1, pp21–27, 1986.

[22] T. Hikichi and F. Itakura, "Investigation on the influence of acoustic environmental change to the acoustic transfer function using 1/3 scale model", Proc. of Acoustic Society of Japan, 2-6-2, pp. 625-626, Sep., 1994.

[23] G. Iliev and N. Kasabov, "Adaptive filtering with averaging in noise cancellation for voice and speech recognition," Proc. ICONIP/ANZIIS/ANNES '99 Workshop, Dunedin, New Zealand, Nov. 22-23, pp. 71–74, 1999.

[24] Julie E. Greenberg, "Modified LMS algorithms for speech processing with an adaptive noise canceller" IEEE Trans. SAP, vol. 6, no. 4, pp. 338–351, 1998.

[25] http://julius.sourceforge.jp/

[26] T. Noguchi, T. Sakai, K. Ohta and M. Yanagida, "Effect of the noise or reverberation on speech recognition rate," Proc. Kansai-Section Joint Convention 2005, pp. 345, Kyoto, Japan, Nov., 2005.

[27] K. Ohta and M. Yanagida, "Blind dereverberation using temporal and spectral subtraction," Proc. of Forum Acusticum 2005, pp. 2573–2578, Budapest, Hungary, August 27-September 2. 2005.

[28] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics." IEEE Trans. ASSP, Vol. 36, No. 2, pp. 145-152, 1988.

[29] A. Baba, D. Matsumoto A. Lee and K. Shikano, "Recognition of Speech with Dereverberation by Spectrum Subtraction in Home Environment", Proc. of Acoustic Society of Japan, no.1-1-9, pp.17-18, Sep., 2004.

[30] Nakajima, H. et al., 1998. Reverberation suppression under frequency-dependent reverberation time condition. Proc. of ASJ, 1-Q-4, pp. 567-568.

[31] D. Matsumoto, A. Baba, A. Lee, H. Saruwatari and K. Shikano, "Speech Recognition of Distant Talking for Human-robot Speech Interface", Proc. of Acoustic Society of Japan, no.1-Q-4, pp.567-568, Mar., 1998.

[32] M. Yanagida and O. Kakusho, "Isolation of source waves by Generalized Convolutional Inverse Matrix." Trans. IECE of Japan, Vol. E64, No. 7, pp. 499-500, 1981.

[33] M. Unoki et al., "A speech dereverberation model based on the MTF concept." Proc. EUROSPEECH2003, pp. 1417-1420, 2003.

[34] T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano, "Model adaptation based on HMM decomposition for reverberant speech recognition." Proc. ICASSP-97, vol. 2,pp. 827-830, 1997.

[35] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure." Proc. ICASSP-2003, vol. 1, pp. 92-95, 2003.

[36] K. Kinoshita, T. Nakatani and M. Miyosi, "Single Channel Blind Dereverberation using multi-step forward linear prediction", Proc. of Acoustic Society of Japan, pp. 511-512, Mar., 2006.

[37] B. Gillespie, H. Malvar and D. Florencio, "Speech Dereverberation via Maximum-kurtosis Subband Adaptive Filtering", Proc. of ICASSP2001, vol. 6, pp. 3701-3704, May, 2001.

[38] K. Ohta and M. Yanagida, "Removing reflected waves using delay time detected by majority decision on auto-correlation functions." Proc. of FIT2004, vol. 2, pp. 365-366, 2004.

[39] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas and Propagation, vol. 34, pp. 276-280, Mar. 1986.

[40] M. Enoki, K. Ota, M. Ida, H. Nagaoka, S. Matsuda, S. Nakamura, Y. Kida, M. Yanagida, "Introducing Speech Interface to IT Home Appliances in the Next Generation" Proc. of IEICE General Conference, D-14-6, Mar., 2007.

[41] http://www.so-net.ne.jp/tv/

[42] http://chasen.naist.jp/hiki/ChaSen/

[43] T.W.Lee,"Independent Component Analysis", Kluewer, 1998.

[44] A.Hyvärinen, J.Karhunen and E.Oja,"Independent Component Analysis", John Wiley, New York, 2001.

[45] V. Krishnamurthy and J. B. Moore, "On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure," IEEE Trans. on SP, Vol. 41, No. 8, pp. 2557-2573, Aug. 1993.

[46] M. Afify and O. Siohan, "Sequential Estimation with Optimal Forgetting for Robust Speech Recognition," IEEE Trans. on SAP, Vol. 12, No. 1, pp. 19-26, Jan. 2004.

[47] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise Adaptive Speech Recognition Based on Sequential Noise Parameter Estimation," Speech Communication, Vol. 42, Issue 1, pp. 5-23, Jan. 2004.

[48] T. A. Myrvoll and S. Nakamura, "Online Cepstral Filtering Using A Sequential EM Approach with Polyak Averaging and Feedback," Proc. ICASSP '05, Vol. I, pp. 261-264, March, 2005.

[49] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," IEEE Trans. SP, Vol. 50, No. 2, pp. 174.188, Feb. 2002.

[50] K. Yao and S. Nakamura, "Sequential noise compensation by sequential Monte Carlo method," Proc. NIPS '01, pp. 1205-1212, Dec. 2001.

[51] M. Fujimoto and S. Nakamura, "Particle Filtering and Polyak Averaging-based Non-stationary Noise Tracking for ASR in Noise", Proc. ASRU '05, pp. 337-342, Nov. 2005.

[52] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura,"Evaluation of blind signal separation method using directivity pattern under reverberant conditions", Proc. ICASSP2000, pp3140-3143, Istanbul, Turkey, June, 2000.

[53] S.Ikeda and N.Murata,"A method of ICA in time-frequency domain",Proc WS on Independent Component Analysis and Blind Signal Separation (ICA'99), pp.365-371, Aussios, France, Jan., 1999.

[54] H. Sawada, R. Mukai, S. Araki and S. Makino, "A Robust Approach to the Permutation Problem of Frequency-Domain Blind Source Separation", in Proc. of ICASSP2003, Vol. 5, pp.381-384, Apr. 2003.

[55] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation", Trans. of IEEE SAP., Vol. 12, pp. 530-538, Sept, 2004.

[56] R. Mukai, H. Sawada, S. Araki and S. Makino, "Near-Field Frequency Domain Blind Source Separation for Convolutive Mixtures", in Proc. of ICASSP 2004, vol.IV, pp.49-52, May 2004.

[57] L. Di Persia, K. Ota and M. Yanagida, "A Method for Solving the Permutation Problem in ICA", Technical Report of IEICE, Vol.105 No.686, pp. 53-58, Mar, 2006.

[58] L. Di Persia, T. Noguchi, K. Ota and M. Yanagida, "Performance of Permutation-Free ICA", Technical Report of IEICE, Vol.106 No.78, pp.1-6, May, 2006.

[59] T. Kim, H. Attias, S-Y. Lee, and T-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," IEEE Transactions on Speech and Audio Processing, vol. 15, no. 1, 2007.

[60] E. Robledo-Arnuncio, H. Sawada, and S. Makino, "Frequency domain blind source separation of a reduced amount of data using frequency normalization, " in Proc. ICASSP2006, pp. V-837 - V-840, May 2006.

[61] D. Pham, P. Garrat and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach", Proc. EUSIPCO, pp. 771–774, 1992.

[62] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", Neural Computation 7, pp. 1129–1159, 1995.

[63] S. Amari, A. Cichocki and H. H. Yang, "A new learning algorithm for blind source separation", Advances in Neural Information Processing Systems 8, MIT press, pp. 757–763, 1996.

[64] S. Amari, "Neural learning in structured parameter spaces - naturaol Riemannian gradient", Advances in Neural Information Processing Systems 9, MIT press, pp. 127–133, 1997.

[65] M. Gaeta and J. Lacoume, "Source separation without prior knowledge: the maximum likelihood solution", Proc. EUSIPCO'90, pp. 621–624, 1990.

[66] P. Comon, "Independent component analysis: A new concept?", Signal Proc., Vol. 36, pp. 287-314, 1994.

[67] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis", IEEE Trans. on Neural Networks, 10 (3), pp. 626–634, 1999.

[68] J.F.Cardoso, and A.Souloumiac, "Blind beamforming for non Gaussian signals", IEEE Proceeding-F, vol. 140, no 6, pp.362-370, Dec 1993.

[69] K. Matsuoka, "Minimal distortion principle for blind source separation", Proceedings of the 41st SICE Annual Conference, vol. 4, pp. 2138–2143, 2002.

[70] N. Wiener, "Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications." MIT Press, 1949.

[71] "Perceptual Evaluation of Speech Quality, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ", ITU-T Recommendation P. 862, Feb, 2001.

[72] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata and T. Morita, "Real-Time Implementation of Two-Stage Blind Source Separation Combining SIMO-ICA and Binary Masking," Proceedings of 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC2005), pp.229 — 232, September 2005.

[73] L. J. Griffith, and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propagation, vol.30, no.1, pp.27–34, 1982.

[74] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to nonparametric problems.," Annals of Statistics, 2, pp.1152–1174, 1974.

[75] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems.," Annals of Statistics 1, pp.209–230, 1973.

[76] M. D. Escobar and M. West, "Bayesian Density Estimation and Inference using Mixtures," Journal of the American Statistical Association, Vol. 90, No. 430, 1995.

[77] A. Kottas, "Dirichlet Process Mixtures of Beta Distributions, with Applications to Density and Intensity Estimation.," Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods, 23rd ICML, 2006.

[78] F. Caron, M. Davy, A. Doucet, E. Duflos, P. Vanheeghe. "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures.", International Conference on Information Fusion (FUSION'06), Florence, Italia, July 10-13, 2006.

[79] F. Caron, "Inférence bayésienne pour la détermination et la sélection de modèles stochastiques.," Doctral thesis of école centrale de Lille.

[80] S. Julier and J. Uhlmann, "A new extension of teh Kalman filter to nonlinear systems.", Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, Orlando, Fl. 1997.

[81] S. Julier, J. Uhlmann and H. Durrant-whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators.", IEEE Transactions on Automatic Control, Vol. 45, No. 3, pp. 477–482, 200.

[82] E. Wan and R. Van Der Merwe, "The unscented Kalman filter for Nonlinear estimation.", IEEE symposium 2000 (AS-SPCC), Lake Louise, Alberta, Canada, 2000.

[83] A. Doucet, N. de Freitas and N. Gordon, "Sequential Monte Carlo Methods in practice.," Springer-Verlag, 2001.

[84] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado,"Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks," Proc. EuroSpeech '01, Vol. I, pp. 221-224, Sept. 2001.

[85] D. Gamerman and H. F. Lopes, "Markov Chain Monte Carlo", Chapman & Hall/CRC, 2006.

[86] M. Fujimoto and S. Nakamura, "Sequential non-stationary noise tracking using particle filtering with switching dynamical system," Proc. ICASSP2006, pp. 769-772, May 2006.

[87] D. Blackwell and J. B. MacQueen, "Ferguson distributions via polya urn schemes.," Annals of Statistics, 1:353–355, 1973.

[88] K. Ota, E. Duflos, P. VanHeeghe and M. Yanagida, "Bayesian Inference for Speech Density Estimation by the Dirichlet Process Mixture", Journal of SIC., vol. 16, No. 3, pp. 227–244, 2007.

[89] http://tosa.mri.co.jp/sounddb/indexe.htm

[90] Jont B. Allen and David A. Berkley, "Image Method for efficiently simulating small-room acoustics", Journal of ASA, vol.65, No.4, pp943–950, Apl, 1979.

[91] http://www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html

[92] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[93] S. Amari, "Neural gradient works efficiently in learning", Neural Computation, 10 (2), pp. 251–276, 1998.

[94] A. Gelman, J. Carlin, H. Stern and D. Rubin, "Bayesian data Analysis.", Chapman and Hall, 1995.

Appendix A

Grammer and vocabulary for speech recognition

Table A.1 shows the grammer for speech recognition. "S" and "INPUT" in Table A.1 denote nonterminal sysmbols, respectively and they are expanded as right-hand side. Other symbols except for "S" and "INPUT", e.g. "NS_B" and "TITLE" etc., are terminal sysmbols which are definded in vocabulary dictionary shown in Table A.2.

Table A.1: Grammer for speech recognition

S	: NS_B INPUT NS_E
INPUT	: TITLE
INPUT	: CHANNEL_NAME
INPUT	: CHANNEL_NAME TV
INPUT	: BANGOU CHANNEL
INPUT	: BANGOU
INPUT	: ZENGO NO CHANNEL
INPUT	: ZENGO
INPUT	: DENGEN ONOFF
INPUT	: TV ONOFF
INPUT	: ONOFF
INPUT	: VOLUME UPDOWN
INPUT	: COMMAND

Table A.2: Example of vocabulary for speech recognition

% CHANNEL_NAME	% CHANNEL	% TITLE
NHK 総合	チャンネル	∞のギモン
NHK 教育		ＥＲＸ緊急救命室
毎日	% VOLUME	ＥＴＶ特集
ABC	音量	ぐっさんの声が生まれた
関西		ぴったんこカン・カン
読売	% UPDOWN	ぷぷっぴ１０
テレビ大阪	UP	ぷぷっぴ１０　ウィークエンド
KBS 京都	DOWN	『ぷっ』すま
NHK 衛星第一		ぷるぷる　アンタッチャブル
NHK 衛星第二	% ZENGO	アートエンターテインメント　迷宮美術館
	前	ああわが家
% TV		あいくるしい
テレビ	% NO	アイシールド２１
	の	愛してるぜベイベ★★
% BANGOU		ＩＺＵＭＯ・猛き剣の閃記
1	% DENGEN	愛ってなに
2	電源	愛のエプロン
3		あいのり
4	%ONOFF	アクターズ・スタジオ・インタビュー
5	ON	あさイチ！
6	OFF	朝だ！生です旅サラダ
7		あさパラ！
8	% COMMAND	あざーっす！
9	消音	あしたをつかめ・平成若者仕事図鑑
10		
11	% NS_B	
12	silB	
34		
	% NS_E	
	silE	

Appendix B

Convergence of the algorithm for estimating the path amplitude

Shown in Appendix B is the convergence property of the algorithm for estimating the path amplitude. Appendix is divided into four sections.

B.1 Characteristic properties of $\Delta R_i\left(l_{ij}\right)$

The purpose of this section is to prove that the solution, or the estimated path amplitude, of equation $\Delta R_i\left(l_{ij}\right)=0$, exists in the interval [-1,1] and the algorithm described in 3.2.2 converges to the solution.

The first step of the algorithm is to estimate $\hat{r_i}\left(k\right)$, the expected kth sampled value for microphone $\#i$ whose principal reverberations are to be removed. Then, its ACF $\hat{R}_i\left(\tau\right)$ at delay $\tau=l_{ij}$ is calculated, where $\hat{R}_i\left(l_{ij}\right)$ is normalized by ACF at null delay.

Next, calculated is the difference $\Delta R_i\left(l_{ij}\right)$ between $\hat{R}_i\left(l_{ij}\right)$, the estimated ACF of the signal received by microphone $\#i$, and $\bar{R}_i^*\left(l_{ij}\right)$, the expected average for $R_i\left(l_{ij}\right)$, and is expressed as follows:

$$\hat{R}_i(l_{ij})-\bar{R}_i^*(l_{ij})=\frac{\sum_{k=0}^{N-1}(r_i(k)-\alpha_{ij}r_i(k-l_{ij}))(r_i(k+l_{ij})-\alpha_{ij}r_i(k))}{\sum_{k=0}^{N-1}(r_i(k)-\alpha_{ij}r_i(k-l_{ij}))^2}-\bar{R}_i^*(l_{ij}) \quad \text{(B.1)}$$

$$=\frac{\alpha_{ij}^2 \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij}) - \alpha_{ij}\left(\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) + \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij}) + \sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij})\right) + R_i(l_{ij})}{\alpha_{ij}^2 \sum_{k=0}^{N-l_{ij}-1} r_i^2(k) - 2\alpha_{ij} \sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij}) + R_i(0)} - \bar{R}_i^*(l_{ij}) \tag{B.2}$$

$$=\left(\frac{\frac{1}{R_i(0)}\left\{\alpha_{ij}^2 \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij}) - \alpha_{ij}\left\{\sum_{k=0}^{N-l_{ij}-1} r_i^2(k) + \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij}) + \sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij})\right\} + R_i(l_{ij})\right\}}{\frac{1}{R_i(0)}\left\{\alpha_{ij}^2 \sum_{k=0}^{N-l_{ij}-1} r_i^2(k) - 2\alpha_{ij} \sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij}) + R_i(0)\right\}}\right) - \bar{R}_i^*(l_{ij}) \tag{B.3}$$

$$\Delta R_i(\alpha_{ij}) = \frac{B\alpha_{ij}^2 - (A+D+E)\alpha_{ij} + R_i(l_{ij}/0)}{A\alpha_{ij}^2 - 2C\alpha_{ij} + 1} - \bar{R}_i^*(l_{ij}) \tag{B.4}$$

where the average ACF $\bar{R}_i^*(\tau)$ at $\tau = l_{ij}$ is independent of α_{ij}, so $\bar{R}_i^*(\tau)$ can be regarded as a constant, and (B.3) is normalized by the ACF at null delay to avoid truncation errors, which may occur by calculation as there is a large difference between absolute value A and others. Each component in (B.3) normalized by $R_i(0)$ is replaced with notations defined as follows:

$$A = \frac{1}{R_i(0)} \sum_{k=0}^{N-l_{ij}-1} r_i^2(k), \quad B = \frac{1}{R_i(0)} \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+l_{ij})$$

$$C = \frac{1}{R_i(0)} \sum_{k=0}^{N-l_{ij}-1} r_i(k)r_i(k+l_{ij}), \quad D = \frac{1}{R_i(0)} \sum_{k=0}^{N-2l_{ij}-1} r_i(k)r_i(k+2l_{ij})$$

$$E = \frac{1}{R_i(0)} \sum_{k=0}^{l_{ij}-1} r_i(k)r_i(k+N-2l_{ij}), \quad R_i(l_{ij}/0) = \frac{R_i(l_{ij})}{R_i(0)}$$

where these are constants and satisfy the following inequalities

$$-1 \le A,\ B,\ C,\ D,\ E,\ R_i(l_{ij}/0),\ \bar{R}_i^*(l_{ij}) \le 1$$

Substituting these constants into (B.3), we can obtain a simple form (B.4) for (B.3). Here we introduce plausible conditions to prove that there exists a solution α_{ij} for $\Delta R_i(l_{ij}) = 0$ within interval [-1,1].

$A \ge 0$ because A is the squared sum of the input signal. Comparing the definitions of A and C, we can see that they are product sums over the same interval, where A is a squared sum, while C is not. Based on the property of the ACF, we have a relation between A and C as $-1 < -A < C < A < 1$. Similarly, we have a relation between A and $D+E$ as $-1 < -A < D+E < A < 1$, or $0 < A+D+E$. These inequalities yield $C^2 < A$. Now, the denominator of (B.4) is concave as quadratic coefficient A is positive. The discriminant of the denominator of (B.4) $d = C^2 - A$ is negative because $C^2 < A$. So, the denominator of (B.4) has no

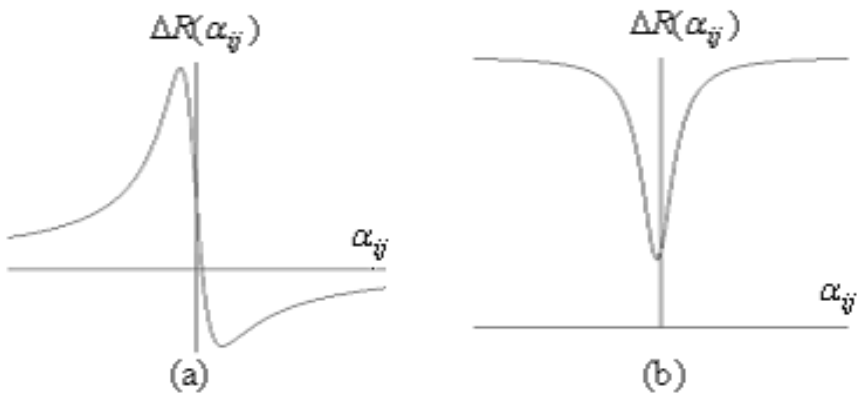

Figure B.1: Function $\Delta R_i(\alpha_{ij})$ where (a) is in general and (b) in special case

solution taking positive values for any α_{ij}. It can be concluded that the function $\Delta R_i(\alpha_{ij})$ is continuous over all region for α_{ij}.

In the case of $\alpha_{ij} = 0$, $\Delta R_i(0)$ becomes $R_i(l_{ij}/0) - \bar{R}_i^*(l_{ij})$, and is positive because $R_i(l_{ij}/0)$ is larger than $\bar{R}_i^*(l_{ij})$. l_{ij} has been chosen at which the difference between the ACF of microphone $\#i$ and the average ACF is positive maximal. As α_{ij} goes to $\pm\infty$, $\Delta R_i(\alpha_{ij})$ asymptotically reaches $B/A - \bar{R}_i^*(l_{ij})$.

B.2 The shape of $\Delta R_i(\alpha_{ij})$ for $-1 \leq \alpha_{ij} \leq 1$

Assuming that target sources are located apart from walls, l_{ij} the time delay that yields the maximum difference between the ACF of the signal received by microphone $\#i$ and the average ACF, is sufficiently larger than unity. This assumption leads to inequalities $A \gg D + E$, $A \gg B$, $A \gg R_i(l_{ij}/0)$ and $A \gg C$. Here, α_{ij} that gives the extremum of the function $\Delta R_i(\alpha_{ij})$ is calculated as

$$\alpha_{ij} = \frac{B - AR_i(l_{ij}/0)}{A(A+D+E) - 2BC} \pm \frac{\sqrt{(B - AR_i(l_{ij}/0))^2 - \{A(A+D+E) - 2BC\}\{2CR_i(l_{ij}/0) - (A+D+E)\}}}{A(A+D+E) - 2BC}$$

Considering relative values for α_{ij} and continuity of $\Delta R_i(\alpha_{ij})$, we can see that $\Delta R_i(\alpha_{ij})$ takes the shape shown in Fig. B.1(a) for the general case and (b) in special case though (b) can be convex or concave.

B.3 Existence of a solution for α_{ij} between -1 and 1

Here, we will confirm that the product $\Delta R_i(-1)R_i(1)$ is negative, in order to show that a solution of the equation $\Delta R_i(\alpha_{ij}) = 0$ exists within the interval [-1,1],

under situation that $\Delta R_i(\alpha_{ij})$ is continuous for $-\infty \leq \alpha_{ij} \leq \infty$. The product $\Delta R_i(-1)R_i(1)$ yields

$$\Delta R_i(-1)\Delta R_i(1) = \left(\frac{B+(A+D+E)+R_i(l_{ij}/0)}{A+2C+1} - \bar{R}_i^*(l_{ij})\right)\left(\frac{B-(A+D+E)+R_i(l_{ij}/0)}{A-2C+1} - \bar{R}_i^*(l_{ij})\right)$$

Dividing denominators and numerators of the fraction part of the above equation by A, we get the result expressed as

$$\Delta R_i(-1)\Delta R_i(1) = \left(\frac{\frac{B}{A}+(1+\frac{D}{A}+\frac{E}{A})+\frac{R_i(l_{ij}/0)}{A}}{1+\frac{2C}{A}+\frac{1}{A}} - \bar{R}_i^*(l_{ij})\right)\left(\frac{\frac{B}{A}-(1+\frac{D}{A}+\frac{E}{A})+\frac{R_i(l_{ij}/0)}{A}}{1-\frac{2C}{A}+\frac{1}{A}} - \bar{R}_i^*(l_{ij})\right)$$

$$\simeq \left(0.5-\bar{R}_i^*(l_{ij})\right)\left(-0.5-\bar{R}_i^*(l_{ij})\right)$$

where fractions B/A, C/A, $(D+E)/A$ and $R_i(l_{ij}/0)/A$ are approximately null and $1/A$ is approximately unity. Let us consider that the product $\Delta R_i(-1)R_i(1)$ is dependent on the average ACF $\bar{R}_i^*(l_{ij})$. The product $\Delta R_i(-1)R_i(1)$ is a quadratic function of the average ACF $\bar{R}_i^*(l_{ij})$. Furthermore, the quadratic coefficient is positive, so this function is concave. Hence, if the average ACF $\bar{R}_i^*(l_{ij})$ satisfies $-0.5 < \bar{R}_i^*(l_{ij}) < 0.5$, the product $\Delta R_i(-1)R_i(1)$ is negative.

On the other hand, it is obvious that $A \gg R_i(l_{ij}/0)$ and $\bar{R}_i^*(l_{ij}) - R_i(l_{ij}/0) < 0$. So, the average ACF $\bar{R}_i^*(l_{ij})$ is approximately null at $\tau = l_{ij}$ and then it satisfies $-0.5 < \bar{R}_i^*(l_{ij}) < 0.5$. It can be concluded that a solution of the equation $\Delta R_i(\alpha_{ij}) = 0$ exists within the interval [-1,1].

B.4 Convergence of the proposed algorithm

Let us verify that the estimation algorithm converges to the solution of the equation $\Delta R_i(\alpha_{ij}) = 0$ that definitely exists within the interval [-1,1]. Here, we consider the initial value for α_{ij}. It is clear that the extremum of the function $\Delta R_i(\alpha_{ij})$ exist only once within the interval [-1,1]. So, we can easily make the algorithm converge using Newton-Raphson method with zero initial value.

Appendix C

Voice-controlled TV system

C.1 System architecture

The TV control system [40] consists of following 6 Sub Systems (SSys.):

- Signal processing SSys.
- Speech recognition SSys.
- Dictionary management SSys.
- Retrieval SSys.
- Command conversion SSys.
- TV control SSys.

A structure of the TV control system is depicted in Fig. C.1. The signal processing SSys. extracts speech segments, reduces noise and suppresses reflected waves. The processed signal is sent to the speech recognition SSys. "Julian" is employed as the speech recognition SSys. here. A speech recognition result obtained by the speech recognition SSys. is sent to the command conversion SSys. and the retrieval SSys. according to the speech recognition result if necessary. A speech recognition result is converted into a TV set activation command by the command conversion SSys., then sent to the TV control SSys. The behaviors of TV are controlled by the TV control SSys. using TV activation commands. The TV program retrieval SSys. tries to find them by TV programs based on a program title itself, a category name or names of performers referring to stored information obtained from an online TV program system. Retrieval results are sent to the TV control SSys. and the TV control SSys. changes the channel if the retrieved program is currently available.

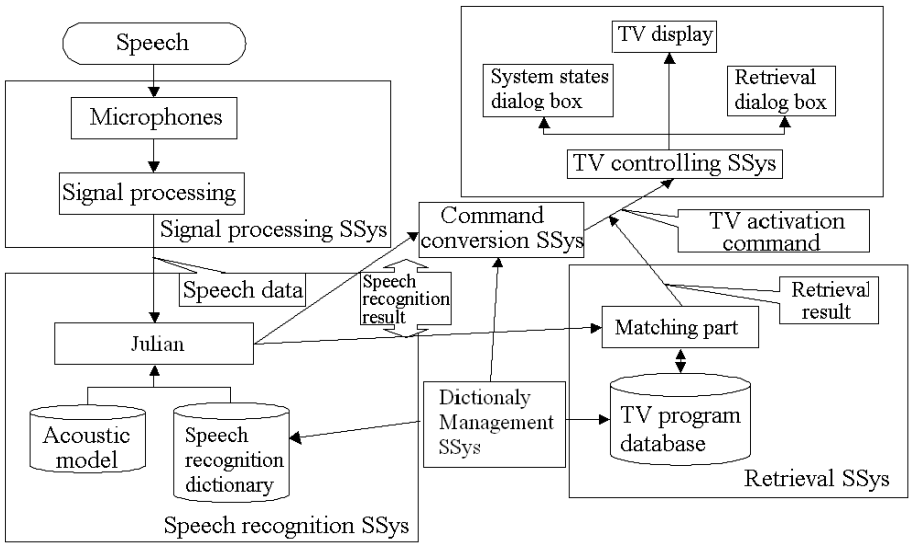

Figure C.1: Structure of TV control system

A speech input device

Figure C.2 shows an input device for speech input for the TV control system. This device is designed as a substitute for the conventional remote controller. It has four microphones mounted on a line with 5cm spacing. It is designed to be put on a table in front of a user expecting 50 $\sim$ 100cm distance to the user.

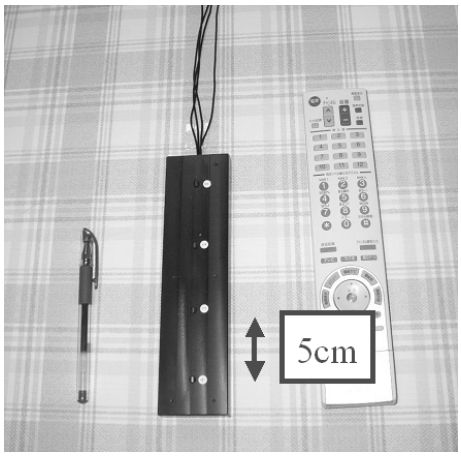

Figure C.2: Interface for speech input

Signal processing SSys.

The signal processing SSys. processes detection of speech segments, estimation of the signal arrival direction, delay-and-sum or phase-rotation and sum processing, known noise reduction and dereverberation. These processes can be combined depending on situations.

Speech recognition SSys.

Julius/Julian[25] developed by continuous speech recognition consortium is employed in the proposed system as the speech recognition decoder. The vocabulary required for speech recognition here is restricted within that for controlling TV set or TV program retrieval. The language model for speech recognition is described as a finite state automaton. Employed as the acoustic model for speech recognition is a phoneme level HMM developed on clean speech data. A vocabulary dictionary is updated periodically by a dictionary management SSys. referring to the online TV program table, e.g. [41].

Dictionary management SSys.

The dictionary management SSys. consists of information extraction part, vocabulary management part, command table generation part and TV program retrieval database generation part.

i) Information extraction part

Information extraction part acquires information about TV programs from online TV program tables available on Internet. **ii) Vocabulary management part**

The vocabulary management part creates the reading using "chasen ver.2.3.3" [42] for each data obtained by the information extraction part. Then each reading is converted to a sequence of phoneme and is added to the speech recognition dictionary.

iii) Command generation part

A command table is referred to when a speech recognition result is to be changed into a TV activation command. The command generation part modifies the command consisting of titles of TV program, broadcasting time, category name, names of performers, channel number extracted by the information extraction part.

iv) TV program retrieval database generation part

TV program names are retrieved by referring to the TV program retrieval database. A row of this database consists of titles of TV program, channel number, broadcasting time, category name and names of performers.

Command conversion SSys.

The command conversion SSys converts a speech recognition result transferred from a speech recognition SSys or a TV program retrieval result transferred from a re-

trieval SSys into a TV activation command and sends them to the TV control SSys.

Retrieval SSys.

The retrieval SSys. performs and/or retrieval using genre names and performer names. Moreover, this SSys. can refine search procedures. To perform the retrieval, we modify the TV program retrieval database using information extracted by the dictionary management SSys. The TV program retrieval database consists of a program title, channel number, broadcasting time, genre, performers and so on.

TV control SSys.

The TV control SSys. controls a TV set with TV control commands transferred from the command conversion SSys. The TV control SSys. selects the specified TV program and displays dialog boxes for showing the system status and/or retrieval results. The dialog box for showing the system status shows the current processing status of the system, e.g. waiting for a user command or during processing etc., together with speech recognition results, the name of the TV station, channel number and sound volume. The other dialog box for retrieval shows the retrieval word and retrieval results, or the number of TV programs matching with the retrieval word and information about the TV program. Figure C.3 shows an example of the output of the system.

The dialog box for showing the system status is placed at the upper left of Fig. C.3 and the dialog box for retrieval is placed at the lower left of Fig. C.3.

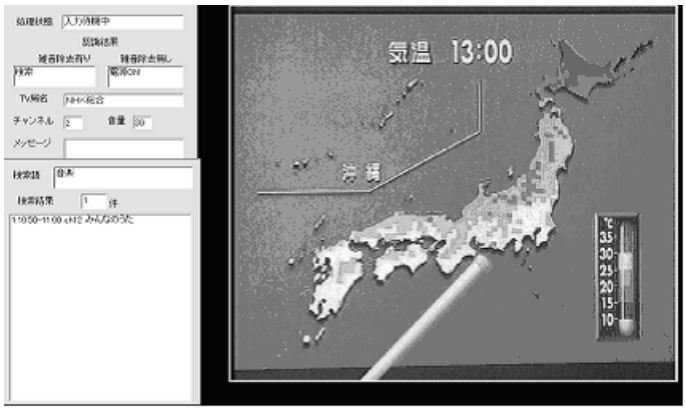

Figure C.3: An expamle of screen shot of TV control system

Appendix D

Supplements on statistics

D.1 Wishart distribution

Let $x_{1:n} = [x_1^T, \cdots, x_n^T]^T$ denote a $n \times p$ matrix which $x_1, \cdots, x_n \in \mathbb{R}^p$ are drawn according to

$$x_1, \cdots, x_n \sim \mathcal{N}(0, \Lambda)$$

A $p \times p$ matrix $A = x_{1:n}^T x_{1:n}$ follows a Wishart distribution.

$$A \sim \mathcal{W}(n, \Lambda)$$

where parameters n and Λ are a degree of freedom and a scale parameter, respectively. The dimension of A ($p \times p$) is not explicitly represented in the notation, but it is determined by Λ. The Wishart distribution can define a distribution on a set of positive definite matrices. The probability density is represented by [94]

$$\mathcal{W}(A; n, \Lambda) = \left(2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i-1}^{p} \Gamma(\frac{n+1-i}{2}) \right)^{-1} |\Lambda|^{-\frac{n}{2}} |A|^{\frac{n-p-1}{2}} \exp\left[-\frac{1}{2} trace(\Lambda^{-1} A) \right]$$

The mean of this distribution is

$$\mathrm{E}[A] = n\Lambda$$

In case dimension $p = 1$, the Wishart distribution is the χ^2 distribution.

D.2 Inverse Wishart distribution

Let Λ denote a $p \times p$ matrix and $n \in \mathbb{N}^*$. A matrix Σ is distributed according to an inverse Wishart distribution ($\Sigma \sim \mathcal{IW}(\cdot; n, \Lambda)$) if Σ^{-1} is distributed according to a

Wishart

$$\Sigma^{-1} \sim \mathcal{W}(n, \Lambda^{-1})$$

The probability density of Σ is given by [94]

$$\mathcal{IW}(\Sigma; n, \Lambda) = \left(2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i-1}^{p} \Gamma(\frac{n+1-i}{2})\right)^{-1} |\Lambda|^{-\frac{n}{2}} |\Sigma|^{-\frac{n-p-1}{2}} \exp\left[-\frac{1}{2} trace(\Lambda\Sigma^{-1})\right]$$

The mean of inverse Wishart distribution is

$$\mathrm{E}[\Sigma] = \frac{\Lambda}{(n-p-1)}$$

D.3 Normal inverse Wishart distribution

Let $\theta = (\mu, \Sigma)$. θ follows a normal inverse Wishart distribution if μ and Σ follow a normal Gaussian distribution

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \frac{\Sigma}{\kappa_0})$$

where hyperparameters $\mu_0 \in \mathbb{R}^p$ and $\kappa_0 > 0$ are known and fixed. Moreover, Σ follows a inverse Wishart

$$\Sigma^{-1} \sim \mathcal{W}(\nu_0, \Lambda_0^{-1})$$

where hyperparameters $\nu_0 > p$ and $\Lambda_0 \in \mathcal{M}_{p\times p}$ are known and fixed. We can simply represent as follows:

$$\theta = (\mu, \Sigma) \sim \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$$

The distribution for θ is given by

$$p(\theta) \propto |\Sigma|^{-\frac{\nu_0+p+2}{2}} \exp\left[-\frac{1}{2} trace(\Lambda_0\Sigma^{-1} - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0))\right]$$

D.4 Dirichlet distribution

A Dirichlet distribution is a distribution on a set of discrete probabilities of the dimension K. Let p_0 denote a vector of the dimension k where $\forall j = 1, \cdots, K$, $p_0(j) \geq 0$ and $\Sigma_{j=1}^{K} p_0(j) = 1$. p_0 defines a set of discrete probabilities of the dimension K. Let $\alpha \geq 0$ denote a scalar coefficient. The Dirichlet distribution $\mathcal{D}(\cdot; \alpha p_0)$ of a vector p_0 and a scale parameter α is defined by

$$Pr(p|p_0, \alpha) = \mathcal{D}(p; \alpha p_0) = \frac{\Gamma(\Sigma_{j=1}^{K} \alpha p_0(j))}{\prod_{j=1}^{K} \Gamma(\alpha p_0(j))} \prod_{j=1}^{K} [p(j)]^{\alpha p_0(j)-1}$$

where p is a vector of the dimension K representing a set of discrete probabilities ($\forall j = 1, \cdots, K$, $p(j) \geq 0$ and $\Sigma_{j=1}^{K} p(j) = 1$) and Γ is a gamma function[1]. The vector p_0 give the mean of a distribution

$$\mathrm{E}[p(j)] = p_0(j).$$

A parameter α is a factor regulating a variance around the mean

$$Var[p(j)] = \frac{(1 - p_0(j))}{(1 + \alpha)}.$$

[1] In case the dimension $K = 2$, the Dirichlet distribution is equivalent to a beta distribution.

RESUMÉ DE LA THÈSE EN FRANÇAIS

Les technologies de la reconnaissance de la parole ont des performances acceptables si l'on utilise un micro dans des environnements calmes. Si des micros se situent à une position distante d'un locuteur, il faut développer des techniques de la soustraction de bruits et de réverbération. Une technique pour réduire des sons émis par les appareils environnants est proposée. Bien que l'annulation adaptative du bruit (ANC) soit une solution possible, l'excès de soustraction peut causer la distorsion de la parole estimée. Le système proposé utilise la structure harmonique des segments vocaliques que les ANCs conventionnels n'a pas prise en compte directement. La méthode de déréverbération conventionnelle provoque l'excès de soustraction car on suppose que la caractéristique de fréquence, est plate. Il faut donc estimer le temps réel de réverbération pour résoudre ce problème. On propose une méthode de dé-réverbération aveugle utilisant un micro avec des fonctions d'auto-corrélation sur la séquence de composants à chaque fréquence. Une technique pour échapper au problème de permutation qui se provoque lorsqu'on utilise l'analyse en composantes indépendantes (ICA) dans le domaine de fréquence, est également proposée : le Multi-bin ICA. Enfin, ce travail propose une technique pour estimer les spectres de bruit et de parole sans développer de modèle de gaussienne à mélange (GMM). Le spectre de la parole est modélisé à l'aide mélange de processus de Dirichlet (Dirichlet Process Mixture : 'DPM') au lieu du GMM.

RESUMÉ DE LA THÈSE EN ANGLAIS

Speech recognition technology reaches almost a practical level if we use a close contact microphone in quiet environments. However, in case microphones are located at a distant position from a speaker, it is necessary to develop noise reduction and dereverberation techniques. A technique for reducing obstructive sounds emitted by the target apparatus to be controlled is proposed. The proposed system uses harmonic structure of voiced segments which conventional ANCs does not directly take into account. A new dereverberation technique considering the frequency characteristics on reflective surfaces is also proposed. Over-subtraction occurs in conventional dereverberation in case of flat frequency characteristics. So, it is required to estimate the actual reverberation time assuming the frequency characteristics of reflection. Proposed is a single channel blind dereverberation technique using auto-correlation functions on the time sequences of frequency components. A technique to escape from the permutation problem which appears in frequency-domain Independent Component Analysis (ICA) is also proposed : the Multi-bin ICA (MB-ICA). Finally, a technique to estimate speech spectrum using a particle filter with a single microphone is proposed. This technique consists in estimating noise and speech spectra using a model based on Dirichlet Process Mixture (DPM) instead of the Gaussian Mixture Model (GMM). It is thus expected to develop a method to estimate the spectrum adaptively.