雑音・残響環境下における
信号音抽出法に関する研究

課題番号:10680374

平成 10－12 年度科学研究費補助金（基盤研究(C)(2)）

研究成果報告書

平成 13 年 3 月

研究代表者　赤木　正人
北陸先端科学技術大学院大学情報科学研究科

目次

[1] 水町、赤木(1999). ”マイクロホン対を用いたスペクトルサブトラクションによる雑音除去法”、電子情報通信学会論文誌、J82-A, 4, 503-512.

[2] Mizumachi, M. and Akagi, M. (2000). "The auditory-oriented spectral distortion for evaluating speech signals distorted by additive noises," J. Acoust. Soc. Jpn. (E), 21, 5 251-258.

[3] Unoki, M. and Akagi, M. (1999). "A method of signal extraction from noisy signal based on auditory scene analysis," Speech Communication, 27, 3-4, 261-279.

[4] 鵜木、赤木(1999). ”聴覚の情景解析に基づいた雑音下の調波復号音の一抽出法”、電子情報通信学会論文誌、J82-A, 10, 1497-1507.

[5] Unoki, M. and Akagi, M. (2001). "A computational model of co-modulation masking release," in Computational Models of Auditory Function, NATO ASI Series, IOS Press, Amsterdam. (in printing)

[6] Ito, K. and Akagi, M. (2000). "A computational model of auditory sound localization based on

ITD," In Recent Developments in Auditory Mechanics, World Scientific Publishing, 483-489.
[7] Itoh, K. and Akagi, M. (2001). “A computational model of auditory sound localization,” in Computational Models of Auditory Function, NATO ASI Series, IOS Press, Amsterdam. (in printing)
[8] Akagi, M., Mizumachi, M.,Ishimoto, Y., and Unoki, M. (2000). "Speech enhancement and segregation based on human auditory mechanisms", Proc. IS2000, Aizu, 246-253.
[9] Mizumachi, M. and Akagi, M. (1998). “Noise reduction by paired-microphones using spectral subtraction,” Proc. ICASSP98, II, 1001-1004.
[10] Mizumachi, M., Akagi, M. and Nakamura, S. (2000). "Design of robust subtractive beamformer for noisy speech recognition," Proc. ICSLP2000, Beijing, IV-57-60.
[11] 石本、鵜木、赤木(2000). ”周期性と調波性を考慮した雑音環境における基本周波数推定”、音響学会聴覚研究会資料、H-2000-81.
[12] Unoki, M. and Akagi, M. (1999). "Segregation of vowel in background noise using the model of segregating two acoustic sources based on auditory scene analysis", Proc. CASA99, IJCAI-99, Stockholm, 51-60.
[13] Akagi, M., Iwaki, M. and Sakaguchi, N. (1998). “Spectral sequence compensation based on continuity of spectral sequence,” Proc. ICSLP98, Sydney, Vol.4, 1407-1410.
[14] Ito, K. and Akagi, M. (2000). "A computational model of binaural coincidence detection using impulses based on synchronization index." Proc, ISA2000 (BIS2000, Invited Session), Wollongong, Australia.

1. はしがき

1.1 研究組織

研究代表者:赤木正人
（北陸先端科学技術大学院大学情報科学研究科教授）
研究分担者:岩城　護（平成 10 年度）
（北陸先端科学技術大学院大学情報科学研究科助手）

1.2 研究経費

平成 10 年度	1,400 千円
平成 11 年度	1,200 千円
平成 12 年度	500 千円
計	3,100 千円

1.3 研究発表

(1) 学会誌等

・雑音抑圧

[1] 水町、赤木(1999). ”マイクロホン対を用いたスペクトルサブトラクションによる雑音除去法”、電子情報通信学会論文誌、J82-A, 4, 503-512.
[2] Mizumachi, M. and Akagi, M. (2000). "The auditory-oriented spectral distortion for evaluating speech signals distorted by additive noises," J. Acoust. Soc. Jpn. (E), 21, 5 251-258.

・音源分離

[3] Unoki, M. and Akagi, M. (1999). "A method of signal extraction from noisy signal based on auditory scene analysis," Speech Communication, 27, 3-4, 261-279.
[4] 鵜木、赤木(1999). ”聴覚の情景解析に基づいた雑音下の調波復号音の一抽出法”、電子情報通信学会論文誌、J82-A, 10, 1497-1507.
[5] Unoki, M. and Akagi, M. (2001). "A computational model of co-modulation masking release," in Computational Models of Auditory Function, NATO ASI Series, IOS Press, Amsterdam. (in printing)

・音源方向推定

[6] Ito, K. and Akagi, M. (2000). "A computational model of auditory sound localization based on ITD," In Recent Developments in Auditory Mechanics, World Scientific Publishing, 483-489.

[7] Itoh, K. and Akagi, M. (2001). “A computational model of auditory sound localization,” in Computational Models of Auditory Function, NATO ASI Series, IOS Press, Amsterdam. (in printing)

(2) 国際会議

・総合

[1] Akagi, M., Mizumachi, M.,Ishimoto, Y., and Unoki, M. (2000). "Speech enhancement and segregation based on human auditory mechanisms", Proc. IS2000, Aizu, 246-253.

・雑音抑圧

[2] Mizumachi, M. and Akagi, M. (1998). “Noise reduction by paired-microphones using spectral subtraction,” Proc. ICASSP98, II, 1001-1004

[3] Mizumachi, M. and Akagi, M. (1999). "Noise reduction method that is equipped for robust direction finder in adverse environments," Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 179-182.

[4] Mizumachi, M. and Akagi, M. (1999). "An objective distortion estimator for hearing aids and its application to noise reduction," Proc. EUROSPEECH99, 2619-2622.

[5] Mizumachi, M. and Akagi, M. (2000). "Noise reduction using a small-scale microphone array under non-stationary signal conditions," Proc. WESTPRAC7, 421-424.

[6] Mizumachi, M., Akagi, M. and Nakamura, S. (2000). "Design of robust subtractive beamformer for noisy speech recognition," Proc. ICSLP2000, Beijing, IV-57-60.

[7] Ishimoto, Y. and Akagi, M. (2000). "A fundamental frequency estimation method for noisy speech," Proc. WESTPRAC7, 161-164.

・音源分離

[8] Unoki, M. and Akagi, M. (1998). “A computational model of co-modulation masking release,” Computational Hearing, Italy, 129-134.

[9] Unoki, M. and Akagi, M. (1998). “Signal extraction from noisy signal based on auditory scene analysis,” ICSLP98, Sydney, Vol.5, 2115-2118.

[10] Unoki, M. and Akagi, M. (1999). "Segregation of vowel in background noise using the model

of segregating two acoustic sources based on auditory scene analysis", Proc. CASA99, IJCAI-99, Stockholm, 51-60.
[11] Unoki, M. and Akagi, M. (1999). "Segregation of vowel in background noise using the model of segregating two acoustic sources based on auditory scene analysis", Proc. EUROSPEECH99, 2575-2578.

・連続聴効果
[12] Akagi, M., Iwaki, M. and Sakaguchi, N. (1998). “Spectral sequence compensation based on continuity of spectral sequence,” Proc. ICSLP98, Sydney, Vol.4, 1407-1410.

・音源方向推定
[13] Itoh, K. and Akagi, M. (1998). “A computational model of auditory sound localization,” Computational Hearing, Italy, 67-72
[14] Ito, K. and Akagi, M. (1999). "A computational model of auditory sound localization based on ITD," Abstracts of Symposium on Recent Developments in Auditory Mechanics, Sendai, Japan, 29P01, 156-157.
[15] Ito, K. and Akagi, M. (2000). "A study on temporal information based on the synchronization index using a computational model," Proc. WESTPRAC7, 263-266.
[16] Ito, K. and Akagi, M. (2000). "A computational model of binaural coincidence detection using impulses based on synchronization index." Proc, ISA2000 (BIS2000), Wollongong, Australia.

(3) 口頭発表

・雑音抑圧
[1] 水町、赤木(1999). ”マスキング特性を考慮した低品質音声に対する歪み評価尺度の提案”、平成 11 年春季音響学会講演論文、3-2-2.
[2] 水町、赤木(1999). ”スペクトルサブトラクションの最適化に関する検討”、平成 11 年秋季音響学会講演論文、3-2-13.
[3] 水町、赤木(2000). ”加法性雑音に対する客観的歪み音声評価尺度”、音響学会聴覚研究会資料、H-2000-4.
[4] 水町、赤木(2000). ”小規模マイクロホンアレーによる音声認識の雑音耐性向上”、平成 12 年春季音響学会講演論文、1-8-5.
[5] 石本、赤木(2000). ”雑音が付加された音声の基本周波数推定と雑音抑圧”、電子情報通信学会技術報告、SP99-169.
[6] 石本、赤木(2000). ”雑音中の音声基本周波数推定法の提案”、平成 12 年春季音響学会講

演論文、2-7-7.
[7] 石本、赤木(2000). ”周期性と調波性を考慮した雑音環境における基本周波数推定”、平成12 年秋季音響学会講演論文、1-Q-7.
[8] 石本、鵜木、赤木(2000). ”周期性と調波性を考慮した雑音環境における基本周波数推定”、音響学会聴覚研究会資料、H-2000-81.
[9] 石本、鵜木、赤木(2001). ”周期性と調波性を考慮した雑音環境における基本周波数推定法の改良”、平成 13 年春季音響学会講演論文、1-6-8.

・音源分離
[10] 鵜木、赤木(1998). “共変調マスキング解除の計算モデルの提案”、音響学会聴覚研究会資料、H-98-51
[11] 鵜木、赤木(1998). ”聴覚の情景解析に基づいた雑音化の定常母音の分離抽出”、平成 10 年秋季音響学会講演論文、2-8-10.
[12] Unoki, M. and Akagi, M. (1999). "Vowel segregation in background noise using the model of segregating two acoustic sources", Proc. 7th meeting of Special Interest Group on AI Challenge, JSAI Tech. Report, SIG-Challeng-9907, 7-14.
[13] 鵜木、赤木(1999). ”聴覚の情景解析に基づいた二波形分離モデルの提案”、電子情報通信学会技術報告、SP98-158.
[14] 鵜木、赤木(1999). ”聴覚の情景解析に基づいた二波形分離モデルの提案”、平成 11 年春季音響学会講演論文、1-2-6.

・音源方向推定
[15] 伊藤、赤木(1999). ”聴覚モデルを用いた音源方向定位に関する一考察”、平成 11 年春季音響学会講演論文、3-2-13.
[16] 伊藤、赤木(2000). ”位相同期特性を考慮した時間差検出モデルについて”、平成 12 年春季音響学会講演論文、1-10-15.
[17] 伊藤、赤木(2000). ”位相同期特性の時間差検出に与える影響について”、平成 12 年秋季音響学会講演論文、2-2-4.
[18] 伊藤、赤木(2001). ”位相同期特性に基づく多重入力と時間差検出に関する一考察”、音響学会聴覚研究会資料、H-2001-5
[19] 伊藤、赤木(2001). ”神経細胞への多重入力の時間的分布と位相同期特性との関係について”、平成 13 年春季音響学会講演論文、1-5-7.
[20] 小林、西田、赤木(2001). ”雑音と反射音に対してロバストな話者方向推定法”、平成13年春季音響学会講演論文、2-7-9.

(4) リサーチレポート

[1] Unoki, M. and Akagi, M. (1998). “A method of signal extraction from noisy signal based on auditory scene analysis,” JAIST Tech. Report, IS-RR-98-0005P.
[2] Unoki, M. and Akagi, M. (1998). “A computational model of co-modulation masking release,” JAIST Tech. Report, IS-RR-98-0006P.

2. 概要

2.1 研究の位置付け

2.1.1 研究の背景

機械による音声認識においては、雑音等で汚れていないきれいな音声ではほぼ実用レベルに達してはいるものの、周囲雑音が存在する場合には認識率の著しい低下は免れない。一方、人間は周囲雑音が大きく、複数の話者が存在しているような状況、あるいは、残響のある環境においてさえも、左右2つの耳で目的とする話者の音声を選択的に聴取することができる。この能力は、音環境にほとんど影響を受けず頑健である(カクテルパーティ効果と呼ばれる)。

そこで、心理学・生理学の知見を基に人間(動物)の優れた聴覚能力を解明してこれをモデル化し、このような問題に適用することを目的とした研究が各所で行なわれている。そして、その一つとして、1990 年代に入り、聴覚を能動的な環境把握システムの一環としてとらえ、その計算理論を構築することにより、上記のような優れた聴覚の働きを他のシステムへ応用する研究(計算機による聴覚情景解析:Computational Auditory Scene Analysis: CASA)が行なわれるようになってきた。

計算機による聴覚情景解析(CASA)に基づいた聴覚の選択的聴取機構に関する研究は、それ自体が 1990 年代に入って行なわれるようになった比較的新しい分野であり、現在、英国 Shefield 大学、米国 MIT、日本では NTT 研究所、ATR 研究所などで行なわれている。

ところが、システムはいくつか発表されてはいるものの、雑音・残響が存在する実音場において、CASA の考え方に則った高性能なシステムは未だに提案されていない。それは、今までのシステムが、観測される音から目的とする音を選択的に聴取する聴覚機能の数理的本質を解明することなく、上辺だけをモデル化しているためである。

2.1.2 研究の目的

本研究の目的は、人間の聴覚機構に存在する選択的聴取機構について、CASA の考え方に沿って、以下に示す研究を行なうことにある。

(1) それぞれの音源あるいは音が聴覚内でどのように記述されているのかの検討、および、人間の聴覚が実現している選択的聴取機能がどのような環境の拘束条件に対応したものであるかの検討を通して、選択的聴取機構の数理的本質を明らかにする。

(2) 選択的聴取機構の数理的記述を基にこの機構を計算機上に実現し、雑音・残響が存在する実音場においてさえも目的信号音を忠実に抽出できるシステムを実現する。

申請者は、既に、

(a) 位相情報を考慮した音源分離法(WASP 法:マイクロホン 1 本使用)

(b) マイクロホン対間の時間差を考慮した雑音抑圧法(NORPAM 法:マイクロホン 2 本使用)

の二つの音源分離・雑音抑圧法について基礎検討を行なってきた。また、論文および国際会議（参考文献[1]-[4]）において成果の発表を行なってきた。

しかし、上記音源分離・雑音抑圧法は聴覚における選択的聴取機構の一部を工学的に実現したに過ぎず、(a) 聴覚における選択的聴取機構の数理的本質の解明、および、(b) 選択的聴取機構の数理的記述を基にした機構の計算機上での実現、が未だに不十分であった。そこで、最新の心理物理学的研究成果および生理学モデルの知見も含めてより高度な信号抽出法を提案する。

2.1.3 研究の意義

提案する研究は CASA の範疇に入る研究ではあるが、他の研究機関が主に top-down 的に音源分離のためのルールを記述し適用しているのとは異なり、あくまでも生理学・心理学的知見にのっとって、聴覚内での音源あるいは音の表現、および、環境の拘束条件と選択的聴取機構の関係を数理的に記述することを試みる。このため、目的信号音の選択的抽出機構の計算理論を関数解析的に記述できるばかりでなく、従来の信号処理技術との融合をはかりながらシステムの構築が可能となる。

現在、雑音・残響抑圧技術として、適応フィルタを用いて雑音を抑圧する方法、マイクロホンアレイを用いて目的とする話者の方向の指向特性を鋭くする方法など、様々な音響的前処理方式が提案されている。しかし、これらの方法は、適応フィルタを計算するための高速な信号処理装置、また、アレイを形作るための多数のマイクロホンを必要とするため、装置は大がかりとなり実用的ではない。

一方、本研究では、聴覚情景解析の研究に関する心理学および生理学から得られた聴覚特性の知見を基にして、雑音・残響抑圧方式を構成することを試みる。人間は、2つの耳から得られる情報のみで音源分離、目的音抽出、雑音・残響抑圧を行なっている。そこで用いられているであろう手法を実現することにより、小規模、実用的な雑音・残響抑圧技術が可能となる。

2.2 成果の概要

2.2.1 まえがき

人間には、雑音が大きい環境であったとしても、簡単に望みの音を聴取できる機能がある。この機能は、人間ばかりではなく他の動物にとっても有益な機能であり、「カクテルパーティ効果」として知られている。もし、この機能がモデル化できたとすれば、音声強調、音源分離、あるいは音声認識、分析に有効な他のアプリケーションを構築する上で、大きな助けになるはずである。

人間のこのような優れた特性をまねてモデル化するためには、工学的知識だけではなく聴覚生理、心理の知見をも必要とする。本報告では、まず始めに、「カクテルパーティ効果」として知られる特性をモデル化するために必要な生理学的、心理学的知見を紹介し、そして次に、これらの知見を用いたモデルを提案する。

カクテルパーティ効果（Cocktail party effect）とは、二つ以上の音メッセージが混在していても一方を選択的に聴取可能であるような聴覚上の効果であり、カクテルパーティのように多数の話者の音声が混在している状況の中で、希望する話者の声を選択して聞くことができることから、この名前が付けられた。カクテルパーティ効果が生じる原因としては、それぞれの音源に対して両耳聴によって知覚される音像の空間的位置（方向と距離）の違い、音の大きさ、ピッチ、音色など音源の特性そのものの違い、また音声の場合には言語的知識、経験などが関係していると見られている。

そこで、本報告では、まず、音源の方向情報、基本周波数、音源が音声とした場合に音源が持つ固有の情報を抽出するための機構のモデル化を試み、そして、抽出された情報をもとにカクテルパーティ効果の数理的モデル化を試みる。

本報告で示すモデルは以下の通りである。

(1) 音声強調

・生理学的知見にもとづいたキャンセレーションの概念の紹介[5][6][7]

・キャンセレーションの概念を用いた空間フィルタによる音声強調

・キャンセレーションの概念を用いた周波数フィルタによる雑音抑圧と基本周波数推定

(2) 音源分離

・Bregman によって提案された「聴覚情景解析」（Auditory Scene Analysis: ASA）と音源分離のための 4 つの発見的規則の紹介[8][9]

・ASA の概念を用いた音源分離モデル

(3) 連続聴効果（音韻修復現象）のモデル

(4) 音源方向推定

・音源方向推定に関わる聴覚末梢系の生理学的機能の紹介

・音源方向推定モデル

2.2.2 音声強調

2.2.2.1 仮定

本報告では、雑音源はN個存在し、時間、周波数、到来方向において局在しているとする。また、受音点の数 M は、N よりも小さいとする。例えば人間では、M=2 であり、音源数 N は非常に大きい。このような条件では、空間フィルタと周波数フィルタが目的の音を取り出すために有効な手段と成り得る。そこで本報告では、空間フィルタと周波数フィルタを構築するための時間的情報として両耳間時間差（inter-aural time difference: ITD）と基本周期を用いて、キャンセレーションの概念にもとづいて二つのフィルタを構築した。

空間フィルタは少数マイクロホンによるマイクロホンアレイで作られており、音声認識、あるいは、補聴システムのための前処理装置として有効に働くことが期待されている。また、周波数フィルタはマイクロホン 1 本での入力が想定されており、雑音が混在した音声からの音声特徴抽出に有効に働くことが期待される。

2.2.2.2 方法

本報告では、音声強調（雑音抑圧）の方法は聴覚生理および聴覚心理の知見を用いることにより構築されている。フィルタ構築のために選ばれた知見は、キャンセレーションの概念である。元々のキャンセレーションは、図 1 に示す神経回路により、周期が T である周期波形を減衰させる働きであり、Durlach [5], Culling & Summerfield [6]は binaural masking level difference (BMLD) をモデル化するために、また de Cheveigné [7]は、基本周波数推定法のモデル化のために用いている。本報告では、空間フィルタのために ITD を遅延 T とし、また、周波数フィルタのために基本周期を遅延 T として用いる。そして、図 2 に示す工学的な回路構成によりキャンセレーションを実現する。

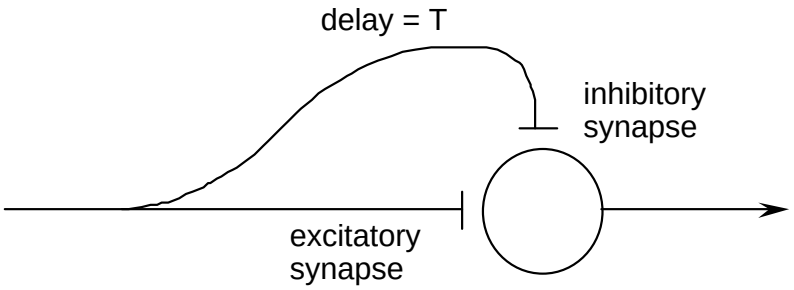

Fig. 1. Basic concept of cancellation model.

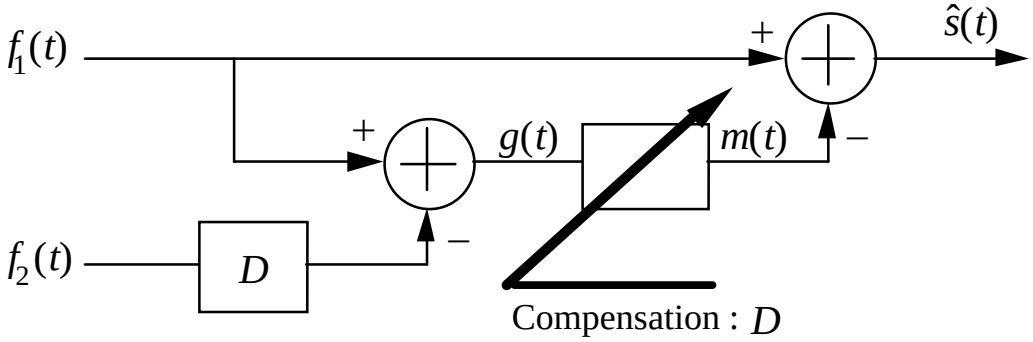

Fig. 2. A circuit of cancellation model.

2.2.2.3 空間フィルタによる音声強調（収録論文[1][9]）

機械による音声認識においては、雑音等で汚れていないきれいな音声では高い認識率を示すものの、周囲雑音が存在する場合には認識率の低下は免れない。このため、雑音環境にロバストな音声認識システムが望まれている。一つの解決法として、音声強調（雑音抑圧）システムを音声認識システムの前処理に用いることが考えられており、雑音抑圧用アダプティブフィルタ[10]とか、大規模マイクロホンアレイ[11]が提案されてきた。しかし、これらの手法は、高性能な DSP を必要としたり、たくさんのマイクロホンを必要とするため、ハンディホンとか音声制御のナビゲーションシステムには不向きである。このような応用には、小規模の前処理システムが必要である。

本報告では、マイクロホン対を用いて信号音以外のある一方向の時間・周波数が局在した雑音を推定し、推定した雑音を引きさることによって信号音を浮かび上がらせる手法を提案する。これと同様の考え方を基にした方法として Grifiths-Jim 型のビームフォーマ[12]があるが、適応フィルタの収束が遅いため、到来方向が変化する雑音とか突発的な雑音には対処できない。また、雑音と信号音は無相関であることを仮定しているため、残響などで相関が存在すれば、信号音が歪むなどの問題がある。一方本手法は、雑音をモデル化しできる限り解析的に抽出することによって、この問題を回避し、雑音を取り去っている。

2.2.2.3.1 定式化

本手法では、等間隔に並んだ 3 本の無指向性マイクロホンによりマイクロホンアレイを構成する。そして、短時間区間ごとに中央のマイクロホン位置での雑音を推定する。雑音抑圧は、推定された雑音を中央マイクロホンで受音した信号から引き去ることによって達成される。

A. 雑音の推定

雑音は、左右一対のマイクロホン（主対）、あるいは中央と左右どちらかのマイクロホンとの対（副対）で得られた信号によって推定される。今、図 3 のように、音声信号がある方向から到来し、雑音は音声以外の方向から到来するものとする。そして、音声信号$s(t)$主対間の時間差が2ζであり、最も大きい雑音$n(t)$の主対間の時間差が2δであるとする。この時、それぞれのマイクロホンで受音される信号は

$$\text{left mic.}: \quad l(t) = s(t-\zeta) + n(t-\delta) \tag{1}$$

$$\text{center mic.}: \quad c(t) = s(t) + n(t) \tag{2}$$

$$\text{right mic.}: \quad r(t) = s(t+\zeta) + n(t+\delta) \tag{3}$$

となる。

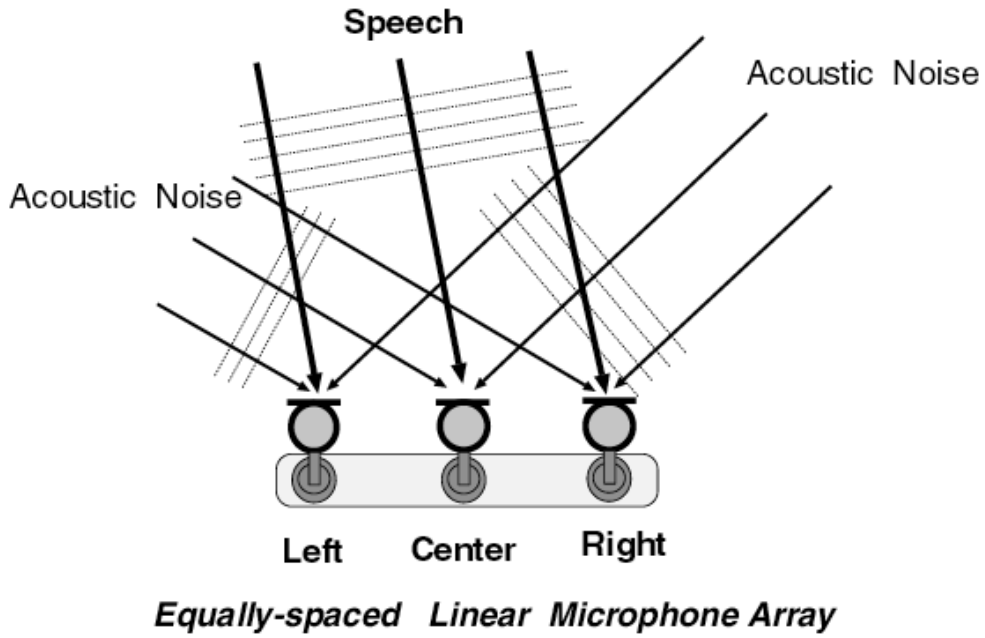

Fig. 3. Relationship between a microphone array and acoustic signals.

説明を簡単にするために$\zeta = 0$として、受音信号の短時間フーリエ変換（STFT）を計算すれば、

$$L(\omega) = S(\omega) + N(\omega)e^{-j\omega\delta} \tag{4}$$

$$C(\omega) = S(\omega) + N(\omega) \tag{5}$$

$$L(\omega) = S(\omega) + N(\omega)e^{j\omega\delta} \quad , \tag{6}$$

となる。ここで、$S(\omega)$は音声信号$s(t)$の STFT、$N(\omega)$は最大雑音$n(t)$の STFT である。

左右マイクロホンでの受音信号$l(t)$と$r(t)$を時間方向に$\pm\tau$シフトし、次のように信号$g_{lr}(t)$を作る。なお、τは任意の$\tau \neq 0$である定数である。

$$g_{lr}(t) = \frac{\{l(t+\tau) - l(t-\tau)\} - \{r(t+\tau) - r(t-\tau)\}}{4}, \tag{7}$$

信号$g_{lr}(t)$は時間領域のビームフォーマーであり、その STFT である$G_{lr}(\omega)$は、

$$G_{lr}(\omega) = N(\omega)\sin\omega\delta\sin\omega\tau. \qquad (8)$$

となる。式(8)から明らかなように$G_{lr}(\omega)$は$S(\omega)$の成分を含んではいない、すなわち、$S(\omega)$はキャンセルされている。

式(8)中の値δは、最大雑音の方向に関連する値であり、短時間区間ごとに雑音方向を推定することにより決定される。雑音方向の推定法は後述する。任意の値τをδと等しくとり、式(8)を$\sin^2\omega\delta$で割ることにより、雑音スペクトル$N(\omega)$が推定できる。ただし、$\omega\delta = n\pi,\ n$: integerである場合には計算できないため、主対の代わりに副対でビームフォーマーを構成する。

$$g_{cr}(t) = \frac{\{c(t+\tau_2) - c(t-\tau_2)\} - \{r(t+\tau_2) - r(t-\tau_2)\}}{4}, \qquad (9)$$

$g_{cr}(t)$の STFT は

$$G_{cr}(\omega) = N(\omega)e^{j\omega\frac{\delta}{2}}\sin\omega\frac{\delta}{2}\sin\omega\tau_2 \quad . \qquad (10)$$

となる。

$\tau_2 = \delta/2$とすれば、式(8), (10)を用いて最大雑音$n(t)$のスペクトラムは

$$\hat{N}(\omega) = \begin{cases} G_{lr}(\omega)/\sin^2\omega\delta, & \sin^2\omega\delta > \varepsilon_1 \\ G_{cr}(\omega)\Big/ e^{j\omega\frac{\delta}{2}}\sin^2\omega\frac{\delta}{2}, & \sin^2\omega\delta \le \varepsilon_1 \ \text{ and } \ \sin^2\omega\frac{\delta}{2} > \varepsilon_2, \\ G_{lr}(\omega)/\varepsilon_2{}^2, & \sin^2\omega\frac{\delta}{2} \le \varepsilon_2 \end{cases} \qquad (11)$$

として求まる。なお、ε_1，ε_2はある小さい値である。

B. 雑音方向の推定

本手法では、雑音方向は次に示す方法によってフレーム毎に自動的に推定される。今、$l(t),\ c(t),\ r(t)$に対して式(9)を用いれば、目的信号を除去された二つの信号が得られる。

$$g_{lr}(t) = \frac{\{l(t+\tau_2) - l(t-\tau_2)\} - \{c(t+\tau_2) - c(t-\tau_2)\}}{4}$$

$$g_{cr}(t) = \frac{\{c(t+\tau_2) - c(t-\tau_2)\} - \{r(t+\tau_2) - r(t-\tau_2)\}}{4}, \qquad (12)$$

なお、τ_2は 0 でない任意の定数である。また、これらの信号の STFT は、式(10)から

$$G_{lc}(\omega) = N(\omega)e^{-j\omega\frac{\delta}{2}}\sin\omega\frac{\delta}{2}\sin\omega\tau_2$$

$$G_{cr}(\omega) = N(\omega)e^{j\omega\frac{\delta}{2}}\sin\omega\frac{\delta}{2}\sin\omega\tau_2 \qquad (13)$$

となる。式(13)では、目的信号の成分が含まれていないため、雑音信号の方向推定に影響を与え

ない。よって、方向推定は雑音成分のみから

$$d(t) = \mathrm{IFFT}\left[\frac{G_{lc}(\omega)G_{cr}{}^{*}(\omega)}{|G_{lc}(\omega)||G_{cr}(\omega)|}\right], \tag{14}$$

$$\delta = \arg\max_{t}[d(t)] \tag{15}$$

として得られる。式(15)で得られたδは、主対に到達する最大雑音の時間差の半分である。

C. 信号の強調

目的信号の強調は、雑音のスペクトル$\hat{N}(\omega)$を推定した後、中央のマイクロホンで受音した信号から引き去ることによって成される。すなわち、強調された信号$\hat{s}(t)$は

$$\hat{s}(t) = c(t) - \mathrm{IFFT}\left[\hat{N}(\omega)\right], \text{ or}$$

$$\hat{s}(t) = \mathrm{IFFT}\left[C(\omega) - \hat{N}(\omega)\right], \tag{16}$$

のように求まる。これは、波形サブトラクション（Wave Subtraction: WS）の方法である。

図 4 に本手法のブロック図を示す。

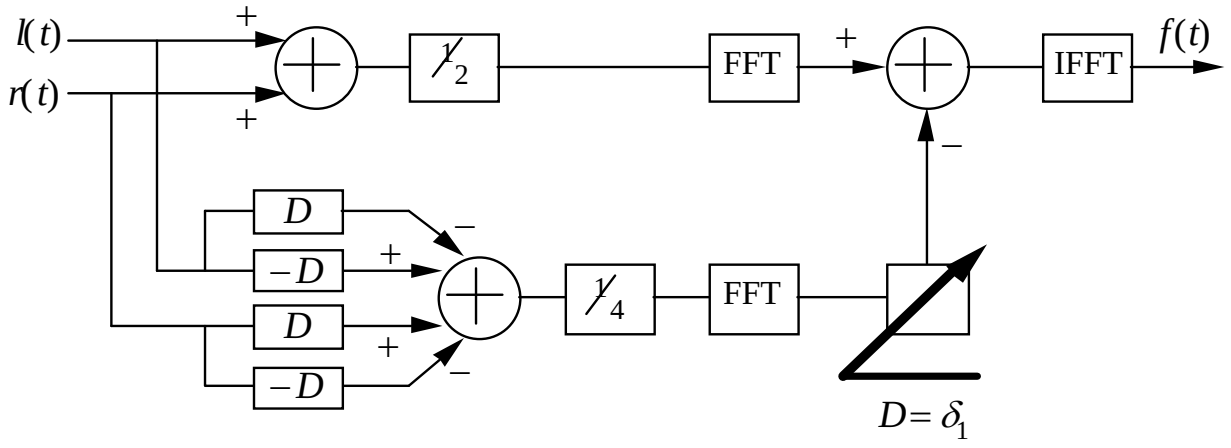

Fig. 4. Block diagram of the proposed method. Delay D is set to half of the ITD, δ.

本手法の目的の一つは、音声認識システムの前処理方式としての雑音抑圧であった。このため、波形そのものではなく、振幅スペクトルのみを対象として音声強調を行なうことも可能である。すなわち、音声強調にスペクトルサブトラクション（Spectral Subtraction: SS）を用いることも可能である。なお、式(11)による雑音スペクトルの近似、あるいは推定ミスのため、雑音スペクトルが入力信号のスペクトルよりも大きくなる可能性があるので、SS には次に示す非線形 SS を用いる。

$$\left|\hat{S}(\omega)\right| = \begin{cases} |C(\omega)| - \alpha \cdot \left|\hat{N}(\omega)\right|, & |C(\omega)| \geq \alpha \cdot \left|\hat{N}(\omega)\right| \\ \beta |C(\omega)|, & \text{otherwise} \end{cases} \quad (17)$$

ここで、αはサブトラクション係数、βはフロアーリング係数である。この式を用いることにより、雑音によって生じた振幅スペクトルの歪みを少なくすることができる。

一般の SS は、雑音スペクトルを無音区間の平均として求めるため、突発雑音などの非定常雑音を抑圧することは困難である。一方本手法は、時々刻々雑音スペクトルを推定するため、非定常雑音に対処可能である。

2.2.2.3.2 評価

A. 音データ

雑音付加音声データとして、2種類の波形データを用意した。両データとも、48 kHz サンプリングの 16 bit データである。

(1) 音データA

音声データ A-I は、ATR 音声データ中の母音/a/に 2 つの狭帯域雑音を付加したものである。図 5 に、原音声(a)、雑音付加音声(b)のスペクトラムを示す。2 つの狭帯域雑音は、中心周波数が 1500 Hz と 2500Hz、帯域幅が 200 Hz であり、両方とも 50 ms 長で音声中に存在する。そして、音声方向が正中、雑音方向が右 30 度となるように計算機上で合成した。

音声データ A-II は、ATR 音声データ中の文音声に広帯域雑音を付加したものである（図 6(b)）。雑音は、中心周波数 1 kHz、帯域幅 1.6 kHz の広帯域雑音が断続的に右 90 度から左 90 度まで移動したとして合成したものである。

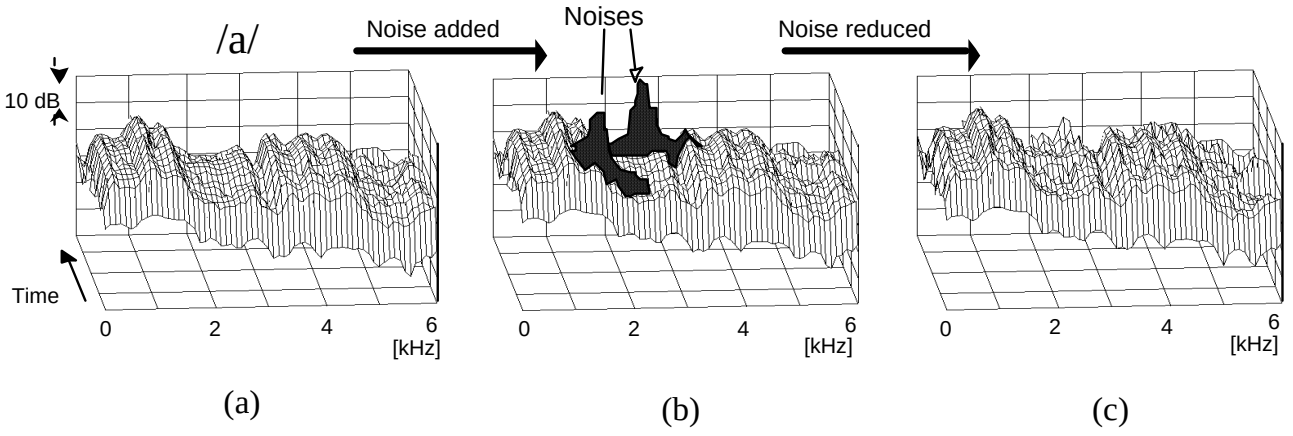

Fig. 5. Simulated results using Sound data A-(I). (a) original noise-free speech wave (vowel /a/), (b)

noise-added speech wave, (c) noise-reduced speech wave.

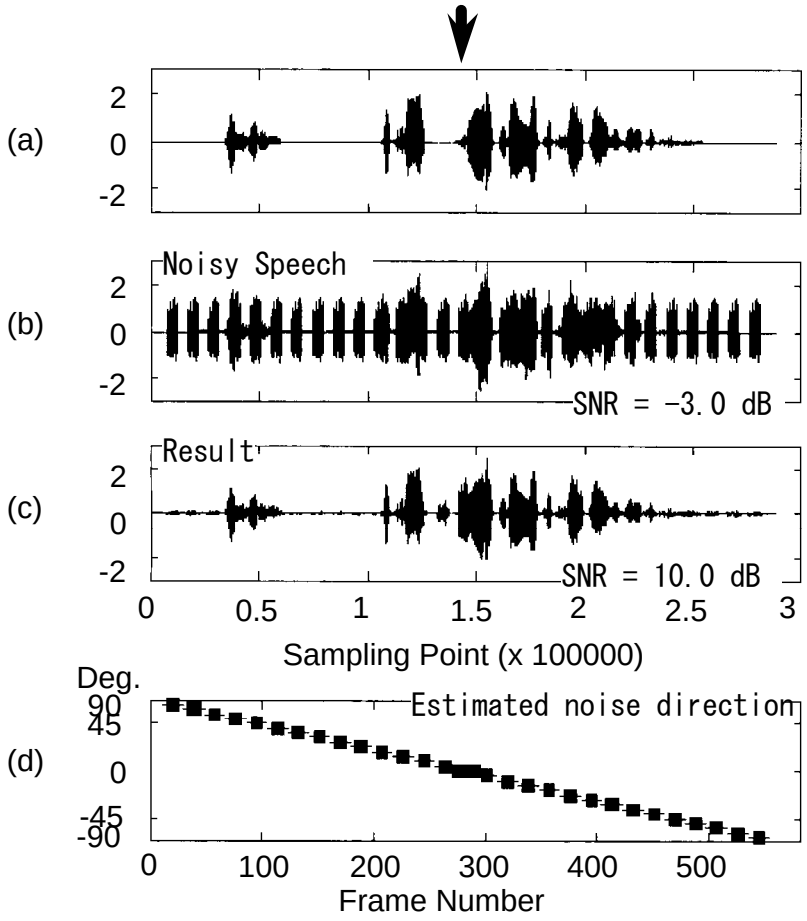

Fig. 6. Simulated results using Sound data A-(II). (a) original noise-free speech wave (ATR, mhtsc101), (b) noise-added speech wave, (c) noise-reduced speech wave, and (d) estimated noise direction.

(2) 音データB

音声データ B は、防音室（残響時間：約 50 ms）内で 2 つのスピーカから放射された音を 3 m 離れたマイクロホンで収録した音データである。音声用スピーカは正面、雑音用スピーカは右 30度に配置した。雑音は 125 Hz から 6 kHz までの周波数成分を含む音であり、SN 比が-10, 0, 10 dB となるように音量を調整した。図 7 に、原音声（ATR 音声データベース、/bunri/）(a)と雑音付加音声(b)を示す。

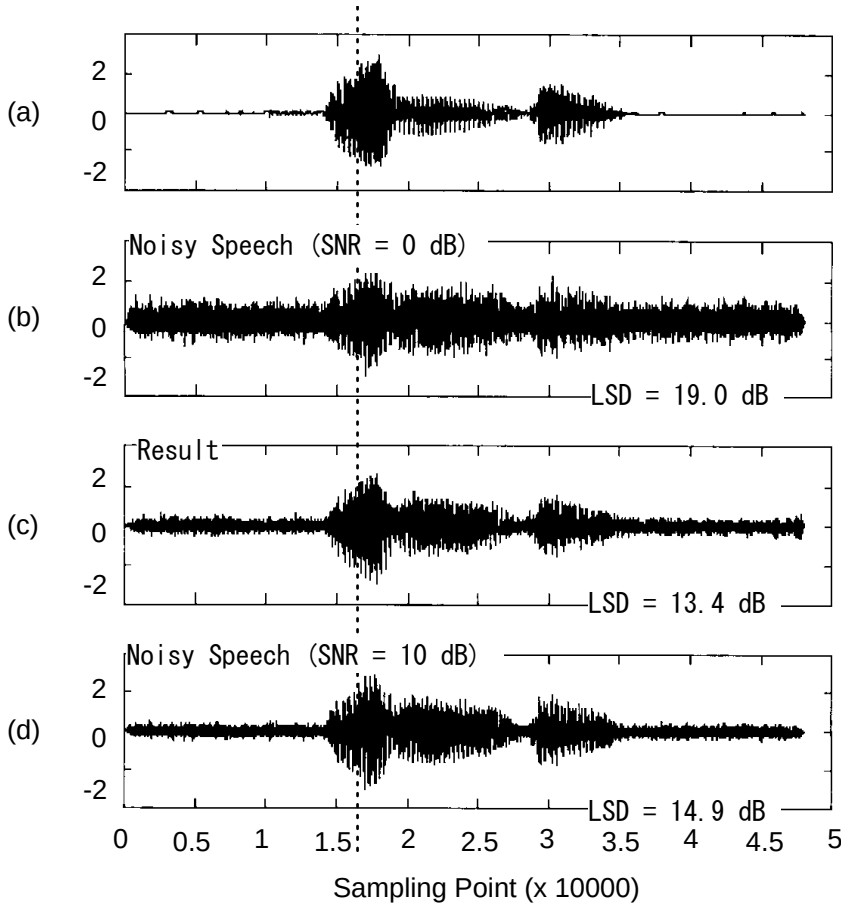

Fig. 7. Simulated results for Sound data B, (a) noise-free speech wave presented by a speaker (ATR, mht14348 /bunri/), (b) noise-added speech wave: SNR = 0 dB, (c) noise-reduced speech wave, and (d) noise-added speech wave: SNR = 10 dB.

B. シミュレーション条件

(1) 波形サブトラクション（WS）

音データ A におけるフレーム長とフレーム周期は 21.3 ms および 10.7 ms であり、音データ B では 85.3 ms および 42.7 ms である。窓関数は、方向推定にハミング窓、雑音抑圧に三角窓を用いている。

(2) スペクトルサブトラクション（SS）

SS による雑音抑圧は、次の条件で行なった。フレーム長 5.3 ms、フレーム周期 2.7 ms、窓関数はハミング窓、ε_1 および ε_2 は 0.6 と 0.2、SS のためのパラメータ α, β は 1 および 0.001。

(3) 評価尺度

評価のために、音データ A 用として SN 比（signal-to-noise ratio: SNR）を用いた。

$$SNR = 10\log_{10}\frac{\sum_n s^2(t_n)}{\sum_n \{s(t_n)-\tilde{s}(t_n)\}^2} \quad (dB), \tag{18}$$

なお、$s(t_n)$は原波形、$\tilde{s}(t_n)$は雑音抑圧波形である。また、音データBのために、次のような対数スペクトラム距離（log-spectrum distance: LSD）を用いた。

$$LSD = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}|S(\omega)| - 20\log_{10}|\tilde{S}(\omega)|\right)^2} \quad (dB), \tag{19}$$

ここで、$20\log_{10}|S(\omega)|$ と $20\log_{10}|\tilde{S}(\omega)|$は、それぞれ原波形と雑音抑圧波形の対数スペクトルであり、フレーム長は1024ポイント、フレーム周期は512ポイント、$W = 6$ kHz である。

C. 結果

(1) 突発雑音：音声データA-I

図5(b)に示した雑音付加音声（雑音部分は黒くしてある）に対して、SSを用いた雑音抑圧をおこなった。結果を図5(c)に示す。図5の(c)と(a)および(b)を比較すれば、本手法によって突発雑音が取り除かれているのは明らかである。

(2) 断続ノイズ：音データA-II

音データA-IIに対して、式(12)-(15)を用いて雑音方向推定を行った結果を図6(d)に示す。雑音方向は右90度から左90度に順序よく移動しており、推定できていることがわかる。推定された方向を用いて、WSにより雑音抑圧を行った結果を図6(c)に示す。図6中の(a), (b), (c)を比較すれば、本手法により雑音成分が取り除かれていることがわかる。式(18)を用いてSN比を計算すると、雑音付加音声で-2.9 dB、雑音抑圧音声で9.6 dBであり、本手法により12 dBの向上が認められた。しかしながら、図6中の矢印の位置では、雑音を抑圧できていない。これは、雑音方向と音声方向が同一となり空間フィルタでは分離ができなかったためである。また、雑音方向が90度に近くなりδが大きくなると、$\omega\delta = n\pi$を満たす各周波数が小さくなり、雑音が低周波数領域に残る。図6(c)の両端に雑音成分が残っているのはこのためである。

(3) 実環境雑音：音データB

音データBに対して、式(12)-(15)を用いて雑音方向推定を行った結果、すべての音区間で右30度が推定できた。この結果をもとに、WSを用いて雑音抑圧を行った。SN比が0 dBのときの結果を図7(c)に示す。図7(b)と(c)を比較すれば、雑音の振幅は減少しており、SN比が10 dBの時の値（図7(d)）とほぼ同じとなっている。

単語/bunri/中の/u/（16000ポイントあたり）の対数スペクトルを図8に示す。原音(a)では、大きなピ

ークとディップが見られるが、雑音付加音声ではスペクトルが平坦となっている。しかし、雑音抑圧音声(c)では、ピークとディップが現れており、雑音を抑圧できていることがわかる。音声認識ではスペクトルの外形を特徴として用いることが多いが、本手法によりピークとディップが再現されることは、本手法が音声認識の前処理として有効であることを示している。

図 9 に式(19)を用いて求めた対数スペクトラム距離を示す。No Process と Processed はそれぞれ雑音付加音声と雑音抑圧音声を表している。図 9 によれば、本手法は、実環境において対数スペクトラム距離を約 5 dB 減少させ、SN 比を 10 dB 以上増加させることがわかる。

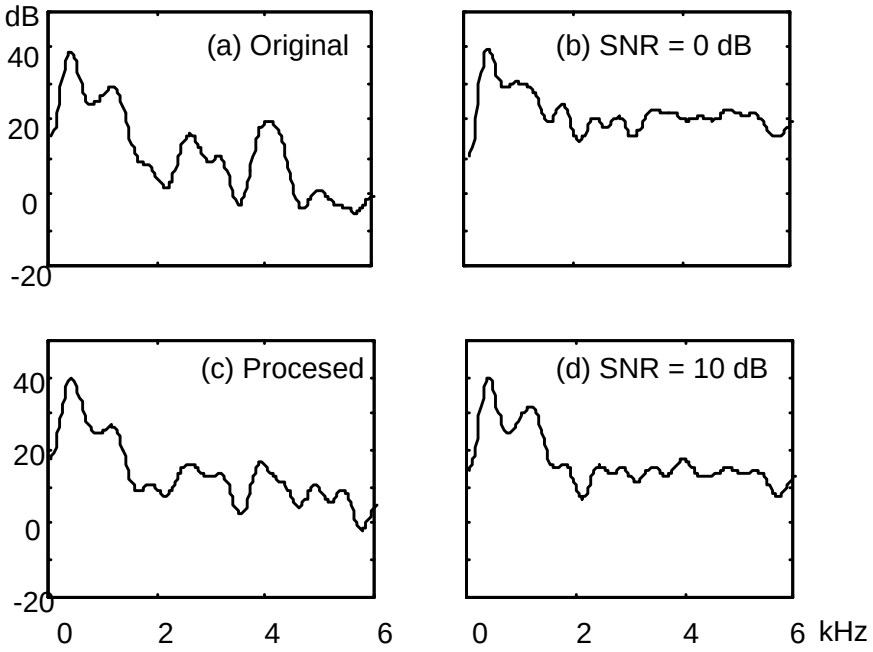

Fig. 8. Log-spectra of vowel /u/ sound at about 16000 points, (a) original sound, (b) noise-added speech wave: SNR = 0 dB, (c) noise-reduced speech wave, and (d) noise-added speech wave: SNR = 10 dB.

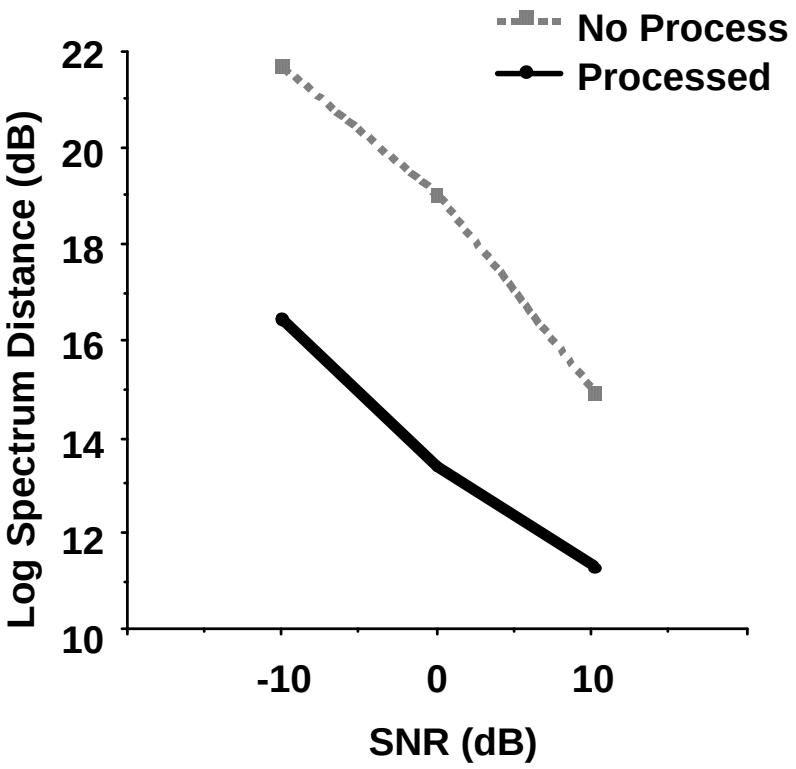

Fig. 9. Mean log-spectrum distances.

D. 音声認識のための前処理（収録論文[10]）

本手法が、雑音付加音声の音声認識に対する前処理に有効であるかどうかを検証した。このために、音声認識法として 12 次の MFCC を特徴量とする話者依存型 HMM を用いた。HMM の学習には ATR 音声データベース中の単語 1048 個を用い、認識には同一話者の音韻バランス単語 216 語を用いて、音韻認識を行った。雑音がない場合の認識率は 84.6%であった。認識用音声には広帯域雑音（125Hz-6000Hz、右 30 度から到来）をさまざまな SN 比で付加した。雑音抑圧には、本手法の他に、典型的な手法として3チャンネルの delay-and-sumビームフォーマー[7]を用いた。なお、delay-and-sum ビームフォーマーには音声方向（正中）をあらかじめ教えてあり、本手法は方向を自動推定している。

音声認識の結果を図 10 に示す。図 10 は、雑音がない場合の音声認識率から何ポイント減少したかを表している。図から明らかなように、本手法では SN 比の減少による認識誤りの増加が抑えられている。一方、典型的な手法である delay-and-sum ビームフォーマーでは、認識誤りを 5 ポイントほどしか減少できていない。SN 比が小さい場合には、本手法との差が顕著である。このことからも、雑音環境での本手法の有効性が示された。

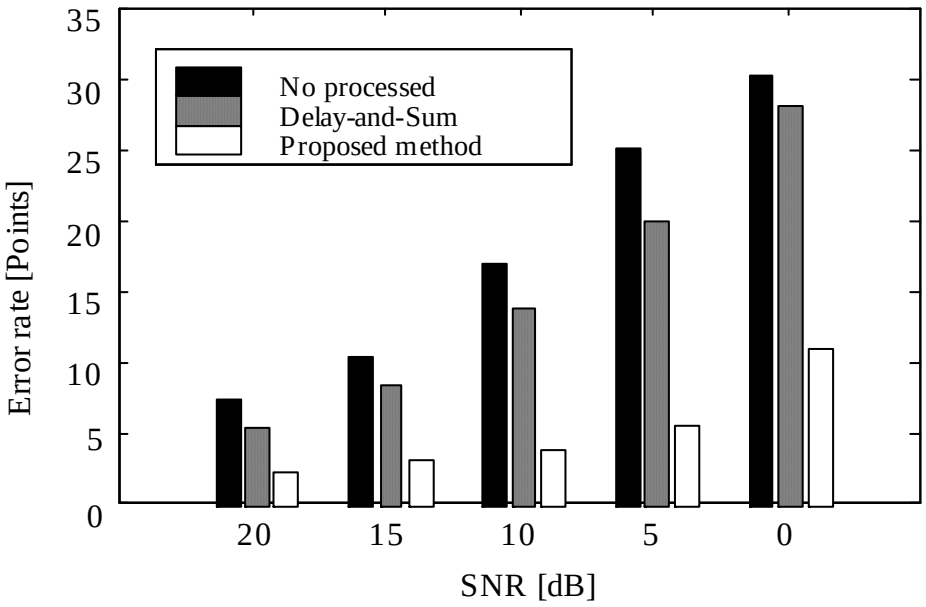

Fig. 10. Phoneme error rates. The three bars correspond to the phoneme error rates of noise-added speech (blackened bar), noise-reduced speech by the optimized delay-and-sum beamformer (grayish bar), and noise-reduced speech by the proposed method (whitened bar).

E. マスキング特性を考慮した低品質音声に対する歪み評価尺度（収録論文[2]）

近年の音声情報処理技術の実用化に伴い、雑音抑圧の需要はますます高まっているが、雑音抑圧に関する研究は、音声認識をターゲットにしたものが多く、雑音により歪んだ音声の聴覚印象の回復を目指した研究は少ない。そこで、本手法を用いて聴感上の印象を向上させることを試みている。

雑音抑圧を行うことによる歪み低減の程度を定量的に評価するためには、客観的な評価尺度が必要となる。このため、同時マスキング、継時マスキングなどの聴覚特性を考慮した歪み評価尺度（Auditory-Oriented Spectral Distortion: ASD）を提案した。ASD により、雑音付加音声の客観的評価が可能となり、補聴器への雑音抑圧法の適用とその客観的評価への道も開けた。

2.2.2.4 周波数フィルタリングによる音声強調（収録論文[11]）

音声分析合成、音源分離など音声情報処理では目的の音声から基本周波数を抽出することが重要である。例えば音声分析合成では基本周波数は音の高さなどを制御する要素であり、音源分離では音源の違いを特徴づける要素として用いられる。しかし、音源分離への応用のように周囲に雑音のある環境において基本周波数を抽出することは、雑音の影響によって目的音声が歪んでしまうために困難である。音声情報処理技術、特に音源分離への実用には、雑音環境にお

いて精度の高い基本周波数を抽出できなければならない。

基本周波数の抽出は古くからの問題であり、これまでにも音声波形の自己相関による推定法やケプストラムから音声スペクトルの包絡特性と微細構造を分離して微細構造を利用する推定法など、様々な方法が考案されている。河原らはフィルタの中心周波数からフィルタ出力の瞬時周波数への写像の不動点を用いて基本周波数を推定する方法(TEMPO2)を提案している[13]。TEMPO2 はクリーンな音声から基本周波数をケプストラム法などよりも高精度で推定することができる。しかし、入力としてクリーン音声のみを考慮しているため、入力音声に雑音が付加された場合はクリーン音声の場合と比べて推定精度が大幅に低下する。そのため、実環境における基本周波数推定法として用いるには耐雑音性の点で充分であるとはいえない。

一方、雑音の存在を考慮した基本周波数推定法として、鵜木らは定 Q gammatone filterbank による瞬時振幅に対して周波数軸上の Comb filtering を行ない、その通過量が最大になるような Comb filter の中心周波数を基本周波数とする推定法を提案した[14]。この手法は耐雑音性が高く、信号対雑音比(SNR) 5 dB 程度の雑音に対しても基本周波数を推定することができるが、クリーンな音声に対する精度は TEMPO2 よりも劣っており、音声情報処理の実用のためにはより精度の高い推定が求められる。

本報告では、連続音声における基本周波数推定のロバスト性を高めるために瞬時振幅の周期性と調波性からそれぞれに対応する基本周波数を推定し、この推定値を用いて音声強調を行い、そして、音声強調を行った音声波形からTEMPO2を用いて信頼性の高い基本周波数推定を行なう方法を提案する。

2.2.2.4.1 基本周波数推定の概要

はじめに、目的音声$s(t)$と雑音$n(t)$が混ざった混合音声$x(t)$に対して、雑音に対してロバストな基本周波数推定法によって、ある程度の精度の目的音声の基本周波数$\tilde{F}0$を推定する(図 11(a))。雑音にロバストな推定法として、瞬時振幅の周期性・調波性を利用した基本周波数推定法を用いる。

次にこの基本周波数に合わせたくし形フィルタによって混合音声から目的音声の調波成分を取り除き推定雑音$\tilde{n}(t)$を取り出す（図 11(c)）。このとき、くし形フィルタの帯域幅を調節することにより、目的音声の調波成分を取り除かないようにする。帯域幅が可変なくし形フィルタの定式化については次節で述べる。$\tilde{n}(t)$を$x(t)$から引き去ると推定音声$\tilde{s}(t)$が得られ、$\tilde{s}(t)$に対して、高精度の基本周波数推定を行なう（図 11(d)）。ここでは、耐雑音性は低いが雑音が小さければ高精度に推定できる河原らの TEMPO2 を用いる。これにより、雑音のある環境において目的音声の高精度の基本周波数推定が可能となる。

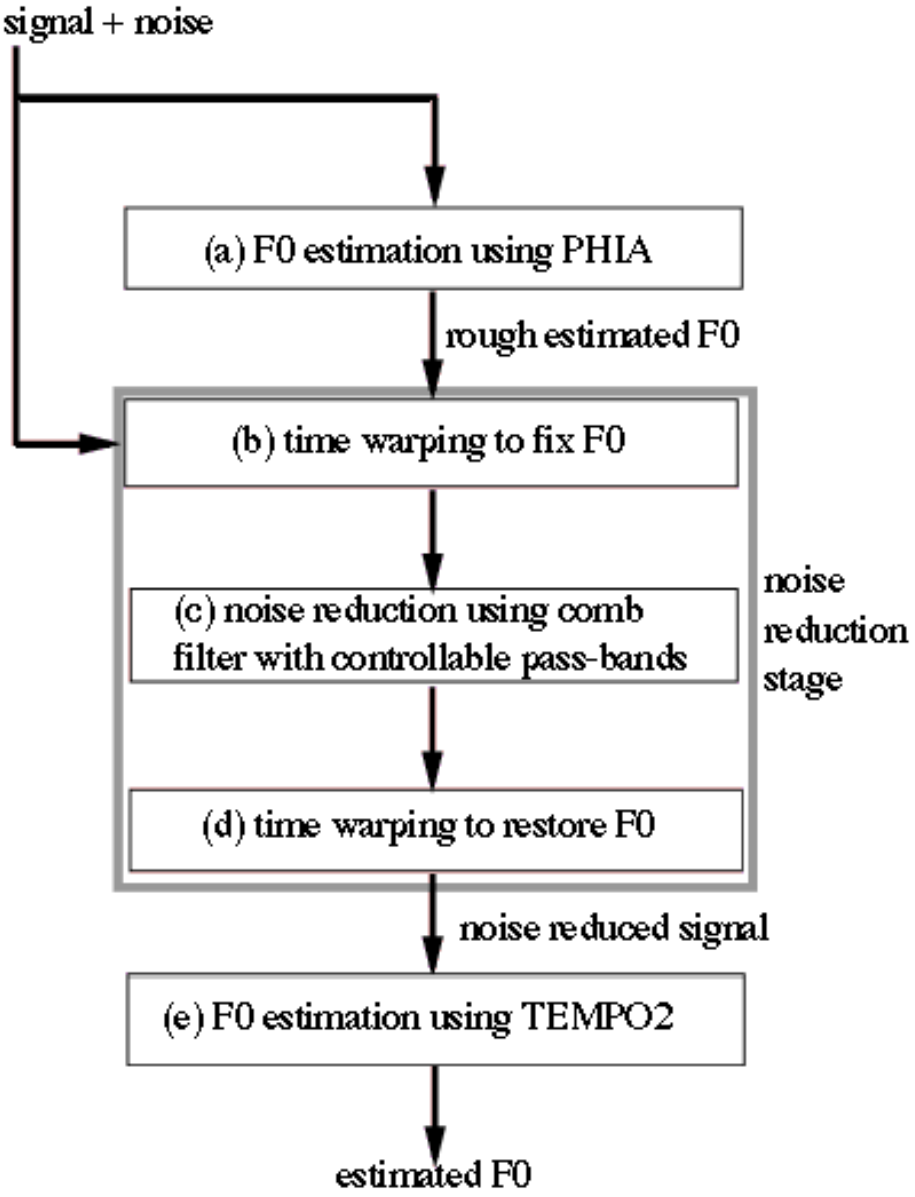

Fig. 11. Flowchart of F0 estimation method in noisy environment.

2.2.2.4.2 瞬時振幅の周期性・調波性を利用した基本周波数推定法

音声波形をフィルタバンクで展開して得られる瞬時振幅には、基本周波数に対応する情報として、時間周波数領域において時間方向に周期性が現われ、周波数方向に調波性が現われる。この周期性と調波性から基本周波数を検出し統合することで、単一の情報から得られる基本周波数よりも信頼性の高い基本周波数を推定できる。この処理の流れを図 12 に示す。

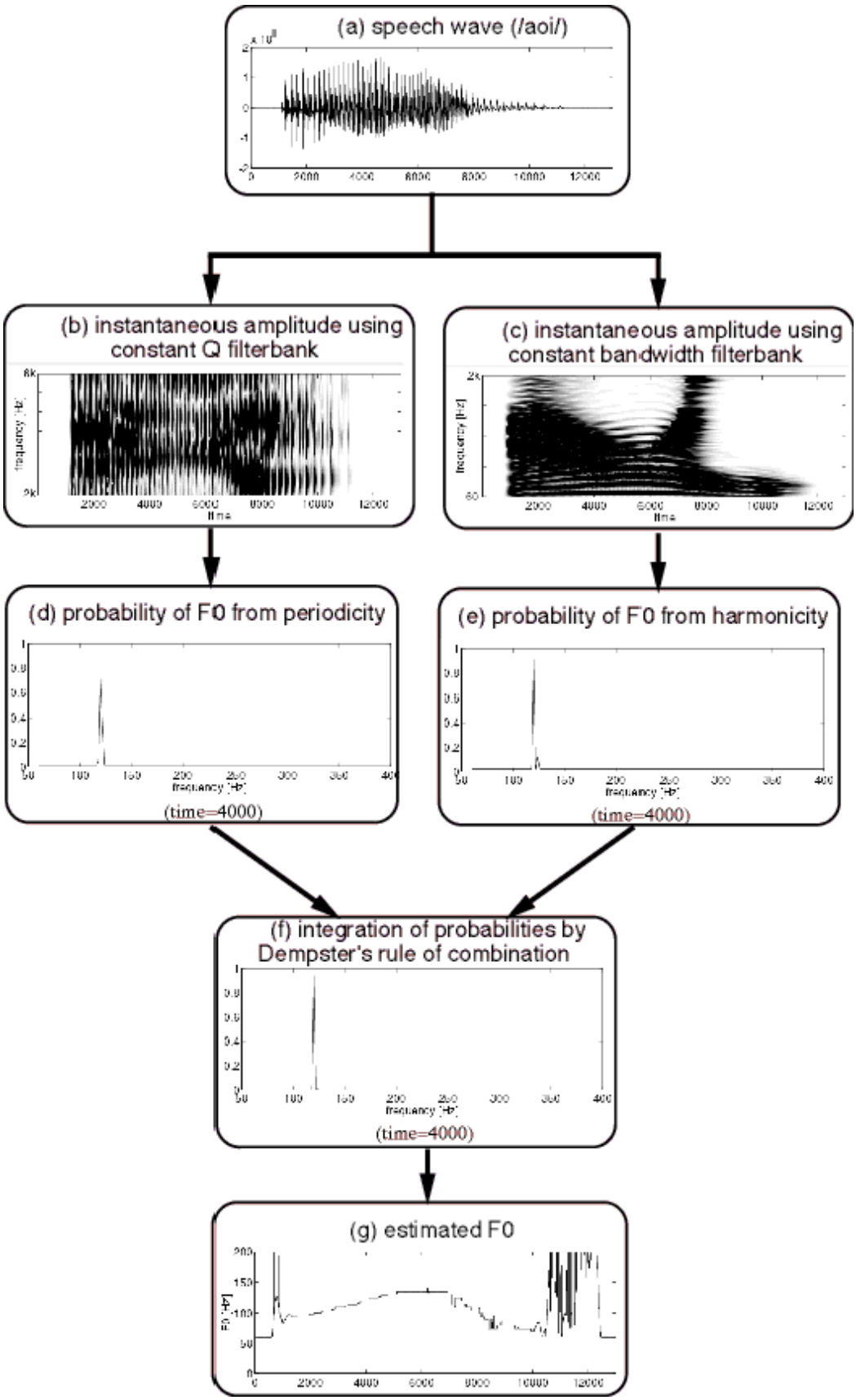

Fig. 12 F0 estimation based on periodicity and harmonicity of instantaneous amplitude (PHIA)

A. 瞬時振幅に現れる周期性・調波性

まず、音声信号に対して定 Q フィルタバンクと定帯域フィルタバンクを用いて瞬時振幅を求める。これは、周期性は定 Q フィルタバンクで展開したときに高周波数帯域に現われ、調波性は定帯域フィルタバンクで展開したときに明確に現われるためである。一例として、図 12(a)の音声波形に対して、図 12(b)に示すように定Qフィルタバンクによる瞬時振幅では時間方向に基本周波数の逆数(1/F0)に対応する振幅変動として周期性が現れ、図 12(c)に示すように定帯域フィルタバンクによる瞬時振幅では周波数方向の振幅変動として調波性が現れていることがわかる。

B. 周期性・調波性の検出

周期性と調波性から基本周波数を検出するために、定 Q フィルタバンクによる瞬時振幅では時間方向への、定帯域フィルタバンクでは周波数方向への自己相関処理を行なう。このとき、時間方向ではチャネル番号ごとに自己相関処理を行ない、各時間における複数の F0 候補を抽出することができる。周波数方向では自己相関のラグ窓長を変えることで同様に複数の F0 候補を抽出する。横軸を周波数、縦軸を頻度とした F0 候補のヒストグラムを周期性・調波性からそれぞれ得られた基本周波数の信頼度と考える。図 12(d)はサンプリング点=4000 における周期性からの信頼度、図 12(e)は調波性からの信頼度を示している。

C. 周期性・調波性による基本周波数の信頼度の統合

周期性と調波性からそれぞれ基本周波数の信頼度が得られるが、これらは独立に評価されたものであるので、信頼度の高さを単純に比較して最も高いものを基本周波数として選ぶことは適切ではない。そこで、Dempster の結合規則によって 2 つの信頼度を統合することにより新たな信頼度を求める。Dempster の結合規則では、m_1, m_2 を独立な証拠に基づく基本確率とし、A_{1i}, A_{2j} $(i, j = 1,2,\ldots)$を各集合要素とするとき、新たな基本確率は、

$$m(A_k) = \frac{\sum_{A_1 \cap A_2 = A_k} m_1(A_{1i}) m_2(A_{2j})}{1 - \sum_{A_1 \cap A_2 = \phi} m_1(A_{1i}) m_2(A_{2j})} \tag{20}$$

となる。

ここでは、周期性と調波性から得られた信頼度を基本確率関数m_1, m_2と考え基本周波数(ヒストグラムの bin)を焦点要素A_{1i}, A_{2j}として信頼度を統合する。統合した信頼度mから最も信頼度の高い基本周波数を選ぶ。図 12(f)は図 12(d)(e)を統合した結果を示し、図 12(g)は統合結果の最大値を各サンプリング点ごと求めることによって得られた推定基本周波数である。

2.2.2.4.3 帯域幅可変くし型フィルタによる雑音抑圧

基本周波数に合わせたくし形フィルタを用いて雑音を抑圧する方法は文献[7]などに述べられているが、本研究で用いるくし形フィルタは、初段の推定法における誤差の影響を小さくし、かつ雑音をなるべく抑圧することが可能であるような、帯域幅を制御することができるフィルタが望ましい。そこで、帯域幅可変くし形フィルタとして次のようにくし形フィルタを定式化する。

A. 帯域幅可変くし型フィルタの定式化

目的音$s(t)$を基本周期$T(t)$の調波複合音として、雑音を$n(t)$とすれば、混合音$x(t)$は

$$\begin{aligned} x(t) &= s(t)+n(t) \\ &= \sum_m a_m e^{j(m\omega_0(t)t+\theta_m)} + \sum_k b_k e^{j(\omega_k t+\theta_k)}, \\ \omega_0(t) &= 2\pi/T(t) \end{aligned} \tag{21}$$

となる。時間とともに変動する基本周期$T(t)$を一定値T $(=2\pi/\omega_0)$と仮定し、式(21)を$\pm T$だけ時間軸でずらしてその差$g(t)$を計算すれば、

$$\begin{aligned} g(t) &= \frac{2x(t)-x(t-T)-x(t+T)}{4} \\ &= \sum_k b_k e^{j(\omega_k t+\theta_k)} \sin^2 \frac{\omega_k}{\omega_0}\pi \end{aligned} \tag{22}$$

となる。$n(t)$の短時間フーリエ変換を$N(\omega_k)$とすれば、$g(t)$の短時間フーリエ変換$G(\omega_k)$は

$$G(\omega_k) = N(\omega_k)\sin^2 \frac{\omega_k}{\omega_0}\pi \quad , \tag{23}$$

となる。式(23)から明らかなように$G(\omega_k)$は目的音の成分を含んではいない、すなわち、目的音はキャンセルされている。よって、雑音スペクトル$N(\omega_k)$は

$$\hat{N}(\omega_k) = G(\omega_k) \Big/ \left\{\sin^2(\omega_k/\omega_0)\pi\right\}. \qquad \sin^2(\omega_k/\omega_0)\pi,\ \ \omega_k/\omega_0 \notin \mathbf{Z} \tag{24}$$

と推定される。$\hat{N}(\omega_k)$を逆フーリエ変換した雑音$\hat{n}(t)$を元の混合音$x(t)$から引き去ることにより雑音が除去され、目的音$s(t)$が推定できる。

しかし、このままでは$\omega_k/\omega_0 \in \mathbf{Z}$であるときに式(24)の雑音スペクトル$N(\omega_k)$が無限大となるため、実際の使用にはある値$\varepsilon$を設定し、

$$\hat{N}(\omega_k) = \begin{cases} G(\omega_k) \Big/ \left\{\sin^2(\omega_k/\omega_0)\pi\right\}, & \left|\sin(\omega_k/\omega_0)\pi\right| \geq \varepsilon \\ G(\omega_k), & \left|\sin(\omega_k/\omega_0)\pi\right| < \varepsilon \end{cases} \tag{25}$$

とする。

図 13 に提案法のフローチャートを示す。図中の遅延Dは基本周期Tである。また、図 14 に本

手法の周波数応答（$T=5$ ms）を示す。図 14 から明らかなように、本手法はεによって帯域幅を制御できるくし型フィルタとなっている。本手法では、εは式(20)における信頼度mから推定される。

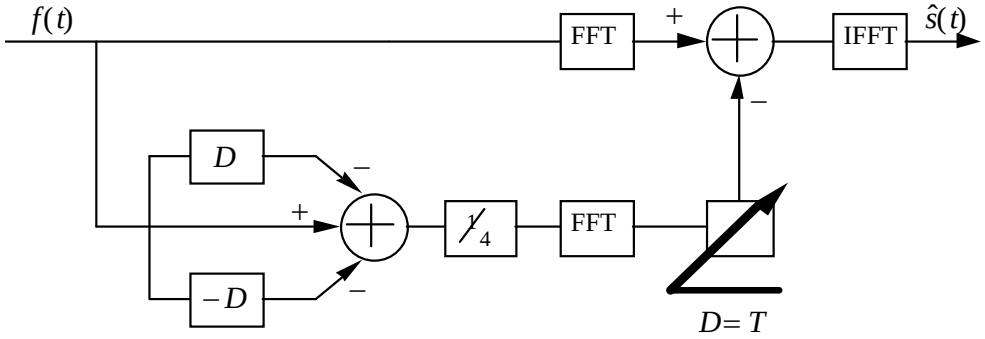

Fig. 13. Block diagram of the proposed frequency filtering.

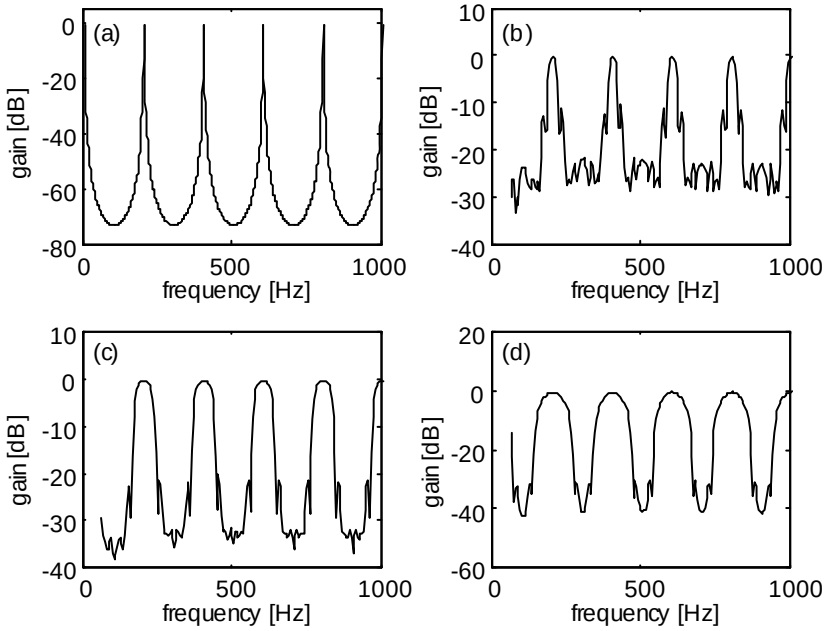

Fig. 14. Frequency response of the model ($T=5$ ms), (a) usual comb filter, (b) $\varepsilon=0.2$, (c) $\varepsilon=0.5$, (d) $\varepsilon=0.8$.

式(22)で基本周期を一定と仮定して計算を行っているが、実音声では基本周期は時間的に変動するため、このままでは推定雑音$\hat{n}(t)$に誤差が生じる。そこで、基本周期を一定にするために音声波形の時間軸での伸縮を行なう。伸縮を行なった後、雑音除去を行い、先に施した波形伸縮操作と逆の操作を行なうことにより、元の基本周波数を持つ音声に戻すことができる。

2.2.2.4.4 推定精度評価実験

A. 実験条件

音声として、音声とEGG(electro glottal graph)波形が同時収録されているデータベース[15]から男女各14名発話の「非常口はどこですか」を48kHzから20kHzにダウンサンプリングして用いた。音声データの例を図 15 に示す。また、雑音として白色雑音とピンク帯域雑音を SNR が 10,5,3,0 dB となるように付加した。なお、SNR は

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_k s^2(k)}{\sum_k n^2(k)}$$

によって計算した。ここで、$s(k)$は信号、$n(k)$は雑音、kは音声区間である。

EGG 波形からTEMPO2によって推定された基本周波数を正解とした。有声区間のうち推定された基本周波数が正解から±5%以内である割合を正解率として評価に用いた。有声区間は音声／EGG データベースに含まれている有声／無声ラベルを参照した。提案法以外に、比較のため、提案法の初段に用いる PHIA のみを用いた場合の結果と、TEMPO2 のみを用いた場合の結果、そしてNollのケプストラム法による結果についても示す。

B. 白色雑音付加音声

推定結果の例として、図15(上)の音声に白色雑音をSNR 3 dBで付加された雑音付加音声の基本周波数推定結果を図 16 に示す。PHIA では SNR 3 dB でも有声区間の多くで基本周波数が推定できている。TEMPO2 では雑音によって推定できない区間が現われており、ケプストラム法はほとんどの区間で推定できない。一方、提案法は、PHIA と同様に雑音にロバストであり、また図15(下)と比較してわかるように、高精度な推定ができている。

白色雑音を付加された音声に対する正解率を図 17 に示す。このグラフは全話者の正解率の平均であり、横軸が SNR、縦軸が正解率となっている。

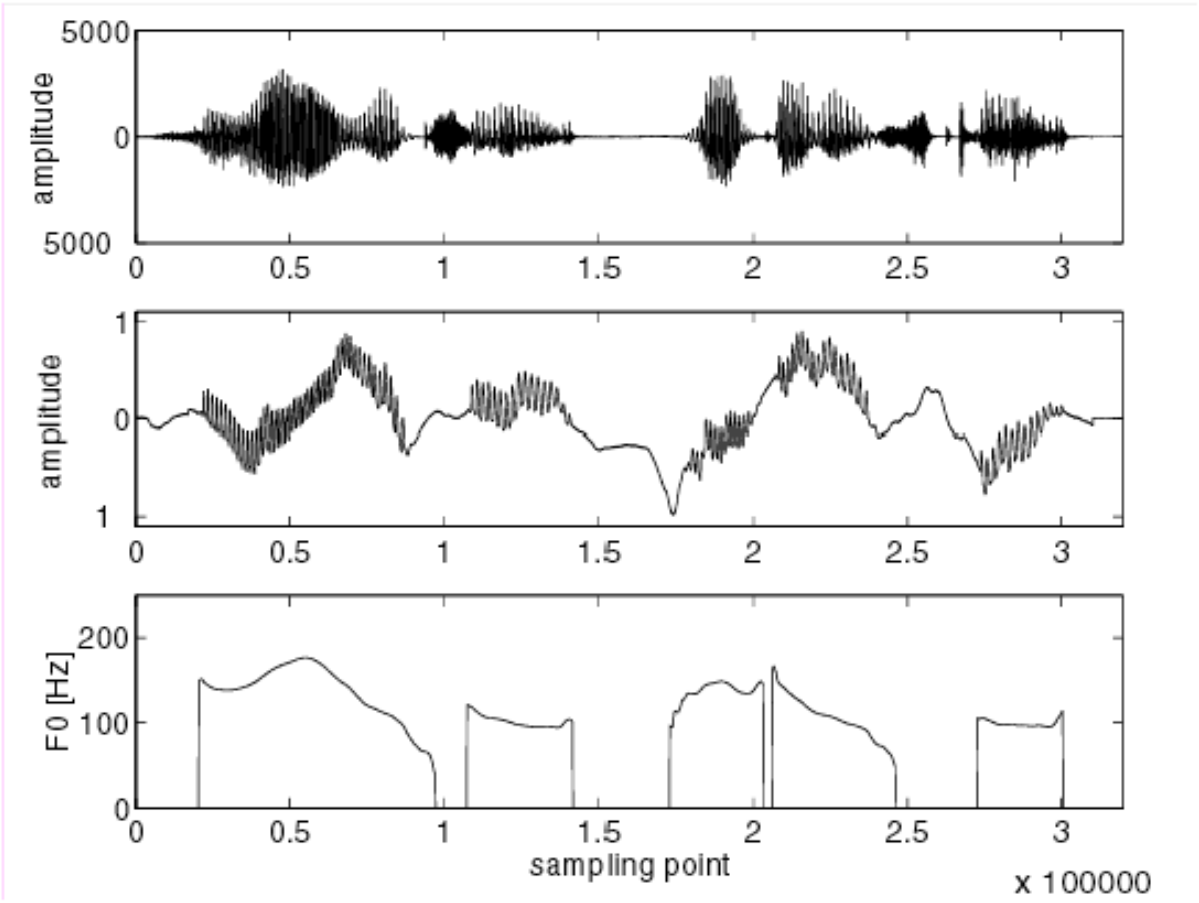

Fig. 15. Speech wave example. Speech wave (upper), EGG wave (middle) and estimated F0 from EGG wave (bottom).

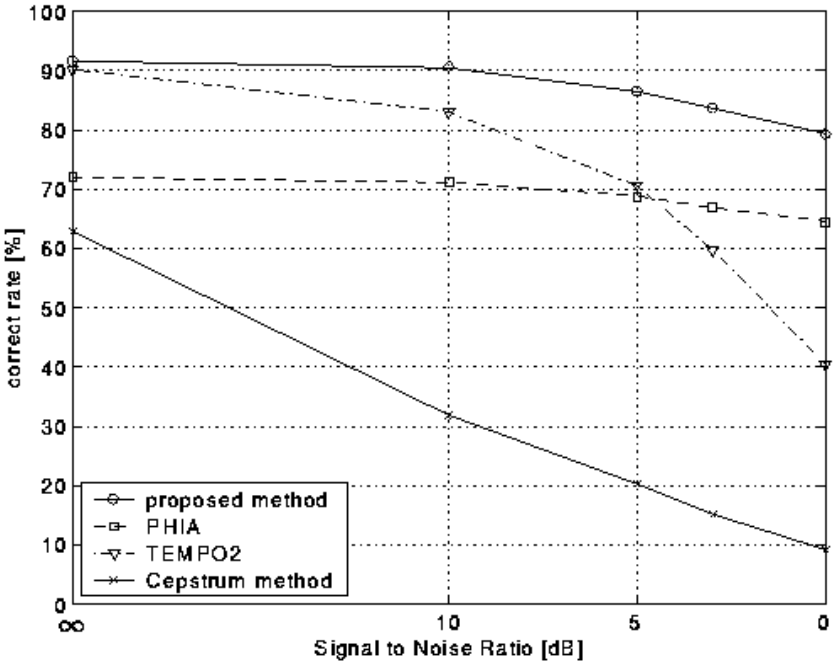

Fig. 17. Correct rates for speech with white noise.

クリーンな音声(SNR=∞)のとき、提案法と TEMPO2 が 90%以上でほぼ同じ正解率であり、PHIA がそれよりも 15%程度低い正解率となっている。すなわち、PHIA 単独では TEMPO2 程の精度は得られないことと、提案法はクリーン音声に対して TEMPO2 のみを用いた場合と同じ高精度な推定が可能であることを示している。ケプストラム法は最も正解率が低かった。

雑音付加音声に対しては、TEMPO2 では雑音が大きくなると正解率が急激に低下し、SNR 0 dB ではクリーン音声の場合より 40%以上低下した。一方、PHIA では、SNR 0 dB でもクリーン音声の場合から 10%程度しか低下しておらず TEMPO2 よりも高い正解率を示していることから、PHIA が高い耐雑音性を有することがわかる。白色雑音では全ての周波数帯域で等しく雑音が加わるため、PHIA は比較的パワーの小さい周期性の情報はあまり利用できないが、調波性の情報を多く利用して基本周波数を推定している。

提案法は PHIA で推定した基本周波数を利用して雑音抑圧を行なうため、PHIA の耐雑音特性に影響を受ける。PHIA の耐雑音性が高いために雑音が大きくなっても提案法の正解率は高く、TEMPO2 のみの場合よりも SNR 0 dB 時の正解率で 30%以上の改善が見られた。

C. ピンク帯域雑音付加音声

ピンク帯域雑音を付加された音声に対する正解率を図 18 に示す。

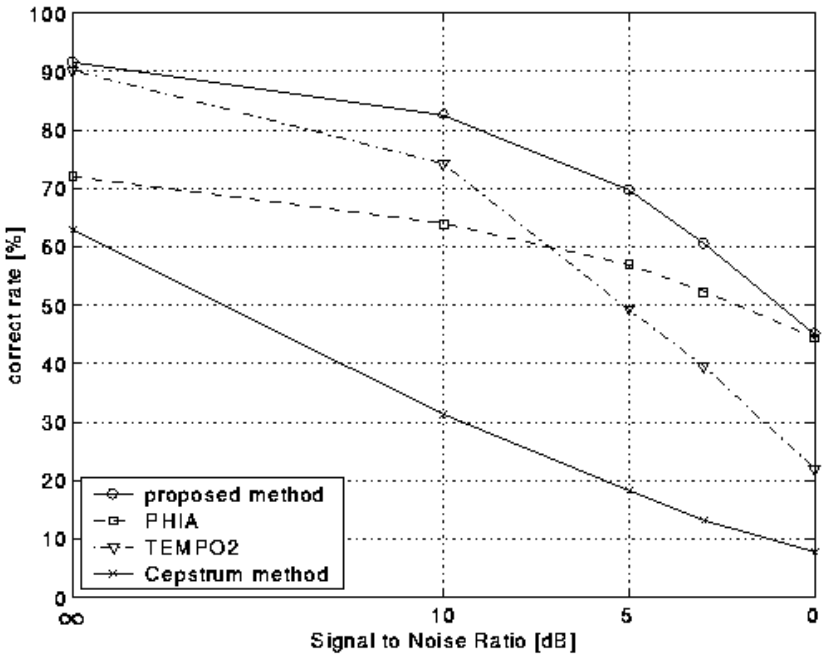

Fig. 18. Correct rates for speech with pink noise.

ケプストラム法は白色雑音の場合と同様、すべての SNR において最低の正解率を示した。TEMPO2、PHIA に関しては白色雑音とほぼ同様の傾向が見られるが、白色雑音の場合よりも雑音の影響を多く受けている。これは、ピンク帯域雑音はパワーが白色雑音よりも低域で大きくなるため、同じ SNR でも基本周波数付近の雑音が白色雑音よりも大きくなることにより、基本波を利用する TEMPO2 により大きな影響を与えているためであると考えられる。PHIA にとってピンク帯域雑音は、低帯域の雑音により調波性の情報が利用しにくくなる上、高帯域にも雑音があるために比較的小さな振幅で現れる周期性の情報も利用が難しくなるような雑音であると考えられ、白色雑音に比べて大きな正解率の低下が見られるが、SNR 0 dB における TEMPO2 よりも 20%以上の正解率を示している。

提案法は、TEMPO2 のみの場合よりも SNR 0 dB のとき正解率で 20%程度改善している。しかし、白色雑音の場合と異なり、雑音が大きくなると正解率が PHIA に近づく。これは、帯域幅可変くし形フィルタでは調波成分と同じ周波数帯域の雑音を除去できないことと、TEMPO2 が低域のパワーの強い雑音に弱いため、TEIPO2 の耐雑音性の低さに影響を受けていることによると考えられる。

2.2.3 音源分離（収録論文[3][4][12]）

本節では、聴覚の情景解析（ASA）にもとづいた聴覚的音分離モデルを提案する。これは、Bregman の発見的規則に関係した制約条件を利用することで、二波形分離問題を解くものである。Bregman は、人間の聴覚系が情景解析を行うために、音響イベントに関係した四つの心理物理的な発見的規則を利用しているということを報告した[9]。仮に、これらの発見的規則を利用して聴覚的な音分離モデルを構築することができれば、ロバストな音声認識システムの前処理だけでなく、様々な信号分離問題へと応用でき、その効果に大きな期待がもてる。

ASA にもとづいた分離モデルは、既にいくつか提案されている。これには主に二つのタイプがある。一つはボトムアップ的なモデル[16]で、もう一つはトップダウン的なモデル[17][18]である。これらのモデルのすべては四つの発見的規則のうちのいくつかを利用しており、音響的な特徴として振幅情報（あるいはパワー）のみを利用している。このため、これらのモデルでは、信号と雑音が周波数領域で重なるような場合、望みの信号を雑音から完全に分離抽出することは困難である。

これに対し、我々は信号と雑音を完全に分離し抽出するためには、振幅だけではなく位相情報も必要であると考え、これらを利用して二波形分離問題の解法に取り組んでいる[3]。この問題は次のように定義されている[3][19]。

はじめに、混合信号 $f(t)$ だけが観測されるものとする。但し、混合信号は二つの信号の和（$f(t) = f_1(t) + f_2(t)$）とする。次に、$f(t)$ はフィルタバンク（K 個のフィルタ）によって周波数分解される。ここで、k 番目のチャンネルの出力を $X_k(t)$ とすれば、これは次式のように表される。

$$X_k(t) = S_k(t)\exp\left(j\omega_k t + j\phi_k(t)\right). \tag{26}$$

また、$f_1(t)$ と $f_2(t)$ に対応する、k 番目の出力を $A_k(t)\exp\left(j\omega_k t + j\theta_{1k}(t)\right)$、$B_k(t)\exp\left(j\omega_k t + j\theta_{2k}(t)\right)$ と仮定すれば、二つの信号の瞬時振幅 $A_k(t)$ と $B_k(t)$ は、次式で表すことができる。

$$A_k(t) = S_k(t)\sin\left(\theta_{2k}(t) - \phi_k(t)\right)\Big/\sin\theta_k t \tag{27}$$

$$B_k(t) = S_k(t)\sin\left(\phi_k(t) - \theta_{1k}(t)\right)\Big/\sin\theta_k t \tag{28}$$

但し、$\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$、$\theta_k(t) \neq n\pi$、 $n \in \mathbf{Z}$ であり、ω_k は k 番目のチャンネルの中心周波数である。また、これらの瞬時位相 $\theta_{1k}(t)$、$\theta_{2k}(t)$ は、次式で表すことができる。

$$\theta_{1k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) - \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) + \cos\phi_k(t)}\right) + \arcsin\left(\frac{A_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right) \tag{29}$$

$$\theta_{2k}(t) = -\arctan\left(\frac{Y_k(t)\cos\phi_k(t) + \sin\phi_k(t)}{Y_k(t)\sin\phi_k(t) - \cos\phi_k(t)}\right) + \arcsin\left(\frac{B_k(t)Y_k(t)}{S_k(t)\sqrt{Y_k(t)^2+1}}\right) \tag{30}$$

但し、$Y_k(t) = \sqrt{\left\{2A_k(t)B_k(t)\right\}^2 - Z_k(t)^2}\Big/Z_k(t)$、$Z_k(t) = S_k(t)^2 - A_k(t)^2 - B_k(t)^2$ である。ここで、$f_1(t)$ と $f_2(t)$ は、すべてのチャンネルにわたって、$A_k(t)$、$\theta_{1k}(t)$ のペア、$B_k(t)$、$\theta_{2k}(t)$ のペアを利用するこ

とで、元の信号に分離、復元可能である。しかし、上記の式をみてわかるように、四つのパラメータ（$A_k(t)$，$B_k(t)$，$\theta_{1k}(t)$，$\theta_{2k}(t)$）は何らかの制約なしに一意な解を求めることができない。すなわち、これは不良設定の逆問題である。

この問題を解くために、我々は、先に、四つの発見的規則に関係した制約条件を利用した基本的な分離モデルを提案した[3]。本節では、実際の音声や雑音を波形レベルで取り扱うことのできる音分離モデルを紹介する。このモデルは、瞬時振幅と瞬時位相、それに基本周波数の連続性の制約を用いている。

2.2.3.1 聴覚的音分離モデル

本節では、目的音 $f_1(t)$ を調波複合音と仮定する。但し、基本周波数を $F_0(t)$ と表すことにする。提案モデルは、$A_k(t)$ と $\theta_{1k}(t)$ の時間微分を制約することで混合信号から望みの信号を分離、抽出する。

提案法は、図 19 に示すように、聴覚的なフィルタバンク、F_0 推定部、分離部、グルーピング部の四部で構成され、表 1 に示すような制約条件を利用する。

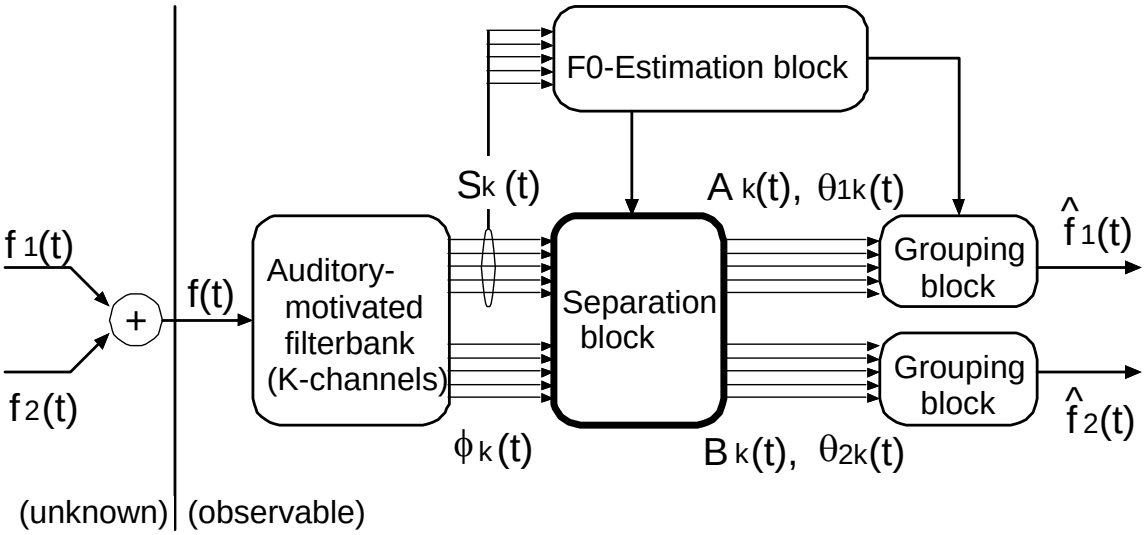

Fig. 19. Auditory sound segregation model.

Table 1. Constraints corresponding to Bregman's psychoacoustical heuristic reguloarities.

Regularity [Bregmann,1993]	Constraint [Unoki,1999]	
(i) common onset/offset	synchronous of onset/offset	$\left\|T_{\mathrm{S}} - T_{k,\mathrm{on}}\right\| \le \Delta T_{\mathrm{S}}, \quad \left\|T_{\mathrm{E}} - T_{k,\mathrm{off}}\right\| \le \Delta T_{\mathrm{E}}$
(ii) gradualness of change (smoothness)	piecewise-differentiable polynomial approximation (spline interpolation)	$dA_k(t)/dt = C_{k,R}(t), \quad d\theta_k(t)/dt = D_{k,R}(t)$ $dF_0(t)/dt = E_{0,R}(t)$ $\sigma_A = \int_{t_a}^{t_b} \left[A_k^{(R+1)}(t)\right]^2 dt \Rightarrow \min$ $\sigma_\theta = \int_{t_a}^{t_b} \left[\theta_k^{(R+1)}(t)\right]^2 dt \Rightarrow \min$
(iii) harmonicity	multiples of the fundamental frequency	$n \times F_0(t), \quad n = 1,2, \ ..., \ N_{F_0}$
(iv) changes occurring in the acoustic event	correlation between the instantaneous amplitudes	$\frac{A_k(t)}{\left\|A_k(t)\right\|} \approx \frac{A_l(t)}{\left\|A_l(t)\right\|}, \quad k \ne l$

2.2.3.1.1 聴覚的なフィルタバンク

聴覚的なフィルタバンクは、観測信号 $f(t)$ を複素スペクトル $X_k(t)$ に分解する。本節では、これをチャンネル数が $K=128$、解析範囲が 60-6000 Hz、サンプリング周波数が 20 kHz の定 Q gammatone フィルタバンクとして実装した。また、瞬時振幅 $S_k(t)$ と瞬時位相 $\phi_k(t)$ は、wavelet 変換で定義された振幅スペクトルと位相スペクトルを利用して決定される。

2.2.3.1.2 F_0 抽出部

F_0 推定部は、目的信号 $f_1(t)$ の基本周波数を推定する。この処理は、周波数方向でみた（チャンネル間の）各瞬時振幅 $S_k(t)$ に Comb フィルタリング処理を施すことで実現されている。$X_k(t)$ のチャンネル数は有限でかつ離散的（間引き的）に配置されているため、推定された $F_0(t)$ はチャンネルの中心周波数としてマッピングされ、同様に離散的な値を取る。また、推定された $F_0(t)$ は階段状に変動するため、その時間微分はすべての微小なセグメント区間で 0 になる。そこで、本論文では、各セグメント区間に対し、表 1(ii)の $E_{0,R}(t)$ を $E_{0,R}(t)=0$ と仮定する。また、上記のセグメントの長さを $T_h - T_{h-1}$ とする。これは、階段状に変動する $F_0(t)$ の区分的に連続な部分に対応し、T_h はその開始点に相等する。

2.2.3.1.3 グルーピング部

グルーピング部は、表 1 の制約条件(i)と(iii)を利用して、目的音の存在する時間周波数領域を決定し、その後で逆 wavelet 変換を利用して、分離された瞬時振幅と瞬時位相からその時間波形に再構成する[19]。$\hat{f}_1(t)$ と $\hat{f}_2(t)$ は、それぞれ $f_1(t)$ と $f_2(t)$ を再構成した信号である。制約条件(i)は、推定された $F_0(t)$ に対応するチャンネル出力 $X_{\hat{l}}(t)$ での立上り/立下り時間（T_S，T_E）と各チャンネル出力での立上り/立下り時間（$T_{k,\mathrm{on}}$，$T_{k,\mathrm{off}}$）を比較する処理として実装された。但し、両者の時間差として、立上りで、$\Delta T_S = 25$ ms、立下りで $\Delta T_E = 50$ ms とした。制約条件(iii)は、$F_0(t)$ の整数倍に対応するチャンネル番号を決定する処理として実装された。

2.2.3.1.4 波形分離部

波形分離部は、二つの音が同時に存在する時間周波数領域において、制約条件(ii)と(iv)を利用して観測された瞬時振幅 $S_k(t)$ と瞬時位相 $\phi_k(t)$ から、$A_k(t)$，$B_k(t)$，$\theta_{1k}(t)$，$\theta_{2k}(t)$ を決定する。これらを求める際、本処理では、初めに $A_k(t)$ と $\theta_{1k}(t)$、$F_0(t)$ の連続性の制約を考慮している。これは、制約条件(ii)に対応しており、表 1(ii)の $C_{k,R}(t)$ と $D_{k,R}(t)$ を一次多項式（$R=1$）として実装されている。この仮定では、$A_k(t)$ と $\theta_{1k}(t)$ の時間的変動の許容が 2 次多項式として

$$A_k(t) = \int C_{k,1}(t)dt + C_{k,0}, \text{ and}$$

$$\theta_{1k}(t) = \int D_{k,1}(t)dt + D'_{k,0}.$$

という制約を受けていることになる。次に、$dA_k(t)/dt = C_{k,R}(t)$を式(27)に代入すると、位相差$\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$に関する1階線形微分方程式を得る。これを解くと、次のような一般解が得られる。

$$\theta_k(t) = \arctan\left(\frac{S_k(t)\sin\left(\phi_k(t) - \theta_{1k}(t)\right)}{S_k(t)\cos\left(\phi_k(t) - \theta_{1k}(t)\right) + C_k(t)}\right), \tag{31}$$

但し、

$$C_k(t) = -\int C_{k,R}(t)dt - C_{k,0} = -A_k(t).$$

である。

$E_{0,R}(t)$で決定されるセグメント$T_h - T_{h-1}$間において、$A_k(t)$，$B_k(t)$，$\theta_{1k}(t)$，$\theta_{2k}(t)$は次の処理手順で決定される。初めに、セグメント内での唯一の解、$A_k(t)$と$\theta_{1k}(t)$が存在すると考えられる二つの推定範囲 $\hat{C}_{k,0}(t) - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)$ と $\hat{D}_{k,0}(t) - Q_k(t) \le D_{k,1}(t) \le \hat{D}_{k,0}(t) + Q_k(t)$ を Kalman filter を利用して決定する。但し、$\hat{C}_{k,0}(t)$と$\hat{D}_{k,0}(t)$は推定値、$P_k(t)$と$Q_k(t)$は推定誤差である。次に、ある$D_{k,1}(t)$に対する$C_{k,1}(t)$の候補を先の推定範囲内で spline 補間されたものを候補として選択し、その候補内から真の解と考えられる$\hat{C}_{k,1}(t)$を次式で決定する。

$$\hat{C}_{k,1}(t) = \underset{\hat{C}_{k,0}(t) - P_k(t) \le C_{k,1}(t) \le \hat{C}_{k,0}(t) + P_k(t)}{\arg\max} \frac{\left\langle \hat{A}_k, \hat{\hat{A}}_k \right\rangle}{\left\| \hat{A}_k \right\| \left\| \hat{\hat{A}}_k \right\|}, \tag{32}$$

但し、$\hat{A}_k(t)$は spline 補間で得られた瞬時振幅であり、$\hat{\hat{A}}_k(t)$はチャンネル間で制約条件(iii)を満たす瞬時振幅の平均である。最後に、$\hat{D}_{k,1}(t)$を次式で決定する。

$$\hat{D}_{k,1}(t) = \underset{\hat{D}_{k,0}(t) - Q_k(t) \le D_{k,1}(t) \le \hat{D}_{k,0}(t) + Q_k(t)}{\arg\max} \frac{\left\langle \hat{A}_k, \hat{\hat{A}}_k \right\rangle}{\left\| \hat{A}_k \right\| \left\| \hat{\hat{A}}_k \right\|}. \tag{33}$$

$\theta_{1k}(t)$と$\theta_k(t)$は、上記で決定された$\hat{D}_{k,1}(t)$と$\hat{C}_{k,1}(t)$に基づいて決定されるため、$A_k(t)$，$B_k(t)$，$\theta_{2k}(t)$を式(27)と(28)、$\theta_{2k}(t) = \theta_k(t) + \theta_{1k}(t)$の関係式により求めることができる。

2.2.3.2 シミュレーション

提案モデルが混合信号 $f(t)$から目的音 $f_1(t)$ を波形レベルで分離、抽出できることを示すために、次のような 3 種類の信号を利用して三つのシミュレーションを行った。

(a) 雑音が付加された AM-FM 調波複合音 [14]

(b) 雑音が付加された単母音（/a/、 /i/、 /u/、 /e/、 /o/）

(c) 雑音が付加された連続母音（/aoi/）

但し、雑音はピンク雑音であり、付加雑音の SNR は、5 から 20 dB まで 5-dB 刻みに調整された。音声信号は、ATR 音声データベースにある日本語母音（男性 2 名、女性 2 名）とした。

提案法の分離特性を評価するために、次式で定義される分離精度を利用した。

$$10\log_{10}\frac{\int_0^T f_1^2(t)dt}{\int_0^T \left(f_1(t)-\hat{f}_1(t)\right)^2 dt} \quad \text{(dB)}. \tag{34}$$

また、表 1 の各制約条件の有効性を示すために、次のような条件での提案法の性能も評価した。

(1) Comb フィルタを利用して調波成分を抽出し、Kalman フィルタを利用して、その成分内にある瞬時振幅 $A_k(t)$ と瞬時位相 $\theta_{1k}(t)$ を決定する場合

(2) Comb フィルタを利用して調波成分のみを取り出す場合

(3) 何も処理をしない場合

ここで、条件(1)は制約条件(ii)のなめらかさを利用しない場合、条件(2)は制約条件(ii)全部と制約条件(iv)を利用しない場合、条件(3)は制約条件すべてを利用しない場合に対応している。

2.2.3.2.1 処理の概要

図 20 に提案モデルの 2 番目のシミュレーションの処理概要を示す。はじめに、パネル A に示される雑音が付加された母音/a/の $f(t)$（$f(t)$の SNR は 10 dB）は、パネル B と C にそれぞれ示されるような瞬時振幅$S_k(t)$\$と瞬時位相 $\phi_k(t)$に分解される。次に、基本周波数 $F_0(t)$ がパネル D に示されるように推定される。目的音 $f_1(t)$ と雑音が同時に存在する時間周波数領域は、パネル E と F にそれぞれ示されるように、制約条件(i)と(iii)を利用して決定される。最後に、二波形の瞬時振幅と瞬時位相は制約条件(ii)と(iv)を利用して、$S_k(t)$と$\phi_k(t)$に基づいて決定される。$A_k(t)$と$\theta_{1k}(t)$はそれぞれパネル I と J に示されるようなり、分離抽出された信号 $\hat{f}_1(t)$ はパネル K のようになる。

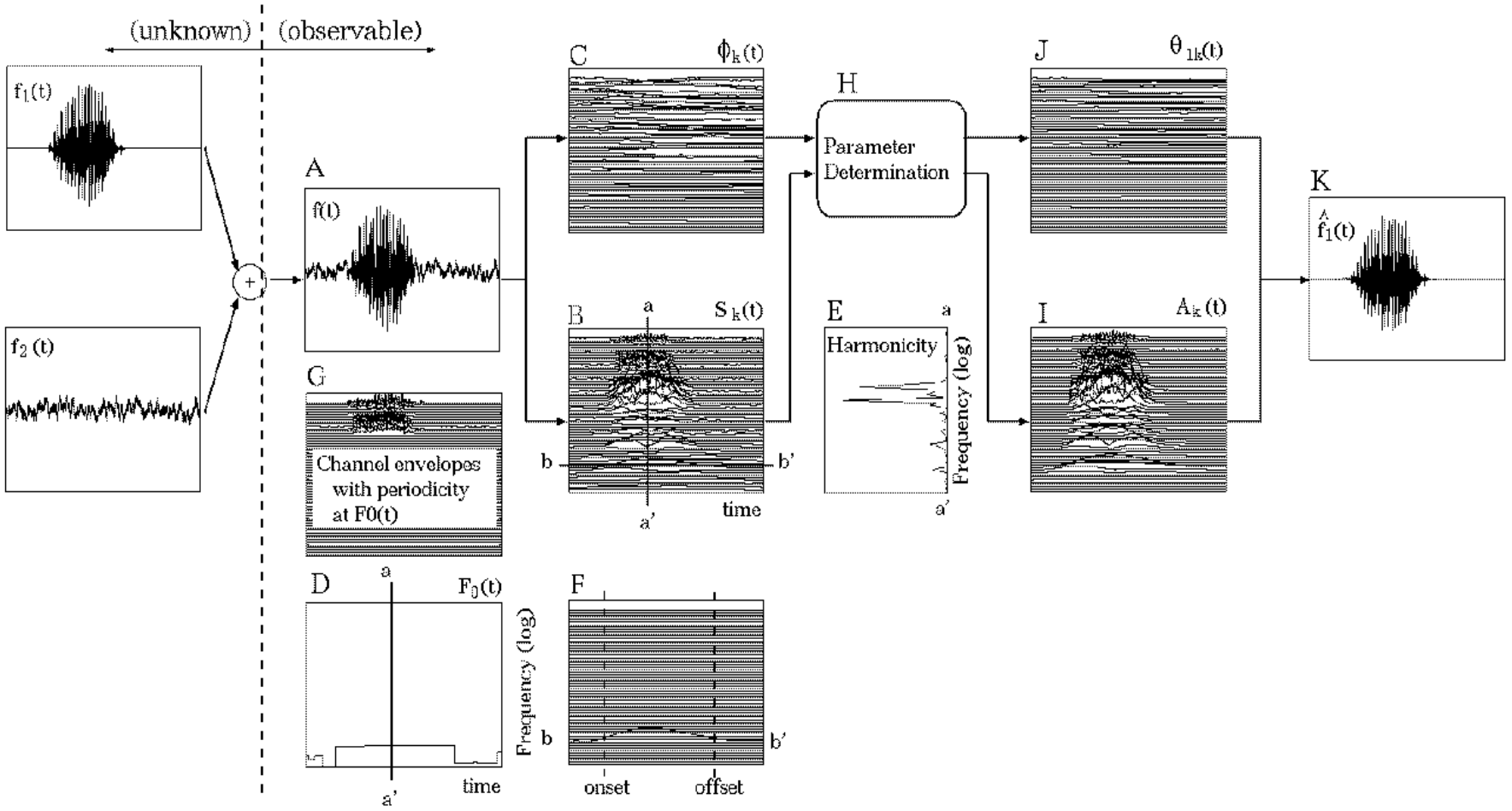

Fig. 20. Overview of signal processing of the proposed model.

2.2.3.2.2 結果と考察

上記に示した三つのシミュレーションと四つの評価条件での分離精度を図 21 に示す。

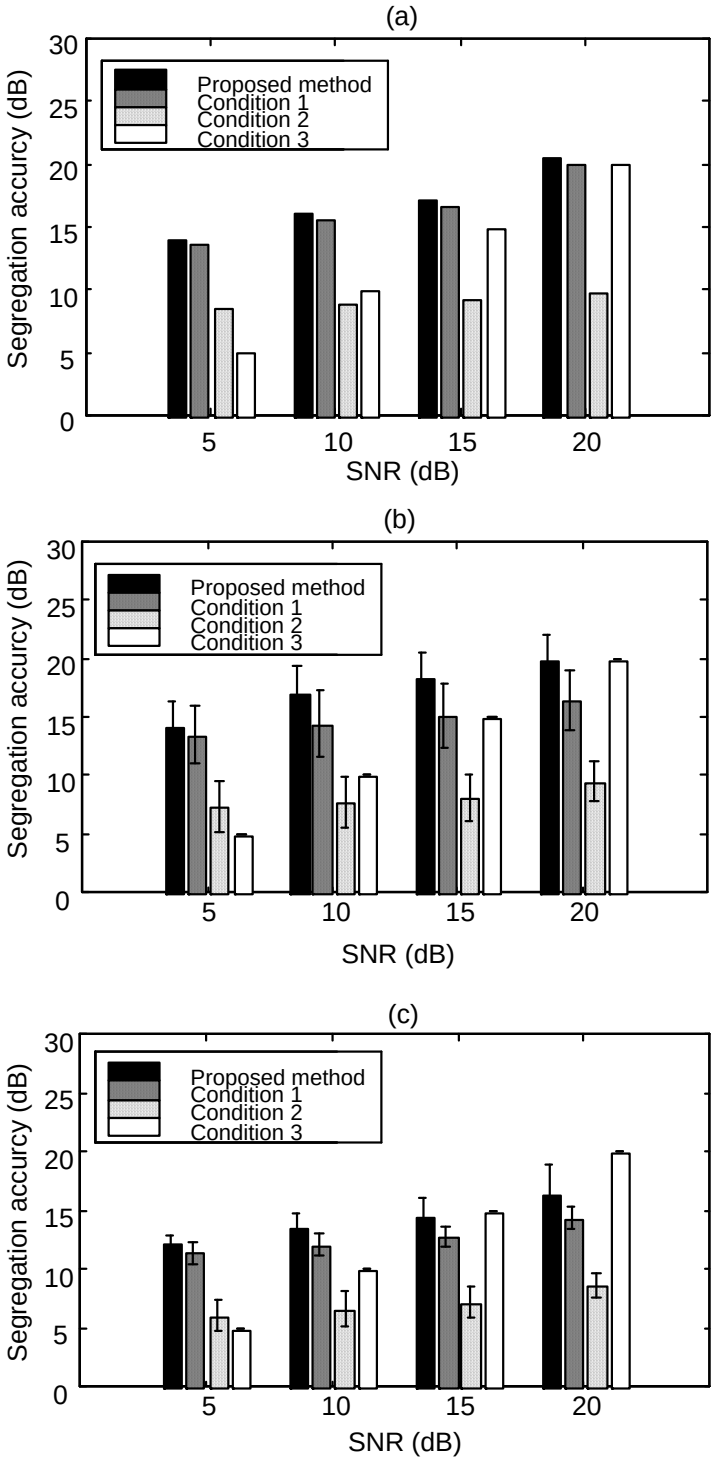

Fig. 21. Segregation accuracies for simulations. (a) AM--FM complex tone, (b) vowel, (c) continuous vowel.

これらの図で、縦棒はすべてのシミュレーションデータに関する分離精度の平均値、エラーバーはそのときの標準偏差を示す。これらの結果は、本提案法が他の三つの方法よりも分離精度が良好であることを示している。また、本提案法が波形レベルにおいても雑音が付加された音声から正確に分離抽出できることを示している。提案法と評価条件(2)を比較すると、瞬時振幅と瞬時位相を利用した分離法が位相を利用しないものよりも優れていることがわかる。最後に、分離精度の改善量を SNR 雑音からの原信号の分離能力とすれば、本提案法は雑音レベルが SNR=5 dB のとき、三つのシミュレーションでそれぞれ 10、9、7 dB であった。

2.2.3.2.3 Co-modulation Masking Release のモデル化（収録論文[5]）

聴覚系の周波数選択制の研究において、マスキングの現象を説明するモデルとしてマスキングのパワースペクトルモデル[20]が広く受け入れられている。このモデルでは、聴取者が、背景雑音中で特定の中心周波数を持つ正弦波信号を見知しようとするとき、信号周波数付近で中心周波数をもち、信号対雑音比が最も高くなる単一の聴覚フィルタの出力を利用するものと仮定している。また、刺激は長時間パワースペクトルとして表現されており、信号のマスキング閾値は、聴覚フィルタを通過する雑音の量によって決定されると仮定している。これらの仮定により、パワースペクトルモデルは、同時マスキングなどの多くの現象をよく説明できるが、成分音間の相対位相やマスカーの短時間変動を無視しているため、説明できないマスキング現象もいくつか存在した。

Hall ら[21]は、聴覚フィルタ間の比較によって、振幅包絡が変動する雑音にマスクされた正弦波信号の検出が容易になるという可能性を示した。このような検知能力の向上が生じるための決定的な条件は、異なる周波数帯域間で振幅包絡の変動が一致しているか、あるいは相関があるということであった。Hall らは、この異なる周波数帯域間の振幅包絡の一致を”共変調”と呼び、これによる検知能力の向上、すなわちマスキングの解除を共変調マスキング解除（Co-modulation Masking Release: CMR）と呼んだ。この現象については、多くの心理実験で確かめられている。しかし、CMR が生じるための条件が知られているにもかかわらず、この条件を利用した計算モデルはほとんど報告されていない。

そこで本稿では、同時マスキング現象の説明に利用されてきたマスキングのパワースペクトルモデル（モデル A）と、前節で提案した二波形分離モデル（モデル B）の二つのモデルを利用し、これに二つのモデルの結果を選択する処理を付加することで、CMR の計算モデルを構築する。

モデルの評価のために、Hall らによる共変調マスキング解除の実験を想定したシミュレーションを二つのモデルについてそれぞれ行なった。モデル A では、マスカーの種類に関係なく、マスカー帯域幅の増加とともにマスキング閾値が増加した。一方、モデル B では、マスカーの種類によってマスキング閾値に変化があった。ランダム帯域雑音の場合にはマスカー帯域幅の増加に関係なく閾値は変化しなかったが、共変調された帯域雑音ではマスカー帯域幅の増加とともにマスキ

ング解除が起こるという結果が得られた。この結果に対し、選択処理は二つのモデルの結果から分離抽出した純音のマスキング閾値の低い方を選択することで、Hall らが示した CMR の結果と同様の傾向を示す特性が得られた。このときのマスキング解除量は最大約 8 dB であった。

これらの結果から、音源分離モデルの応用として、Co-modulation Masking Release のモデル化が可能となったことが示された。

2.2.4 連続聴効果（音韻修復現象）のモデル化（収録論文[13]）

人間は、まわりに大きな雑音がり、複数の話者が存在しているような状況においてさえも、目的とする話者の音声を選択的に聴取することができる。これは、人間に雑音によってかき消された音を予測・補正する能力が備わっていることが一因である、と言われている。そして、この能力は、連続聴効果（Illusion of Continuity）あるいは音韻修復現象（Phonemic Restoration）と呼ばれている。もし、この能力がモデル化できるならば、モデルは雑音付加音声からきれいな音声を取り出す、また、雑音によりかき消されて聞こえない音声を予測することが可能となる。

本稿では、連続聴効果のモデルとして、スペクトルピークの軌跡追跡法と、これをもとにしたスペクトル系列の予測追跡法を提案する。この目的には、2 次系システムによるスペクトルピーク軌跡の外挿、IFIS によるスペクトル補間など、すでにさまざまな手法が提案されているが、これらの手法はスペクトルのピークのみの予測追跡法であったり、補間スペクトルに歪みを生じる。これらの欠点を補うために、提案法では次のようなスペクトル表現、追跡、外挿法を用いている。

(1) 音声波形をケプストラム不偏推定法によって対数スペクトル系列に変換する

(2) 対数スペクトル系列自体を外挿するのは難しいので、対数スペクトルのピークをまず 4 つのパラメータで表現する。4 つのパラメータとは、Shamma らが提案した聴覚一次野のモデル[22]をもとにして、ピークの振幅、周波数、帯域幅、対称性を、Gabor 関数を母関数としたウェーブレット変換で抽出したものである。

(3) 4 つのパラメータの軌跡を 2 次系システムで記述する。そして、このシステムを用いて、雑音のない区間ではこれらのパラメータを追跡し、雑音の区間ではこれらのパラメータを外挿する。

(4) 追跡あるいは外挿された 4 パラメータから、ウェーブレット逆変換を用いて予測スペクトルを計算する。

モデルの性能を評価するために、連続母音の音韻変化部を白色雑音で置き換えた音声波形を用意した。そして、提案モデルにより、雑音置換部の予測を行なった。なお、この音声は連続聴効果を生じることは確認済みである。

処理の結果、提案手法は、雑音置換部分であっても 4 パラメータを予測追跡し、連続母音スペクトル系列を修復した。

2.2.5 音源方向推定（収録論文[6][7][14]）

2.2.5.1 はじめに

両耳間時間差(ITD: interaural time difference)に基づく音源方向定位は、音源からの音波が左右の耳に到達する時の時間差を用いて音源の方向を知覚する聴覚の機能である。この機能は一般に、両耳間相互相関を基礎とする時間差検出回路モデルよって表現され[23]、幾つかの計算機モデルが提案されている[24]。本稿では、神経発火やシナプス伝達などの生体内部で行なわれる情報伝達の信号パターンを計算機上に表現し、それらを時間差検出回路モデルに適用するとともに、実際の聴神経に現れるような時間的なゆらぎをもつ神経インパルスを模擬した信号列を入力として用い、信号伝達の時間的な冗長性や時間的なゆらぎが両耳間時間差の検出に与える影響について検証する。

2.2.5.2 両耳間相互相関

時間差検出機構は一般に、左右の耳から送られるインパルス信号の相関に基づく両耳間相互相関モデルで表現されている[23][24]。相互相関モデルは、図 22（上）に示すように、刺激音の特定の位相角に同期して発火する神経インパルス列を利用して ITD の計算を行なう。但し実際の聴神経は、図 22（下）に示すように、常に正確な刺激音の位相角で発火する訳ではなく、神経インパルスは時間的なゆらぎを含む[25]。このため、この時間的なゆらぎは、時間差検出に対するノイズとして捉えられてきた。

2.2.5.3 時間差検出回路モデル

神経インパルスやシナプス伝達の信号の時間長は、人が知覚できる微細な時間差（約 10 μs）[26]に対して非常に大きく冗長的である。この冗長性が時間差検出に与える影響を調べるために、神経インパルスやシナプス伝達の信号を計算機上に表現し、時間差検出回路モデルへ適用する。

時間差検出回路内の信号伝達を計算機上へ表現するために、Rothman らの膜電位の方程式[27][28]を利用した。

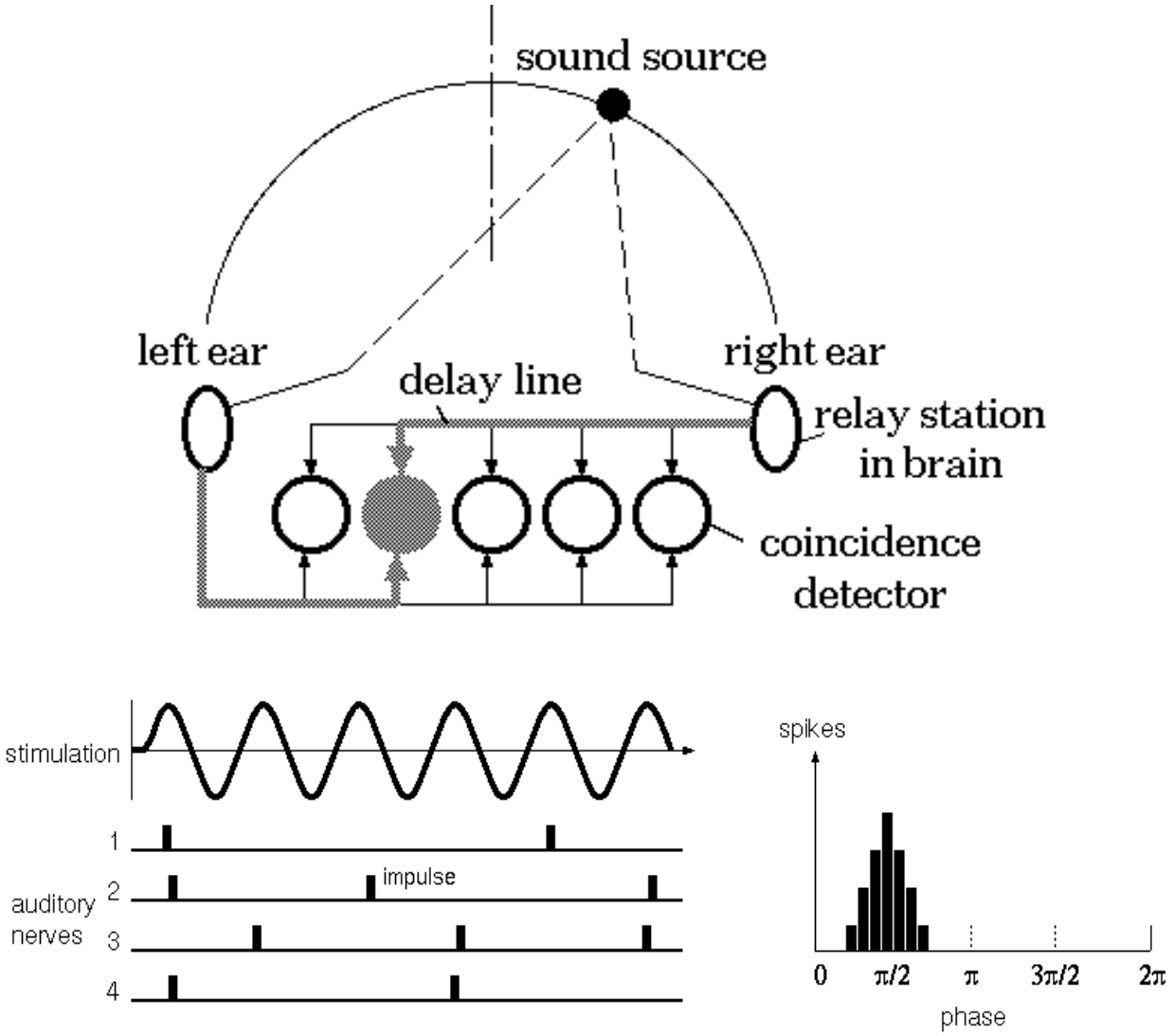

Fig. 22. Jeffress's coincidence detector circuit (upper) and spike fluctuation in the auditory nerve fibers (bottom).

$$C_m \frac{dV(t)}{dt} = -G_{\mathrm{Na}}\left(V(t) - E_{\mathrm{Na}}\right) - G_{\mathrm{K}}\left(V(t) - E_{\mathrm{K}}\right) - G_{\mathrm{Cl}}\left(V(t) - E_{\mathrm{Cl}}\right) - G_{\mathrm{L}}\left(V(t) - E_{\mathrm{m}}\right), \tag{35}$$

$$G_n(t) = A_n \frac{t - t_n}{\tau_p} \exp\left[1 - \frac{t - t_n}{\tau_p}\right],$$

但し、活動電位に至る関数を分離し、シナプス後電位(PSP)の時間推移のみを表現した。

さらに、閾値の関数を別に設け、PSP が予め決められた閾値電位を越えた時、スパイクを出力するよう設定した。閾値は回路内の全ての検出器において均一であると仮定した。

$$V_{\mathrm{threshold}}(t) = \beta \cdot e^{-t - t_r / \tau_r} + E_{\mathrm{threshold}}, \tag{36}$$

時刻tにおける閾値電位は$V_{\mathrm{threshold}}(t)$である。$t_r$は最新の発火時刻であり、$E_{\mathrm{threshold}}$は閾値の基準

電位を示す。βは最大閾値電位を表し、τ_rは相対不応期や順応に応ずる時定数とする。

シナプス伝達の表現にこれらの式を用いることは充分でないかもしれないが、信号の時間的冗長性を表わすための現段階での一手段とする。

これらの信号モデルを適用した一致検出回路モデルに、左右一対のインパルス信号を時間差無しで入力した時の、回路内のシナプス後電位(PSP)の時間的変化を図 23 に示す。ITD 0 μs の地点の他に広い範囲で大きな電位が発生し、ITD の軸上に PSP のピークが現れている。しかし、そのピークはなだらかで、微細な時間差の検出が可能なほどの明確な相関性は見出せない。

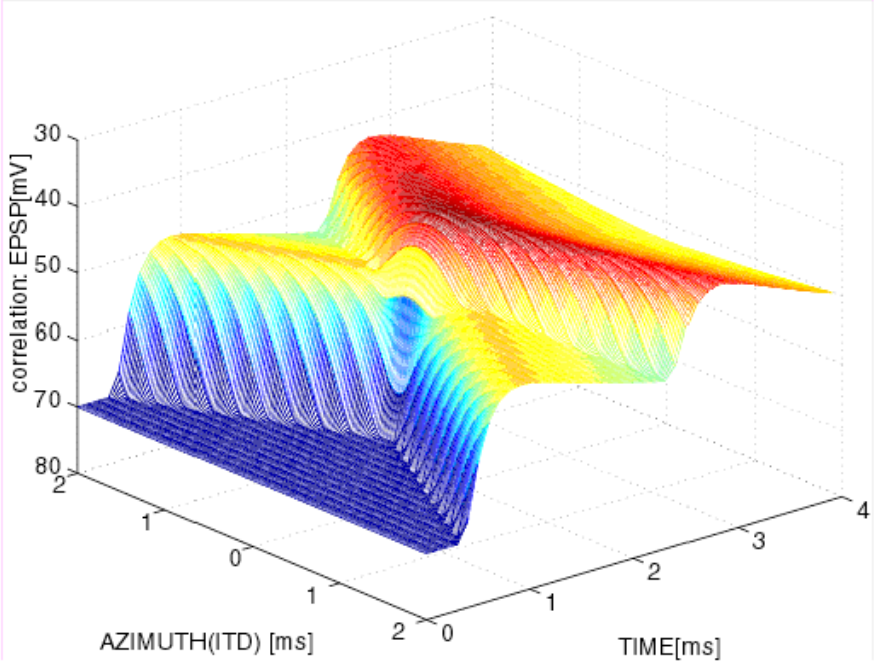

Fig. 23. Postsynaptic potentials in a coincidence detector circuit. This graph indicates the temporal transition of PSPs in the model after an impulse is applied to each input of the circuit.

2.2.5.4 時間差検出機構における抑制性

近年、時間差検出機構において抑制性の働きの存在が示唆されている[29][30]。その働きは未だ明確ではないが、Funabiki ら[31]は、鶏胚の層状核細胞において、時間分解能を向上させる抑制性入力の役割について報告している。それには、細胞体への GABA 作動性の抑制性入力が細胞の入力抵抗を減少させ、EPSP の時間経過を短くすることにより時間分解能が向上することが示された。そこで、この抑制の働きをモデル上に表現するために、膜時定数を変化させて、異なる EPSP の時間経過の特性を表現した。

2.2.5.5 ITD 出力機構

PSP が予め決められた閾値電位を越えた時、その検出器は発火し、スパイクを出力する。全ての検出器が同じ閾値電位を持つと仮定すると、PSP が閾値電位を越えた範囲の検出器だけが発火するため、非線形な出力を与える(図 24)。

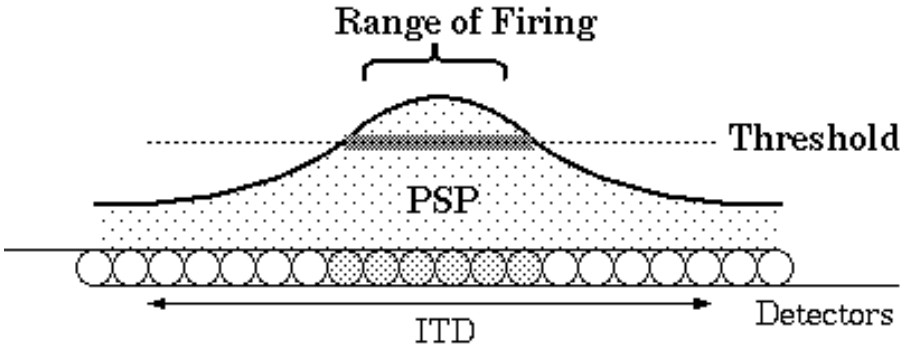

Fig. 24. Threshold level at the peak of the potential envelope.

図 25 の上図は、CF300Hz で、刺激音の特定の位相角に同期して発火する神経インパルス列の周期ヒストグラムを示す。このインパルス列を左右差(ITD) 0 μs で 0.3 秒間入力した時の本モデルの出力を図 25 下図に示す。非線形な出力機構により、ITD 軸上の特定の範囲にスパイクが出力し続け、ITD は一意に決まらないことが分かる。抑制性のモデルを導入しても、azimuth 軸上での幅は狭くなるものの、ヒストグラムの上部は平らであり、ITD は決まらない。

続いて、時間的なゆらぎを含むインパルス列を入力した時の、非線形な出力機構の動向をみる。図 26(A)のように、入力されるインパルス信号のゆらぎに従って PSP の丘の現れる位置は ITD 軸上を変動し、それと共に発火する範囲も軸上を変動する。

しかし、閾値電位の値によっては、PSP の丘の出現位置の変動に関わりなく、常に発火を持続するオーバーラップする部分が生ずることがある。入力されるインパルス信号のゆらぎが正規分布に従うならば、このオーバーラップする部分の中に正しい ITD が含まれる可能性が高い。この発火範囲の変動をスパイクヒストグラム上に積み重ねていくと、オーバーラップする部分のヒストグラムはより早く増加していく（図 26(B)）。

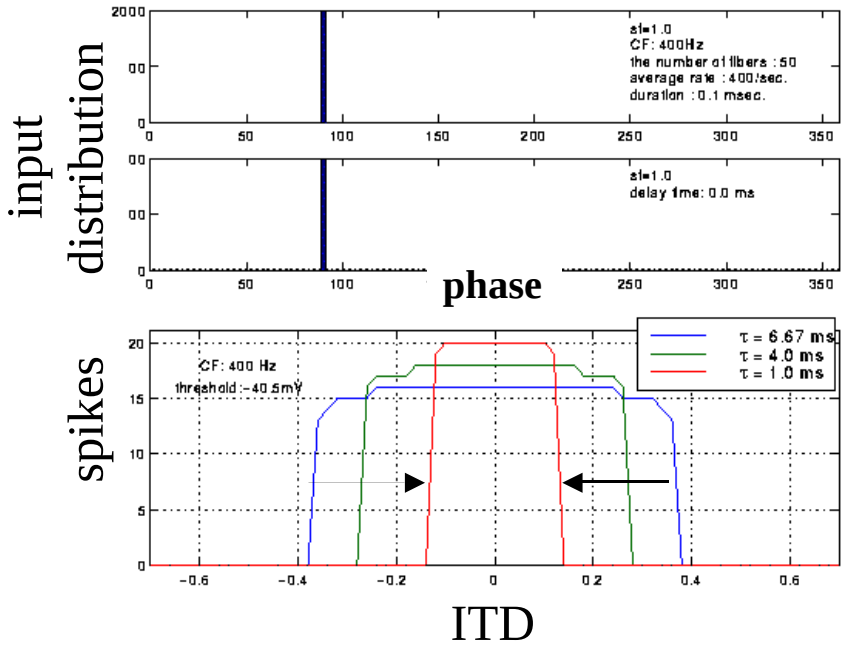

Fig. 25. Period histogram of the impulse train firing in synchronization with a certain phase of stimuli and the spike histogram obtained by the nonlinear output mechanism (ITD = 0 μs). The envelope of the spike histogram looks so square that it is difficult to determine the ITD.

このように、時間的なゆらぎを含むインパルス列を入力する場合、完全に位相同期するインパルス列を入力する場合に比べ、出力されるスパイクヒストグラムの包絡は、求めるべき ITD の辺りにピークを形成する。抑制性のモデルを導入すれば、ピークはより鮮明になる。(図 27)

2.2.5.6 考察

本稿では、両耳間時間差に基づく音源方向定位の計算機モデルとして、生体内部で行なわれる情報伝達の信号パターンを計算機上に表現し、時間差検出回路モデルに適用すると共に、時間的なゆらぎを含むインパルス信号列を入力として用いた。その結果、時間的なゆらぎを含むインパルス信号列を用いるとき、ITD を指し示すヒストグラムのピークがより顕著に現れることが示された。それは、インパルス信号の時間的なゆらぎが、時間的に冗長な信号伝達の過程や非線形な出力機構において、ITD の抽出に貢献する可能性を示唆するものであった。また、モデルにおいて、位相同期特性の比較的良好なインパルス列の入力に対し、抑制の働きによって EPSP の時間経過を短くすることは、Funabiki らの示唆したとおり、その時間分解能を改善し、一致検出を先鋭化することが分かった。

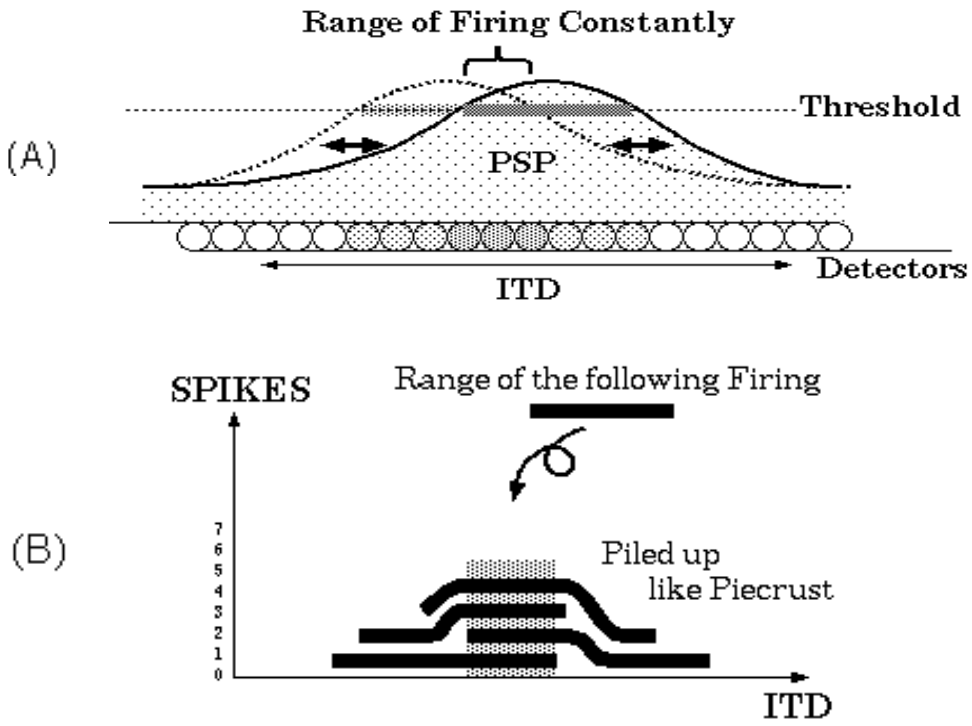

Fig. 26. Nonlinear output mechanism.

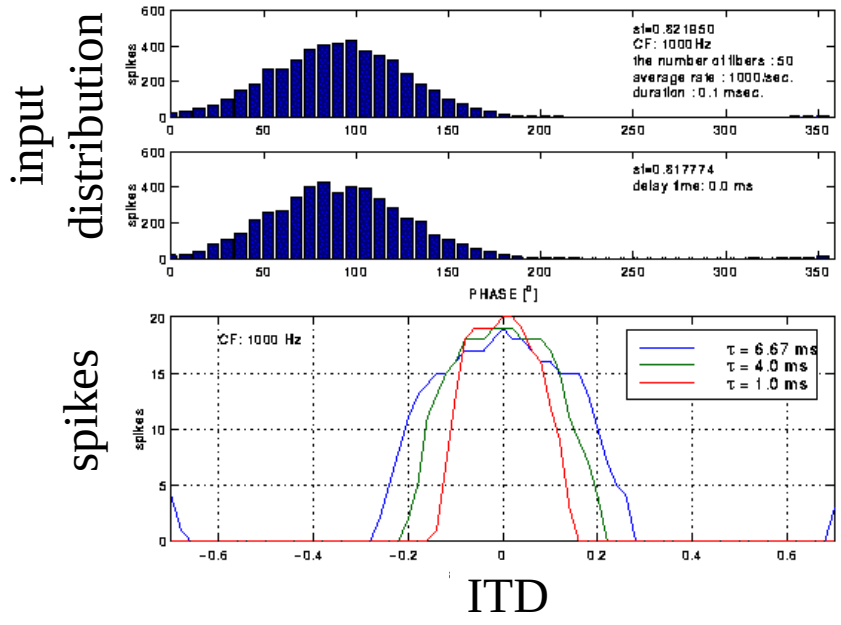

Fig. 27. Period histogram of the impulse train with a fluctuation in time and the spike histogram obtained by the simulation (ITD = 0 μs). The peak of the envelope of the spike histogram indicates the ITD.

2.2.6 まとめ

本研究の目的は、カクテルパーティ効果と呼ばれる人間の聴覚機構に存在する選択的聴取機構について、計算機による聴覚情景解析の考え方にそって、選択的聴取機構の数理的本質を明らかにし、そして、これを計算機上に実現することによって、雑音・残響が存在する実環境においてさえも目的信号音を忠実に再現できるシステムを実現することである。

そこで、本研究では、

(1) 音声強調:少数マイクロホンによる目的音と雑音の推定方法の検討
(2) 音源分離:雑音中の音声を抽出するための手がかりの検討
(3) 聴覚末梢の生理モデルによる音源方向の推定

を行なった。得られた結果は以下の通りである。

1. 少数マイクロホンによる目的音と雑音の推定

聴覚内に存在すると考えられているキャンセレーションをモデル化し、これを空間フィルタリング、周波数フィルタリングに適用して、音声強調を行なった。

まず、空間フィルタリングについては、3 本のマイクロホンのみを用いて、雑音中から音声スペクトルを抽出する方法を提案した。提案法は、目的音方向に零点を作り、雑音のみを取り出して雑音スペクトルを推定し、元の混合波形のスペクトルから雑音スペクトルを引き去ることによって雑音除去を行なう。このため、突発雑音でも除去が可能である。この方法により、従来の delayed-sum 法で 6 本のマイクロホンを用いた場合と同等の結果が得られた。そして、その有効性について、音声認識前処理実験、実環境での雑音除去実験を通して検討を行った。これより、従来の delayed-sum 法あるいは Griffith-Jim 法以上の雑音抑圧結果が得られた。また、雑音抑圧音声の客観評価を目的として、継時マスキングを考慮した評価尺度の検討を行った。

さらに、周波数フィルタリングについて、1 本のマイクロホンのみを用いて、雑音中の音声の特徴(基本周波数)を抽出する方法を提案した。雑音にはロバストであるが精度が十分でない方法と、雑音には敏感であるが高精度の方法を、雑音抑圧技術によりつなぎ合わせることにより、雑音にロバストでしかも高精度な方法とした。雑音中では従来成し得なかった高精度の基本周波数推定が可能となった。基本周波数は、人間の聴覚機構に存在する選択的聴取機構で用いられている重要な手がかりであり、これを高精度で推定できたことは、今後の研究に大いに貢献するもとと考えられる。

2. 雑音中の音声の抽出

Bregman によって提唱された、聴覚の情景解析(ASA)のための四つの発見的規則を、物理的制約条件として捕らえ直すことにより、調波復号音と雑音を分離する二波形分離モデルを提案した。また、入力位相に関する制約条件を再考し、これを正確に求める方法を考案することで、波形

レベルで正確に雑音中の母音を分離抽出できるモデルを新たに提案した。このモデルを用いれば、雑音中からきれいな音声を抽出でき、音声認識率を向上させる可能性がある。

また、二波形分離モデルを CMR のモデル化に適用することにより、CMR の計算機上での再現が可能となった。

3. 聴覚の生理モデルによる音源方向の推定

神経発火やシナプス伝達などの生体内部で行なわれる生理的情報伝達の信号パターンを計算機上に実現し、それらを時間差検出回路モデルに適用することで音源方向定位機能を構築した。そして、実際の聴神経に現れるような時間的なゆらぎを持つ神経インパルスを模擬した信号を入力として用いて、信号伝達の時間的な冗長性や時間的ゆらぎが両耳間時間差の検出に与える影響について検証した。その結果、信号伝達の時間的な冗長性や時間的ゆらぎが、信号伝達の過程や非線型な出力機構において、両耳間時間差の抽出に貢献する可能性が示された。

上記 3 研究の関係を図示すれば、図 28 のようになる。左右耳から入った音により音源方向推定が行なわれ、その情報をもとに空間フィルタリングによる音声強調（雑音抑圧）が行なわれる。また、周波数フィルタリングされた音声から基本周波数が抽出され、方向推定・音源分離の情報として使われる。最後に、雑音抑圧された音声から、基本周波数情報を用いて音声が分離される。

カクテルパーティ効果が生じる原因としては、それぞれの音源に対して両耳聴によって知覚される音像の空間的位置（方向と距離）の違い、音の大きさ、ピッチ、音色など音源の特性そのものの違い、また音声の場合には言語的知識、経験などが関係していると見られている。上記 3 研究では、空間的位置（方向と距離）の違い、ピッチ、音色など音源の特性そのものの違いに焦点をあててモデル化を行なった。本研究で明らかにし、構築した部分はここまでである。言語的知識についてはまったく手がつけられていない。しかし今後、図 28 右端に点線で示した知識の利用が可能となれば、より高度の選択的聴取機構のモデルが実現するであろう。

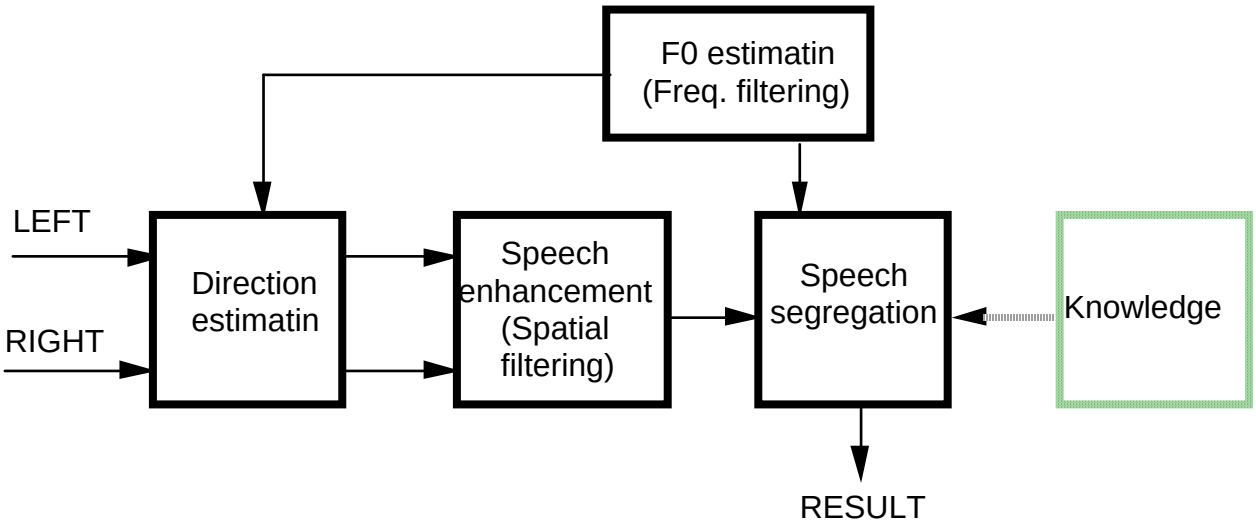

Fig. 28. Schematic graph of relationship between the proposed models.

REFERENCES

[1] Akagi, M. and Mizumachi, M. (1997). "Noise Reduction by Paired Microphones", Proc. EUROSPEECH97, 335-338.

[2] Unoki, M. and Akagi, M. (1997). "A method of signal extraction from noisy signal based on auditory scene analysis", Proc. CASA97, IJCAI-97, Nagoya, 93-102.

[3] Unoki, M. and Akagi, M. (1997). "A method of signal extraction from noisy signal", Proc. EUROSPEECH97, 2587-2590.

[4] 鵜木、赤木(1997). ”雑音が付加された波形からの信号波形の一抽出法”、電子情報通信学会論文誌、J80-A, 3, 444-453.

[5] Durlach, N. L. (1963). "Equalization and Cancellation Theory of Binaural Masking-Level Difference," J. Acoust. Soc. Am., 35, 8, 1206-1218.

[6] Culling, J. F. and Summerfield, Q. (1995). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am., 98, 2, 785-797.

[7] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am., 93, 6, 3271-3290.

[8] Bregman, A.S. (1990). " Auditory Scene Analysis", Academic Press.

[9] Bregman, A.S. (1993). "Auditory Scene Analysis: hearing in complex environments," in Thinking in Sounds, pp. 10--36, Oxford University Press, New York.

[10] Kaneda, Y. and Ohga, J. (1986). "Adaptive microphone-array system for noise reduction," IEEE trans. ASSP, 34, 6, 1391-1400.

[11] Flanagan, J. L. et.al. (1991). "Autodirective microphone systems," Acoustica, 73, 2, 58-71.

[12] Griffiths, L. and Jim, C. (1982). "An Alternative Approach to Linearly Constrained Adaptive Beamforming," IEEE AP-30, 1, 27-34.

[13] Kawahara, H., Katayose, H., de Cheveigne, A., and Patterson, R.D. (1999). "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 andPeriodicity," Proc. EUROSPEECH99, 2781-2784.

[14] Unoki, M. and Akagi, M. (1998). "Signal Extraction from Noisy Signal based on Auditory Scene Analysis," Proc. ICSLP'98, vol. 4, pp. 1515--1518.

[15] 阿竹, 入野, 河原, 陸, 中村, 鹿野 (2000). ”調波成分の瞬時周波数を用いたピッチ推定法の検討”、信学技報, SP99-170.

[16] Cooke, M. P. and Brown, G.J. (1993), "Computational auditory scene analysis : Exploiting principles of perceived continuity," Speech Communication, vol. 13, pp. 391-399.

[17] Ellis, D. P. W. (1996). "Prediction-driven computational auditory scene analysis," Ph.D. thesis, MIT Media Lab.

[18] Nakatani, T., Okuno, H. G., and Kawabata, T. (1994). "Unified Architecture for Auditory Scene Analysis and Spoken Language Processing," Proc. ICSLP '94, 24, 3.

[19] Unoki, M. and Akagi, M. (1999). "Signal Extraction from Noisy Signal based on Auditory Scene Analysis," Speech Communication, vol. 27, no. 3 , pp. 261--279.

[20] Patterson, R. D. and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," In Frequency Selectivity in Hearing (ed. B. C. J. Moore), Academic Press.

[21] Hall, J. W. Haggard, M. P. and Fernandes, M. A. (1984). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am., 76, 50-56.

[22] Wang, K. and Shamma, S. A. (1995). "Spectral shape analysis in the central auditory system," IEEE Trans. Speech & Audio Process., 3, 5, 382-395.

[23] Jeffress, L. A. (1948). "A Place Theory of Sound Localization," J.Comp.Physiol.Pychol.

[24] Stern, R. M. and Trahiotes, C. (1995). "Models of Binaural Interation in Hearing," Ed. B.C.J. Moor, Academic Press, 347-386.

[25] Johnson, D. H. (1980). "The relationship between spike rate and synchrony in responses of auditory nerve fibers to single tones," J. Acoust. Soc. Am.68, 1115-1122.

[26] Moore, B. C. J. (1997). "Psychology of Hearing," Academic Press

[27] Rothman, J. S., Young, E. D., and Manis, P. B. (1993). "Convergence of Auditory Nerve Fibers onto Bushy Cells in the Ventral Cochlear Nucleus: Implications of a Computational Model," J. Neuro Physiol., Vol.70, No.6.

[28] Brughera, A. R., Stutman, E. R., Carney, L. H., and Colburn, H. S. (1996). "A Model with Excitation and Inhibition for Cells in the Medial Superior Olive," Auditory Neuroscience, Vol.2, 219-233.

[29] Lachica, E. A., Rubsamen, R., and Rubel, E. W. (1994). "GABAergic Terminals in Nucleus Magnocellularis and Laminaris Originate From the Superior Olivary Nucleus," J.Comp.Neurol., 348, 403-418.

[30] Pena, J. L., Viete, S., Albeck, Y., and Konishi, M. (1996). "Tolerance to Sound Intensity of Binaural Coincidence Detection in the Nucleus Laminaris of the Owl," J. Neuroscience, 16(21), 7046-7054.

[31] Funabiki, K., Koyano, K., and Ohmori, H. (1998), "The role of GABAergic inputs for coincidence detection in theneurons of nucleus laminaris of the chick, " J. Physiol., 508.3, 851-869.

3. 研究成果