HOKKAIDO UNIVERSITY

Title	テレビ会議における発話区間検出と音声雑音除去への応用に関する研究
Author(s)	佐々木, 一人
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12641号
Issue Date	2017-03-23
DOI	https://doi.org/10.14943/doctoral.k12641
Doc URL	https://hdl.handle.net/2115/65729
Type	doctoral thesis
File Information	Kazuto_Sasaki.pdf



Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

博士論文

テレビ会議における発話区間検出と音声雑音除去への応用に関する研究

Study on Speech Segment Detection and Its Application to Audio Noise Removal for Video Conferences

北海道大学 大学院情報科学研究科
メディアネットワーク専攻

佐々木　一人

2017 年 3 月

目 次

第 1 章　序論　1

1.1　本研究の背景 1

1.2　本研究の目的 3

1.3　論文の構成 5

第 2 章　映像と音響信号の協調利用に基づいた発話区間検出と音声雑音除去への応用　6

2.1　はじめに 6

2.2　提案手法 6

2.2.1　発話区間の推定 6

2.2.1.1　映像における参加者の口の動きの取得 7

2.2.1.2　音響信号の分割・分類手法の利用 8

2.2.1.3　発話区間の推定 8

2.2.2　音響信号のノイズ除去 8

2.3　実験結果 9

2.3.1　SS 法によるノイズ除去 9

2.3.2　NLMS 法によるノイズ除去 10

2.4　まとめ 11

第 3 章　複数の参加者によるテレビ会議における発話区間抽出とノイズ除去への応用　16

3.1　はじめに 16

3.2　提案手法 16

3.3　実験結果 17

3.3.1 SS 法によるノイズ除去 17
3.3.2 NLMS 法によるノイズ除去 18
3.4 まとめ 19

第 4 章 Decision Level Fusion に基づく発話区間検出と音声雑音除去への応用 23

4.1 はじめに 23
4.2 提案手法 24
4.2.1 画像・音響特徴に注目した発話区間検出 25
4.2.1.1 画像特徴量の抽出 25
4.2.1.2 音響特徴量の抽出 27
4.2.1.3 各特徴に基づいた分類 28
4.2.2 複数の分類結果の DLF による統合 30
4.2.2.1 各アノテーターの分類精度と識別モデル 31
4.2.2.2 学習段階 32
4.2.2.3 テスト段階 35
4.3 実験結果 36
4.3.1 発話区間検出の性能評価 37
4.3.2 発話区間検出結果のノイズ除去への応用 40
4.4 まとめ 40

第 5 章 結論 56

謝辞 58

著者の研究業績 63

第1章　序論

1.1　本研究の背景

近年，カメラやマイクロフォン，スピーカー等を搭載したスマートフォンやタブレットPCなど，モバイル機器の普及拡大とその家庭内への浸透が着実に進んでいる．内閣府が2016年4月8日に発表した2016年3月実施の消費動向調査（本調査は2人以上の世帯を対象に4038世帯から回答を得たもので，単身世帯は含んでいない）によると，2015年度のスマートフォンの世帯あたりの普及率（以後、世帯普及率という）は，前年度比6.8ポイント増の67.4％，100世帯当たりの保有台数は135.4台，スマートフォン以外の従来型携帯電話の世帯普及率は64.3％，100世帯当たりの保有台数は100.5台となり，初めてスマートフォンがそれ以外の携帯電話を逆転するという新旧携帯電話の交代劇も起きている．また，スマートフォンとそれ以外の携帯電話の両者をあわせた2015年度の全携帯電話の世帯普及率は，前年度比0.9ポイント増の95.3％，100世帯当たりの保有台数は235.9台と、２台持ち世帯の増加傾向も継続している。さらに、パソコン（PC）の世帯普及率は，前年度比1.1ポイント増の79.1％，100世帯当たりの保有台数は128.0台となり，タブレットPCに代表されるタブレット型端末の世帯普及率は、前年度比3.7ポイント増の32.0％，100世帯当たりの保有台数は41.6台に達し，一般家庭におけるカメラやマイクロフォン，スピーカー等のマルチメディア機能を搭載した端末の常態化を物語っている．

また，こうした端末の普及動向と相まって，それらを利活用した多彩なサービスの進展も進んでいる．特に，画像や映像，音声やテキストが連携したマルチメディア型のサービスは，数多く提供されるようになり，例えば，テレビ電話サービスに代表される画像・映像付きの音声通話サービスは，もはやごく普通に利活用されるようになり，映像とテキストや音声が連携した見守り系のサービスや、映像付きのビデオレターやビデオメッ

セージサービス等、多くのマルチメディア型サービスが登場してきている．

さらに，こうしたサービスを，利用者が，いつでも，どこでも，好きな時に，利用できる，いわゆるユビキタス環境の整備も図られてきている．インターネットという技術革新を背景に，有線系を主体に展開してきた日本の情報通信基盤が，無線系技術の技術進歩と，それを利活用した各種サービスの高度化により，無線系単独での発展展開も含めて，有線・無線の有機的な結合によるサービス基盤として進展し，また，一般化してきていると言える．見方を変えると，ビジネスサイドでの普及・利用はもとより，コンシューマサイドでのマルチメディア型サービスの普及・展開が，社会・経済・国民生活のありとあらゆる活動において，無くてはならないものへと確実に進化し，また，定着してきているとも言える．

一方，こうしたユビキタス環境下におけるマルチメディア型のサービスは，利用時の環境条件に大きく左右されやすいのも事実である．例えば，ノイズの発生状況次第では，極端なサービス低下を招くことも少なくないし，場合によってはサービスとして利用できない状況も発生しかねない．こうした実情を踏まえると，音声が中核的なサービス価値であるマルチメディアサービスにおいては，音声雑音の除去は，非常に重要な課題になる．

このような背景から, これまで様々な音声雑音除去手法 [1, 2, 3, 4, 5] が提案され，その性能改善についても，多くの研究者によって研究がなされている．音声雑音除去の代表的な手法の一つとしては，音響信号から発話区間を検出し，残りの非発話区間からノイズの正確な情報を取得することによって，発話区間に重畳されたノイズを除去するという方法がある．この手法においては，いかに発話区間を高精度に検出できるかが，音声雑音除去の良し悪しを決める鍵となる．実際にこれを実現するため，音響信号から発話区間を検出する研究は活発に行われている [6, 7]．また，近年，マルチメディア信号処理の研究分野では，複数のモダリティを協調利用する方法が提示され，各モダリティを単体で用いた場合よりも，様々なアプリケーションにおいて高精度な処理結果が得られることが報告されている [8]．また，複数モダリティを協調利用する手法は，発話区間の検出においても，その高精度化が期待できるとの報告もなされている [9]．すなわち，具体的にテレビ会議における音声雑音除去を対象とした場合を想定すると，次の 2 つの特

徴に着目し，これらを協調利用した取り組みを図ることにより，発話区間の検出とその高精度化が期待できるということである．その一つ目の特徴としては，テレビ会議の場合，サービスを利用している者同士で会話が行われることになるため，その中には，利用者が話をしている発話区間と，話をしていない非発話区間が含まれているという点である．二点目の特徴としては，テレビ会議においては，音響情報のみならず，相手方に示す映像情報（例えば，カメラで撮像される話者の顔の映像）が存在し，こうした映像情報と他の情報（音響情報）との協調利用が可能になるという点である．つまり，一番目の特徴を活用することによって，発話区間と非発話区間を分類・分離・分割し，得られる非発話区間における音響信号から雑音情報の取得が可能となることから，その結果を利用することで，発話区間における雑音の除去が可能になる．次に，二つ目の特徴である，音響情報と映像情報との協調利用により，発話区間をより高精度に検出することが期待できるということである．

しかしながら，これを実現するためには，達成しなければならない課題もある．従来の発話区間検出手法においては，各モダリティを協調利用し統合する際に，精度を考慮した統合が行われていない．そのため，各モダリティを単独に用いた場合と比較した際に，必ずしも高精度な結果が得られるとは限らないのである．前述した通り，この手法においては，音声雑音除去の良し悪しを決定する鍵は，話者の発話区間を，より高精度に検出できるかにかかっていると言って良い．折角，複数モダリティの協調利用を図っても，検出の高精度化が実現できなければ、また、高精度化が図れないケースが存在してしまっては，音声雑音除去の性能向上を図るという最終目的が達成できなくなってしまう．したがって，最終目的を達成するためには，複数モダリティの協調利用の際，精度を考慮した統合手法を構築し，それを実現することが必要になる．

1.2　本研究の目的

本研究は，異種特徴の統合解析に基づいたテレビ会議における発話区間検出と音声雑音除去への応用について研究した成果をまとめたものである．

背景でも述べた通り，今日，モバイル機器の普及と家庭への浸透が進展する中，ユビキタス環境下におけるマルチメディア型のサービスは，社会・経済・国民生活のありと

あらゆる活動にとって，無くてはならないものへと確実に進化し，定着し続けてきている。こうした中，コミュニケーションの基本かつ中心的な役割を果たす音声を中核に据えたマルチメディアサービスにおいては，音声雑音除去が重要課題となっており，その解決に向けて，これまで多くの研究が行われ，近年では，マルチメディア信号処理の研究分野において，複数のモダリティを協調して利用することで，それらを単体で用いた場合よりも高精度な処理の結果が得られ，また，複数のモダリティの利用により発話区間検出の高精度化が可能になることが報告されている．しかし，発話区間検出のより高精度化を実現するためには，協調利用の統合時に各モダリティから得られた結果の精度を考慮した統合手法を構築・実現する必要がある．

このような状況を考慮し，本研究では，テレビ会議における発話区間検出に関し，映像情報と音響信号の情報の両方を協調的に用いることで，それぞれの情報を単独で用いた場合よりも高精度に発話区間検出を可能にする手法の構築を目指すこととする．具体的には，各々の情報から検出された結果を協調利用する際に，精度を考慮した高精度な統合手法の実現を目指し，発話区間の確率を用いる統合手法の構築や，分類精度を考慮して統合を可能とする手法 Decision Level Fusion(DLF) を導入することで，より高精度な発話区間の推定を可能にすること目指す．また，これらの統合手法によって得られた発話区間の検出結果から，音声雑音除去への応用を行うことを本研究の目的とする．

なお、以上の目的を実現するため，本論文ではまず，映像中の利用者の口領域を抽出し，その動きの特徴に基づいて発話区間の検出を可能とするだけでなく，音響信号の音響クラスへの分類を行うことによる発話区間検出を可能にすることに取り組む．こうして得られた，それぞれ全く別の異種特徴から取得された検出結果を統合することで，最終的な発話区間の推定を可能とする．このとき，本論文では，各々のモダリティにおける発話区間の確率を統合する手法に加え，各々のモダリティにおける分類精度を考慮する新しい統合手法 (Decision Level Fusion: DLF) を導出し，これに基づいて，各々のモダリティを用いた場合よりも高精度な推定を可能とする手法の構築を目指す．また，発話区間のより高精度な検出結果から，残りの非発話区間による正確なノイズ情報が取得可能となることから，その情報を用いることで，簡便な処理により発話区間に重畳したノイズの高精度な除去が可能になることを明らかにする．

1.3　論文の構成

本論文は，5 つの章で構成されている．まず，第 1 章では，研究の背景と目的，及び，本論文の構成について述べる．次に，第 2 章では，映像中の利用者の口領域の動き特徴及び音響信号の分類結果から，発話区間である確率をそれぞれ算出し，それらを統合した確率に基づいて最終的な発話区間の推定を可能にする手法を提案している．また，その結果に基づいた音声信号の雑音除去についても説明を行っている．さらに，第 3 章では，第 2 章で有効性が確認された手法について，複数人が発話する環境下においても，発話区間を高精度に検出可能とする手法を提案している．具体的に，複数の利用者が撮像された映像情報の中から，真に発話を行っている利用者の口領域を選定することで，正確な発話区間を検出し，その結果，高精度な音声雑音除去が可能になることを示している．第 4 章では，DLF に基づいた異種特徴量の統合による高精度な発話区間検出手法の提案を行い，その有効性の確認を行っている．具体的には，顔の口領域に着目した視覚的な特徴量と音響信号に基づく特徴量を用いて，正確な発話区間検出を実現するための試みについて論じている．本手法では，利用者の口領域の動き特徴及び音響信号の分類結果 [10] から得られる特徴に対して，Support Vector Machin [11] により構築される識別器に基づいて分類を行い，複数の分類結果を得るとともに，さらに，各々の分類精度を考慮した DLF [12]，具体的には，Supervised Learning from Multiple Experts (SLME) を導入し，これによる分類結果統合から，高精度な発話区間の検出を可能としている．これまでも，分類結果の統合手法については，様々な提案がなされている [13, 14, 15, 16] が，本論文では，各分類結果の精度を考慮することが可能な統合手法に注目し，その統合手法を用いた際の効果について，実際に評価実験を行い，提案する手法が各々のモダリティよりも高精度に発話区間を検出可能であることを示している．また，従来のマルチモーダル統合手法と比較しても，高精度に発話区間検出を実現していることを明らかにしている．最後に, 第 5 章では, 本研究の成果について要約し, 論文全体のまとめとしている．

第2章　映像と音響信号の協調利用に基づいた発話区間検出と音声雑音除去への応用

2.1　はじめに

本章では，音響信号にノイズが重畳した環境下のテレビ会議において，カメラで撮像された映像と得られた音響信号の協調的利用を図ることによって，発話区間の抽出を可能とする手法を提案する．さらに，取得された発話区間に基づいて効率的にノイズ除去を可能にする手法について検討する．さらに，本章の最後では，実験により提案手法を用いることで，実際に音声雑音の低減が可能であることを確認する．本提案手法を用いることで、様々な SN 比の信号に対して, ノイズ低減を行うことが可能となる．

2.2　提案手法

本節では，映像と音響信号の両方を用いた発話区間検出手法と音響信号からのノイズ除去への応用について示す．図 2.1 には，提案手法の概要を示す．提案手法では，まず，テレビ会議の参加者の発話区間抽出を口の動きと音響信号の分類結果に基づいて行う．さらに，これらの発話区間から推定可能となる非発話区間より抽出されたノイズ情報に基づき，ノイズの除去を可能とする．以下では，提案手法の具体的な処理について詳述する．

2.2.1　発話区間の推定

本節では，提案手法における発話区間の推定方法を説明する．提案手法では，発話区間の推定を，映像に撮像されている参加者の口の動きの取得および音響信号の分割・分類により行う．以下では，それぞれの方法について説明する．

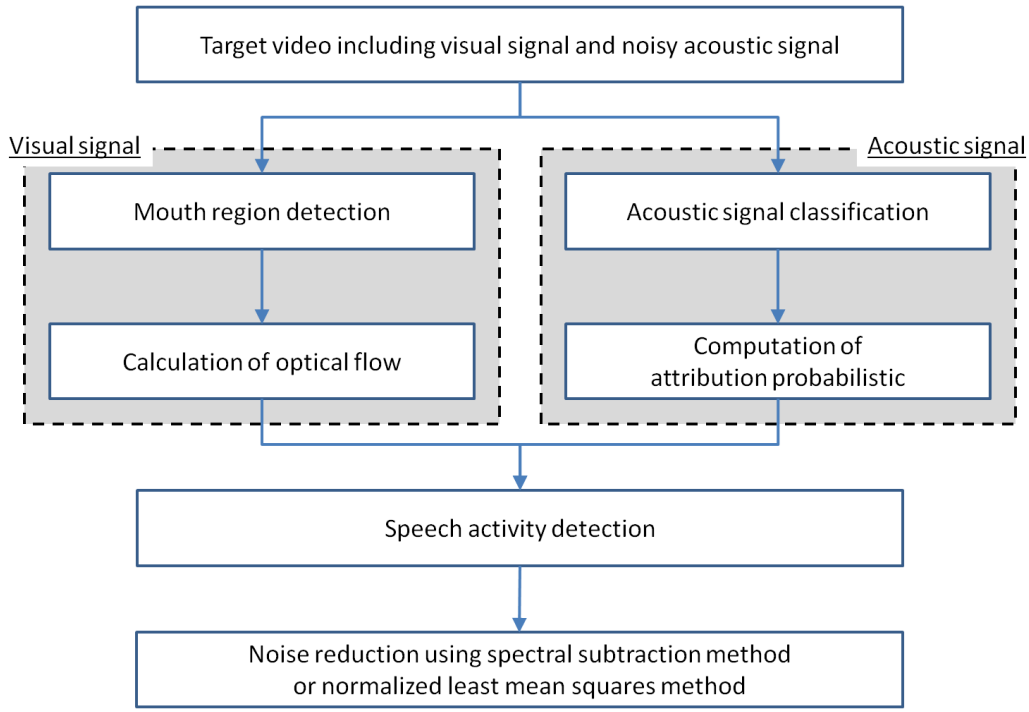


図 2.1: 本章で提案する映像と音響信号の協調利用に基づいた発話区間検出手法のブロック図

2.2.1.1　映像における参加者の口の動きの取得

提案手法では，発話区間を推定するため，映像における参加者の口領域を検出し，口の動きを取得する．具体的には，以下の手順に示す.

1) Adaboost[17] による顔領域検出に基づいた口領域の検出

2) 検出された口領域からの勾配法によるオプティカルフロー算出

3) 算出されたオプティカルフローの平均振幅値 x より，$1-\exp(-\frac{x^2}{\alpha})$(ただし，$\alpha$ はパラメータ) を用いて発話区間である確率を算出

なお、提案手法における、手順 1) Adaboost による口領域の検出については, OpenCV のライブラリを利用し，映像から口領域を検出した．具体的に，この関数は，Haar-like 特徴を入力にとる決定木を基本分類器としている [17]．また，学習には，OpenCV のサン

プルに付属するものを用いている．以上により，提案手法では，上記により算出される確率の大きな区間を発話区間である可能性が高いと考える．

2.2.1.2　音響信号の分割・分類手法の利用

提案手法では，発話区間を推定するため，文献 [10] に従い，音響信号から短時間あたりの各オーディオクラス (無音，音声，音楽，雑音) の帰属確率を算出する．この結果より，音声のオーディオクラスの帰属確率が高い区間は，発話区間である可能性が高いと考えられる．

2.2.1.3　発話区間の推定

提案手法では，手順 1）および 2） の結果を利用して，映像から発話区間を推定する．具体的には，口の動きから取得された確率値と音声のオーディオクラスの帰属確率をかけた値を指標とし，その値がしきい値以上の区間を発話区間として推定する．この処理を全ての時刻の映像・音響区間に対して適用することで，発話区間の抽出が可能となる．

2.2.2　音響信号のノイズ除去

本節では，提案手法における音響信号のノイズ除去手法を説明する．提案手法では，伝統的なノイズ除去手法である以下の 2 つの一方を用いる．

- Spectral Subtraction 法 [1](以降，SS 法)

  SS 法は，ノイズが付加された音響信号のパワースペクトルから，別途推定したノイズのパワースペクトルを差し引き，そのパワースペクトルを利用してフーリエ逆変換することでノイズを除去した音響信号を復元する手法である．これによって，非発話区間において推定されたノイズ情報に基づいて，発話区間からノイズを除去することにより，元の信号の推定が可能となる．

- Normalized Least Mean Squares(NLMS) 法 [2]

  NLMS 法は，フィルタ出力と所望の信号との二乗誤差を最小とするように適応フィルタの係数を更新し，所望の特性を実現するフィルタを作成する手法である．こ

れによって，所望の信号からあらかじめ学習させたフィルタを利用することで，音響信号のノイズ除去を可能とする．

提案手法では，前述の発話区間の推定により，得られる非発話区間から抽出したノイズの情報を利用し，SS 法もしくは NLMS 法によって音響信号からノイズを除去する．

2.3　実験結果

本節では，提案手法の有効性を検証するために行った実験結果を示す．本実験では，スマートフォンで撮像を行った 480 × 360 画素，30 fps，44.1 kHz の 10 秒の映像を利用する．なお、本実験で使用したスマートフォンは，iphone4s である．本映像には，単一の話者による会話のみが含まれている．さらに，音響信号に対してノイズを重畳し，提案手法を適用して得られるノイズ音響信号を用いてノイズ除去を行う．以下では，得られた実験結果を示す.

2.3.1　SS 法によるノイズ除去

提案手法により発話区間を推定し，SS 法によりノイズを除去した際の実験結果を示す．表 2.1 は本実験によるノイズ除去前の SN 比およびノイズ除去後の SN 比を示している．Noise 1 の SN 比は，提案手法を適用した際に SN 比の向上が確認された最大の値であり，ノイズ除去前の SN 比 18.78 に対し，ノイズ除去後の SN 比は 21.48 となった．また，Noise 4 の SN 比は，提案手法により発話区間が正しく推定された際の最小の値であり，ノイズ除去前の SN 比は -7.26，ノイズ除去後の SN 比は 4.62 になっている．この表より，提案手法によりノイズ除去が可能となった SN 比の範囲およびノイズ除去の精度が確認できる．

次に，各 Noise(1〜4) を付加した信号に対して SS 法によるノイズ除去を行った際の信号の振幅を図 2.2〜2.5 に示す．図 2.2〜2.5 に示すとおり，提案手法によって様々な SN 比の信号からノイズ除去が可能となっていることが確認できる．

表 2.1: ノイズ除去前の SN 比および SS 法によるノイズ除去後の SN 比

	Noise 1	Noise 2	Noise 3	Noise 4
ノイズ除去前の SN 比	18.78	12.74	0.09	-7.26
ノイズ除去後の SN 比	21.48	19.47	10.70	4.62

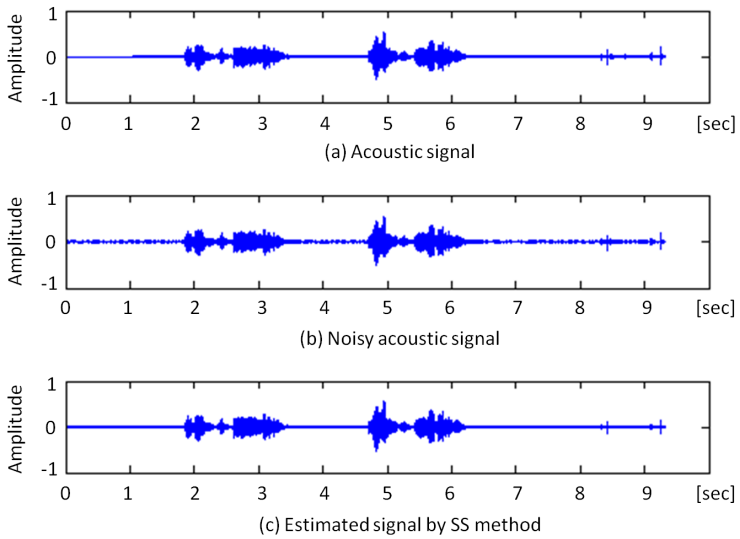


図 2.2: 表 2.1 に示す Noise 1 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

2.3.2　NLMS 法によるノイズ除去

あらかじめ用意された，実験対象となる映像の撮像された場所と同一の場所で作成された学習データから，発話区間を推定し，NLMS 法によりノイズを除去した際の実験結果を以下に示す．最初に，ノイズ除去前の SN 比およびノイズ除去後の SN 比について，表 2.2 に示す．Noise 1 の SN 比は，提案手法を適用した際に SN 比の向上が確認された最大の値であり，ノイズ除去前の SN 比 18.75 に対し，ノイズ除去後の SN 比は 18.98 となった．Noise 4 の SN 比は，提案手法により発話区間が正しく推定された際の最小の値

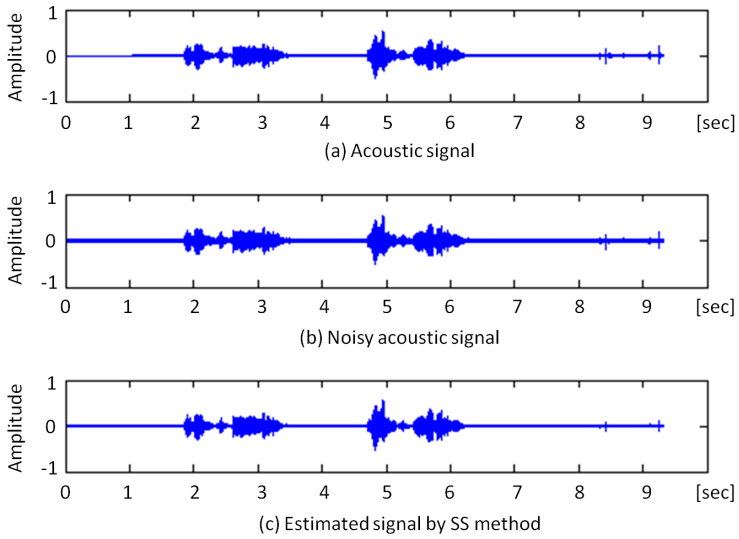


図 2.3: 表 2.1 に示す Noise 2 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

であり，ノイズ除去前の SN 比は -7.26，ノイズ除去後の SN 比は 1.32 になっている．この結果から，提案手法によりノイズ除去が可能となった SN 比の範囲およびノイズ除去の精度が確認できる．

次に，各 Noise(1〜4) を付加した信号に対して NLMS 法によるノイズ除去を行った際の振幅を，図 2.6〜2.9 に示す．図 2.6〜2.9 に示すとおり，提案手法によって様々な SN 比の信号からノイズ除去が可能となっていることが確認できる．

2.4　まとめ

本節では，ビデオ会議における映像，音響信号の両方を用いることで，発話区間を検出し，その結果から音響信号のノイズ除去を行う手法について検討を行った．実際に，検出結果に対して，SS 法，NLMS 法を適用することにより，SN 比の向上が得られることを確認した．ただし，これらの手法では，発話区間検出を行う際に統合する二つのモダリティの検出精度について十分に考慮した処理を行うことが実現されていない．この

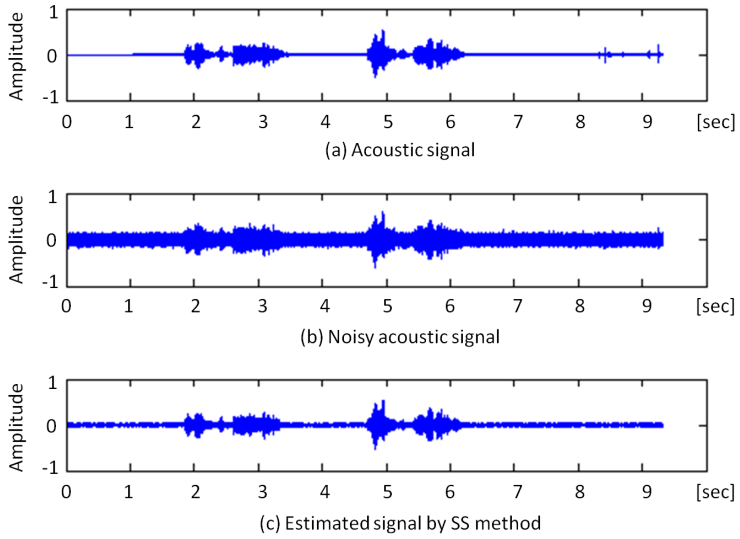


図 2.4: 表 2.1 に示す Noise 3 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

表 2.2: ノイズ除去前の SN 比および NLMS 法によるノイズ除去後の SN 比

	Noise 1	Noise 2	Noise 3	Noise 4
ノイズ除去前の SN 比	18.78	12.74	0.09	-7.26
ノイズ除去後の SN 比	18.98	14.50	4.31	1.32

課題を解決するためには，各モダリティによる発話区間の検出がどのような場合に的確に行われるかを判断可能な統合手法の実現が必要である．

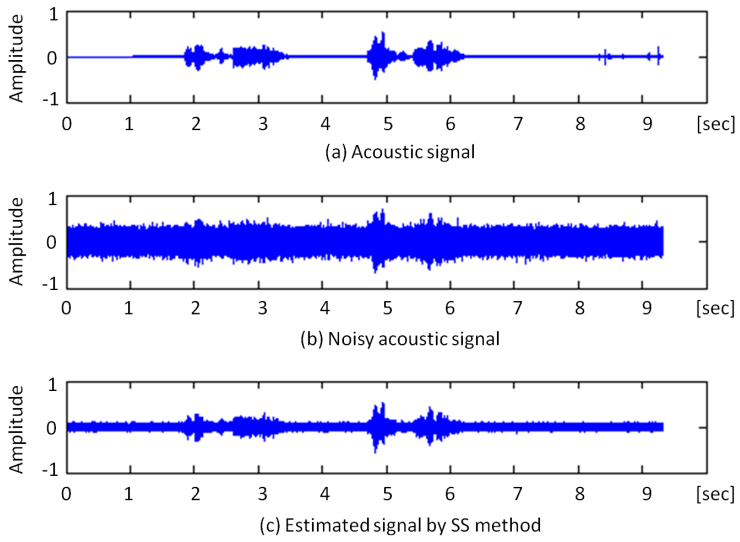


図 2.5: 表 2.1 に示す Noise 4 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

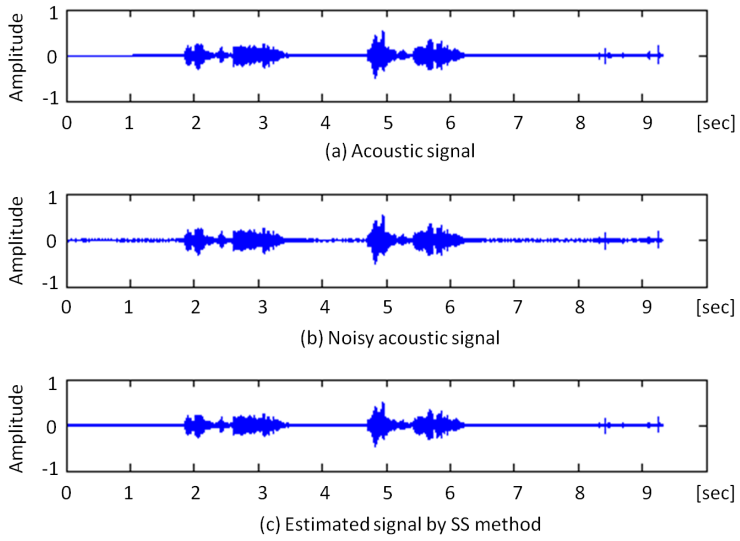


図 2.6: 表 2.2 に示す Noise 1 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

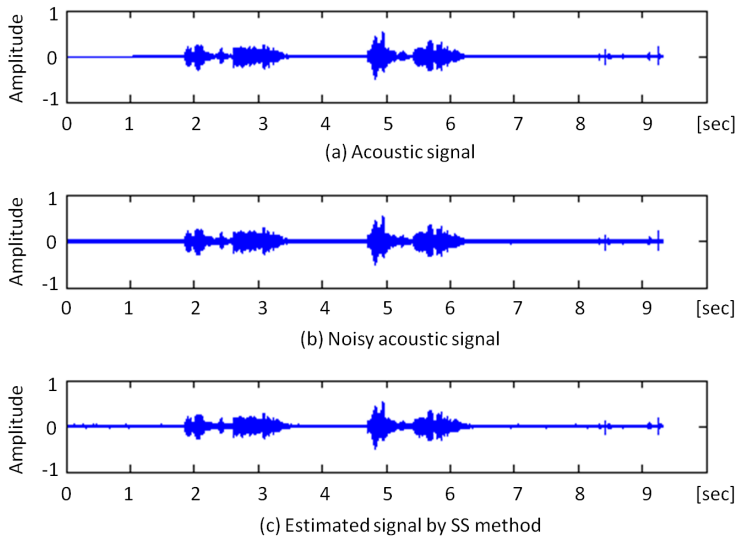


図 2.7: 表 2.2 に示す Noise 2 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

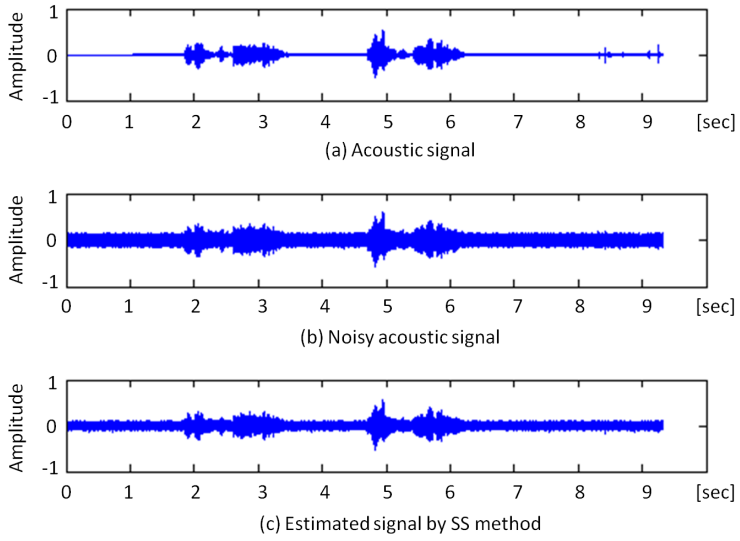


図 2.8: 表 2.2 に示す Noise 3 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

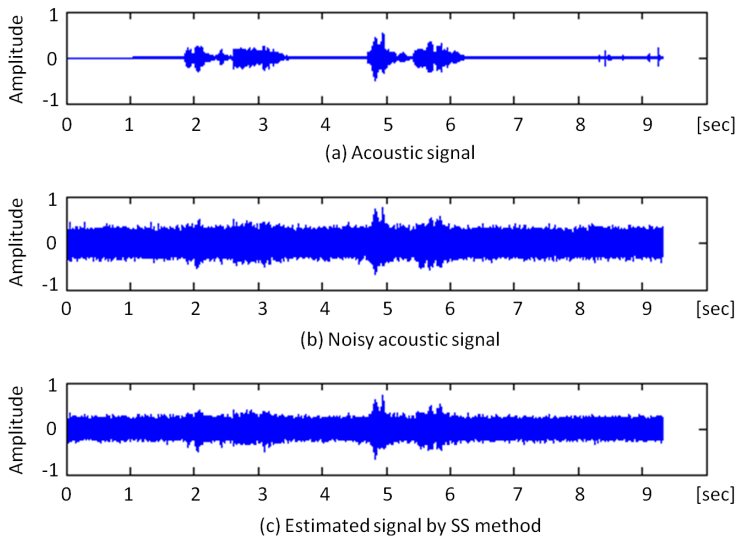


図 2.9: 表 2.2 に示す Noise 4 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

第3章　複数の参加者によるテレビ会議における発話区間抽出とノイズ除去への応用

3.1　はじめに

複数人が発話するノイズ環境下のテレビ会議において，カメラで撮像された発話者各々の映像と，得られた音響信号の協調的利用により，効率的にノイズ除去を可能にする手法について検討する．また，音響信号のノイズ除去については，伝統的なノイズ除去手法を用い，実験により提案手法が実現可能であることを示す．本提案手法を用いることにより, 複数人が発話するテレビ会議において, 様々な SN 比の信号に対してノイズ除去を行うことが可能となる．

3.2　提案手法

図 3.1 に提案手法の概要を示す．提案手法における発話区間の推定方法を説明する．提案手法では，発話区間の推定を，映像に撮像されているテレビ会議参加者各々の口の動きの取得および音響信号の分割・分類により行う．具体的に，映像における参加者各々の口の動きの取得による発話区間の抽出及び音響信号の分割・分類手法による発話区間抽出は，2.2 と同様の方法で取得することが可能である．さらに，各映像に撮像されたテレビ会議参加者各々の口の動きから取得された確率値と, 音声のオーディオクラスの帰属確率を各々かけた値を指標とし，いずれかの参加者の指標がしきい値以上であれば，その区間をその参加者の発話区間として推定する．

提案手法では，テレビ会議における特定の発話者における 前述の発話区間の推定結果により得られる非発話区間から抽出したノイズ情報を利用して，SS 法もしくは NLMS 法によって音響信号からノイズを除去する．

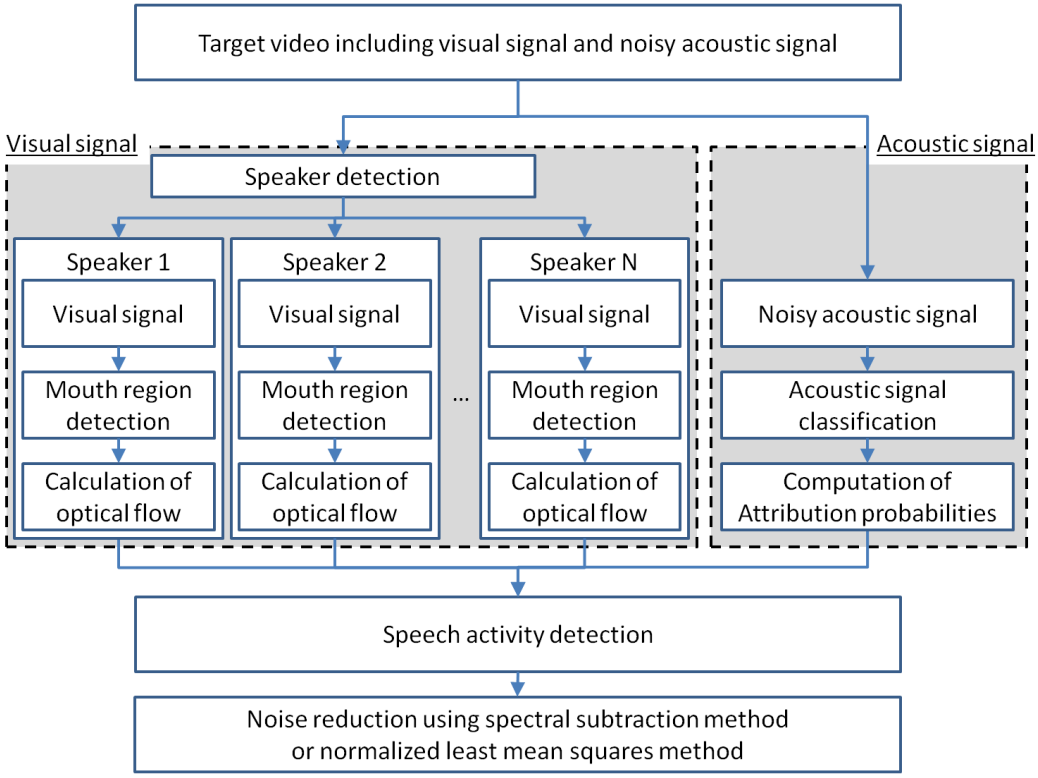


図 3.1: 本章で提案する複数の参加者によるテレビ会議における発話区間抽出手法のブロック図

3.3　実験結果

提案手法の有効性を示すため，実験結果を示す．本実験では，8 秒の時間長の映像 (480 × 360 画素，30fps，44.1kHz，iphone4s で取得) を用いた．この映像の音響信号に対して，ホワイトノイズを付与することで，ノイズの重畳した音響信号を作成し，この信号に対して提案手法の適用を行った．以下では，本実験で得られた結果について示す．

3.3.1　SS 法によるノイズ除去

提案手法により発話区間を推定し，SS 法によりノイズを除去した際の実験結果を示す．まず，ノイズ除去前の SN 比およびノイズ除去後の SN 比を表 3.1 に示す．Noise 1 の SN 比は，提案手法を適用した際に SN 比の向上が確認された最大の値であり，ノイズ除去

表 3.1: ノイズ除去前の SN 比および SS 法によるノイズ除去後の SN 比

	Noise 1	Noise 2	Noise 3	Noise 4
ノイズ除去前の SN 比	19.73	13.71	1.03	-6.30
ノイズ除去後の SN 比	21.20	19.31	10.37	4.75

前の SN 比 19.73 から，ノイズ除去後の SN 比は 21.20 になっている．また，Noise 4 の SN 比は，提案手法により発話区間が正しく推定された際の最小の値であり，ノイズ除去前の SN 比が -6.30 だったものが，ノイズ除去後の SN 比は 4.75 になっている．表 3.1 より，提案手法によるノイズ除去が可能となった SN 比の範囲およびノイズ除去の精度が確認できる．

次に，各 Noise(1〜4) を付加した信号に対して SS 法によるノイズ除去を行った際の信号の振幅を図 3.2〜3.5 に示す．図 3.2〜3.5 に示すとおり，提案手法を用いることで，様々な SN 比の信号で，SS 法によるノイズ除去が可能となっていることが確認できる．

3.3.2　NLMS 法によるノイズ除去

あらかじめ用意した，本実験対象となる映像が撮像された場所と同一の場所で作成された学習データを用い，発話区間を推定し，NLMS 法によりノイズを除去した際の実験結果を以下に示す．まず表 3.2 に，ノイズ除去前の SN 比およびノイズ除去後の SN 比の比較表を示す．Noise 1 の SN 比は，提案手法を適用した際に SN 比の向上が確認された最大の値であり，ノイズ除去前の SN 比 13.72 から，ノイズ除去後の SN 比は 14.33 になっており、Noise 3 の SN 比は，提案手法により発話区間が正しく推定された際の最小の値で，ノイズ除去前の SN 比が -6.30 だったものが，ノイズ除去後の SN 比は -1.13 になっている．この表より，提案手法によりノイズ除去が可能となった SN 比の範囲およびノイズ除去の精度が確認できる．

次に，各 Noise(1〜3) を付加した信号に対して，NLMS 法によるノイズ除去を行った際の振幅を，図 3.6〜3.8 に示す．図 3.6〜3.8 に示すとおり，提案手法を用いることで，様々な SN 比の信号で，NLMS 法によるノイズ除去が可能となっていることが確認できる．

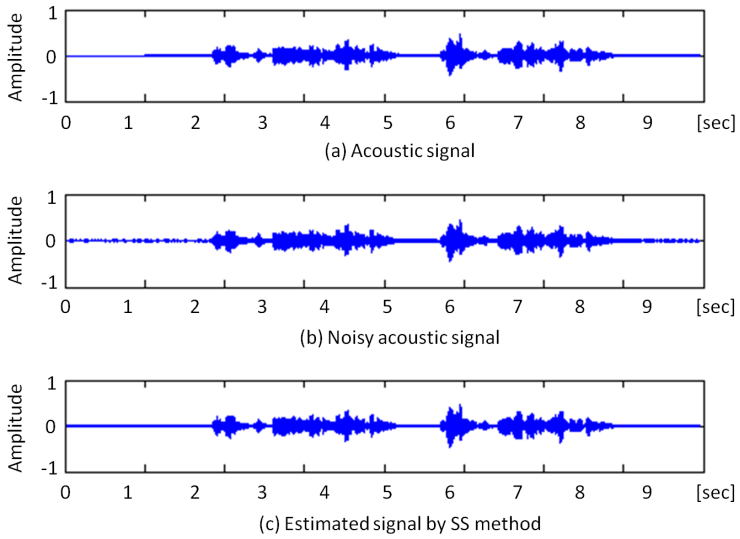


図 3.2: 表 3.1 に示す Noise 1 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

表 3.2: ノイズ除去前の SN 比および NLMS 法によるノイズ除去後の SN 比

	Noise 1	Noise 2	Noise 3
ノイズ除去前の SN 比	13.72	1.03	-6.30
ノイズ除去後の SN 比	14.33	2.57	-1.13

3.4　まとめ

本節では，複数人が発話するノイズ環境下のテレビ会議において，カメラで撮像された発話者各々の映像と，得られた音響信号の協調的利用により，効率的にノイズ除去を可能にする手法について検討を行った．また，テレビ会議における単独発話者の際の音響信号のノイズ除去とほぼ同水準の効果が得られることも確認できた。

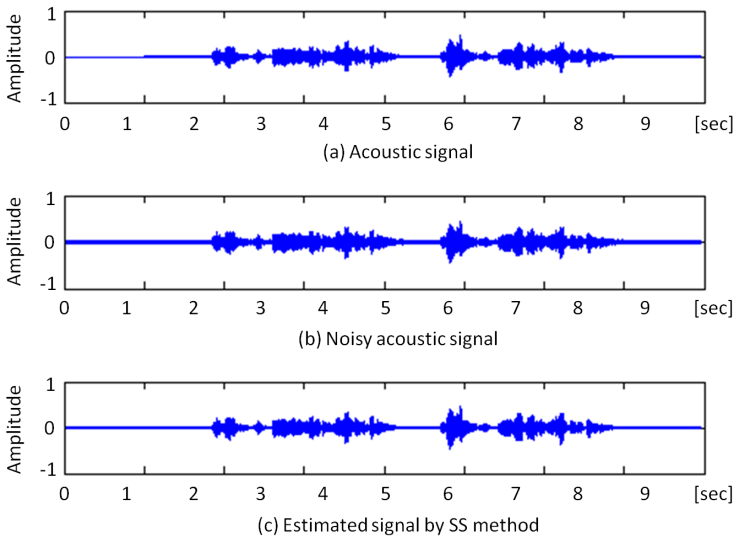


図 3.3: 表 3.1 に示す Noise 2 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

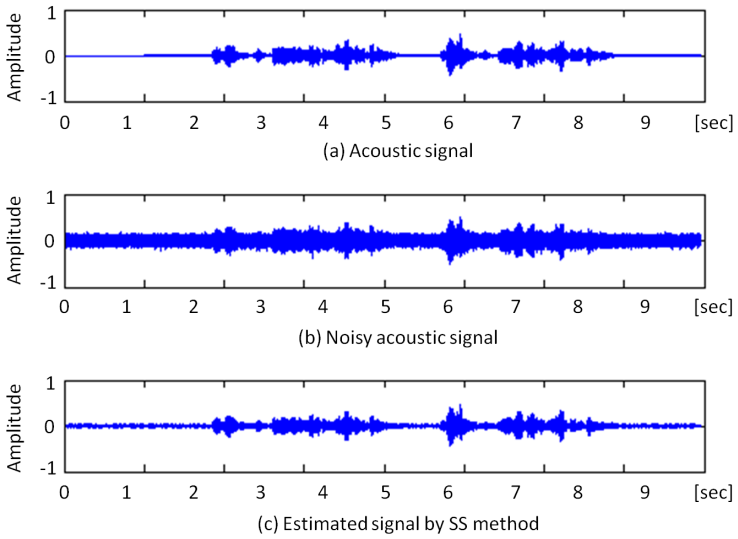


図 3.4: 表 3.1 に示す Noise 3 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

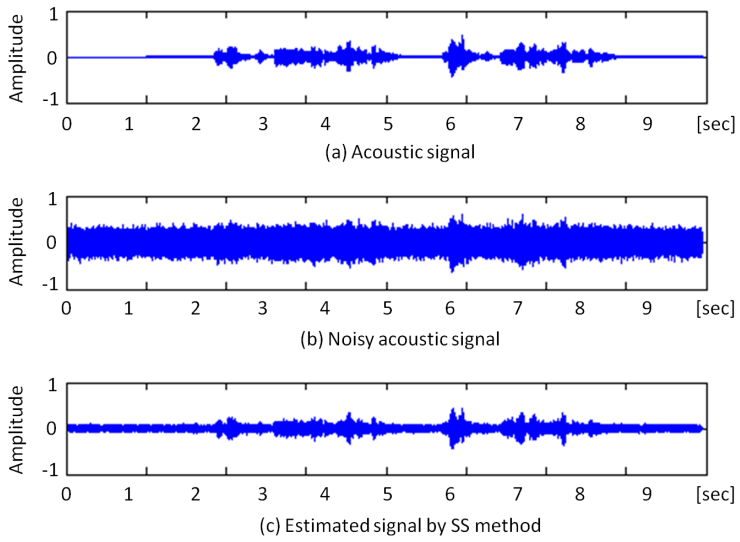


図 3.5: 表 3.1 に示す Noise 4 を付加した信号に対して SS 法によるノイズ除去を行った際の各信号の振幅

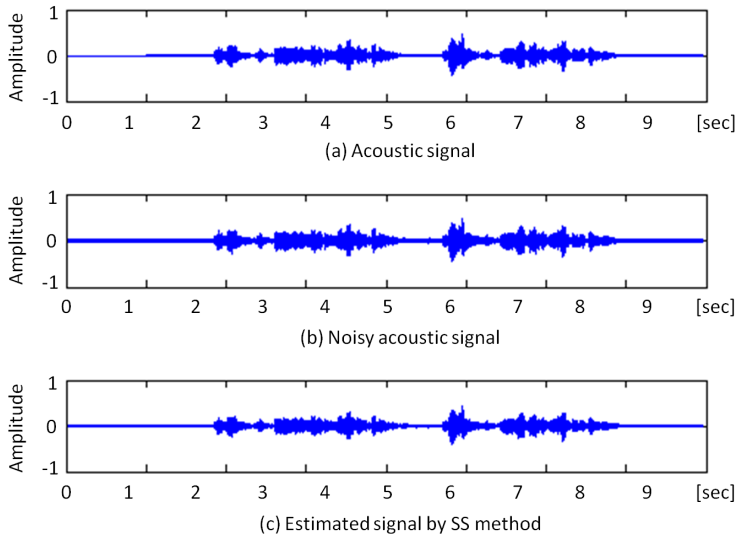


図 3.6: 表 3.2 に示す Noise 1 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

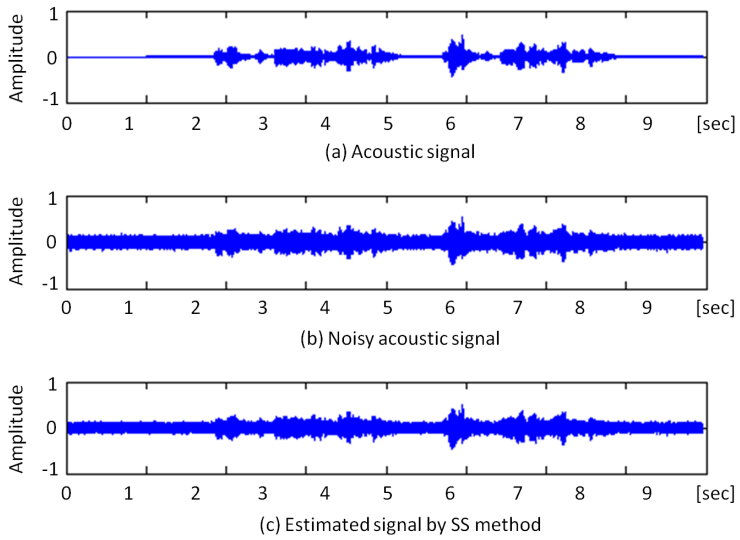


図 3.7: 表 3.2 に示す Noise 2 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

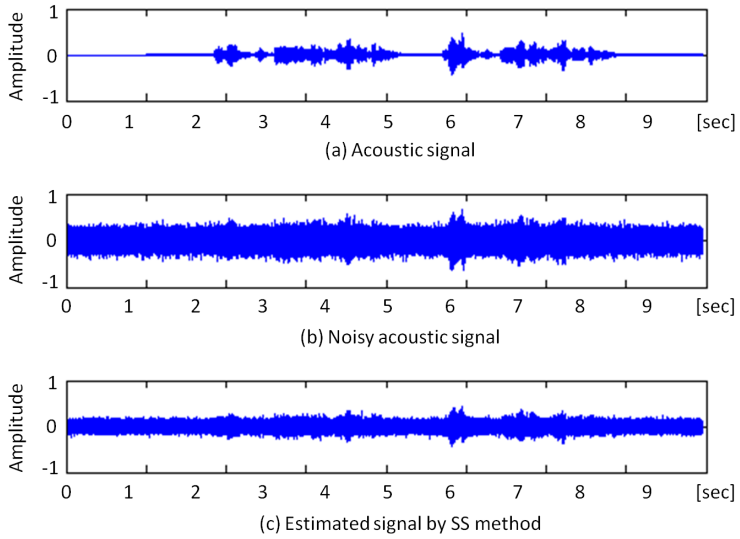


図 3.8: 表 3.2 に示す Noise 3 を付加した信号に対して NLMS 法によるノイズ除去を行った際の各信号の振幅

第4章　Decision Level Fusionに基づく発話区間検出と音声雑音除去への応用

4.1　はじめに

本章では，Decision Level Fusion (DLF) に基づいたテレビ会議における発話区間検出手法とその音声雑音除去への応用について提案する．提案手法ではまず，テレビ会議における参加者の口領域を抽出し，その動きの特徴から Support Vector Machine (SVM) [11] に基づいて，発話区間検出を行う．さらに，文献 [10] により提案されているオーディオクラスの分類手法により得られる 4 つのクラスの帰属度を用いた SVM による発話区間検出も合わせて行う．これらの 2 つの手法により得られる複数の発話区間検出結果から，我々は，各々の検出結果の性能を考慮して高精度に結果の統合を可能とする Supervised Learning from Multiple Experts [12] を導入し，DLF に基づく発話区間検出結果の統合を実現する．提案手法では、これらの検出結果の性能を考慮しながら，各モダリティの分類結果と音響画像特徴間の関連性を学習することが可能である．したがって，提案手法が導入する DLF に基づいた発話区間検出結果の統合では，各モダリティが，どの程度の精度で分類が可能になるかに関して考慮することが可能となり，統合による発話区間検出精度の向上が期待できる．以上のようにして得られる発話区間検出結果に基づいて，提案手法では音声雑音の除去を可能とする．具体的に，残りの非発話区間における音響信号の情報を用いることで，提案手法ではそれらの純粋なノイズ情報の取得が可能となり，その情報を用いることで，発話区間におけるノイズの除去が可能になる．

提案手法における最大の新規点は，DLF アプローチを導入した音響画像信号のマルチモーダル処理による発話区間検出という点にある．一方，これまで，Bagging [13] や Boosting [14]，さらには，Random Forest [15]， Bayesian Network [16] などが ensemble learning の手法として提案されている．その中でも，提案手法は，以下の理由により，上

記に示す DLF アプローチ，すなわち，Raykar らによって提案された手法に基づく手法 [12] を採用している．

- 各参加者，各ノイズ環境に対して，大量の学習サンプルを準備することは困難であり，限られた学習サンプルから高精度に複数の分類結果を統合可能とする手法が必要となる．
- 互いに分類精度が異なる SVM に基づく分類結果に対して，それらの分類精度と識別器に入力される特徴との間の関連性を考慮可能な統合手法が必要である．

Raykar らによって提案されている DLF アプローチ，すなわち，Supervised Learning from Multiple Experts に基づく手法は，上記二点を満たしていることから，提案手法において採用した．本手法は，医用画像処理の分野において，読影の能力が異なる放射線技師の読影結果を高精度に統合するために実現された computer-aided diagnosis (CAD) システムを実現するために提案されている．また，この手法では，各放射線技師の読影の能力と特徴量の両方を考慮して DLF による統合を行うことが可能である．したがって，複数人の放射線技師を，複数の SVM に基づいた識別器とみなすことで，本 DLF アプローチは本研究が対象とする問題に適用可能となる．

4.2　提案手法

本節では，DLF に基づいた発話区間検出手法について説明する．提案手法では，対象の映像内の各セグメントを二つのクラス，すなわち，発話区間と非発話区間に分類することで，発話区間検出を実現する．具体的に，提案手法は二つのフェーズにより構成される．まず，第一フェーズにおいて，我々は発話区間と非発話区間を分類する二種類の識別器を構築する．これらの二つの識別器は映像シーケンスより得られる画像特徴量と音響信号から得られる音響特徴量からそれぞれ構築される．これらの二種類の識別器を用いることにより，二種類の分類結果を得ることが可能となる．次に，第二フェーズでは，DLF を利用し，上記二種類の識別器の分類精度を考慮しながら，分類結果の統合を行う．各フェーズの詳細については，以下の節においてそれぞれ説明する．また、提案

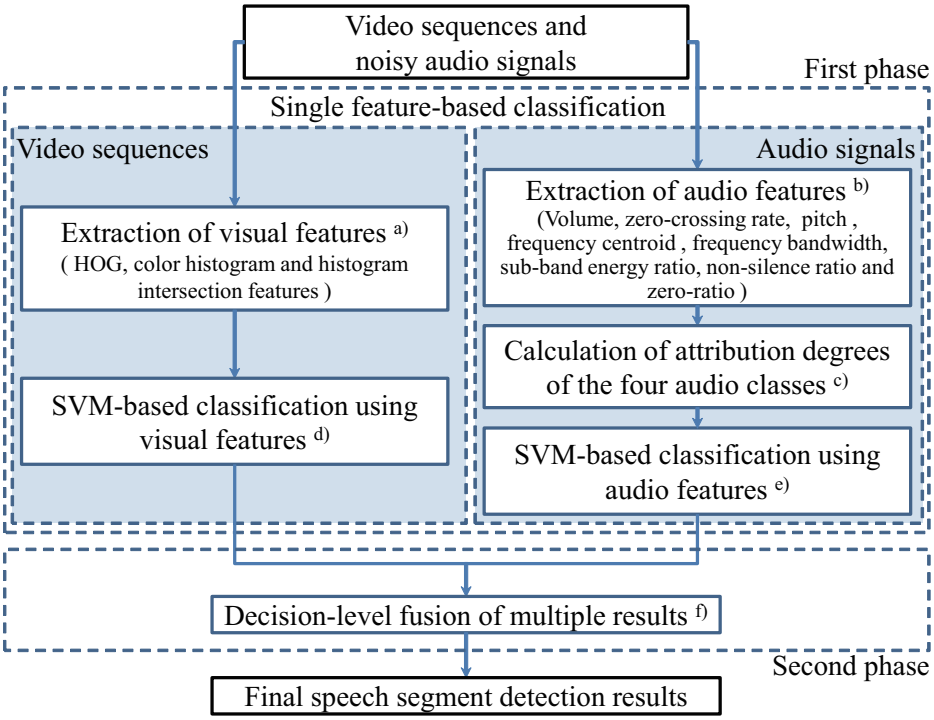


図 4.1: 本章で提案する発話区間検出手法のブロック図

手法のブロック図を図 4.1 に示す.

4.2.1　画像・音響特徴に注目した発話区間検出

本節では，提案手法の第一フェーズについて説明を行う．具体的にはまず，提案手法において用いられる画像特徴量および音響特徴量について説明を行い，さらに，各々の特徴量に基づいた発話区間検出手法について説明を行う.

4.2.1.1　画像特徴量の抽出

映像シーケンスの各フレーム ($X \times Y$ 画素) より，提案手法では参加者の口領域から画像特徴量の抽出を行う．まず，提案手法では，Adaboost [17] に基づき参加者の顔画像領

域の検出を行い，図 4.2 に示すように，得られる顔領域の下部を口領域 ($X_{\mathrm{m}} \times Y_{\mathrm{m}}$ 画素) としてｓ抽出する．次に，得られる口領域より，局所的な画像の変化や照明変化にロバストな HOG 特徴量 [18] を算出する．具体的に，提案手法では $\frac{X_{\mathrm{m}}}{2} \times \frac{Y_{\mathrm{m}}}{2}$ 画素のブロック，$\frac{X_{\mathrm{m}}}{6} \times \frac{Y_{\mathrm{m}}}{6}$ 画素のセルを利用し，各ブロックは 3×3 のセルにより構成される．また，上記のブロックは，$\frac{X_{\mathrm{m}}}{6} \times \frac{Y_{\mathrm{m}}}{6}$ 画素の間隔 (セルのサイズと同様の間隔) で得られており，口領域から 16 ブロックが抽出されている．さらに，各ブロックでは，9 ビンの勾配ヒストグラムが各セルにおいて算出されている．したがって，提案手法では各ブロックから $3 \times 3 \times 9 = 81$ の特徴量が算出され，$16 \times 81 = 1296$ 次元の HOG 特徴量が得られる．

以上の HOG 特徴量に加え，提案手法では，以下の二種類の色ヒストグラムに基づく特徴量を抽出された口領域から算出する．まず，提案手法では，抽出された口領域から色ヒストグラムを算出する．このとき，色ヒストグラムのビン数，すなわち，特徴量の次元は，各 RGB チャネル毎に 8 とし，合計 512 に設定されている．次に，提案手法では，上記で得られた色ヒストグラム特徴ベクトルと非発話区間から合成された色ヒストグラム特徴ベクトル間で得られる 512 次元の色ヒストグラムインタセクションベクトルを求める．このとき,「非発話区間から合成された色ヒストグラム特徴ベクトル」は，音響特徴量に基づいて検出された非発話区間の色ヒストグラムベクトルからベクトルメディアンを求めることにより算出される．なお，音響特徴量に基づいた発話区間/非発話区間の検出方法は，本節の以降で説明する．一般に，音響特徴量に基づく発話区間検出の精度は，画像特徴量に基づく検出精度に比べて高いことから，提案手法では，音響特徴量に基づいて検出された非発話区間から得られる色ヒストグラム特徴ベクトルを，非発話区間を代表する合成色ヒストグラムベクトルを算出するために利用する．以上に示す色ヒストグラムインタセクションベクトルの算出は，背景差分特徴量を求める手法と類似した役割となる．

以上のようにして，2320(= 1296 + 512 + 512) 次元の画像特徴量を得ることが可能となる．提案手法では，口が開いているか，閉じているかを検出するため，唇の形状とその色の分布をそれぞれ表現可能な HOG 特徴およびカラーヒストグラムを利用している．さらに，口の動きを検出するため，ヒストグラムベースの特徴量と非発話区間における特徴量との間の差分を見るためのヒストグラムインタセクション特徴量を算出している．我々

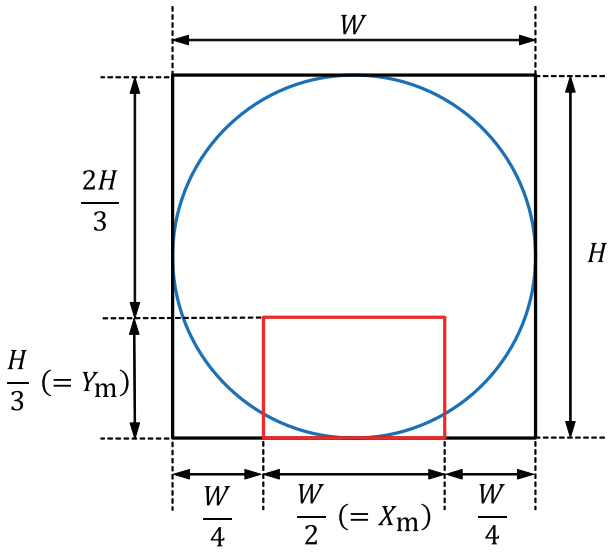


図 4.2: 提案手法における口領域の抽出．青い円は，Adaboost [17] によって検出される顔領域であり，黒い四角は青い円を包含する領域であり，赤い四角 (口領域) は，黒い四角の領域に基づいて抽出される．

は，口の動きを表現するための特徴量として，色ヒストグラムに基づく特徴量を用いているが，これは提案手法の計算時間を削減するためであり，この理由から，HOG 特徴量に基づくインタセクション特徴量を算出していない．ただし，発話区間検出精度をより高精度に画像特徴量から求めるためには，ダイナミックテクスチャ[19] や CHLC (Cubic Higher-order Local Auto Correlation) [20] 等の時空間特徴量を導入することが有効であると考えられるため，その高度化については，さらなる検討が必要である．

4.2.1.2　音響特徴量の抽出

提案手法ではまず，音響信号の分割を等間隔で行う．このとき，分割間隔は，映像シーケンスの連続するフレーム間隔と同一とする．また、各音響セグメントに対して，音響信号分類手法 [10] を適用することで，音響特徴量を抽出する．具体的に，各音響セグメントに対し，音響信号の分類手法を適用し，我々は無音，スピーチ，音楽，ノイズの 4 つのクラスに属する度合いを算出する．文献 [10] ではまず，volume, zero-crossing rate,

pitch, frequency centroid, frequency bandwidth, sub-band energy ratio, non-silence ratio and zero-ratio に基づく音響特徴量を求め，さらに，それらに主成分分析を適用することで，次元削減を行う．次に，マハラノビス距離に基づいて対象の音響信号の事前クラスタリングを行い，各クラスへの帰属確率を算出する．最後に，得られる帰属確率から音響信号の分類をファジーアルゴリズム [21],[22],[23] に基づいて行うことにより，4 つのクラス (無音，スピーチ，音楽，ノイズ）への帰属度を算出可能とする．このようにして得られる 4 次元の帰属度のベクトルを，提案手法では音響特徴ベクトルとして，各音響セグメントごとに取得する．

4.2.1.3　各特徴に基づいた分類

本節では，前節までに得られる画像特徴量および音響特徴量をそれぞれ用いた発話区間/非発話区間の分類手法 (第一フェーズ) について説明する．前節まで示してきたとおり，提案手法では画像特徴量および音響特徴量を算出することが可能となる．このようにして算出された画像特徴量に対して，我々は，Max-Relevance and Min-Redundancy (mRMR) アルゴリズム [24] を適用し，特徴選択を行うことで，分類に有効な画像特徴量のみを抽出する．一方，4 次元の音響特徴量については，提案手法では，特徴選択を行わず，そのまま利用する．提案手法では，低次の画像特徴量を求めているが，これらのすべてが発話区間/非発話区間の分類に有効であるとは限らない．したがって，提案手法では，特徴選択アルゴリズムの中でも代表的な mRMR アルゴリズムを導入し，不要な画像特徴量の除去を試みている．一方，音響特徴量は音響信号分類手法に基づいたソフトインデキシングの結果であり、音響特徴量は画像特徴量に比べて高次の特徴量となるため，mRMR アルゴリズムを適用していない．

提案手法では，画像特徴ベクトルおよび音響特徴ベクトルの各々を用いた 2 種類の識別器を別々に構築する．今，N 個の特徴ベクトル $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$ $(i = 1, 2, \cdots, N)$ により構成される学習データセットが与えられ，それらに対応するラベル $y_i^{\mathrm{svm}} \in \{1, -1\}$ が与えられているとき，特徴ベクトル $\tilde{\boldsymbol{x}}_i$ は，カーネル関数が Gaussian カーネルの非線形写像により高次元の特徴空間へ射影され，$\psi(\tilde{\boldsymbol{x}}_i) \in \mathbb{R}^{d'} (d' \gg d)$ が得られる．提案手法では，上記の学習データセットを利用し，SVM 識別超平面の算出を文献 [11] に基づいて行う．発

話区間と非発話区間の 2 つのクラスの分割は，重みベクトル $\boldsymbol{v} \in \mathbb{R}^{d'}$ とバイアス $\beta \in \mathbb{R}$ により定義される最適な超平面を求めることで可能となる．具体的に，最適な超平面は以下のように定義される．

$$f^{\mathrm{svm}}(\tilde{\boldsymbol{x}}) = \boldsymbol{v}^{\top}\psi(\tilde{\boldsymbol{x}}) + \beta \tag{4.1}$$

このとき， $\tilde{\boldsymbol{x}}$ は，そのクラスラベル (発話区間/非発話区間) が未知の入力ベクトルである．さらに，$\tilde{\boldsymbol{x}}$ のクラスラベルは，次式にしたがって決定することができる．

$$y^{\mathrm{svm}} = \mathrm{sign}[f^{\mathrm{svm}}(\tilde{\boldsymbol{x}})] \tag{4.2}$$

このとき，y^{svm} が正のとき，$\tilde{\boldsymbol{x}}$ はラベルが 1 の正例クラスに属し，負であるときには，ラベルが -1 の負例クラスに属する．このようにして，新しい入力ベクトル $\tilde{\boldsymbol{x}}$ に対して，SVM に基づく分類が可能となる．提案手法では，限られた数の学習データに対して汎化性を有する代表的な分類手法の一つである SVM を導入することで，高精度な発話区間/非発話区間の分類を可能とする．

SVM では，マージン最大化と誤差最小化の二つの指標により構成されるコスト関数を最小化する最適な超平面を求める．ラグランジュ未定乗数法に基づく式 (4.1) の最終結果は，次式のように与えられる [11].

$$f^{\mathrm{svm}}(\tilde{\boldsymbol{x}}) = \sum_{i=1}^{N} \alpha_i y_i^{\mathrm{svm}} \kappa(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}) + \beta, \tag{4.3}$$

上式において，$\kappa(\cdot,\cdot)$ はカーネル関数であり，特徴空間 $\mathbb{R}^{d'}$ における内積を与える．なお，提案手法では，ガウシアンカーネルを利用する．次に，係数 α_i は，次式の convex

Quadratic Programming (QP) 問題の解を求めることにより算出することができる．

$$\begin{aligned} &\text{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i^{\mathrm{svm}} y_j^{\mathrm{svm}} \kappa(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j), \\ &\text{subject to} \quad \sum_{i=1}^{N} \alpha_i y_i^{\mathrm{svm}} = 0, \\ &\qquad\qquad 0 \leq \alpha_i \leq C, \ (i = 1, 2, \cdots, N), \end{aligned} \tag{4.4}$$

ただし，α_i はラグランジュ未定乗数であり，C は学習データセットにおける誤分類数と最大マージンにおける分離性能との間のトレードオフを決定するパラメータである．

提案手法では，対象の映像の各セグメンから得られる画像特徴ベクトルと音響特徴ベクトルをそれぞれ $\tilde{\boldsymbol{x}}$ とみなし，それらを式 (4.3) に基づいて得られる SVM の識別器に入力する．このとき，画像特徴ベクトルに基づいた分類においてインパルス状の未検出が生じる可能性がある．したがって、提案手法では，非発話区間を中心とする 30 セグメントの長さのスライディングウィンドウを利用し，その半数以上が発話区間である場合，対象のセグメントを発話区間として補正する．以上のようにして，各セグメントごとに 2 種類の分類結果を得ることが可能となる．

4.2.2　複数の分類結果の DLF による統合

本節では，提案手法の第二フェーズ，すなわち，前節により得られる二種類の分類結果の DLF に基づく統合手法について説明を行う．提案手法では，Supervised Learning from Multiple Experts [12] を用いて二種類の分類結果の統合を行う．

提案手法において利用する DLF は computer-aided diagnosis (CAD) の分野において提案されており，元々，複数の人間のアノテーター (複数の放射線技師) による読影結果の統合を行っている．一方，本節において，我々は画像・音響特徴量に基づいて構築される二つの識別器をアノテーターとみなす．これら二つの分類結果を統合するため，我々は各アノテーターの分類精度に注目し，より高い分類精度を有するアノテーターの分類結果に対して高い重みを付与する．提案手法の具体的な詳細については，以下で説明する．

なお，提案手法では参加者の顔が検出できなかった場合，画像特徴量を求めることができず，SVM に基づく分類結果も得ることはできない．このような場合は，提案手法では以下の DLF の処理を行わずに，単純に音響特徴量に基づく分類結果を最終結果として出力する．

4.2.2.1　各アノテーターの分類精度と識別モデル

以下に，各アノテーターの分類精度と提案手法で用いる DLF の識別モデルについて説明する．今，$y^j \in \{0,1\}$ をアノテーター $j \in \mathcal{J}$ により付与された特徴ベクトル $\boldsymbol{x}$ に対するラベルとする．ただし，$\mathcal{J} = \{\mathrm{v}, \mathrm{a}\}$ はアノテーターの集合を表し，“v” と “a” は，画像と音響のアノテーターに対応している．さらに，$y^j = 0$ および 1 は，それぞれ発話および非発話に対応している．特徴ベクトル $\boldsymbol{x}$ の詳細については，以降で説明を行う．真のラベル $y \in \{0,1\}$ が与えられたとき，二つのアノテーターの分類精度として定義される P^j_{se} (sensitivity) and P^j_{sp} (specificity) は次式で算出される．

$$P^j_{se} \quad := \Pr[y^j = 1 | y = 1] \tag{4.5}$$

$$P^j_{sp} \quad := \Pr[y^j = 0 | y = 0] \tag{4.6}$$

提案手法では，具体的な識別モデルを次式のように定義する．

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} \tag{4.7}$$

ただし，$\boldsymbol{w}$ は重みベクトルである．提案手法の第二フェーズにおいて，特徴ベクトル $\boldsymbol{x} \in \mathbb{R}^r$ は各々の要素が画像及び音響特徴量を用いた分類における式 (4.3) の SVM 出力値であり，2 次元の特徴ベクトル ($r = 2$) となる．これまで、SVM の出力値に注目した

様々な手法が提案されており，さらに，SVM 出力値は SVM 識別器における分類結果の信用度として利用することが可能であることが報告されている [25]．画像特徴量および音響特徴量に基づく SVM 識別結果の信用度は，DLF を行う際の重要な特徴量となることから，提案手法では，それらの値を要素とするベクトルを式 (4.7) の $\boldsymbol{x}$ として採用する．得られるベクトルに基づいた最終的な分類結果 $\hat{y}$ は，次式により与えられる．

$$\hat{y} = \begin{cases} 1 & f_w(\boldsymbol{x}) \geq Th \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

ただし，Th は事前に決定される閾値である．先に述べた通り，$\hat{y} = 1$ および 0 は，それぞれ発話区間および非発話区間を表す．したがって，これらは，前節までにおいて示されたラベル $+1$ および -1 にそれぞれ対応する．

4.2.2.2　学習段階

今，二つのアノテーターによる分類結果が付与された N 個の学習データ $\mathcal{D} = \{\boldsymbol{x}_i, y_i^{\mathrm{v}}, y_i^{\mathrm{a}}\}_{i=1}^{N}$ が，それらの真のラベル y_i $(i = 1, 2, \cdots, N)$ とともに与えられているとする．識別関数 $f_w(\cdot)$ を得るため，我々は重みベクトル $\boldsymbol{w}$ を，式 (4.5) および (4.6) に示す各アノテーターの分類精度を用いることで，学習データから求める必要がある．以下では，その具体的な処理を説明する．

学習データ $\mathcal{D}$ から，$\boldsymbol{w}$ に関する尤度は次式により与えられる．

$$\Pr\left[\mathcal{D}|\boldsymbol{w}\right] = \prod_{i=1}^{N} \Pr\left[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|\boldsymbol{x}_i, \boldsymbol{w}\right] \tag{4.9}$$

さらに，各アノテーターから得られる sensitivity $\boldsymbol{P_{se}} = \{P_{se}^j | j \in \mathcal{J}\}$ および specificity $\boldsymbol{P_{sp}} = \{P_{sp}^j | j \in \mathcal{J}\}$ から，上式は以下のように書き換えられる．

$$\begin{aligned} \Pr[\mathcal{D}|\boldsymbol{w}] \quad = \quad & \prod_{i=1}^{N} \Big\{ \Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 1, \boldsymbol{P_{se}}] \Pr[y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}] \\ & + \Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 0, \boldsymbol{P_{sp}}] \Pr[y_i = 0|\boldsymbol{x}_i, \boldsymbol{w}] \Big\} \end{aligned} \tag{4.10}$$

ここで，各アノテーター $j \in \mathcal{J}$ は互いに独立であると仮定すると，$\Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 1, \boldsymbol{P_{se}}]$ は以下のように書き換えられる．

$$\begin{aligned}\Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 1, \boldsymbol{P_{se}}] &= \prod_{j\in\mathcal{J}} \Pr[y_i^j|y_i = 1, P_{se}^j] \\ &= \prod_{j\in\mathcal{J}} [P_{se}^j]^{y_i^j}[1 - P_{se}^j]^{1-y_i^j}\end{aligned} \tag{4.11}$$

同様にして，$\Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 0, \boldsymbol{P_{sp}}]$ も以下のように書き換えられる．

$$\begin{aligned}\Pr[y_i^{\mathrm{v}}, y_i^{\mathrm{a}}|y_i = 0, \boldsymbol{P_{sp}}] &= \prod_{j\in\mathcal{J}} \Pr[y_i^j|y_i = 0, P_{sp}^j] \\ &= \prod_{j\in\mathcal{J}} [P_{sp}^j]^{1-y_i^j}[1 - P_{sp}^j]^{y_i^j}\end{aligned} \tag{4.12}$$

その結果，式 (4.10) に示す尤度は，

$$\Pr[\mathcal{D}|\boldsymbol{w}] = \prod_{i=1}^{N} [a_i p_i + b_i(1 - p_i)] \tag{4.13}$$

となる．ただし，

$$\begin{aligned}p_i &= \Pr[y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}] \\ &= \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i) \\ &= \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x}_i)}\end{aligned} \tag{4.14}$$

であり，

$$a_i = \prod_{j\in\mathcal{J}} [P_{se}^j]^{y_i^j}[1 - P_{se}^j]^{1-y_i^j} \tag{4.15}$$

$$b_i \quad = \quad \prod_{j \in \mathcal{J}} [P_{sp}^j]^{1-y_i^j} [1 - P_{sp}^j]^{y_i^j} \tag{4.16}$$

である．上記に示す尤度最大化の解は，以下に示す対数尤度の最大化を行うことで取得することが可能となる．

$$\hat{\boldsymbol{w}}_{ML} = \arg\max_{\boldsymbol{w}} \Big\{ \ln \Pr[\mathcal{D}|\boldsymbol{w}] \Big\} \tag{4.17}$$

なお， Raykar らによって提案されている DLF 手法 [12] は，教師なし学習アプローチであるが，提案手法の学習段階では，真の分類結果が存在し，これを用いることが可能であることから，我々は定式化の変更を行う．具体的に，Raykar らによって提案されている DLF 手法 [12] では，真の分類結果が未知の条件で，複数の分類結果の統合を可能とするものである．一方，提案手法では，ラベル y_i を直接利用し，それらを真値として利用することで，教師あり学習を実現する．今，$\boldsymbol{y} = [y_1, \cdots, y_N]$ を真値の集合としたとき，対数尤度は次式の通り記載できる．

$$\ln \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{w}] = \sum_{i=1}^{N} \big\{ y_i \ln p_i a_i + (1 - y_i) \ln(1 - p_i) b_i \big\} \tag{4.18}$$

提案手法では，ベクトル $\boldsymbol{w}$ を上記の尤度が最大となるように推定する．具体的に，

$$\frac{\partial}{\partial \boldsymbol{w}} \{ \ln \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{w}] \} = 0 \tag{4.19}$$

を解くことで，解を求めため，次式の更新により最終的な $\boldsymbol{w}$ の最適解を得る．

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \boldsymbol{H}^{-1} \boldsymbol{g} \tag{4.20}$$

式 (4.20) において，$\boldsymbol{g}$ は勾配ベクトル，$\boldsymbol{H}$ はヘシアン行列であり，η は学習係数である．勾配ベクトル $\boldsymbol{g}$ とヘシアン行列 $\boldsymbol{H}$ はそれぞれ次式のように与えられる．

$$\boldsymbol{g} = \sum_{i=1}^{N} [y_i - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)]\boldsymbol{x}_i \tag{4.21}$$

$$\boldsymbol{H} = -\sum_{i=1}^{N} [\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)][1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)]\boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{4.22}$$

ただし，$\sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$ は式 (4.14) に示される通りである．提案手法では，式 (4.20) に示す繰り返し処理を $||\triangle \boldsymbol{w}_t||_2^2 < 0.001$ となるまで行う．なお，$\triangle \boldsymbol{w}_t = \boldsymbol{w}_t - \boldsymbol{w}_{t-1}$ であり，$\boldsymbol{w}_t$ は t 回目の繰り返しで得られる $\boldsymbol{w}$ である．

4.2.2.3　テスト段階

今，ラベル y が未知のテストデータ $\{\boldsymbol{x}, y^{\mathrm{v}}, y^{\mathrm{a}}\}$ が与えられたとき，最終的な分類結果は以下のようにして与えられる．具体的にはまず，以下の確率を算出する

$$p = \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x})} \tag{4.23}$$

さらに,

$$a = \prod_{j \in \mathcal{J}} [P_{se}^j]^{y^j} [1 - P_{se}^j]^{1-y^j} \tag{4.24}$$

$$b = \prod_{j \in \mathcal{J}} [P_{sp}^j]^{1-y^j} [1 - P_{sp}^j]^{y^j} \tag{4.25}$$

とするとき，アノテーター j の P_{se}^{j} および P_{sp}^{j} は学習データから求められ，y^{j} はアノテーター j によりテストデータの分類結果として得られる．したがって、提案手法では，各アノテーターの分類精度を考慮しながら，最終的な分類結果を求めることが可能となる．さらに，μ は p, a および b を用いて次式で与えられる．

$$\begin{aligned} \mu &= \Pr[y=1|y^{\mathrm{v}}, y^{\mathrm{a}}, \boldsymbol{x}, \boldsymbol{w}] \\ &\propto \Pr[y^{\mathrm{v}}, y^{\mathrm{a}}|y=1, \boldsymbol{w}]\Pr[y=1|\boldsymbol{x}, \boldsymbol{w}] \\ &= \frac{ap}{ap+b(1-p)}. \end{aligned} \tag{4.26}$$

上式に示す μ は，真値に対する確率的推定値を表している．したがって，提案手法では最終的な分類結果を式 (4.8) に代えて，次式により算出する．

$$\hat{y} = \begin{cases} 1 & \mu \geq T_{\mu} \\ 0 & \text{otherwise} \end{cases} \tag{4.27}$$

ただし，T_{μ} は事前に定められる閾値である．

以上のようにして，提案手法では複数の分類結果を統合することにより発話区間の高精度な検出を可能とする．最後に，画像特徴量を用いた際と同様にスライディングウィンドウを用いた発話区間の未検出の抑制を適用することで，最終的な検出結果を取得する．

4.3　実験結果

本節では，提案手法の性能を評価するため，比較実験を行う．本実験では，10 本の映像を利用し，各々の映像は 30 秒，1280×720 画素，30 fps，44.1 kHz である．したがって，各セグメントの長さは $\frac{1}{30}$ 秒となる．これら 10 本の映像の中で，5 本の映像は男性による会話の映像，残りの 5 本の映像は女性による会話の映像である．各映像は，約 15 秒の発話区間が存在し，残りの約 15 秒が非発話区間に対応する．

以下，発話区間の検出結果について 4.3.1 で示し，さらに，それらの結果に基づいたノイズ除去に関する結果を 4.3.2 に示す.

4.3.1　発話区間検出の性能評価

本節では，提案手法における発話区間検出の性能評価の結果を示す．先に述べた通り，本実験では，提案手法の性能の評価のため，10 本の映像を利用し，これらに対して，leave-one-out 法を適用することにより精度の検証を行う．実験では，比較手法として一方のモダリティのみを用いた手法，すなわち映像特徴量または音響特徴量のみを用いた手法を用いる．したがって，これらの手法では，画像特徴量または音響特徴量を SVM に入力した際の結果を最終的な分類結果として採用する．さらに，提案手法の DLF に基づく統合手法の有効性を確認するため，early fusion (feature-level fusion) に基づく統合手法，majority voting [26] により最終分類結果を決定する late fusion に基づく手法も比較手法として導入する．これらの詳細について，以降で具体的に説明する．提案手法では，各アノテーターの分類精度を考慮した分類結果統合を可能とする手法であり，上記に示す比較手法の導入は妥当であると考えられる．本実験では，全ての手法において SVM のカーネル関数はガウシアンカーネルとし，そのカーネルパラメータはグリッドサーチにより決定した．SVM の実装は，libSVM のライブラリ [27] を用いて行っている．本実験における評価指標として，我々は Recall，Precesion および F-measure を用いた．これらの具体的な定義を次式に示す.

$$\mathrm{Recall} = \frac{N_c}{N_m}, \tag{4.28}$$

$$\mathrm{Precision} = \frac{N_c}{N_a}, \tag{4.29}$$

$$\text{F-measure} = \frac{2 \times \mathrm{Recall} \times \mathrm{Precesion}}{\mathrm{Recall} + \mathrm{Precision}}, \tag{4.30}$$

ただし，N_c は発話区間として正しく検出されたセグメントの数であり，N_m は手動により真の発話区間として判定されたセグメントの数である．さらに，N_a は各手法により自動的に検出された発話区間のセグメントの数である．

本実験では，対象の映像の音響信号に対して，ホワイトノイズ，エアコンのノイズ，コンピュータサーバーのノイズを，その SN 比を変更しながら付与を行った．このようにして得られるノイズが重畳した音響信号を含む映像に対して，実際に発話区間検出を行い，提案手法および比較手法により得られる結果から，Recall，Precision および F-measure を求める．図 4.3〜4.10 に，発話区間検出を行った例を示す．さらに，音響信号の SN 比と発話区間検出精度の関係を図 4.11，4.12 および 4.13 に示す．

先に述べた通り，本実験では以下の比較手法を準備し，提案手法の発話区間検出結果との比較を行った．

(i)　画像特徴量のみを用いた発話区間検出手法

(ii)　音響特徴量のみを用いた発話区間検出手法

(iii)　Early fusion に基づく発話区間検出手法

(iv)　Late fusion (majority voting [26]) に基づく発話区間検出手法

比較手法 (i) と (ii) は，4.2.1 のみを用いることにより実装されている．比較手法 (iii) は，画像特徴量と音響特徴量を合わせて統合したベクトルを用意し，これを SVM に入力することで，最終分類結果を取得する．したがって，本実験ではこの feature-level fusion 手法を early fusion の手法として利用する．比較手法 (iv) は，SVM 出力 [27] から算出される発話区間の確率を比較し，最大の確率の結果を出力する手法である．

Recall と Precision はトレードオフの関係があることから，提案手法とその他の比較手法の性能を比較するため，F-measure に注目する．図 4.3〜4.10，4.11，4.12 および 4.13 に示す結果から，提案手法が従来手法と比較して高精度に発話区間検出が可能であることが確認できる．したがって，DLF に基づく提案手法は，高精度な発話区間検出に有効であると言える．提案手法では，各モダリティの分類性能に注目して，二つの分類結果の統合を行うことが可能であり，統合による高精度化が可能になっている．

表 4.1: 提案手法における各手順の計算時間．各値は，$\frac{1}{30}$ 秒の各セグメントを処理するための計算時間を表している．本表に示す a)–f) は，図 4.1 に示す a)–f) に対応する．本実験では，Intel(R) Core(TM) i7 960 CPU 3.20 GHz および 24.0 Gbytes RAM のパーソナルコンピュータを利用した. さらに，提案手法は C++で実装されている．

Procedure	Computation time (sec)
a) Extraction of visual features	1.246
b) Extraction of audio features	5.327×10^{-3}
c) Calculation of attribution degrees	4.809×10^{-3}
d) SVM-based classification using visual features	0.496×10^{-3}
e) SVM-based classification using audio features	0.129×10^{-3}
f) Decision-level fusion of multiple results	2.014×10^{-3}

図 4.11，4.12 および 4.13 に示すとおり，提案手法の F-measure は対象のノイズがホワイトノイズであるとき，比較手法 (ii) よりも精度が低下する傾向がある．ホワイトノイズの場合，音響信号のみを用いる比較手法 (ii) の性能は，エアコン，コンピュータサーバーがノイズである場合と比較して高くなることが確認できる．提案手法は，各モダリティの分類精度を考慮した DLF を可能とするものの，その識別器は式 (4.7) に示す SVM 出力 (2 次元) を入力とする線形のモデルである．したがって，線形モデルのみを用いた精度の向上には限界が存在するため，比較手法 (ii) と比較して精度向上が得られなかったと考えられる．これらの問題を解決するためには，DLF の分類モデルにおいてカーネル法を導入する，または，新しい特徴量を式 (4.7) の $\mathbf{x}$ に用いることで考えられ，その導入については今後の検討課題である．

最後に，提案手法の計算時間について示す．$\frac{1}{30}$ 秒の各セグメントを処理する際の各手順の計算時間を表 4.1 に示す. 表の結果より，SVM を用いた各々特徴量に基づく分類と DLF において，提案手法は高速に演算が可能であることが確認できる．一方，画像特徴量の算出について計算時間を要することが確認できる．提案手法において，計算時間の削減のため特徴量をシンプルなものに限っているものの，計算時間の削減はより実用性の向上のために必要である．

4.3.2　発話区間検出結果のノイズ除去への応用

本節では，提案手法により得られる発話区間検出結果のノイズ除去への応用を行った結果を示す．前節に示すように取得された発話区間検出結果より，我々は非発話区間からノイズの情報を取得し，音響信号のノイズ除去に応用する．本実験では，提案手法および比較手法の精度を単純に比較するため，Spectral Subtraction 法 [1] を用いてノイズ除去を行う．ここで，前節で示した Recall と Precision はそれぞれ，発話区間からのノイズ除去および非発話区間からのノイズ情報の抽出に大きな影響を与える．具体的に，高精度なノイズ除去を行うためには，高い Recall で発話区間の検出を行う必要がある．したがって，短い誤検出はノイズ除去の精度には大きな影響を与えない．つまり，比較手法 (ii) のような短い誤検出は以降のノイズ除去に影響が少ない．一方，発話区間の未検出は，ノイズの除去を行うことができなくなるため，大きな影響が存在する．

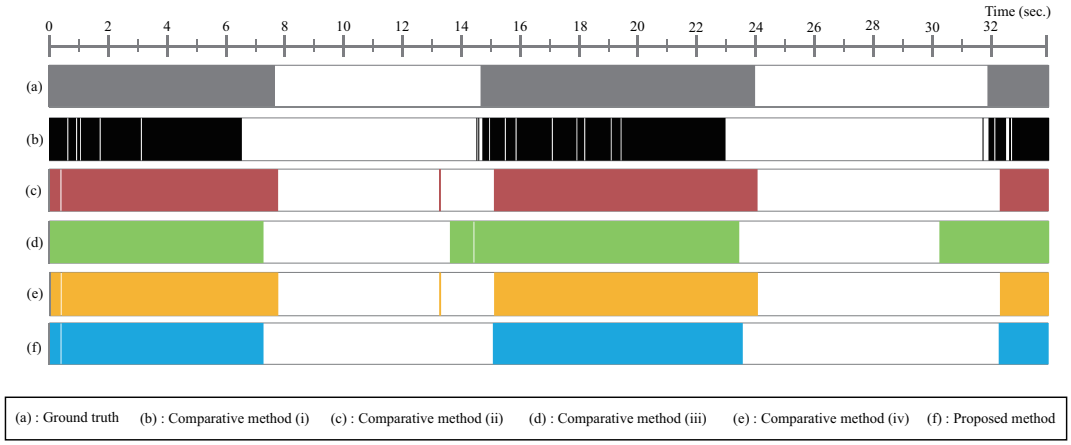
図 4.14〜4.16 に，前節に示す提案手法および比較手法により得られる発話区間の情報を用いたノイズ除去の結果を示す．具体的に，図 4.14〜4.16 では，ノイズを含む音響信号の SN 比と復元された信号の SN 比の関係を示している．なお，本実験では，SN 比を発話区間のみから算出している．

以上により得られる結果から，提案手法の非発話区間の推定の結果を用いたノイズ除去は比較手法と比べて高精度であることが確認できる．また，これらの結果より，前節における発話区間検出結果の精度と大きな関連性を有することが確認できる．したがって，本実験における結果から，高精度な発話区間/非発話区間の検出結果が，高精度なノイズ除去を導くことが明らかとなった．

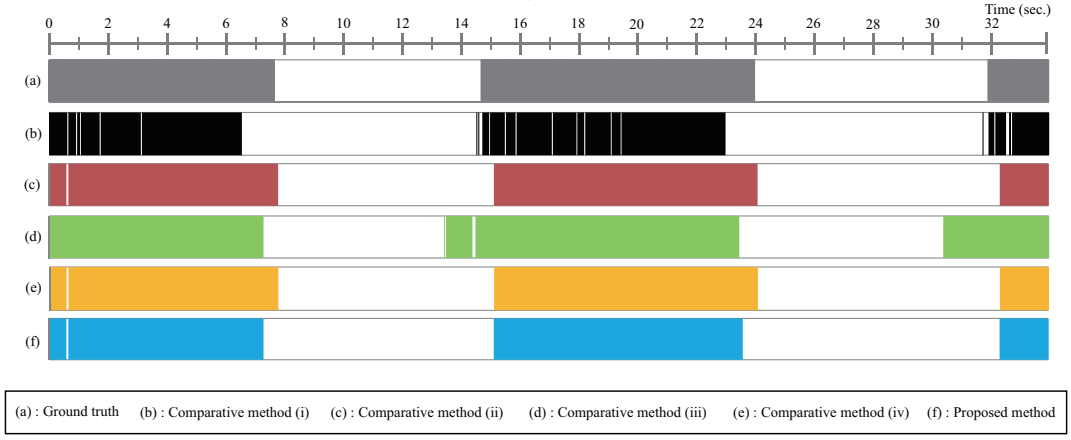
4.4　まとめ

本章では，DLF に基づいたテレビ会議における発話区間検出手法およびその音響信号のノイズ除去への応用について説明を行った．提案手法は，映像から取得される画像特徴量・音響特徴量を用いることで，高精度な発話区間検出を可能とする．したがって，これら複数の特徴量を用いることで，提案手法ではそれぞれ分類結果を取得し，これらを DLF に基づいて統合する．このとき，我々は Supervised Learning from Multiple Experts を DLF に導入することで，複数の特徴量から得られた結果に関する分類精度を考慮した

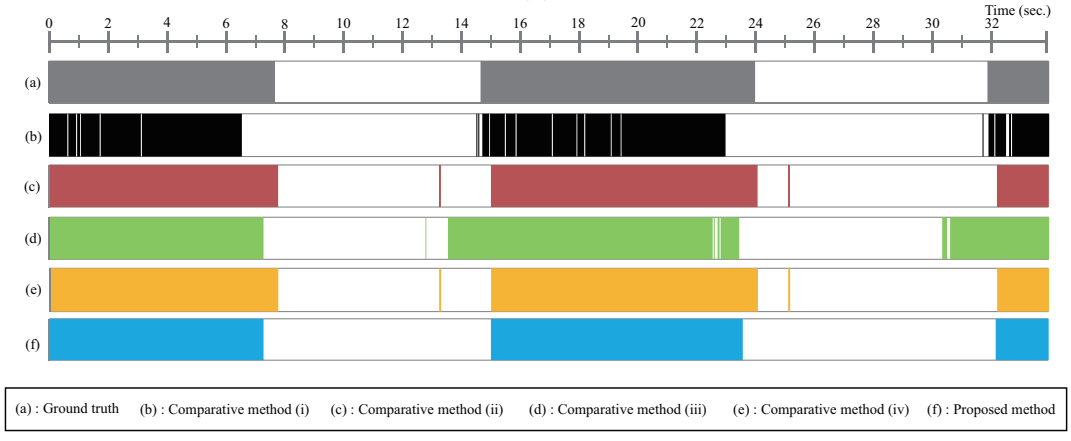
統合を可能とする．さらに，得られた発話区間検出結果を用いることで，提案手法では高精度なノイズ情報の取得が可能となり，結果的に，高精度なノイズ除去を行うことが可能になる．本章の最後では，実験結果を示し，提案手法が従来手法と比較して高精度な発話区間検出が可能であるだけでなく，その結果に基づいた高精度なノイズ除去が可能であることを確認した．



(a)

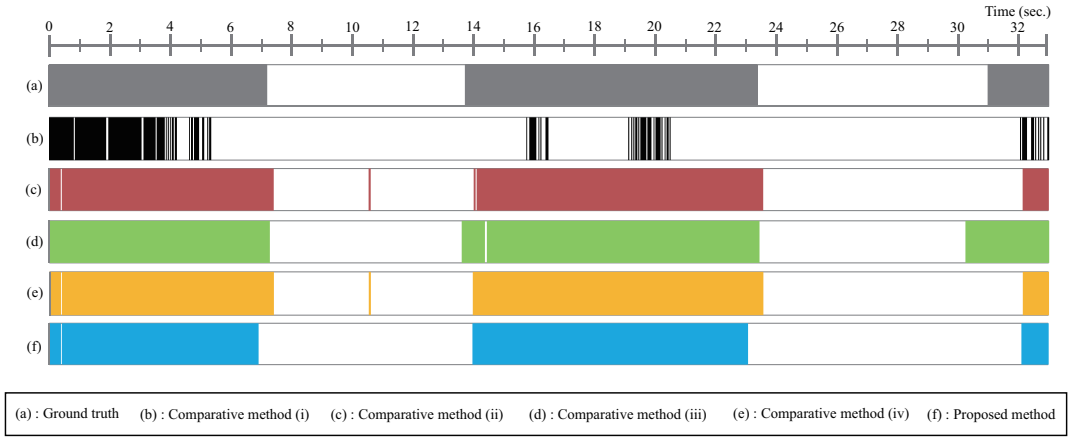


(b)

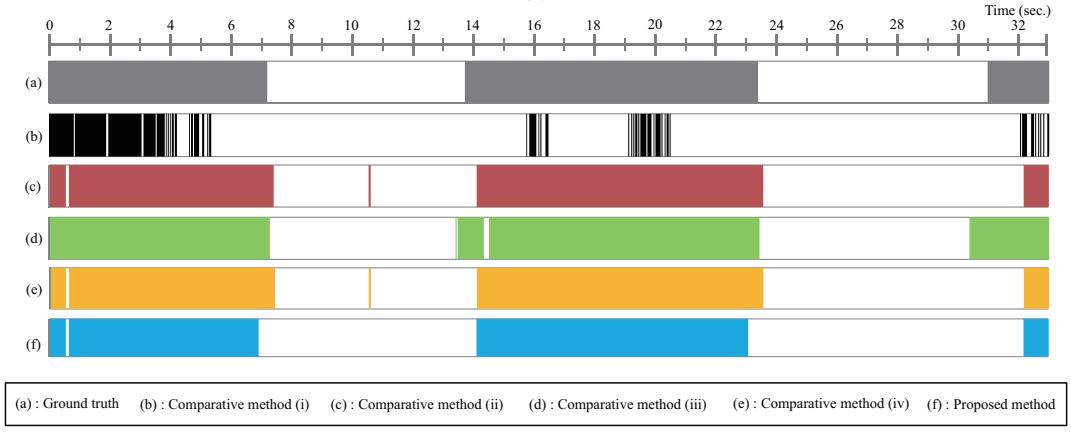


(c)

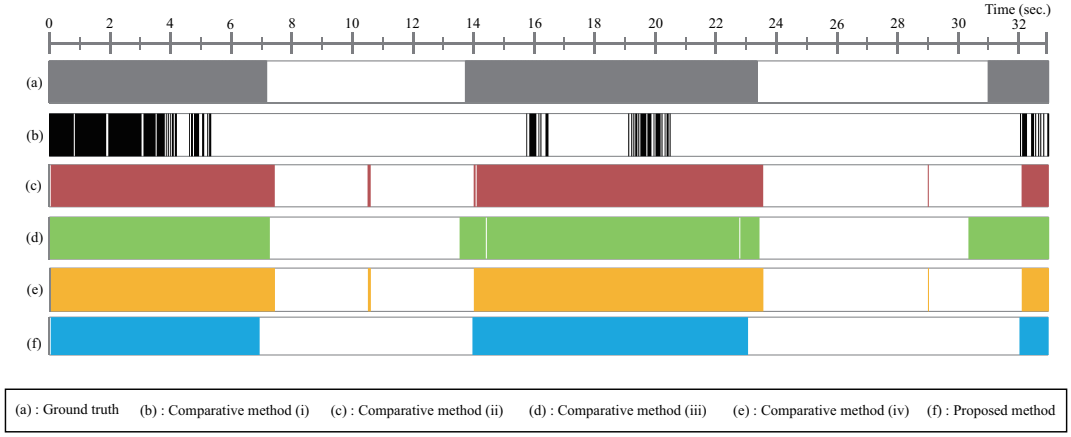
図 4.3: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 1). 音声信号は男性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は 6.0 dB と設定した.



(a)

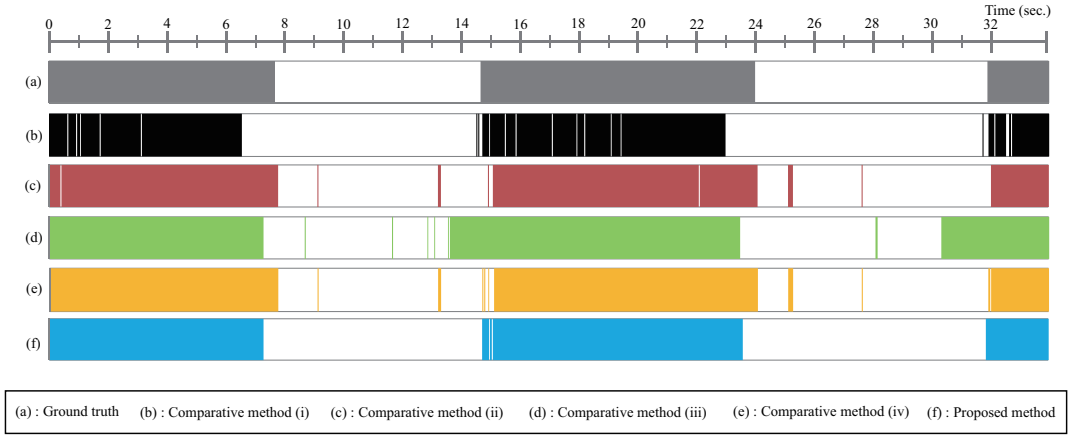


(b)

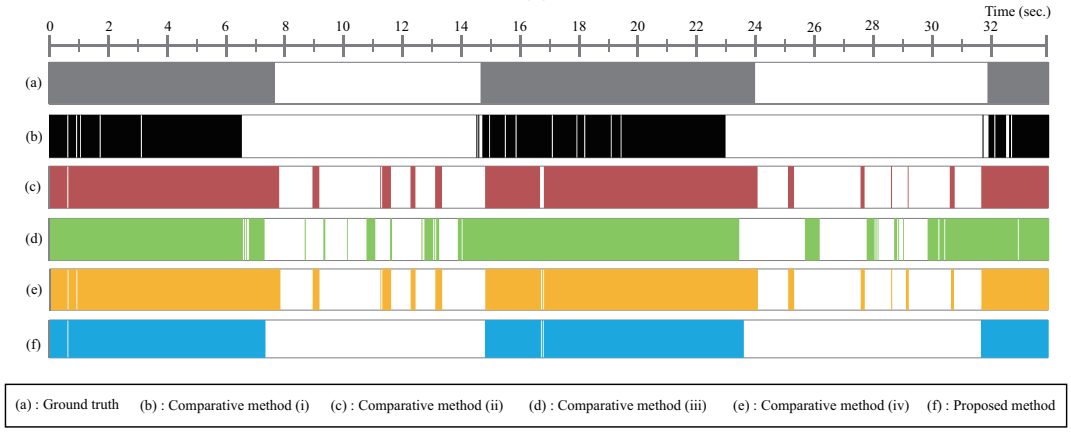


(c)

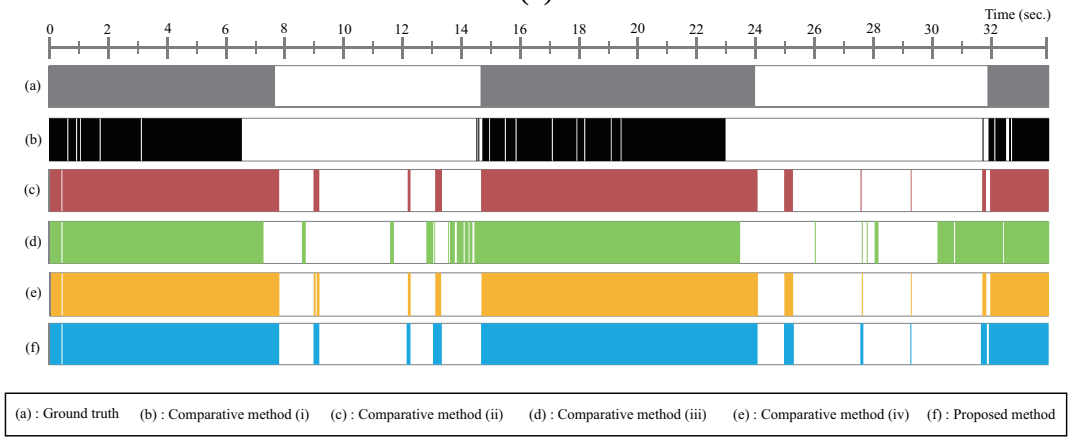
図 4.4: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 2). 音声信号は男性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は 6.0 dB と設定した.



(a)

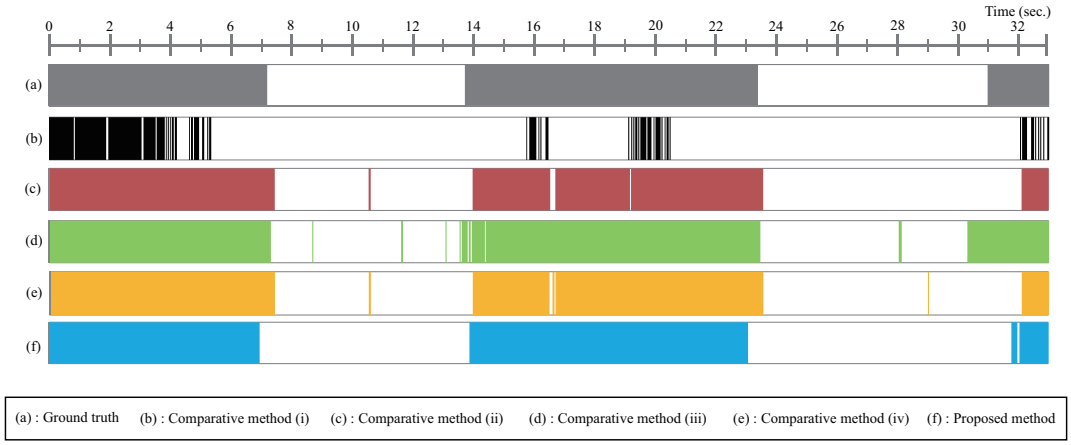


(b)

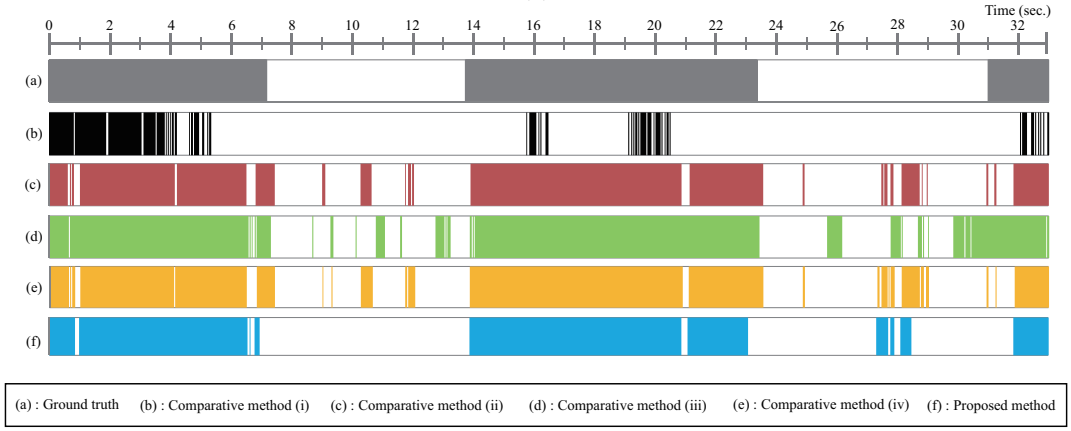


(c)

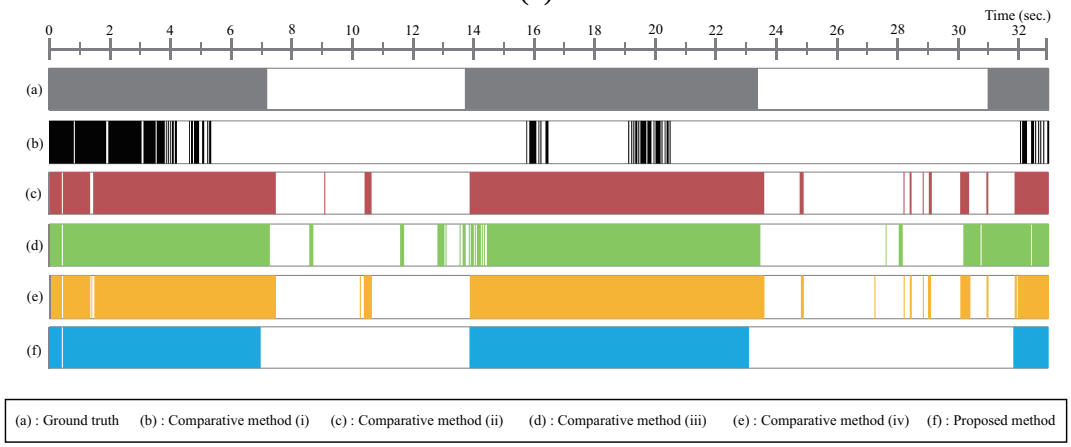
図 4.5: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 1). 音声信号は男性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は –6.0 dB と設定した.



(a)

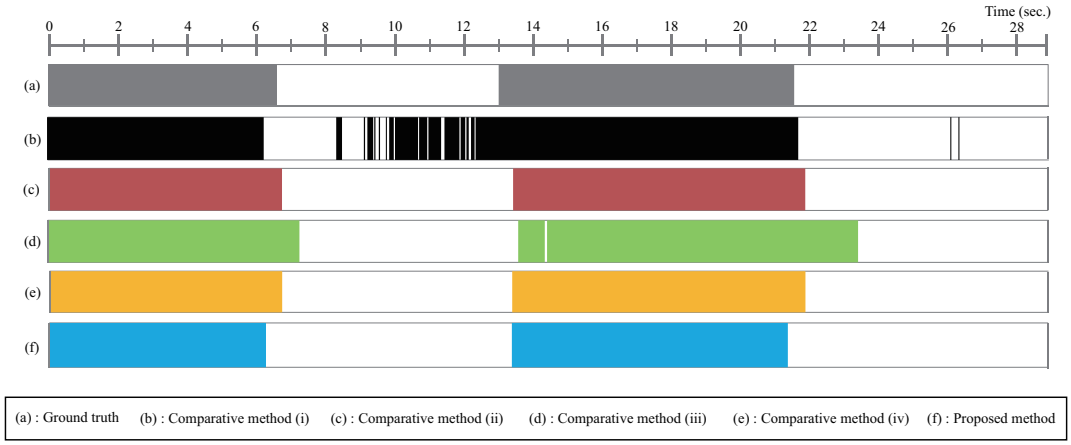


(b)

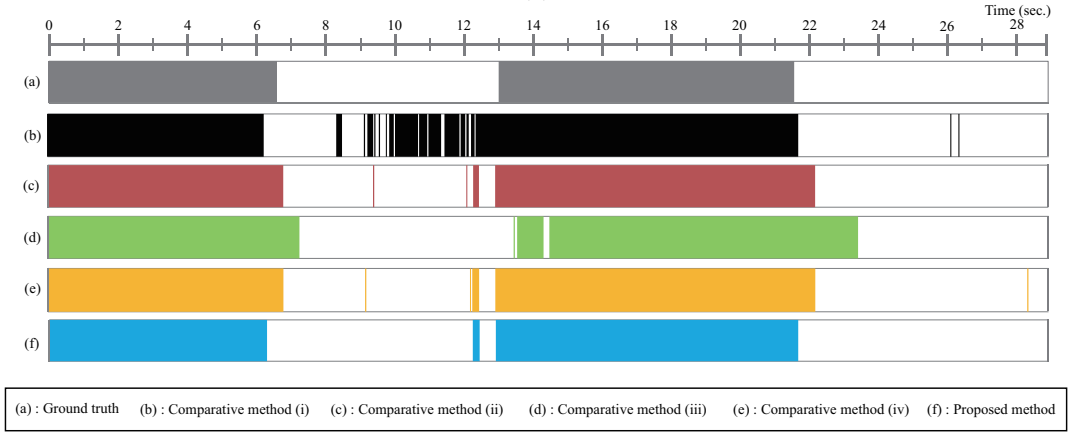


(c)

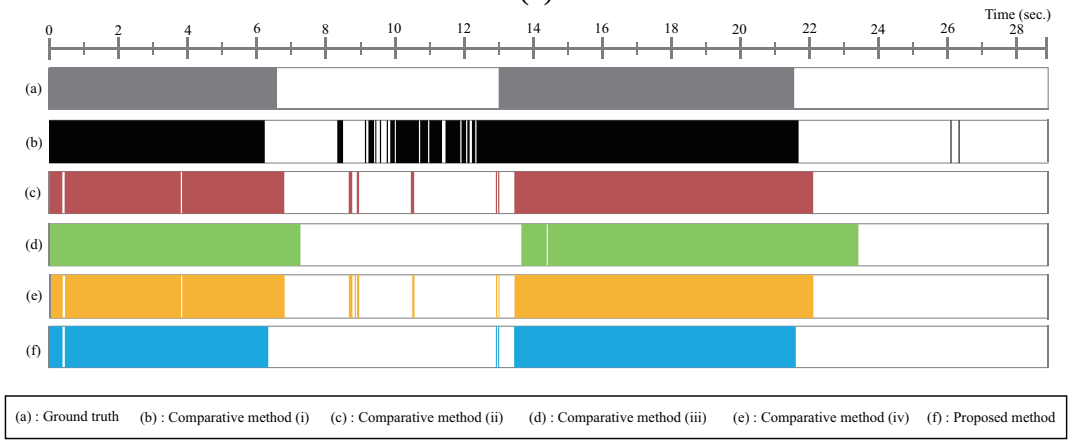
図 4.6: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 2). 音声信号は男性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は −6.0 dB と設定した.



(a)

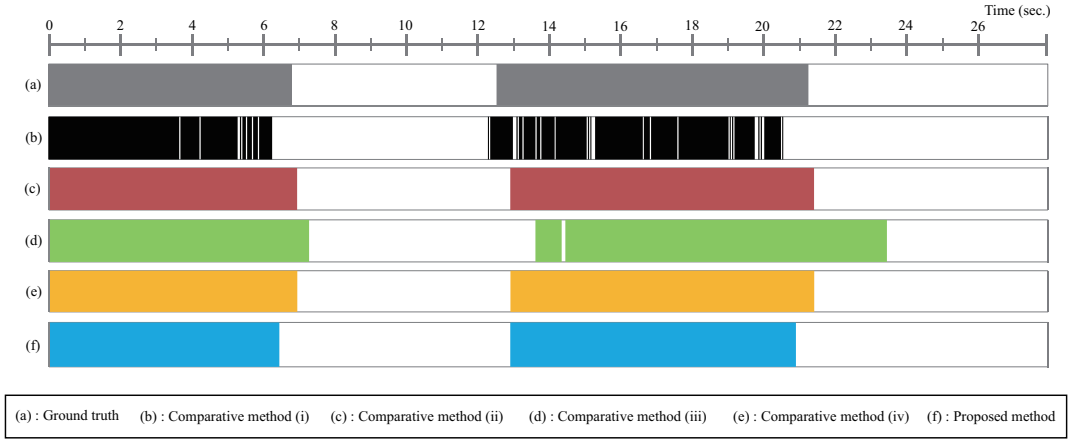


(b)

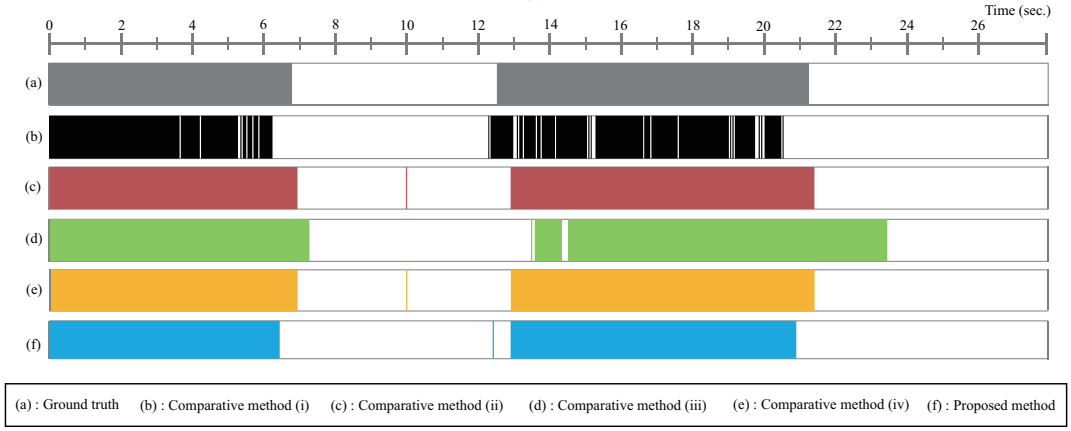


(c)

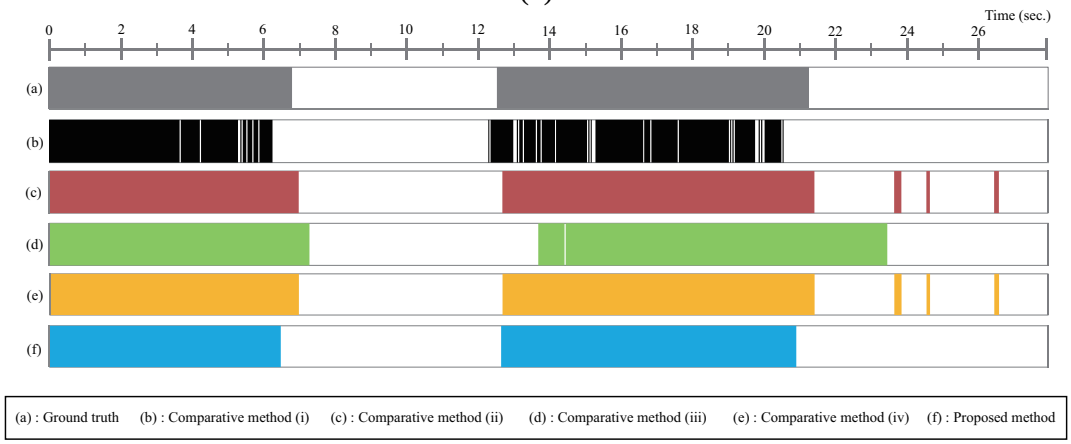
図 4.7: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 1). 音声信号は女性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は 6.0 dB と設定した.



(a)

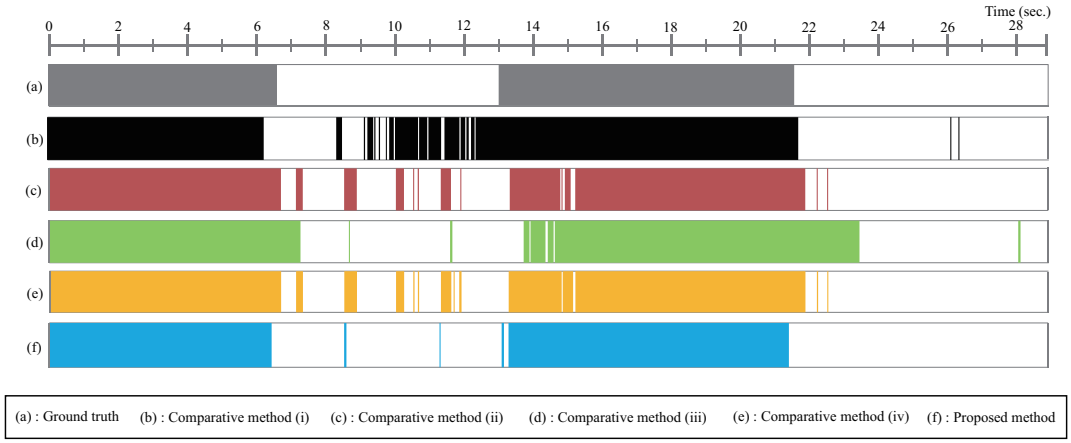


(b)

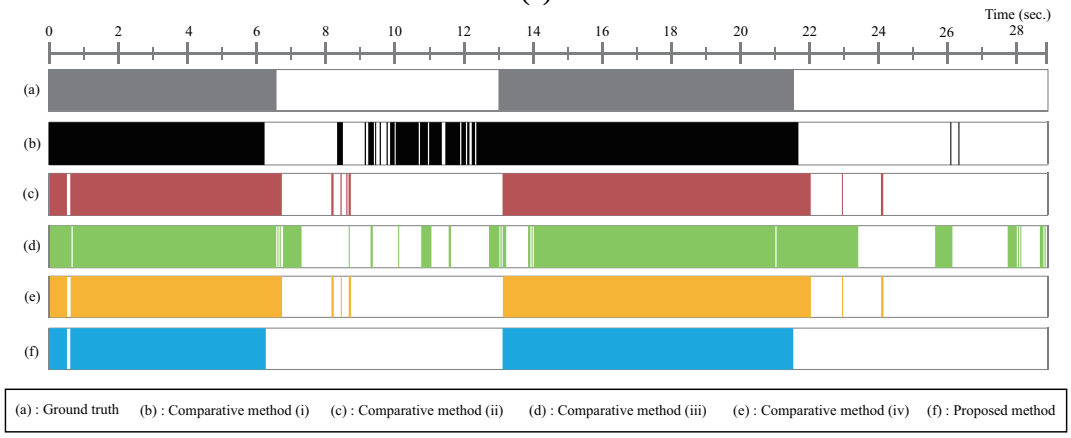


(c)

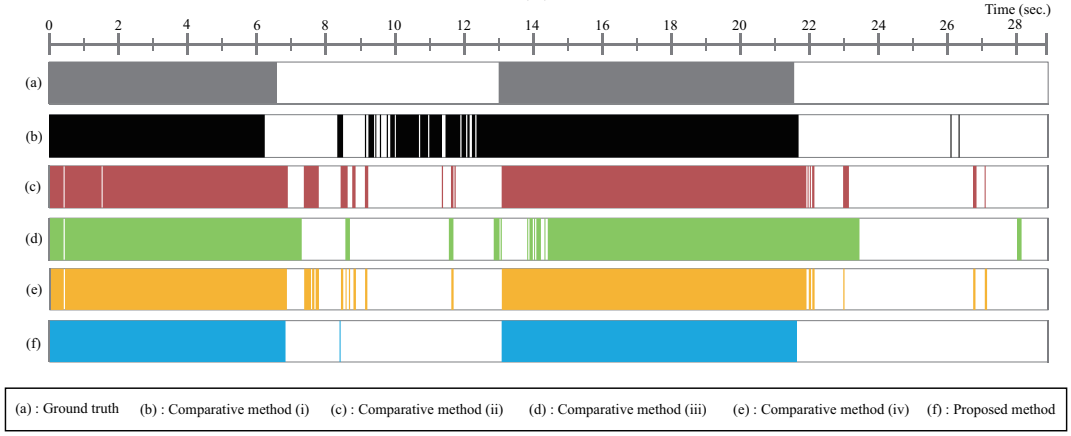
図 4.8: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 2). 音声信号は女性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は 6.0 dB と設定した.



(a)

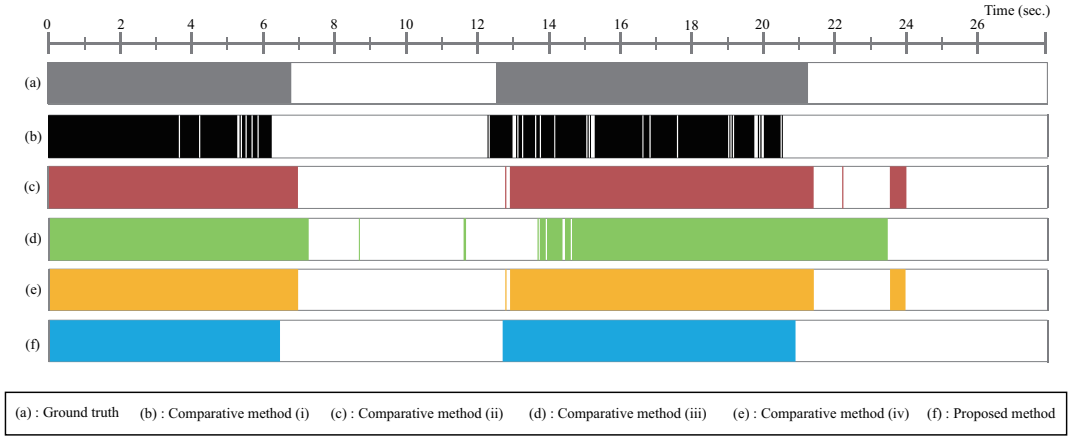


(b)

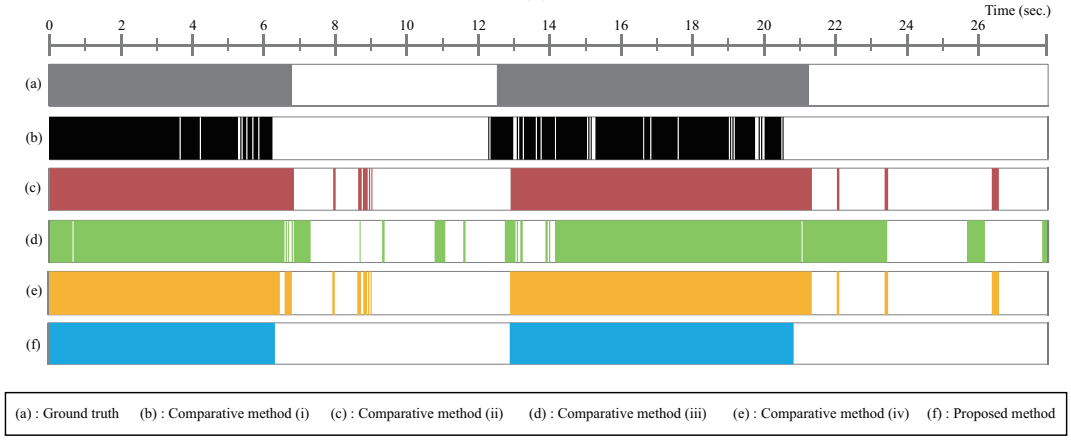


(c)

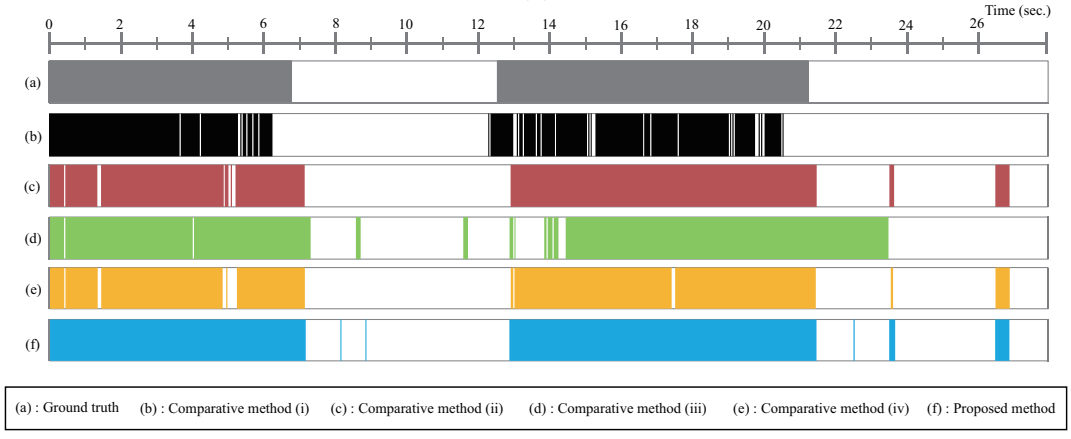
図 4.9: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 1). 音声信号は女性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は –6.0 dB と設定した.



(a)

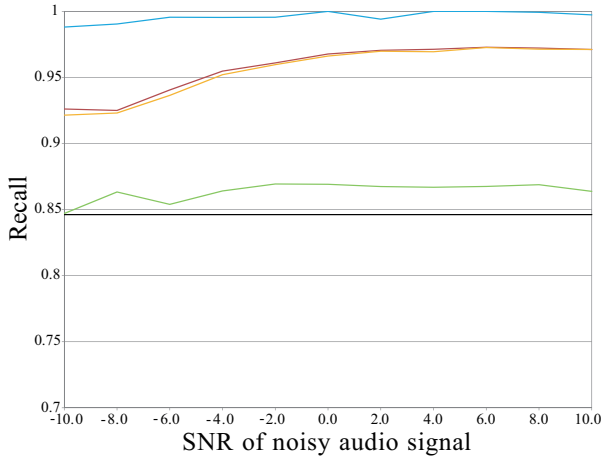


(b)

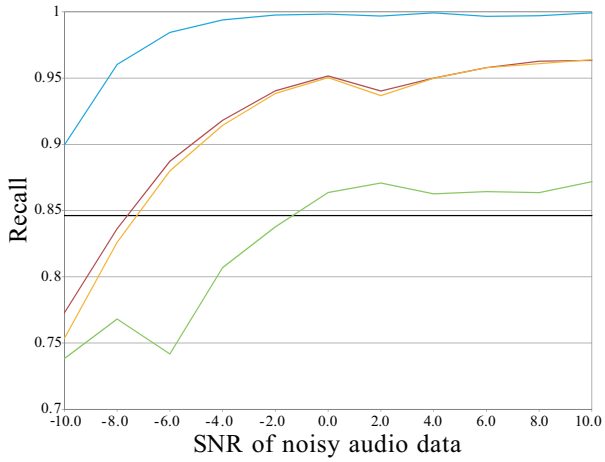


(c)

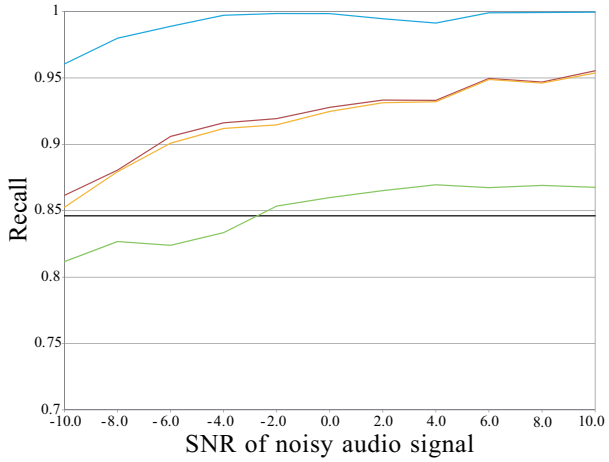
図 4.10: 提案手法および従来手法により発話区間検出を行った結果の例 (サンプル信号 2)．音声信号は女性を対象としており，ノイズの種類は，(a) ホワイトノイズ，(b) エアコンのノイズ，(c) コンピュータサーバーのノイズであり，音響信号の SN 比は −6.0 dB と設定した．



(a)

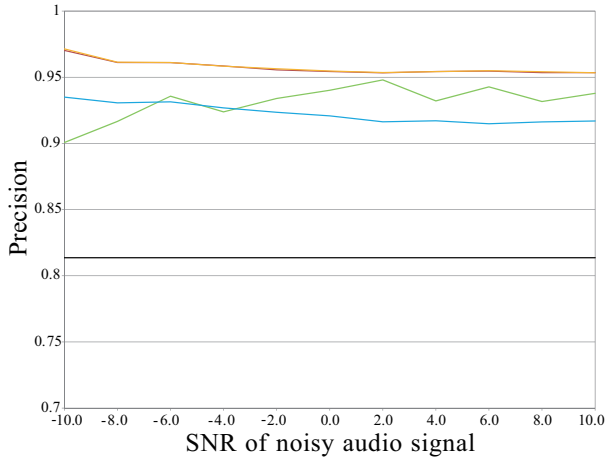


(b)

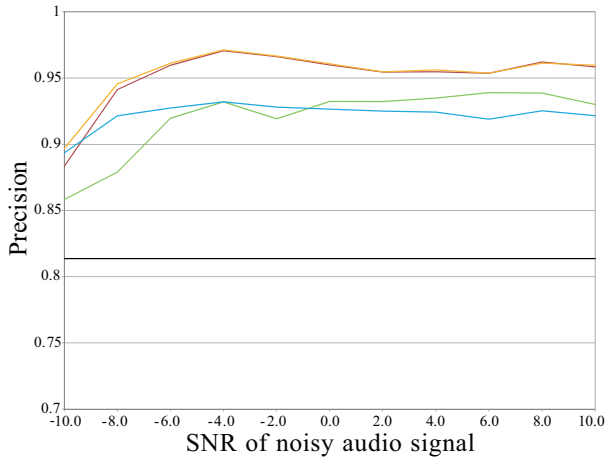


(c)

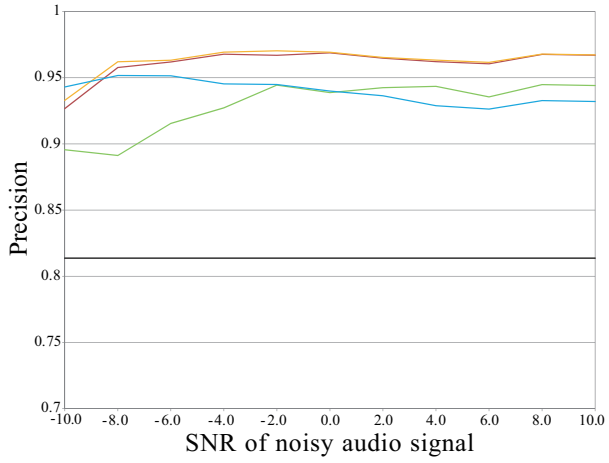
図 4.11: 音響信号の SN 比 (SNR) に対する提案手法および比較手法による発話区間検出結果 (Recall) の比較．本図における (a)〜(c) はそれぞれ，音響信号にホワイトノイズ，エアコンのノイズ，コンピュータサーバーのノイズを付与した例を示している．



(a)

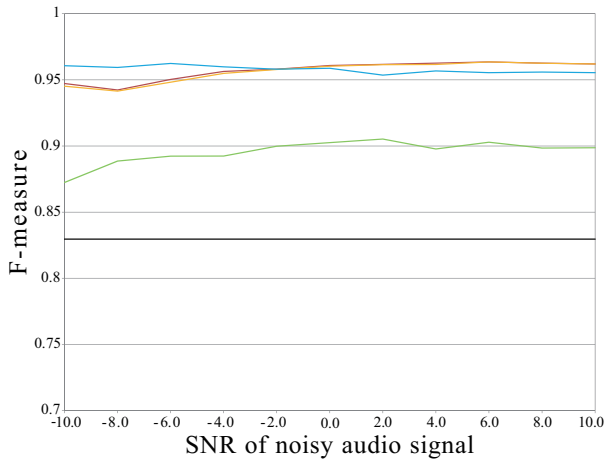


(b)

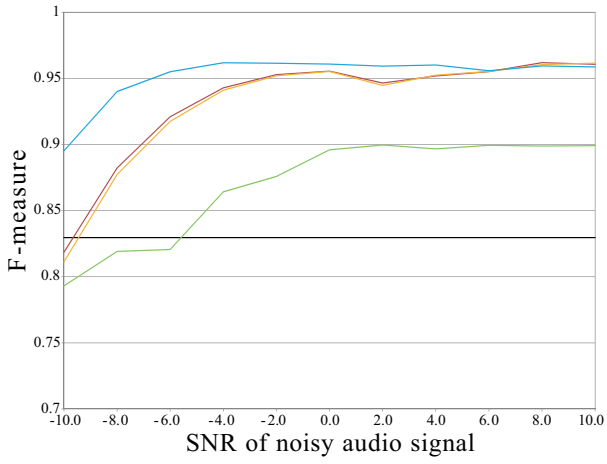


(c)

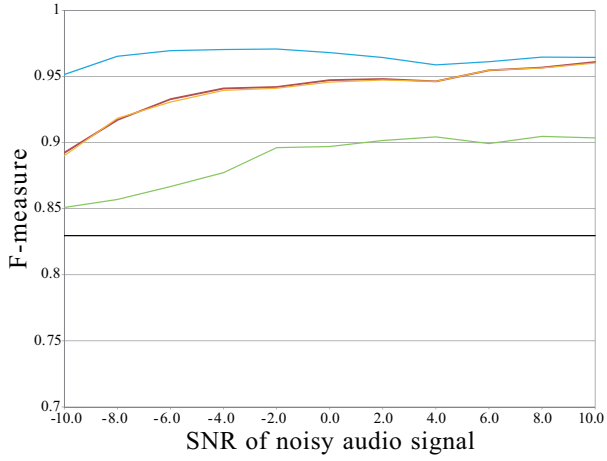
図 4.12: 音響信号の SN 比 (SNR) に対する提案手法および比較手法による発話区間検出結果 (Precision) の比較．本図における (a)〜(c) はそれぞれ，音響信号にホワイトノイズ，エアコンのノイズ，コンピュータサーバーのノイズを付与した例を示している．



(a)



(b)



(c)

図 4.13: 音響信号の SN 比 (SNR) に対する提案手法および比較手法による発話区間検出結果 (F-measure) の比較．本図における (a)〜(c) はそれぞれ，音響信号にホワイトノイズ，エアコンのノイズ，コンピュータサーバーのノイズを付与した例を示している．

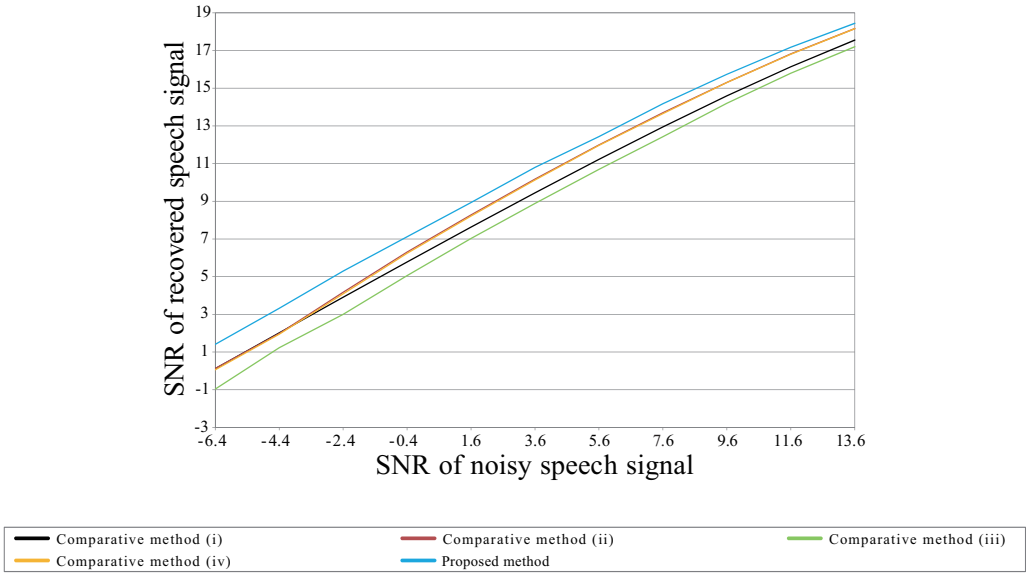


図 4.14: 提案手法及び比較手法に基づく発話区間検出結果を用いたノイズ除去の結果．ノイズの種類は，ホワイトノイズ．本図では,「発話区間の音響信号の SN 比 (SNR)」と「各手法により発話区間を検出した結果に基づいて Spectral Subtraction 手法に基づいてノイズ除去を行った信号の SN 比 (SNR)」の関係を示している．

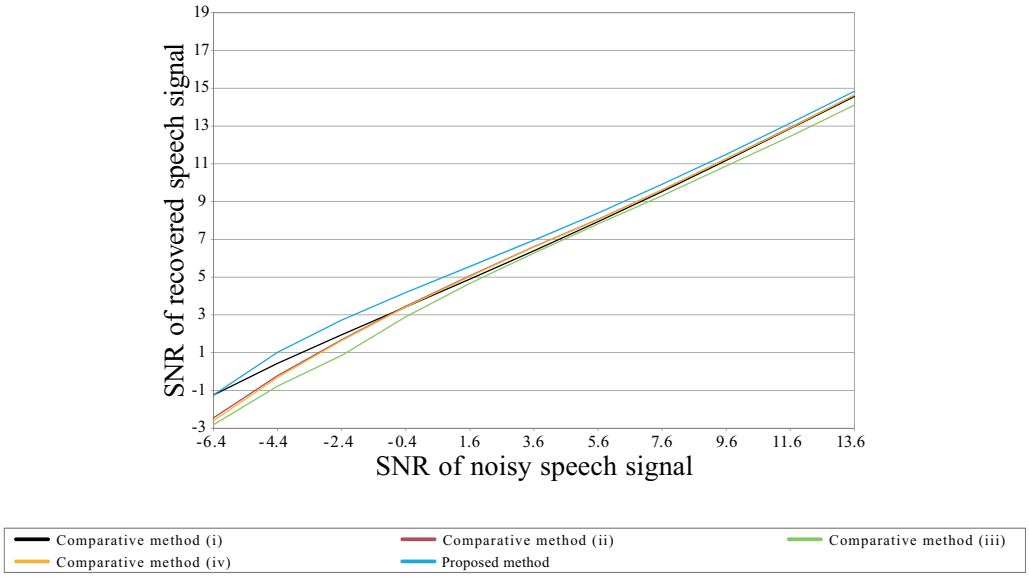


図 4.15: 提案手法及び比較手法に基づく発話区間検出結果を用いたノイズ除去の結果．ノイズの種類は，エアコンのノイズ．本図では,「発話区間の音響信号の SN 比 (SNR)」と「各手法により発話区間を検出した結果に基づいて Spectral Subtraction 手法に基づいてノイズ除去を行った信号の SN 比 (SNR)」の関係を示している．

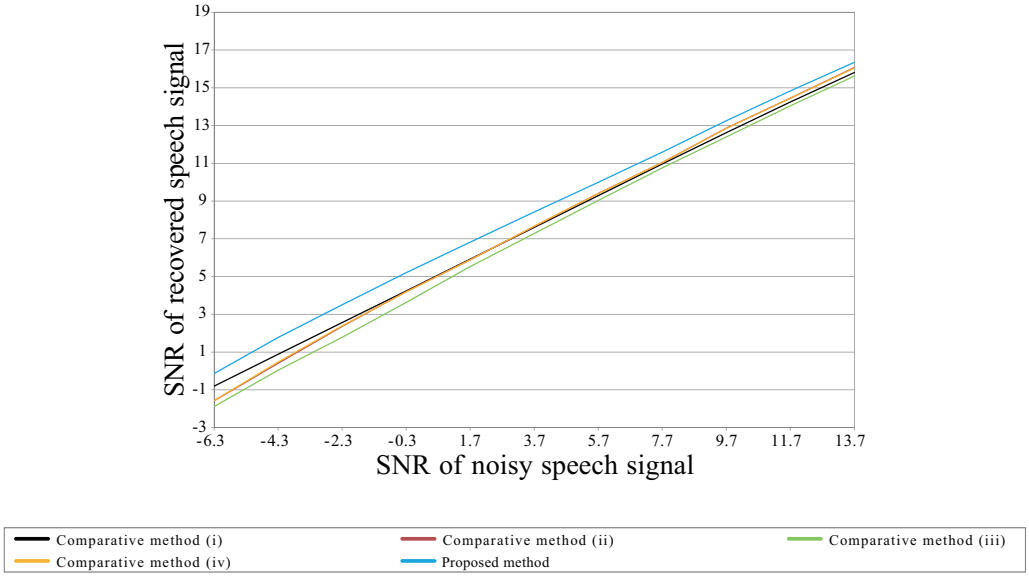


図 4.16: 提案手法及び比較手法に基づく発話区間検出結果を用いたノイズ除去の結果．ノイズの種類は，コンピュータサーバーのノイズ．本図では,「発話区間の音響信号の SN 比 (SNR)」と「各手法により発話区間を検出した結果に基づいて Spectral Subtraction 手法に基づいてノイズ除去を行った信号の SN 比 (SNR)」の関係を示している．

第5章　結論

本研究では，異種特徴の統合解析に基づいたテレビ会議における発話区間検出と音声雑音除去への応用について提案を行った．従来の発話区間検出手法では，音響信号の特徴のみを用いることで，発話区間の検出を行っており，その精度には限界が存在した．また，音響信号だけでなく，映像の特徴を用いることで，協調的に処理を行う手法も存在するものの，それらをどのように高精度に統合すべきかについての課題が存在した．したがって，複数のモダリティにおける分類結果を高精度に統合するための試みが課題として残されていた．

これに対し，本論文では，映像中の口領域を抽出し，その動きの特徴に基づいて発話区間の検出を可能とするだけでなく，音響信号の音響クラスへの分類を行うことで，同様に発話区間の検出を可能とし，これらを各々の信号の特性に基づいて高精度に統合可能とする手法を提案した．具体的に，各々のモダリティにおける発話区間の確率を統合する手法に加え，各々のモダリティにおける分類精度を考慮する新しい統合手法 (Decision Level Fusion: DLF) を導出し，これに基づいて，各々のモダリティを用いた場合よりも高精度な推定を可能とする手法を実現した．特に，各々の検出結果の性能を考慮して高精度に結果の統合を可能とする Supervised Learning from Multiple Experts [12] を導入し，DLF に基づく発話区間検出結果の統合を実現した点が，本研究において最も大きな新規点である．このようにして高精度に取得される発話区間検出結果に基づいて，提案手法では音声雑音の除去を可能とする．具体的に，残りの非発話区間における音響信号の情報，すなわち，ノイズのみを含む音響信号の情報を用いることで，それらの純粋なノイズ情報の取得が可能となり，その情報を用いることで，提案手法では，発話区間に置けるノイズの除去が可能になる．

本論文では，実際に，評価実験により提案する手法が各々のモダリティよりも高精度に発話区間検出が可能であることを示した．また，上記の比較に加えて，従来のマルチ

モダール統合手法と比較しても高精度に発話区間検出が実現されていることを明らかにした．さらに，得られる結果を用いることで，従来と比較して，高精度な音響信号の雑音除去が可能になることも示した．

本論文では主に，発話区間の高精度な検出手法について注目しており，音響信号の雑音除去については簡便な手法を用いたものにとどまっている．音響信号の雑音除去手法についても，これまで様々な研究が行われており，より高精度な原信号の推定が期待できる．したがって，今後は近年の最新のノイズ除去技術との連携について検討を行う予定である．また，DLF に関する改良手法も様々提案されており [28, 29]，それらの導入についても検討を行う予定である．

謝辞

本研究は，著者が北海道大学大学院情報科学研究科メディアネットワーク専攻に，2011年10月から現在まで在学し，途中，本務の都合により研究活動を休止した期間も含め，約5年半に渡って取り組んだものである．

本研究の遂行のみならず，研究活動と企業活動との両立など，公私にわたる様々な悩みに対して親身になって相談に応じて頂き，都度，適切なアドバイスを頂戴するなど，本研究の完遂に向けて，終始ご指導ご鞭撻を頂戴した北海道大学大学院情報科学研究科，長谷山美紀 教授に深く感謝します．

本論文をまとめるにあたり，多くの有益なご助言とご指導を賜るとともに，副査をお引受け頂いた北海道大学大学院情報科学研究科，荒木健治 教授，坂本雄児 教授，山本強 教授に深く感謝致します．

また，本研究の遂行および本論文の執筆にあたり，終始一貫して親身のご指導と多大なるご助言を頂戴いたしました北海道大学大学院情報科学研究科，小川貴弘 准教授に深く感謝致します．高橋翔 特任助教におかれましては，本研究の手法実装に関し，ご指導ご支援を頂き深く感謝致します．

著者の研究期間中に研究活動のご支援とご協力を頂いた北海道大学大学院情報科学研究科 情報メディア学講座 メディアダイナミクス研究室の諸先輩方，並びに，学生の皆さんに感謝致します．

参考文献

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. Nagumo and A. Noda, “A learning method for system identification,” IEEE Transactions on Automatic Control, vol. 12, no. 3, pp.282–287, 1967.

[3] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 2, pp. 109–118, 2002.

[4] J. Ramirez, J.C. Segura, C. Benitez, A. De la Torre and A. Rubio, “An effective subband OSF-based VAD with noise reduction for robust speech recognition,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 6, pp. 1119–1129, 2005.

[5] D. Ying, X. Lu, J. Li, Y. Yan, J. Dang and F. Soong, “Noise estimation using a constrained sequential HMM IN log-spectral domain,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4553–4556, 2012.

[6] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz and C. Jutten, “An analysis of visual speech information applied to voice activity detection,” IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. I601–I604, 2006.

[7] S. Siatras, N. Nikolaidis, M. Krinidis and I. Pitas, “Visual lip activity detection and speaker detection using mouth region intensities,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 1, pp. 133–137, 2009.

[8] M. Haseyama, T. Ogawa, and N. Yagi, “A review of video retrieval based on image and video semantic understanding,” ITE Transactions on Media Technology and Applications, vol. 1, no. 1, pp. 2–9, 2013.

[9] V. P. Minotto, C. B. O. Lopes, J. Scharcanski, C. R. Jung, L. Bowon, “Audiovisual voice activity detection based on microphone arrays and color information,” IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 1, pp. 147–156, 2013.

[10] N. Nitanda, M. Haseyama, “ Audio-based shot classification for audiovisual indexing using PCA, MGD and Fuzzy algorithm, ”IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E90-A, no. 8, pp. 1542–1548, 2007.

[11] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, pp. 273–297, 1995.

[12] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, “Supervised learning from multiple experts: whom to trust when everyone lies a bit,” the 26th Annual International Conference on Machine Learning, pp. 889–896, 2009.

[13] L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[14] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, pp. 119–139, 1997.

[15] L. Breiman, “Random Forests,” Machine Learning, vol 45, no. 1, pp. 5–32, 2001.

[16] I. Ben-Gal, “Bayesian Networks,” Encyclopedia of Statistics in Quality and Reliability, 2008.

[17] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I511–I518, 2001.

[18] N. Dalal and B.Triggs, “Histograms of oriented gradients for human detection,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), vol.1, pp. 886–893, 2005.

[19] G. Doretto, A. Chiuso, Y. N. Wu and S. Soatto, “Dynamic Textures,” International Journal of Computer Vision, vol. 51, no. 2, pp.91–109, 2003.

[20] T. Kobayashi and N. Otsu, “Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation,” The 17th International Conference on Pattern Recognition, vol. 4, pp. 741–744, 2004.

[21] L. Peters, S. Guo and R. Camposano, “A novel analog fuzzy controller for intelligent sensors,” Fuzzy Sets and Systems, vol. 70, pp. 235–247, 1995.

[22] C. C. Lee, “Fuzzy logic in control systems: Fuzzy logic controller,” IEEE Transactions on Systems, Man and Cybernetics, vol. 20, no. 2, pp. 404–435, 1990.

[23] T. A. Runkler and M. Glesner, “A set of axioms for defuzzification strategies towards a theory of rational defuzzification operators,” IEEE International Conference on Fuzzy Systems, vol. 2, pp. 1161–1166, 1993.

[24] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, maxrelevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[25] T. Maeda, T. Yamasaki, and K. Aizawa, “Multi-stage object classification featuring confidence analysis of classifier and inclined local naive Bayes nearest neighbor,” IEEE International Conference on Image Processing, pp. 5177-5181, 2014.

[26] R. Dymitr and G. Bogdan, “Classifier selection for majority voting,” Information fusion, vol. 6, no. 1, pp. 63–81, 2005.

[27] C. -C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” 2011, software available at http://www.csie.ntu.edu.tw/?cjlin/libsvm/.

[28] F. Rodrigues, F. Pereira and B. Ribeiro, “Learning from multiple annotators: distinguishing good from random labelers,” Pattern Recognition Letters, vol. 34, no. 12, pp. 1428–1436, 2013.

[29] Y. Xu and U. Lu, “Adaptive weighted fusion: A novel fusion approach for image classification,” Neurocomputing, vol. 168, pp. 566–574, 2015.

著者の研究業績

論 文

(A) 学会誌

[**A-1**] Kazuto Sasaki, Takahiro Ogawa, Sho Takahashi, Miki Haseyama: “DLF-based speech segment detection and its application to audio noise removal for video conferences,” ITE Transactions on Media Technology and Applications, vol. 4, no. 1, pp. 68-77 (2016)

(B) 国際学会

[**B-1**] Kazuto Sasaki, Takahiro Ogawa, Miki Haseyama: “Audio Noise Removal for Video Conferencing Based on Audio-Visual Analysis,” Proceedings of International Workshop on Advanced Image Technology 2013, pp. 293 - 296 (2013)

[**B-2**] Kazuto Sasaki, Takahiro Ogawa, Miki Haseyama: “Audio Noise Removal for Multi-Speaker Video Conferencing Using Multiple Video Signals,” Proceedings of International Workshop on Advanced Image Technology 2013, pp. 926 - 929 (2013)