



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 [論文内容及び審査の要旨]
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
Doc URL	https://hdl.handle.net/2115/65728
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Kazuto_Sasaki_abstract.pdf, 論文内容の要旨



学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称 博士（情報科学） 氏名 佐々木 一人

学 位 論 文 題 名

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)

本論文は、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ている。

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に代表されるモバイル機器の普及と相まって、それらを活用した多彩なサービスが進展してきている。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、また、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。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サービスは、利用時の環境条件に大きく左右される場合も少なくなく、例えばノイズの発生状況次第では、極端なサービスの低下を招きかねない。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、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で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が、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。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ことで、それら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。また、発話区間の検出においても高精度化が期待できることが報告されている。しかしながら、従来の発話区間検出手法では、各モダリティにおける精度を考慮した統合は行われておらず、結果、各モダリティを用いた場合と比較して、必ずしも高精度な結果が得られるとは限らないという課題が残っている。

本研究は、こうした状況を踏まえ、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したものである。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が困難となる課題を解決するため、発話区間の確率を用いた統合手法の構築や、分類精度を考慮して統合を可能とする Decision Level Fusion(DLF)

を導入することで、より高精度な発話区間推定を可能にしようとしている。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目指している。

本論文は、5つの章から構成されている。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する他、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から得られる特徴に対して、Support Vector Machine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、具体的に、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することが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。

以上、要約すると、本論文は、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関する新たな手法の提案を行い、その有効性について検証を行っている。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を導入することで、より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SLME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は、本研究における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、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、発話区間におけるノイズの除去が可能になる。