送受協調型画像音響センサフュージョンによる音声強調

Vision Referential Speech Enhancement Combining an Acoustical Sensor and a Vision Sensor

松本　光春（Mitsuharu MATSUMOTO, Ph. D.）
電気通信大学　情報理工学研究科　情報学専攻　准教授
(Professor, University of electro-communications, Graduate School of Informatics and Engineering, Department of Informatics)
IEEE　電子情報通信学会　他
受賞：エリクソン・ヤング・サイエンティスト・アワード (2009 年)．FOST 熊田賞（2011 年）Outstanding paper award in IEEE International Conference on Consumer Electronics (2016 年)　Best paper award in 25th International Symposium on Artificial Life and Robotics (2020 年) 他
著書：図解入門 よくわかる最新センサ技術の基本と仕組み, 秀和システム (2020 年)　電子部品が一番わかる(しくみ図解), 技術評論社 (2013 年) 他
研究専門分野：知覚情報処理　知能ロボティクス

あらまし

本研究は公共の場所での講演や集会での利用など騒音環境下での不特定多数に向けた情報伝達を想定した音声強調に関する研究である。本研究は現在広く普及しているスマートフォン/タブレットでの利用を想定し、単一のマイクロホンとカメラを組み合わせたセンサフュージョンによる音声強調を行う。
話し手と聞き手の間でスピーカー・マイクロホンだけでなく、ディスプレイ・カメラをそれぞれ携帯・装着し、マイクロホンからの音と同時に、音声強調のための補助情報を画像情報として聞き手に送信することで、外部雑音の種類に影響されない頑健な音声強調が可能となることが期待される。

1．研究の目的

音声強調にはマイクロホンアレイと呼ばれる複数のマイクロホンを用いた処理系を用いるのが一般的である[(1)-(3)]。しかし、マイクロホンアレイは互いに異なった場所にマイクロホンを置き、その位相差を利用することで処理を実現するため、原理的にシステムの小型化が難しく、処理のために特別な装置を必要とする。災害現場や工事現場など多数の人や騒音源が混在する環境ではマイクロホンアレイでの音声強調技術でしばしば利用される定常性、独立性、スパース性などを仮定できないため、音声などで有用な独立成分分析[(4)*1]やバイナリマスク[(5) *2]を利用した音源分離などを利用することは困難である。従来型の音声強調で前処理として必要となる雑音方向の推定も難しい。

このような状況のなか、話者同士の会話の補助のため、現在、多くの人が利用しているスマートフォンやタブレットが利用できないかと考えたことが本研究の基本的な着想の開始点である。標準的なスマートフォンやタブレットでは従来型の音声協調技術で必要となる多数のマイクロホンが装備されていることはまれであり、１つないし、多くても２つ程度のマイクロホンとカメラが装着されている。このスマートフォンの持つ標準的な装備で実現できる音声協調技術として話し手と聞き手がスピーカーとマイクロホンだけでなく、ディスプレイとカメラをそれぞれ携帯または装着し、マイクロホンからの音と同時に、音声強調のための補助情報を画像情報として聞き手に送信することで、外部雑音の種類に影響されず、一般的なスマートフォンやタブレットで利用可能な音声強調技術の実現を目指している。

2．提案手法の概要

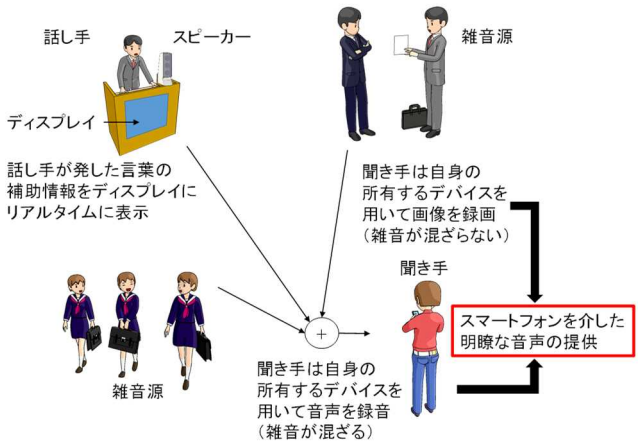

図１　提案手法の概要図

送受協調型画像音響センサフュージョンによる音声強調

Vision Referential Speech Enhancement Combining an Acoustical Sensor and a Vision Sensor

図 1 に本研究で構想した音声強調手法の概念図を示す。この手法では、話し手はスピーカーを通した音声信号だけでなく、ディスプレイを通した画像信号として音声の情報を発信する。聞き手はスマートフォンやタブレットなどに搭載された標準的なカメラとマイクロホンを通して受信した画像、音声信号から目的信号を強調する。この枠組みは従来の音声強調技術では難しかった外部雑音の種類に影響されない音声強調を実現可能にする送受協調型のセンサフュージョンの枠組みである。

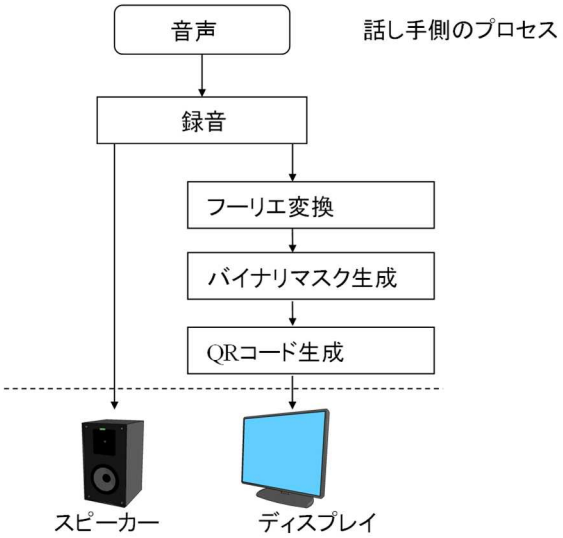

図 2　話し手側の処理プロセス

図 3　聞き手側の処理プロセス

図 2、図 3 に話し手側、聞き手側の処理プロセスをそれぞれ示す。

通常、話し手はマイクを通して録音した音声をスピーカーから音声として流し、聞き手側はマイクロホンを用いてその音声を録音する。この枠組みでは話し手から聞き手に提供される情報は音声情報のみであるため、周囲に雑音が生じている状況では話し手側の音声をうまく聞き取ることが難しくなる。

これに対し、提案手法では話し手側が音声だけでなく画像情報も積極的に利用することで話し手に自分の音声を提供し、聞き手側は得られる音声、画像情報を合成することでより高品質な音声を取得する。

話し手側は図 2 に示すように音声をスピーカーから音声として流すだけでなく、その情報をフーリエ変換し、周波数空間上での音声の有無を検出する。この処理によって得られた周波数空間上での音声の有無に関する情報をバイナリマスクとして記録する。聞き手側にバイナリマスク情報を適切に伝えるため、時間ごとのバイナリマスク情報を QR コードとしてコーディングし、その QR コードをディスプレイ上に表示する。

聞き手側はマイクロホンを通して話し手からの音声を録音するだけでなく、カメラを用いて話し手から提供される QR コードを取得する。次に得られた QR コードをデコードして得られた音声に対応するバイナリマスクを取得する。その後、録音された音声をフーリエ変換し、バイナリマスクをかけた後、逆フーリエ変換して音声を取得することで目的音声が強調される。

公共の場においては話し手の音響信号は不特定多数の雑音によって妨害されるため従来手法を用いて目的信号を復元することは難しい。これに対し、提案手法を用いると画像情報としてマスク情報が提供されるため、聞き手は雑音に妨害されることなく、直接マスク情報を得ることができる[6]。

3．研究の方法

提案手法の有効性を確認するため、複数の強度を持つ白色雑音を用いて音声強調実験を行った。

目的信号には ATR 新聞読み上げデータベースから男性 1 名、女性 1 名の音声を選択した。雑音信号としては 4 種類の白色雑音を用意した。雑音レベルは

送受協調型画像音響センサフュージョンによる音声強調

Vision Referential Speech Enhancement Combining an Acoustical Sensor and a Vision Sensor

dBFS(decibels relative to full scale)を用いて示している。

表 1 に実験に用いた目的信号と雑音信号の詳細を示す。また、図 4 に実験結果の一例を示す。

表 1　目的信号と雑音信号

目的信号	ATR 新聞読み上げデータベース
雑音信号	白色雑音
雑音レベル	-10, -15, -20, -25 (dBFS)

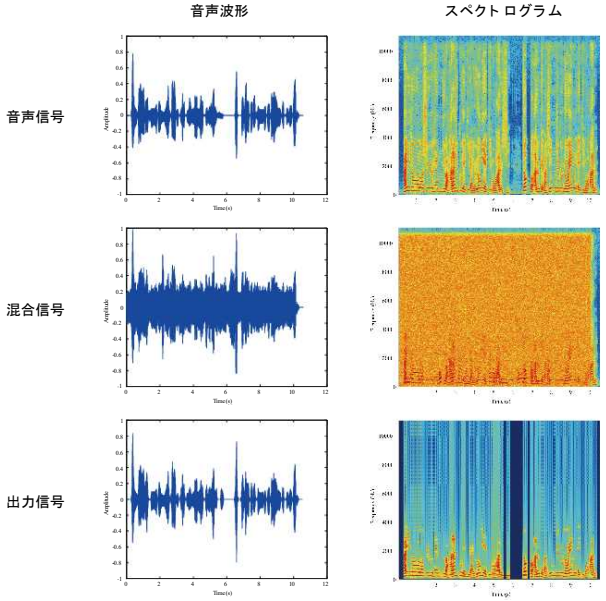

図 4　音声波形とスペクトログラムの例

図 4 左はそれぞれ上から目的音となる音声の波形、混合音の波形、処理後の音声の波形である。また、図 4 右はそれぞれ上から目的音となる音声のスペクトログラム、混合音のスペクトログラム、処理後の音声のスペクトログラムを示す。

図 4 に示すように信号が雑音に埋もれてしまうような大きな雑音が混合された時であっても提案する処理を行うことで目的音の特徴が復元され、雑音が取り除けていることが確認できる。

先に記述した通り、通常のバイナリマスクを用いた手法では周波数軸上で重なりのある雑音に対してバイナリマスクを推定することが難しく、雑音を取り除くことが難しい。これに対し、提案する手法では周波数軸上での音声情報を画像情報として聞き手側に陽に提供するため、周波数軸上で音声と雑音に重なりがあっても重なりのない部分についてバイナリマスクを実施することが可能であることが確認できた。

さらに提案手法で利用する各種パラメータの変化に対する提案手法の頑健性を調査した[7][8]。サンプリング周波数は 22050Hz に設定し、FFT(Fast Fourier Transform)の窓サイズを 128 から 8192 まで変化させた。提案手法の有効性を定量的に評価するため、*NRR*(Noise reduction ratio)[*3] を計算した。

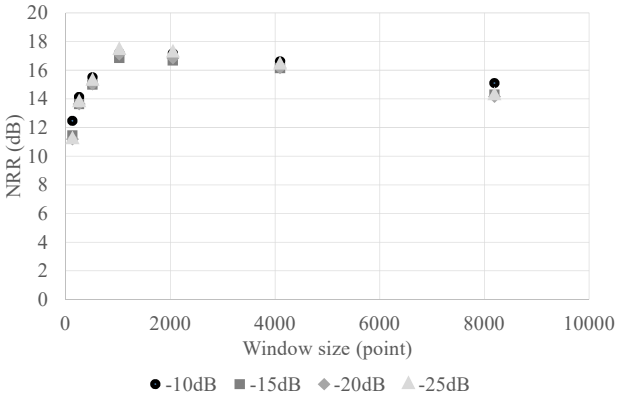

図 5　FFT の窓幅と NRR の関係

図 5 は雑音レベルが-10dBFS、-15dBFS、-20dBFS、-25dBFS の時の窓幅に応じた NRR(Noise reduction ratio)[*3] の値である。図 5 に示すように全体として比較的高い音声強調性能が確認できるが、特に窓幅が 1000-2000 程度のときに音声強調性能が高く、窓幅により音声強調性能に違いがあることが確認できる。

また、提案手法のさらなる性能向上のため、バイナリマスクとスペクトルサブトラクション[*4] を組み合わせた雑音除去の枠組みについても検討した。スペクトルサブトラクションを用いて周波数軸上で信号と雑音が重なっている部分への雑音除去を行うことでさらなる NRR の向上が行えることも確認した[9]。

4．将来展望

異なるセンサを組み合わせることで目標となる音声を強調し、不要な雑音を抑制するセンサフュージョン技術は従来型の雑音抑制技術では難しい様々な問題に対する可能性を秘めている。提案された手法はバイナ

リマスクを用いているが、話し手の時間周波数軸上での音声位置を画像情報として明示的に送信しているため、周波数軸上に重なりのある雑音に対しても音声強調が可能である。また、重なりのある雑音についてもスペクトルサブトラクションのような処理を行うことでさらなる音声強調性能の向上を行うことができる。

現在、本研究の枠組みを応用し、マイクロホンとカメラの組み合わせではなく、マイクロホンと骨伝導マイクの組み合わせにより、雑音除去を行う枠組みにも取り組んでおり、その有効性を示唆する結果を得始めている(10)。

おわりに

本研究は単一のマイクロホンとカメラを組み合わせたセンサフュージョンによる音声強調について検討した。話し手と聞き手がスピーカーとマイクロホンだけでなく、ディスプレイ・カメラをそれぞれ携帯・装着し、マイクロホンからの音と同時に、音声強調のための補助情報を画像情報として聞き手に送信することで、外部雑音の種類に影響されない音声強調の可能性を確認した。

また、同枠組みをマイクロホンと骨伝導マイクの組み合わせに適用する新たな枠組みを提案し、その有効性を確認している。これらの結果を踏まえ、提案手法の応用範囲の拡大に向けさらなる展開を検討している。

用語解説

***1 独立成分分析：混合された複数の信号のそれぞれが統計的に独立であると仮定することで混合信号を混合前の複数の信号に分離するブラインド信号分離技術のこと。**

***2 バイナリマスク：周波数軸上で信号を除去するか否かを決定するために利用される。0、1 の情報を持ち、0 である部分の信号は取り除かれ、1 である部分の信号は残される。というような処理を行うために利用する。**

*3 NRR: Noise reduction ration のこと。ここでは以下のように定義する。

$$NRR = SIR_{after} - SIR_{before} \quad \cdots\cdots (1)$$

ここで SIR_{before}，SIR_{after} はそれぞれ音声強調前，音声強調後の *SIR*(Signal-to-interference ratio) を示す。SIR は以下により定義される．

$$SIR = \frac{\|M(\tau,\omega)S(\tau,\omega)\|}{\|M(\tau,\omega)N(\tau,\omega)\|} \quad \cdots\cdots (2)$$

ここで $N(\tau, \omega)$は雑音信号の和 $n(t)$の時間周波数領域での表現であり以下のように記述される。

$$n(t) = \sum_{i=1}^{N} n_i(t) \quad \cdots\cdots (3)$$

*4 スペクトルサブトラクション:周波数軸上の信号と雑音の混合がその和であると近似し、雑音部分を減算することで目的信号を強調しようとする手法のこと。

参考文献

[1] J. Benesty, M. Souden and Y. Huang, “A Perspective on Differential Microphone Arrays in the Context of Noise Reduction,” IEEE Transactions on Audio, Speech, and Language Processing, Vol.20, No.2 pp.699-704 (2012)

[2] M. Matsumoto, S. Hashimoto “Blind identification of aggregated microphones in time domain,” Journal of the Acoustical Society of America, Vol.121, No.5, pp.2723-2730, (2007)

[3] K. Kumatani, J. McDonough and B. Raj, “Microphone Array Processing for Distant Speech Recognition: From Close-Talking Microphones to Far-Field Sensors,” IEEE Signal Processing Magazine, Vol.29, No.6, pp.127-140 (2012)

[4] C. Ghita, R. D. Raicu and B. Pantelimon, “Implementation of the FastICA algorithm in sound source separation,” Proc. 2015 9th International Symposium on Advanced Topics in Electrical Engineering, pp.19-22 (2015)

[5] S. Richard and O. Yilmaz, On the approximate W-disjoint orthogonality of speech, Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.529-532 (2002)

[6] M. Matsumoto, "Vision-referential speech enhancement of an audio signal using mask information captured as visual data," Journal of the Acoustical Society of America, pp.338-348, Vol.145, No.1, 2019.

[7] M. Matsumoto, "Performance evaluation of speech enhancement of single acoustic signal referring to image information," 2021 Fifth

送受協調型画像音響センサフュージョンによる音声強調

Vision Referential Speech Enhancement Combining an Acoustical Sensor and a Vision Sensor

International Conference on Imaging, Signal Processing and Communications (ICISPC2021), pp.62-66, Online, July 23 - July 25, 2021.
MS 明朝 10/Century10
[8] 松本光春、 "送受協調型画像音響センサフュージョンによる音声強調、" 2019 年 電気学会 電子・情報・システム部門大会、 pp.1181-1182、 沖縄、 2019 年 9 月 4 日-7 日。
[9] M. Matsumoto, "Vision-referential speech enhancement with binary mask and spectral subtraction," The 6th International Workshop on Signal Processing and Machine Learning (SiPML-2020), pp.420-428, Online, Oct. 28 - Oct. 30, 2020.
[10] 川口純輝、 松本光春、 "骨伝導マイクとマイクロホンの組み合わせによる雑音除去" 計測自動制御学会 システム・情報部門 学術講演会 2021、 pp.154-155、 オンライン、 2021 年 11 月 20 日-22 日。

この研究は、平成３０年度ＳＣＡＴ研究助成の対象として採用され、平成元年～令和３年度に実施されたものです。