ClothTalk: 騒音環境でもGPUなしでリアルタイムに綺麗な声を入力可能な導電布マイクロフォン

平城裕隆[1,2,a)] 金澤周介[2,b)] 三浦貴大[2,c)] 吉田学[2,d)] 持丸正明[2,e)] 暦本純一[1,3,f)]

概要：ワイヤレスなウェアラブルデバイスにおけるリアルタイムでの音声強調は，クリアな会話に重要である．特に，大きな声をはらずとも，高品質な音源分離と背景雑音の低減をモバイル環境で実行する必要がある．本研究では限られた計算資源のデバイスにおいても雑音環境下でクリアな音声を入力可能なワイヤレスマイクである ClothTalk を提案する．ClothTalk は導電布のダイヤフラムによるマイクであり，周囲のノイズを到達する角度によらず除去でき，囁き声まで拾える，低消費電力なデバイスである．ClothTalk は，導電布によるフレキシブルなダイヤフラムのマイクを曲げることによって指向性を高め，装着者の発話を支配的に入力するとともに，周囲の全方位からのノイズを低減を可能にする．実験によりダイヤフラムを湾曲させることで声がより入りやすいこと，マイクの角度によって口元が隠れずとも入力できることを示した．加えて，ダミーヘッドによる統制された実験を行い，ノイズを分離する指標である Si-SNR において既存の手法を上まった．本研究は GPU のような豊富な計算資源がない端末でもリアルタイムに音声強調が可能であり，周囲の人の声や雑音がある環境でも音声入力や会話が可能である．

1. はじめに

クリアな音声入力は，オンライン会議などの音声コミュニケーションや，スマートデバイスへの音声入力など，音声を利用したアプリケーションにとって重要である．しかし，背景雑音や干渉する周囲の声がある騒がしい環境においてユーザーの声のみを正確に取得することは，従来のマイクロホンにとって依然として大きな課題である．

これまで音声分野で提案されてきた従来のソフトウェアベースの手法 [1], [2], [3] は，ノイズ特性に関する事前知識に依存することが多く，実世界の環境では利用することが難しい．この事前知識への依存を軽減するために，深層学習によるリアルタイムなノイズ除去の手法 [4], [5], [6] が提案されている．しかしこれらの手法は，競合する周囲の話者の音声のラベルを，ターゲットの音声として誤って出力する問題は依然として解決していない．

この問題に対処するため，HCI 分野の研究では，複数の音声チャンネルやマルチモーダルな手法に基づいたハードウェアでのセンシング手法を提案してきた [7], [8], [9], [10], [11]．しかし，これらの手法は中規模から大規模なモデルの深層学習モデルに依存しているため，GPU がないデバイスでのリアルタイムな利用が制限される．また，特に [8], [11] のようなマルチモーダルな計測によって既存のヘッドセットを拡張する手法は，マイクに近い角度など特定の方向からノイズを受けやすいが，角度によるノイズの入りやすさが考慮されていない．

ClearBuds[12] や In-Ear-Voice[13] のようなイヤホンのマイクによる入力は，ビームフォーミングを利用してユーザーの音声を強調することで，ハンズフリーな音声強調を可能にしている．しかし，これらの方法はユーザーの音声と前方から発せられる対面の話者の音声を区別するのが困難である．

より口元に近い場所での入力手法として，マスク型のマイクである WhisperMask[14] が提案されている．これは騒音環境下でユーザーのささやき声を取得できるが，マスクが皮膚に接触しない距離での録音には対応していない．またマスク越しの会話は表情伝達を妨げており，医療従事者など特定の用途での利用に制限される上，繰り返し使うには衛生上の問題がある．

本研究ではフレキシブルな導電布によるダイアフラム

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[2] 産業技術総合研究所
AIST, Kashiwa, Chiba 277–0882, Japan
[3] ソニーコンピューターサイエンス研究所
Sony CSL Kyoto, Kyoto 600–8086 , Japan
[a)] hiraki-uts1@g.ecc.u-tokyo.ac.jp
[b)] kanazawa-s@aist.go.jp
[c)] miura-t@aist.go.jp
[d)] yoshida-manabu@aist.go.jp
[e)] m-mochimaru@aist.go.jp
[f)] rekimoto@acm.org

© 2025 Information Processing Society of Japan.

図 1 回路とダイアフラムを耳掛け型のデバイスとして統合した．耳掛けの部分とダイアフラム間の部分はフレキシブルなクラフトワイヤーで繋ぎ，使用者がマイクの位置を自由に調整可能である．

を用いた耳掛け型のマイクである ClothTalk を提案する．ClothTalk はノイズの角度依存性がなく，リアルタイムにハンズフリーで GPU を用いずに音声強調が可能であるウェアラブルデバイスである．ClothTalk の設計は，フレキシブルなダイヤフラムを持つマイクロフォンの特性を活用している．具体的には，マイクロフォンの湾曲を強めることで指向性がより強調されることや，角度によらず周囲のノイズを低減できることを明らかにした．また，口の真正面といった理想的な入力でなく多少ずれても同程度の入力が可能であり，顔を隠すことなく入力が可能であることを示した．提案手法のノイズ除去の性能を示すために，実空間での合成データセットを作成して評価を行った．音声強調や音源分離のソフトウェアのタスクでは，クリアな音声を収録し，スピーチノイズや背景ノイズを加えたデータを作成して，元のクリアな音声を推定するタスクを行うが，提案手法はハードウェアのみで音声強調を行うため，統制された環境での実験が必要である．実験の結果，ターゲットとなるユーザーの音声をノイズから分離する指標である Si-SNR において，既存の深層学習による音声強調ソフトウェアの結果を優位に上回った．

本論文の貢献は以下のとおりである．

- 耳掛け型のノイズキャンセルマイクの開発: (約)80 dB の環境ノイズの中で，ささやき声も入力可能なウェアラブルマイクロフォン，ClothTalk を開発した．
- フレキシブルマイクの特性評価：フレキシブルなマイクを曲げることにより音声が入力されやすく，面積が大きい方がより効果的であることを示した．また装着位置についても評価し顔の表情を妨げることなく入力可能であることを示した．
- ノイズの角度依存性の評価：周囲から入力されるノイズが角度によらずおおよそ一定であることを示した．
- Si-SNR での評価：複数話者での対話を想定した騒音環境を実際の空間で作成し，Si-SNR において既存の深層学習を用いてリアルタイムノイズ除去ソフトウェアよりも優れた性能であることを示した．

2. 関連研究

2.1 マイク技術とハードウェアによる音声強調

単一指向性マイクロフォンは，ピンマイクやヘッドセットに一般に用いられるノイズ低減技術の一つである．これらのマイクロフォンでは，ダイアフラムの振動を片側に制限することで指向性を実現し，集音パターンを 180 度に限定している [15]．このような特性は，パネルディスカッションなど，複数の話者が同時に存在する場面において特に有効である．しかしながら，これらのマイクは背景ノイズの減衰が効果的でなく，80dB など高いノイズ環境では追加の信号処理を求められる．

咽頭マイクや非可聴つぶやき（NAM）マイクロフォンは，発話した際に首の表面に表れる振動を圧電素子を用いて音声に変換する手法である [16], [17]．首の表面から装着者の声のみを収集することができ，周囲の環境音は圧電素子を十分に振動させないため背景ノイズへの耐性が高い．しかし，首元に密着して装着する必要があるため，首を振ったり頷くなどの動作によってノイズが生じる．また．音声は体内の皮膚組織を通過した後の声であるため，音声認識に重要なフォルマントが欠損 [18], [19] し，聞こえやすい音声や正確な音声認識を行うための後処理を必要とする．

マイクロフォンを両耳に搭載したイヤホンは，ビームフォーミングを用いて話者の音声を収録することが可能である．例えば，ClearBuds[12] では左右のイヤホンを低遅延に同期することで 2ch のマイクアレイ入力を用いてリアルタイムに音声強調することができる．しかし，イヤホン型マイクによる音声強調は，正面から左右に届く声の到達時間距離 (TDoA) を区別することが難しく，角度による依存性を克服できていない．

WhisperMask[14] は，コンデンサマイクのダイヤフラムを大きくすることで感度が下がり装着者の声を周囲の雑音よりも大きくするマスク型のマイクを提案している．しか

© 2025 Information Processing Society of Japan.

表 1 ウェアラブルなノイズ除去手法の比較

	原理	表情伝達	ハンズフリー	囁き声	GPU 不要	ノイズの角度依存性
UltraSpeech(2022)[9]	マルチモーダル (超音波)	✓		✓		
Clearbuds(2022)[12]	ビームフォーミング	✓	✓		✓	
In-Ear-Voice(2023)[13]	骨伝導マイク+マイク	✓	✓		✓	
EarSE(2024)[11]	マルチモーダル (超音波)	✓	✓	✓		
WhisperMask(2024)[14]	フレキシブルマイク		✓	✓	✓	
ClothTalk(提案手法)	**フレキシブルマイク**	✓	✓	✓	✓	✓

しマスクタイプのデザインは会話中の自然な表情を妨げており，医療従事者など特定の用途での利用に制限される上，指向性や距離に依存した評価がなされていない．

2.2 シングルチャンネルによる音声強調ソフトウェア

音声や信号処理の分野では，音声からノイズを減らしクリアに音を届けることは長年の課題である．wiener fileter や calman filter など信号の特性を用いたものや，IVA[20], MNF[2] など音声の特徴を用いたものが提案されてきた．これらは電源ノイズや救急車のサイレンなど既知のノイズに関しては一定の効果を得ることができる．

また，近年は深層学習による音声強調が提案されており，複数話者の発話や背景のノイズを合成したノイズ付きのデータセットから取り出したいターゲットの音を強調・分離する手法が提案されている．ConvTasNet[4], Denoiser[21], Sepformer[5] などは GPU 環境においてリアルタイムに動作し，GPU がある端末において利用可能である．しかし，これらは音源分離において競合する音源をターゲット音声のラベルにしてしまうパーミュテーション問題が解決できていない．

2.3 マルチモーダルな音声強調

目的とする音声波形以外の，発話に関連するデータを用いて音声を強調する手法も重要なアプローチとして注目されている．音声を用いたものでは，骨伝導マイクや咽頭マイクを用いて，体の表面から得られた自身の音を元に発話区間などの発話情報を取得して強調する手法が提案されている [22]．このような接触型の音声強調は複数チャンネルで音声を扱うことで発話区間のような声の情報を計測できるため優れているが，接触位置の変化によるノイズがあり首振りや頷きなどによるノイズが生じるため制限される．Schilk ら [13] は，イヤホンに取り付けた骨伝導マイクを用いた音声強調を提案している．しかし，接触による安定性の問題があるほか，囁き声など小さい声を取得することが難しい．

超音波やアコースティックイメージング，レーダーなど音声以外のモダリティを用いて，音の反射から発話中における唇の微細な変化を捉えて音声を強調する手法が提案されている．Ozturk らは，ミリ波センシングと深層学習を組み合わせた音声強調と分離システム，RadioSES[7] を提案している．RadioSES は音声と異なるチャンネルで話者の声帯の振動を捉えることで音声を強調するが，移動する人に適用できずモバイルなシナリオでの利用が難しい．UltraSE[8] や UltraSpeech[9] ではスマートフォンでの入力において，発話中における超音波のからの反射をプロファイルとして声と同時に計測し，ノイズでない時のプロファイルを再現するように音声を強調する．しかし，スマートフォンに向かって話す必要があるため，運転中などハンズフリーなシナリオでの利用が難しい．中でも EarSE[11] は市販のセンサーやデバイスのみをヘッドセットに取り付け，超音波の反射によって発話時の皮膚や舌の形状の変化を計測する手法を提案している．これによりハンズフリーな入力を可能にしている．しかし，EarSE は Transformer アーキテクチャを用いた音声強調の推論を行なっており，5 秒の音声クリップの処理に最大で約 15.7 秒かかるなど GPU が付いた環境での利用に制限され，GPU のないスマートフォンやタブレットなどと組み合わせた利用が難しい．

3. ClothTalk: 導電布によるフレキシブルなエレクトレットコンデンサマイク

本研究では，導電布のダイアフラムからなるフレキシブルエレクトレットコンデンサマイクロフォン（ECM）である ClothTalk を提案する．製作過程は，振動を取得するダイアフラムと周辺回路の 2 つの主要コンポーネントから構成される．提案手法により，様々なウェアラブル応用に適した，高い柔軟性と軽量性を持つノイズに強い音声入力デバイスを実現する．

3.1 ダイヤフラムの製作過程

従来のコンデンサマイクは，一般的に円筒形の構造を採用しており，円形のディスクのようなダイヤフラムによってコンデンサを形成している．音圧によってディスク部分が振動し，コンデンサ間の距離が変化することで生じる電圧変動を増幅することで入力信号としている．本研究では，このコンデンサ部分をフレキシブルな導電布で製作した．このコンデンサを実現するため，誘電体としてペルフルオロアルコキシアルカン（PFA）フィルムを採用した．PFA は優れた電気絶縁性を保ちながら柔軟に曲げられることか

© 2025 Information Processing Society of Japan.

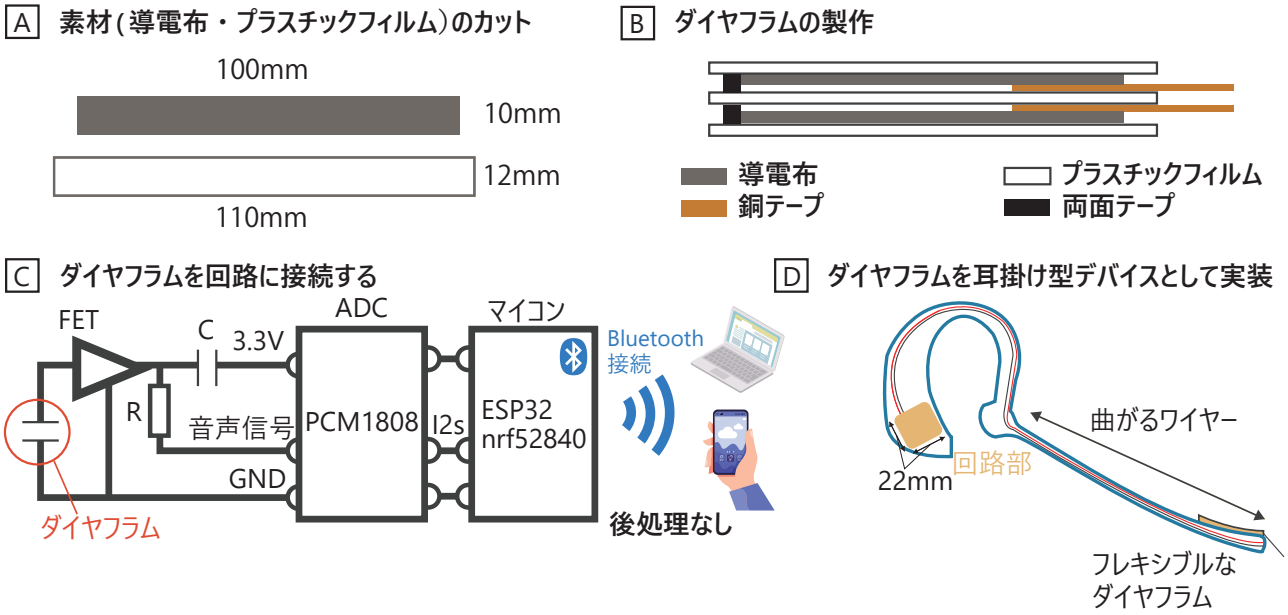

図 2 (a) 材料のカット：導電布 2 枚とプラスチックフィルム 3 枚をレーザーカッターでカットする．(b) 導電布とプラスチックフィルムを積層してダイアフラムを組み立てる．(c) ダイアフラムを回路部分に組み込む．これらのコンポーネント間には ADC（アナログデジタル変換器）が組み込まれている．(d) システム全体を耳にかけるデバイスとして統合する．

ら，フレキシブルなマイクに適した特性を有している．

(1) 厚さ 200μm の銀メッキ導電性布地を 10cm × 1cm のサイズで 2 枚カットした．PFA フィルムは厚さ 12.5μm，寸法 10.5cm × 1.2cm にカットし，導電布同士の接触を防止した．また，5cm × 1cm の布地片と対応するサイズの PFA フィルムを用いて，より短いダイアフラム (5cm × 1cm) も製作し，異なるダイアフラムの長さにおける性能評価を可能にした．

(2) 次に，2 枚の導電性布地の間に PFA フィルムを挟み込んでコンデンサを組み立てた．この積層構造をテープで固定し，風などの環境要因の影響を軽減するため，2 枚の導電布とその間の PFA からなる三層を追加の PFA フィルム層で被覆した．

(3) 周辺回路との接続を可能とするため，コンデンサの電極部分に銅テープを貼付した．

(4) 発話時の呼気によるファブリックコンデンサの擾乱を防止するため，テープ部分の上から最終層として PFA フィルムを被覆した．

以上のプロセスを経て，提案するフレキシブルなダイヤフラムの製作を行った．

3.2 周辺回路

ClothTalk は ECM の原理に基づき動作し，微小な電圧変化を増幅して音声信号を生成する．これをオペアンプを用いて 100 倍に増幅した．増幅されたアナログ信号は，PCM1808PWR アナログデジタル変換器によってデジタル形式に変換され，I2S（Inter-IC サウンド）フォーマットで出力される．このデジタル出力により，ノイズが少ない音声伝送とマイコンの音声処理システムとの互換性を確保している．I2S 出力は，Seeed nrf52840 のマイコンに接続した．このマイコンは 2cm × 2cm の大きさで，Bluetooth 機能を搭載しておりウェアラブルデバイスに利用できる．デバイスの耳掛けの部分にマイコンを固定してコンパクトに設計している．

3.3 耳掛け型デバイスへの統合

提案するデバイスは，耳掛け型のワイヤレスマイクであり，イヤーフックとダイアフラム間の接続には，軽い力で湾曲可能な塩化ビニル被覆軟鋼線を使用した．ワイヤーは半径 2mm と細径であり，目立たない外観を実現している．ワイヤーの先端には，湾曲した ClothTalk ダイアフラムを固定するための治具を 3D プリンタにより製作した．治具は前面（使用者側）と背面の両方向からの音声受信を可能とするため，両面開放型の設計とした．試作した ClothTalk の総重量は約 20g であり，一般的な補聴器等と比較して軽量である．この設計により，長時間での装着が可能となっている．

4. Clothtalk の音響特性

ノイズ低減においては，声の入りやすさとノイズの入りにくさが重要である．本節では ClothTalk の性能について評価する．ClothTalk はフレキシブルなダイヤフラムであ

© 2025 Information Processing Society of Japan.

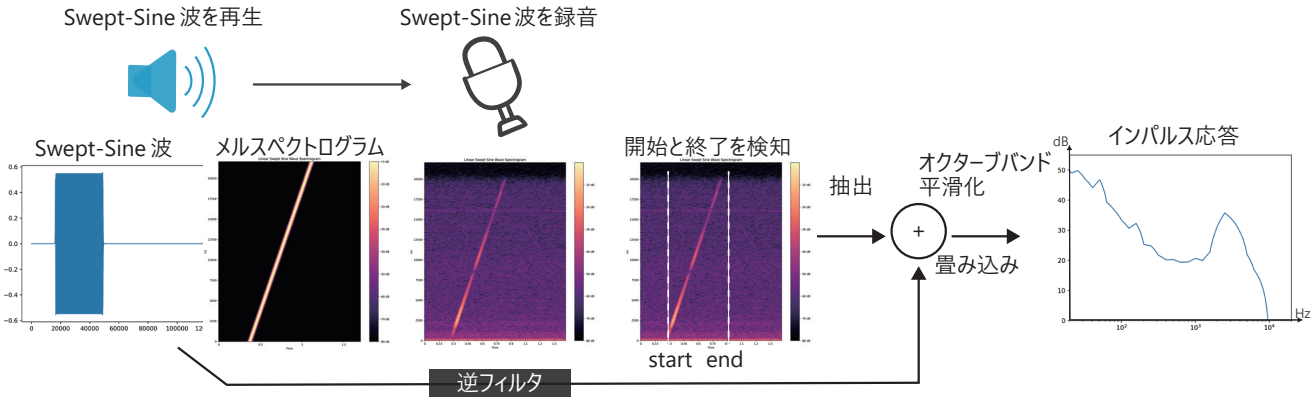

図 3 Swept-sine 法 [23] によるインパルス応答の測定．チャープ信号とその逆フィルタを事前に用意し（左），録音した音声 (中央) に逆畳み込みを適用することでインパルス応答（右）を得る．

り，その曲げによる性能変化が重要であると考えられる．

さらに，ヘッドセットなどのデバイスにおいては，音声入力品質に影響を与える要因として，垂直方向の角度 (口と耳を繋ぐ水平面に対する仰俯角) も重要となる．この角度は，顔を遮ることなく効果的な音声入力が可能かどうかを判断する上で，ユーザーの快適性と実用性に関わる重要な要素である．

また，マイクの性能に対するノイズの影響は，ノイズ源がマイクに対してなす角度によって変化することが知られている．例えば，マイクの正面やマイクに近接した位置からのノイズは，他の位置からのノイズとは異なる影響を及ぼすと考えられる．そこで本研究では，ClothTalk の性能に対するノイズの角度依存性について検討を行う．まとめると本節では以下の Research Question について検討する：

- RQ1. ClothTalk の曲率と長さは音声入力にどの程度影響するか
- RQ2. 口部軸に対する垂直角度は ClothTalk の音声入力にどの程度影響するか
- RQ3. ノイズが入射される角度は ClothTalk のノイズ低減にどのような影響を与えるか

4.1 実験条件

これらの Reasearch Question に対する測定は，提案するマイクの周波数応答特性により評価する．マイクの周波数応答は，20 Hz から 20,000 Hz の範囲で連続波 Swept-Sine 波を用いたインパルス応答により決定する [23], [24]. Swept-Sine 波は 65,536 のサンプリング点で構成され，10 回の測定の平均を取った後，1/3 オクターブバンド平滑化 [25] により処理を行う（図 3 参照）．マイクのインパルス応答の計測には，マウスシミュレータを備えたダミーヘッド SAMAR4700M[26] を使用した．SAMAR4700M は，人工口に関する ITU-T Rec P.51[27] 規格に適合した精密音響測定用ダミーヘッドである．特に音声入力において重要な 315-3150 Hz の周波数帯域では，周波数特性の偏差が 1dB 未満となるよう厳格に設計されており，実験の再現性が確保されている．測定環境として，多孔質の電波吸収材を備えた無響室を使用した．測定時には，壁面から 1 m 離れた位置，地上高 1 m にダミーヘッドを設置した．無騒音状態での音圧レベルは 28.8 dB であった．

4.2 ダイヤフラムの曲率と長さ

まず，曲率と長さに関する評価を実施した．長さは製作した二つの CLothTalk ダイヤフラム，10cm × 1cm と 5cm × 1cm を用いた．曲率については，半径 50 mm，半径 150 mm，平面の 3 種類の構成で測定を行った．これらの曲率に対応するフレームを PLA 材料を用いて 3D プリンタで製作し，曲面に沿わせた状態で測定を実施した．距離は 0 mm，10 mm，20 mm，30 mm，40 mm の 5 段階で変化させた．距離の調整には XYZ 軸フラットスチールステージ [28] を使用し，距離の誤差は 0.5 mm 未満に抑えた．

図 4(右) では，ダイヤフラムの曲率と音声入力の関係を示している．本研究では，R5，R15，平面 (R-flat) の 3 種類の曲率の条件で評価した．結果として，曲率を持つダイヤフラムは平面と比較して優れた音声入力感度を示し，その効果は R5，R15，平面の順に高いことが明らかとなった．また，ダイヤフラムの長さが性能に与える影響についても検討を行った．5cm と 10cm の 2 種類の長さを比較した結果，10cm のダイヤフラムが 5cm のものより優れた性能を示し，長さの増加が音声入力の指向性向上に寄与することが示唆された．これらの結果は，曲率と長さがフレキシブルなマイクの設計における重要なパラメータであることを示唆している．

4.3 垂直角度を変えた際の評価

ヘッドセットやインカム型のマイクなど，ユーザーが自身でマイクの位置を調整可能なデバイスにおいては，位置

© 2025 Information Processing Society of Japan.

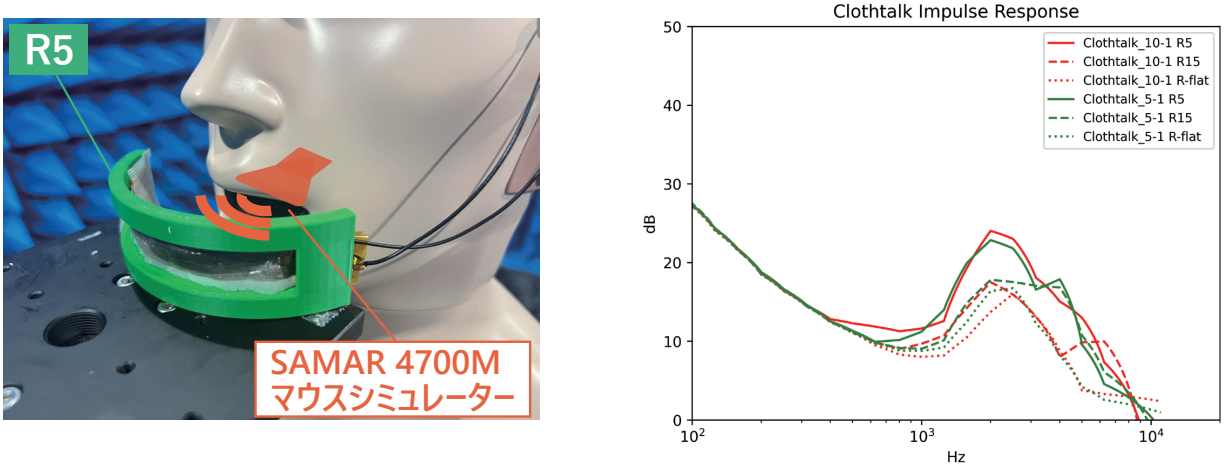

図 4 ダイヤフラムの曲率と長さが音声入力に与える影響．（左）計測状況：ダミーヘッド (SAMAR4700M) から Swept-Sine 波を出力し，曲率を変えた際の入力を評価する．（右）ダイヤフラムの曲率（R5, R15, 平面:R-flat）と長さ（10cm × 1cm: Clothtalk_10-1，5cm × 1cm:Clothtalk_5-1）における音声入力の周波数応答の結果．グラフは，曲率を持つダイヤフラム（R5 および R15）が平面と比較して高い感度を示し，R5 が最も良好な性能を発揮することを示している．また，10cm のダイヤフラムは 5cm のものより優れた性能を示し，長さの増加に伴い指向性が強まることを示唆している．

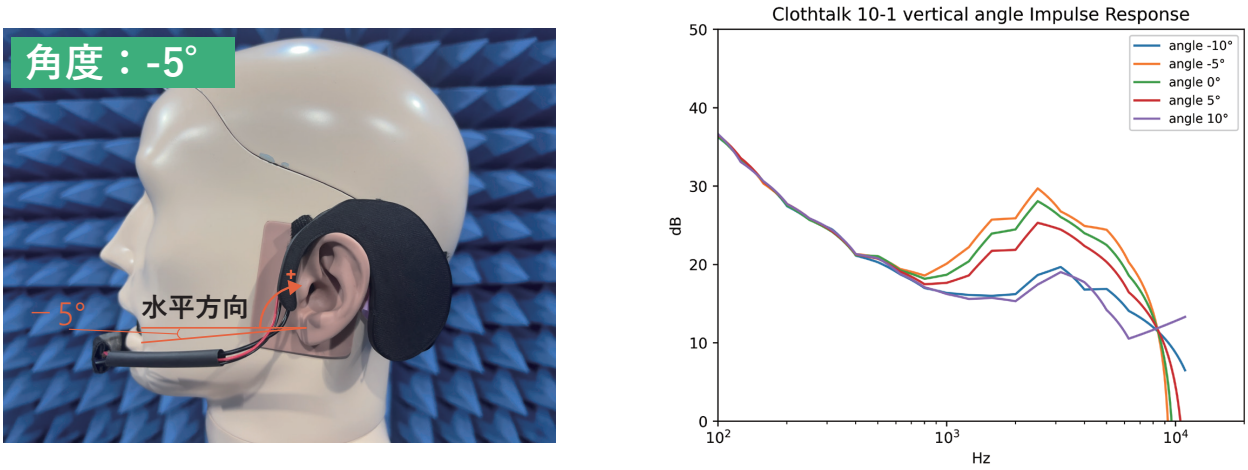

図 5 10cm × 1cm のダイヤフラムにおいて，垂直方向の角度 (口と耳を繋ぐ水平面に対する仰俯角) がインパルス応答に与える影響．（左）水平面に対して-5 度の角度でマイクを配置したダミーヘッドでのセットアップ．（右）異なる垂直角度（-10 度，-5 度，0 度，5 度，10 度）において 100 Hz から 10 kHz の周波数範囲でのインパルス応答の結果．特に 800 Hz から 8 kHz の中周波数帯域において，0 度や± 5 度の位置と，± 10 度の位置では顕著な違いが観察される．

の変化に対する適応性が重要となる．本研究ではこの適応性の程度を定量化した．測定は，マイクが口部正面に位置する 0 度を基準とし，口と耳を繋いだ水平面を基準として-10 度，-5 度，0 度，+5 度，+10 度の 5 点で実施した（頭頂部を正の方向としている）．結果は図 5 に示す通りである．-5 度から+5 度の範囲では入力が一貫して効果的に維持されることが確認された．-5 度の位置では，マイクは口部から約 2cm 下方に位置し，口の表情を視認できる．しかしながら，より極端な角度においては性能の大幅な低下が観察された．具体的には，鼻の位置である 10 度の傾斜により入力感度が最大 10 dB 低下し，音声入力の有効性が損なわれる可能性が示唆された．

4.4 ノイズ源の角度による影響

ノイズの影響はその発生源の方向により変化する．本研究では，30 cm の距離から 94 dB の Swept-sine 波を出力した際にマイクが受ける音の特性を評価した．具体的にはある位置から発せられた Swept-Sine 波のインパルス応

© 2025 Information Processing Society of Japan.

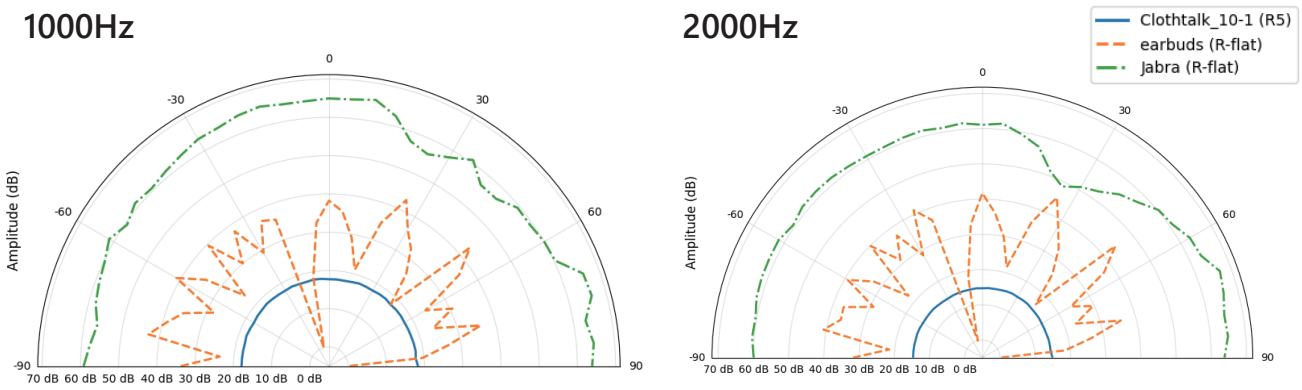

図 6 10cm × 1cm のダイヤフラムを 5 度間隔で回転させて実施した，周囲のノイズの角度による影響の結果を示す．同一音圧レベルにおいて，ClothTalk はノイズに対する感度が低く，環境音に対する耐性が 180 度全方向に対して高いことが示された．

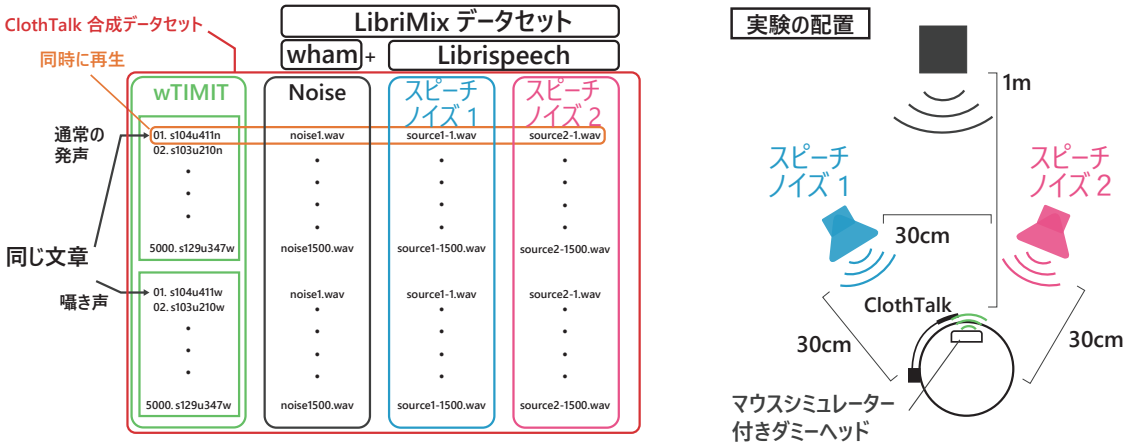

図 7 物理的に音声とノイズを組み合わせることで合成データセットを生成した．音声データは wTIMIT データセットから取得し，ノイズは Librispeech Dataset を含む Wham データセットを用いて生成した．

答を計測し，それぞれの周波数ごとの音圧を計算する．測定は，真正面を 0 度として-90 度から+90 度までの 180 度の円弧上で，5 度間隔で記録を行った．ノイズ源として使用した Sony XB100 スピーカーは，20 Hz から 20,000 Hz の範囲で平坦な周波数特性を示している．計測は 10cm × 1cm のダイヤフラムで行い，比較対象として，ヘッドセット（Jabra Evolve2 40）およびイヤホン型マイクを使用した．イヤホン型マイクは Clearbuds[12] 筐体に MEMS マイクを組み込み，左右 2 チャンネルを有線で同時記録可能な構成とした．この評価の結果は図 6 に示す通りである．

結果として，1000 Hz や 2000 Hz などの代表的な周波数において，提案するマイクはイヤホン型マイクや Jabra のようなヘッドセットに比べてノイズの干渉を受けにくい特性を示した．同様の傾向は他の周波数帯域においても確認された．

5. 音声をノイズから分離する性能の評価

5.1 合成データセット

音声強調タスクは一般的に，クリアな環境で収録された音声に対して収録後にノイズを付加し，信号処理 [2], [3] や深層学習による推論 [4], [5], [21] によって元の音声を抽出する手順が採用される [11]．このソフトウェアでの合成データセットでは，実験参加者のクリアな発話に対してノイズを加えるタイミングを精密に制御することが可能である．しかしながら，提案する ClothTalk システムではフレキシブルなダイヤフラムの曲げによるハードウェアでのノイズ低減を行っているため，録音した時点でノイズがなく，録音後にノイズを付加しても取り除くことはできない．そのためソフトウェアベースの音声強調技術で一般的な，音声信号の修正後のノイズ付加と除去という手順での比較は難しい．さらに，ソフトウェアでは 50msec 程度の区間で推論が可能である [4] が，実験参加者の発話のタイミングはこのような細かい制御は極めて困難であるので，実験参加

© 2025 Information Processing Society of Japan.

表 2 リアルタイム音源分離と音声強調モデルを用いた提案手法との性能比較

デバイス	STOI ↑	PESQ ↑	SI-SNR ↑	Si-SDR ↑
ClothTalk(10cm × 1cm)	0.087	1.061	-32.229	**-32.229**
Jabra+Sepformer[5]	0.178	1.183	-54.032	**-54.032**
Jabra+Denoiser[21]	0.123	1.117	-49.012	-49.012
Jabra+ConvTasNet[4]	0.317	1.066	-41.163	-41.163

者の発話とノイズ出力を同期することは難しく，声とノイズが被らないことが多い．これらの課題に対処するため，本研究ではダミーヘッドを用いた合成データセットによる評価を採用した．この手法により，実環境を模擬しながら制御された実験が可能となる．本節では，合成データセット収集に使用したスピーカーとマイクロフォンの構成について説明し，データセット自体の詳細，評価指標，比較基準について概説する．その後，分析結果を提示する．

5.1.1 スピーカー設定

音の声源となるスピーカーは 4 節で説明したマウスシミュレータ付きダミーヘッド・SAMAR4700M を使用した．会話を想定して，音声のノイズを出力するスピーカーとして Sony XB1002 台をダミーヘッドから 30cm 離して配置した．さらに背景ノイズのスピーカーをダミーヘッドの正面の 1m 先に設置した．30cm の距離は一般的な会話距離を想定して選定し，配置は図 7 に示す通りである．音声ノイズのスピーカー 2 台と背景ノイズのスピーカーの出力レベルは，ダミーヘッドの口部位置で 60dB となるよう調整した．ダミーヘッドの音声出力は，会話音量を想定して 30cm 先で 60 dB となるよう校正した．騒音レベルは RION NA-28 Class 1 騒音計を用いて測定し，精度± 0.1 dB を確保した．全ての音声は 16000 Hz のサンプリングレートで再生した．

5.1.2 計測するマイク

4 節の結果に基づき，測定用の ClothTalk マイクロフォンには曲率半径 5cm，長さは 10cm × 1cm のものを選定した．比較対象として，市販のノイズキャンセリングヘッドセットである Jabra Evolve2 40（以下，Jabra）を使用した．

5.1.3 データセット

音声再生には，同一話者による通常発話と囁き声を含む wTIMIT コーパス [29] を使用した．wTIMIT コーパスは TIMIT コーパスの一部を採用し，29 名の話者からの音声を収録している．ノイズ生成には，音源分離用の WHAM! データセットを使用した．WHAM!データセットは WSJ コーパスから選定された 2 名の話者の音声と様々な背景ノイズサンプルを組み合わせたものである．再生と録音は MacBook Pro 2019 で行い，タイミング制御はマルチプロセスにより管理した．プロセス間遅延は 1 ミリ秒と測定され，十分な同期精度を確保した．wTIMIT データセットから 5,000 サンプルを選定し，同一話者が同一文を通常発声と囁き声の両方で読み上げた事例に焦点を当てた．各通常発声と囁き声のペアに対して，同一のノイズ構成（2 つの音声ノイズ源と 1 つの背景ノイズ源）を適用した．

5.1.4 評価指標

背景ノイズから音声を分離する効果を評価するため，広く採用されている 4 つの指標を使用した．これらの指標はクリーン音声信号の復元品質を評価するものである：

- SiSNR $(-\infty to + \infty)$：スケール不変信号対雑音比．SiSNR は信号電力対ノイズ電力の比を定量化する指標である．従来の SNR が振幅変動の影響を受けるのに対し，SiSNR はスケーリング計算を組み込むことで，異なる測定環境や構成間での一貫した比較を可能とする．
- SiSDR $(-\infty to + \infty)$：スケール不変信号対歪み比．SiSDR は信号の全体的な歪みを評価し，ノイズだけでなく信号自体の歪みも考慮する．信号強度の変動にも対応するため，振幅スケーリングを組み込んで SDR の概念を拡張している．
- STOI (0 to 1)：短時間客観的明瞭度指標．STOI は処理された音声信号の明瞭度と明確さを客観的に評価する指標である．0 から 1 の間のスコアを出力し，1 が最大の明瞭度を表す．1/3 オクターブバンド分析を採用し，人間の聴覚知覚の周波数分解能を近似している．
- PESQ (1 to 5)：音声品質の知覚評価．PESQ は音声品質評価の国際標準である．POLQA の前身として開発され，客観的音声品質測定の基礎を確立した．PESQ の分析範囲は一般的に 3400Hz 程度までをカバーし，従来の電話帯域幅を反映している．

5.1.5 ベースライン手法

比較評価のため，以下の 3 つのベースライン手法を用意した：

- ConvTasNet[4]：時間領域で直接音声分離を行う全畳み込み型ニューラルネットワークである．周波数領域ではなく時間領域で音源分離を直接実行することで，リアルタイム処理を実現している．
- Denoiser[21]：ConvTasNet と同様に時間領域で直接音声強調を行うニューラルネットワークである．U-Net 構造に類似したスキップ接続を持つ LSTM 層を採用しており，CPU でのストリーミング処理を可能としている．
- Sepformer[5]：RNN を使用せず Transformer のみを用

© 2025 Information Processing Society of Japan.

いた音声分離モデルである．この設計により並列処理が可能となり，従来の RNN ベースの手法と比較して，より高速かつメモリ効率の良い実行を実現している．

4 つの音源分離モデル（ClothTalk，Jabra Sepformer，Jabra Denoiser，Jabra ConvTasNet）について，上記の 4 つの性能指標を用いて評価した．

結果は表 2 に示す通りである．結果から得られた主な知見は以下の通りである：STOI：Jabra ConvTasNet が最も良い STOI スコア（0.316858）を達成し，優れた音声明瞭度を示した．一方，ClothTalk は最低の STOI スコア（0.087090）となった．PESQ：Jabra Sepformer が最も良い PESQ スコアを示し，最良の知覚音声品質を実現した．ClothTalk は最低の PESQ スコア（1.061325）を示したが，モデル間の PESQ スコアの差は比較的小さかった．SI-SNR および SDR：ClothTalk は両指標において他のモデルよりも良い性能（-32.228974）を示し，次点の Jabra ConvTasNet（-41.162827）と比較して顕著な改善を達成した．

性能トレードオフ：ClothTalk はノイズ低減と歪み抑制（SI-SNR と SDR スコアで示される）において優れた性能を示したが，音声明瞭度（STOI）と知覚品質（PESQ）では他のモデルと比較して低い性能となった．一貫性：Jabra Sepformer と Jabra Denoiser は全指標において安定した性能を示し，概して Jabra Sepformer が優位な結果となった．これらの結果は，各モデルが異なる強みと弱みを持つことを示唆している．ClothTalk はノイズ低減と歪み抑制において特に効果的である一方，Jabra ConvTasNet や Jabra Sepformer はそれぞれ音声明瞭度と知覚品質において優位性を示している．

6. 議論と今後の課題

本研究では ClothTalk が大規模な深層学習を用いた従来手法を上回るノイズ除去性能を示したことを明らかにした．しかし，ユーザーの音声明瞭度（STOI スコア）や音声品質（PESQ スコア）に関しては既存のヘッドセットと比較して劣っており，特に低い音声強調能力において改善が必要である．現在は提案手法において単に録音した状態と，ノイズキャンセリングヘッドセットに音声強調モデルを適用した場合を比較しているため，提案手法のイコライザを調節することや低遅延な音声強調モデルを組み合わせることでより良い音質を実現できると考えている．また，実際の使用環境での長期評価や多様なユーザーに対するテストが不足しており，今後はより広範なユーザー層への適用性を評価することが課題である．さらに，本研究の実験では固定された条件でのノイズと音声入力の評価を行ったが，動的に変化する環境での性能評価が不足している．例えば，ユーザーが歩行中や車内などの移動中のシナリオにおいても安定した音声入力が可能であるかについて検討する必要がある．これらの動的環境下での性能を評価することで，ClothTalk の実用性をさらに向上させるための改善点を明らかにすることが期待される．

7. 結論

本研究では，GPU が不要でハンズフリーなウェアラブル音声入力デバイス ClothTalk を提案し，全方向からのノイズに対応しながらユーザーの声を強調することに成功した．実験結果から，ClothTalk は騒音環境下においても既存の深層学習ベースの音声強調手法と比較して優れたノイズ除去性能を示し，特に実際の音響環境でのノイズ耐性を持つことが確認された．

しかし，音声明瞭度と音声品質においてはまだ改善の余地があり，特にユーザーの快適性を損なわずに明瞭な音声を収録するためのさらなる設計改善が必要である．今後は，より広範な環境および多様なユーザーシナリオでの評価を通じて ClothTalk の実用化を進める予定である．

謝辞

本研究は JST ACT-X グラント番号 JPMJAX23KG, JST ムーンショット型研究開発事業グラント番号 JPMJMS2012, 国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）委託業務（JPNP23025）の支援を受けたものです．

参考文献

[1] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

[2] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[3] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[4] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.

[5] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25, 2021.

[6] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[7] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang,

© 2025 Information Processing Society of Japan.

Min Wu, and K J Ray Liu. RadioSES: mmWave-Based audioradio speech enhancement and separation system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1333–1347, 2023.

[8] Ke Sun and Xinyu Zhang. Ultrase: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, MobiCom '21, pages 160–173, New York, NY, USA, September 2021. Association for Computing Machinery.

[9] Han Ding, Yizhan Wang, Hao Li, Cui Zhao, Ge Wang, Wei Xi, and Jizhong Zhao. Ultraspeech: Speech enhancement by interaction between ultrasound and speech. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), sep 2022.

[10] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3):1–30, September 2021.

[11] Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. Earse: Bringing robust speech enhancement to COTS headphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4):1–33, January 2024.

[12] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M. Seitz. Clearbuds: Wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '22, page 384–396, New York, NY, USA, 2022. Association for Computing Machinery.

[13] Philipp Schilk, Niccolò Polvani, Andrea Ronco, Milos Cernak, and Michele Magno. In-Ear-Voice: Towards Milli-Watt audio enhancement with Bone-Conduction microphones for In-Ear sensing platforms. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, IoTDI '23, pages 1–12, New York, NY, USA, May 2023. Association for Computing Machinery.

[14] Takahiro Miura Manabu Yoshida Masaaki Mochimaru Jun Rekimoto Hirotaka Hiraki, Shusuke Kanazawa. Whispermask: a noise suppressive mask-type microphone for whisper speech. In *The Augmented Humans International Conference (AHs 2024)*, pages 1–16, Melbourne, VIC, Australia, 2024.

[15] Benjamin B Bauer. A century of microphones. *Proceedings of the IRE*, 50(5):719–729, 1962.

[16] Shota Shimizu, Makoto Otani, and Tatsuya Hirahara. Frequency characteristics of several non-audible murmur (nam) microphones. *Acoustical Science and Technology*, 30(2):139–142, 2009.

[17] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 5, pages V–708, 2003.

[18] Amritha Vijayan, Bipil Mary Mathai, Karthik Valsalan, Riyanka Raji Johnson, Lani Rachel Mathew, and K. Gopakumar. Throat microphone speech recognition using mfcc. In *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*, pages 392–395, 2017.

[19] Junki Kawaguchi and Mitsuharu Matsumoto. Noise reduction combining a general microphone and a throat microphone. *Sensors*, 22(12), 2022.

[20] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis: An extension of ica to multivariate components. In *International conference on independent component analysis and signal separation*, pages 165–172. Springer, 2006.

[21] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. In *Proc. Interspeech 2020*, pages 3291–3295, 2020.

[22] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–2517, 2012.

[23] angelo farina. simultaneous measurement of impulse response and distortion with a swept-sine technique. *journal of the audio engineering society*, february 2000.

[24] Nivedita Arora, Steven L. Zhang, Fereshteh Shahmiri, Diego Osorio, Yi-Cheng Wang, Mohit Gupta, Zhengjun Wang, Thad Starner, Zhong Lin Wang, and Gregory D. Abowd. Saturn: A thin and flexible self-powered microphone leveraging triboelectric nanogenerator. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2), jul 2018.

[25] K Elenius. Long time average spectrum using a 1/3 octave filter bank, 1980.

[26] Southern Acoustics. Samar 4700 (artificial mouth compatible) - if you are looking for acoustic measurement equipment or dummy heads, southern acoustics, 2022. Accessed: 2022-10-26.

[27] International Telecommunication Union. Artificial mouth. Technical report, ITU-T Recommendation P.51, 1996.

[28] SigmaKoki Corporation. Xyz axis steel extended contact translation stages tsd-1005s. `https://jp.optosigma.com/en_jp/tsd-1005s.html` (Accessed: 2024-09-08).

[29] Boon Pang Lim. Computational differences between whispered and non-whispered speech. *PhD Thesis UIUC*, 2010.

© 2025 Information Processing Society of Japan.