WhisperMask: 騒音環境で音声入力可能なマスク型マイク

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：騒がしい環境や複数人が同時に話す場合，意図しない音声の干渉により，音声入力を使用することは困難である．既存のマイクは遠くの音や皮膚に接触するノイズも拾ってしまうほか，既存のノイズ除去ソフトウェアは囁き声のような小さな声をノイズの中から分類することが難しく，声を強調して話す必要がある．本研究では騒音環境で利用可能なコンデンサマイク，WhisperMask を提案する．WhisperMask は，コンデンサマイクの振動板を導電布で作成することで，装着者の声のような導電布を駆動させるような音しか入力されない．特に，200 Hz から 5 kHz 以下の周波数帯域では，装着者による約 80 dB の音声は，周囲の雑音の入力に対して約 10 dB 大きく入ることを明らかにし，騒音環境下における囁き声の音声認識が既存のマイクより優れていることを示した．提案するマイクは，騒音環境以外においても音声の後処理なくユーザーの声のみをクリアに拾えるため，通話や音声入力，音声変換など様々な音声インタラクションに応用可能である．

1. はじめに

スマートデバイスの普及と音声インタフェースの進化にともなって，外出先での連絡は身近なものとなった．特に Apple Airpods は 2020 年には 100 万台売れている [1] ほか，smart watch[2], [3] や smart ring[4] などにもマイクが搭載されてきている．このように音声通話だけでなくスマートアシスタントの対話的な入力 [5] でも期待されており，ウェアラブルな環境での音声入力はますます重要になっている．これらのインタフェースは便利になっているにもかかわらず，外出先での利用には，他の話者の声や車・電車などの環境音によって装着者の音声が干渉されてしまうといった技術的な課題がある [6]．そこで本研究では新たなマイクの振動板の形状を提案することで，利用者の音声のみを取得するマスク型のインタフェースを探求する．

周囲のノイズなしに音声入力を行うことは積年の課題であり，ハードウェア，ソフトウェアに渡って何十年にもわたって取り組みがなされている．ハードウェアにおいては，単一指向性マイク [7] に始まり，咽喉マイク [8]，NAM マイク [9]，イヤホン型マイク [10], [11] と進化しており，マイクが音を拾う原理と信号処理と組み合わせることでノイズの低減を行なっている．これらはさまざまな特性があり，騒音環境で，ウェアラブルに，接触のノイズがなく，囁き声のような小さい声まで拾うことが難しい．

一方で，ソフトウェアを用いて周囲の環境から目的とする自身の声を取得することは，音声信号処理の分野では古くからある問題であり，Blind Source Separation (BSS)[12] や Speech Enhancement[13], [14] による手法が提案されてきている．これらは，無指向性の単一のマイクや上記の特殊なマイクで得られた音声に信号処理や機械学習を用いた後処理を行うことでノイズを除去するが，特定のノイズデータに対して学習してしまうことで囁き声や小さい声などノイズに埋もれやすい声に対応することが難しい．

本研究では，マイクの振動構造を変更することにより，ハードウェアのみでノイズへの耐性を高めるマスク型のマイクを提案する．これはピンマイクと同様のエレクトレットコンデンサマイクであるが，ダイヤフラムの素材と大きさが異なっており，従来のコンデンサマイクより柔らかい素材である導電布で構成され，数十倍の振動面積を持つ．提案するマイクはマイク単体で，ソフトウェアの後処理なく装着者の音声を支配的に入力できることを確認した．特に騒音環境において優れた Signal-noise Ratio(SNR) を得ており，周囲の雑音が (約)60 dB 以下の環境では外部のノイズが入らない．装着者が歩行している場合においても音声入力が正しく動作する点において優れている．また，音

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

声を録音するだけで装着者の声のみを取得できるため既存のソフトウェアによるノイズ低減処理に比べて後処理が不要である．

本論文の貢献は以下のとおりである．

- マスク型マスクの開発: (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，WhisperMask を開発した．
- マイクの音響特性の評価：WhisperMask のマイクは，200 Hz から 5 kHz の周波数帯域において，外部ノイズに対して装着者の声の入力が 10 dB 程度大きく入力されることを明らかにした．
- SNR での評価：WhisperMask は，周囲のノイズが 60 dB 以下の環境では周囲のノイズが 30 dB の時と変化せず，70 dB のノイズ環境において，既存のマイクの SNR を 10 dB 上回ったことを示した．
- 音声認識の比較： 本装置は，ソフトウェアによるノイズ抑制システムと比較して，30%pt 高い音声認識率を達成し，ノイズの多い状況下でも通常の音声とささやき声の両方を入力可能であることを示した．

2. 関連研究

声によるインタラクションは人が行える重要なモダリティの一つであり，電話やオンライン通話や音声コマンド入力，スマートアシスタントとの対話的操作など，様々な用途で用いられている．そのため，明瞭な音声入力データを取得することは重要な課題であり，他の人が話している環境や地下鉄や工事現場といった周囲で騒音が発生しているような環境でも安定して利用できることが重要である．これを実現する手段として，音声，信号処理，インタラクションなどの分野で長年に渡って研究が行われてきており，ハードウェアとソフトウェアの両面から提案がなされてきた．特に，マイクの長さが長くなることや個数が増えることで，マイクの指向性が上がり，利用者の声を選択できることが知られている [15] が，これらは大規模になるためウェアラブル環境での利用が想定されていない．そのため，インタラクションに重要な，ウェアラブルな構成でかつリアルタイムに動作するものを中心にまとめる．

2.1 ノイズを低減するウェアラブルマイク

ハードウェアにおいては，単一指向性マイク [7] に始まり，咽喉マイク [8], NAM マイク [9], イヤホン型マイク [11] など，マイクが音を拾う原理と信号処理と組み合わせることでノイズの低減を行なっている．

2.1.1 単一指向性マイク

ピンマイクやヘッドセットなどに見られる単一指向性マイクは，今日において最も手軽に利用されるノイズ低減手法の一つである．これはマイクの中の振動板であるダイアフラムの振動方向を裏側から塞ぎ，片側に制限することで指向性を 180 度に絞ることができる [7]．これによってマイクに対して限定的な方向の音を強く記録することができ，パネルディスカッションのような複数話者が同時に話す際に重要である．しかし，これらは音の距離を制限することはないため，背景雑音をマイク単体で除去することができず，騒音環境でそのまま利用することが難しい．

2.1.2 アレイマイク

アレイマイクは複数のマイクを搭載しており，発せられた音がそれぞれのマイクに到達する時間差を利用することで，音の到来方向を絞る，beamforming[12] が行われている．これによって発話の話者を選択することができる．しかし，音声の方向を絞り込むにはマイクが円弧上あるいは直線上に複数必要であるため，ウェアラブルな利用が難しい．さらに，音の到来方向を推定できても，背景雑音か否かを判断することはできず，背景ノイズに対応するための後処理を必要とする．

2.1.3 咽頭マイク

咽頭マイクは，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 [8]．首元に取り付けて表面の音のみを取得することで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 [16], [17].

2.1.4 NAM マイク

NAM（Non-audible murmur）マイクは咽頭マイクのように耳の後ろの皮膚に接触させるマイクであるが，無指向性マイクにシリコンを取り付けることで皮膚とマイクの境界をなめらかにつなぐ音声入力デバイスである [9]．NAM マイクは咽頭マイクと同様に周囲の環境音の影響を大幅に低下させるが，頷きなどがノイズになる．また，咽頭マイクとは異なり耳の後ろに装着することで音を取得するが，同様に音声強調などの後処理が必要 [18], [19], [20] である．

2.1.5 イヤホン型マイク

イヤホンに挿入するイヤーピース型のマイクが提案されており，左右両耳にイヤホンが取り付けられている．イヤホン型マイクは運動や生体データのモニタリングによって健康情報取得も行われている [21]．イヤホン型マイクは両耳に搭載されたイヤホンで beamforming を行なっており，装着者の音声を選択的に取得することができ，ハードウェアオープンソースによって開発もより簡単になっている [10]．両耳で取得した音声からノイズを低減する機械学習の手法も提案されている [11]．しかし，これらは囁き声のようなはっきりと話さない声に対しては入力が難しい．

表 1 原理に基づいたマイクの分類

マイクの種類	ノイズ低減の原理	背景雑音の影響	ウェアラブル	接触ノイズ	囁き声
単一指向性マイク	振動方向の制限	△	✓	✓	✓
アレイマイク	ビームフォーミング	△		✓	✓
咽頭マイク	圧電マイク	○	✓		
NAM マイク	接触型マイク	○	✓		
イヤホン型マイク	音源分離	△	✓	✓	
WhisperMask	**導電布の振動板**	◎	✓	✓	✓

2.2 音源分離・音声強調によるノイズ低減

周囲の環境から目的とする自身の声を取得することは，音声信号処理の分野では古くからある問題であり，Blind Source Separation (BSS)[12], [22] や Speech Enhancement[13], [14] という手法が提案されてきている．Blind Source Separation では音源の発生位置や音源の混ざり方が未知である音声を目的の音声とそれ以外に分離する，いわゆるカクテルパーティー問題を解くことが主題である．

BSS ではマイクを増やして声が伝播する空間の分解能をあげること [15] や，複数話者は異なる内容を話すという独立性を手がかりに音源を分離する手法 [23] や，行列を低次元の行列に分解するといった信号処理の手法 [24], [25] が長らく提案されてきた．また，これらと deep learning を組み合わせた手法 [26], [27] が存在する．さらに小規模なモデルで動作する話者分離手法や音声強調の手法が提案されており [28], [29], [30]，リアルタイムで動作する．

しかし，これらの機械学習による手法は，特定の英語話者データセットや，特定のノイズデータセットを元に学習しているため，汎化性能に限界がある．例えば，通常の話し声のデータに最適化されているため，囁き声の入力や，騒音環境下での囁き声など実環境でのデータとの乖離がある．

2.3 マスク型インタフェース

COVID-19 によってマスクの利用が増え，日常的に利用する機会が増えたことでマスクは新たな研究対象として着目されつつある．マスクを身につけることによって表情が隠れてしまうという課題があり，これに対してフォトリフレクタや静電容量センサーで表情を読み取って，LED やディスプレイで提示する手法 [31], [32], [33] が提案されている．

また，マスクが顔に接するインタフェースであることから顔や口に関するインタラクションを行うウェアラブルインタフェースとしても利用されており，呼吸の検知 [34] やアイトラッキング [35]，口の形状認識 [36], [37], マスクの紐などを用いた着脱検知 [38] なども行われている．

また，マスクによって声が遮断されてしまい，音声が聞き取りにくくなる現象が指摘されている [39], [40]．これを解決する手法として，マスクにセンサーを埋め込むデバイスによって，無声発話の入力であるサイレントスピーチインタラクション [41], [42] を行う手法も提案されている [43], [44]．しかし，これらの手法はコマンドを認識して alexa などのスマートアシスタントとコミュニケーションするものであり，幅広い音声認識や会話は難しい．

本研究ではマスクの中にマイクを入れることによって，こもることなく音声を取得でき，さらにそれ自体が周囲のノイズを低減できる手法を提案するため，このようなマスク型インタフェースにおける課題を解決できることが期待できる．

3. WhisperMask

本研究では，騒音環境でも入力可能なマスク型マイクである WhisperMask を提案する．WhisperMask はエレクトレットコンデンサマイクロフォン (ECM) の原理に基づいて，導電布とエレクトレットによって実装している．

3.1 動作原理：エレクトレットコンデンサマイクロフォン

ECM は電源部分と振動板 (ダイヤフラム) の部分があり，振動板はコンデンサを形成する電極とエレクトレットから構成される．音がダイヤフラムを物理的に振動させることでコンデンサの電圧が変化し，Field effect transistor (FET) や AD 変換を経て音になる．本研究では，ECM における新たな振動板を提案し，振動板をデザインすることで話者の声のみを入力することが可能になる．

3.2 WhisperMask の設計

提案するマイクの振動板は，電極である導電布とエレクトレットなプラスチックフィルムからなるエレクトレットコンデンサで構成されている．プラスチックフィルムは PFA(Per Fluoro Alkoxy polymer) であり，厚さ 12.5μm の PFA フィルムを使用した．フィルムの外周に防風用の粘着テープを貼り，その両面に厚さ 200μm の導電布を固定した．これにより，エレクトレットを 2 枚の電極で挟んだエレクトレットコンデンサの構造を形成した．

振動板は Field effect transistor (FET) に接続されて音声信号を出力するが，提案するマイクは市販のコンデンサマイクに比べて電圧値が低いため，Field effect transistor

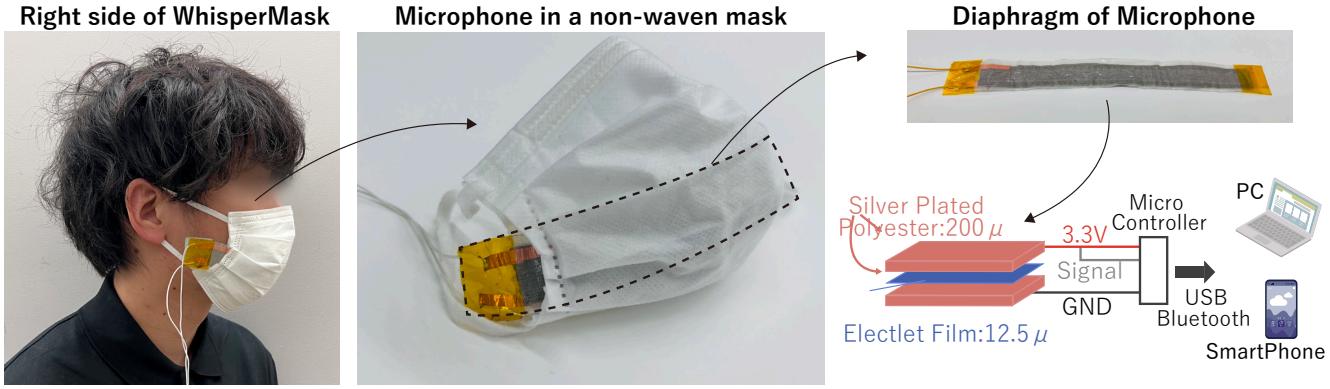

図 1 マスク型マイクである WhisperMask は，ハンズフリーで目立たない音声入力が可能である (左)．マイクは不織布の中に埋め込まれている (中央)．振動板をマイコンに接続し，USB や bluetooth 経由で PC・スマートフォンに接続可能である．

(FET) によって 3 倍に増幅した．また，入力電流はオーディオ用 SoC (System on a Chip) である BM63 (Microchip Technology) を経由して Bluetooth で接続され，60 mm × 50 mm のケースに収まっており重さは 14 g である．このようにして得られたダイヤフラムと計測回路をメッシュタイプの布からなる通気性の良いマスクを作り，その中に実装した．マスクには切れ込みを入れてあり，ダイヤフラムの位置が固定できるようになっている．

3.3 振動板のパターン設計

WhisperMask はマスク型マイクであり，マスクを装着することで利用する．しかし，マイクの振動板が重くなるとマスクがずれ落ち，認識の低下につながる．そこで本研究ではマスクに埋め込められるよう 200 × 40 mm, 200 × 20 mm，100 × 20 mm(図 3) の 3 つを用意し，実装を行なった．(図 1).

4. WhisperMask の周波数特性

マイクにとって周波数特性は最も重要な要素の一つである．これはマイクによっては特定の周波数を拾わなかったり，一部の周波数帯を強調しすぎるためである．例えば音声認識は音声を用いたインタラクションにとって非常に重要であるが，人間の音声の周波数帯は 100 Hz – 5 kHz になっており，この周波数帯を感知しないマイクは音声認識が非常に困難である．また，ノイズに対して信号がどれぐらいはっきりと見えるかが重要であり，SNR によって計測される．SNR の値が低い場合は，得られた信号のうち多くがノイズであるため，話者の音声など目的の信号を明確に拾うことが難しい．

4.1 Swept-Sine によるインパルス応答

4.1.1 Swept-Sine 波の特性

インパルス応答を計測することで周波数特性を計測する．インパルス応答はインパルス波を計測する手法や，ホワイトノイズによって計測する手法など計測手法がいくつか存在するが，低い周波数から高い周波数まで，連続的に変化する正弦波である Swept-Sine を用いた手法が広く用いられている [45]．Swept-Sine 波による計測は一つの信号で幅広い周波数帯域の情報を取得でき，ノイズに強い上に計測が簡易であるためよく用いられている．また，Swept-Sine は信号が決定的であるためホワイトノイズなどによる計測手法と比べて雑音が少なく，繰り返して計測することによってランダムなノイズの影響を低減することができ，正確で再現性のある値を取得することができる．Swept-Sine によって得られた波形はシステムのインパルス応答と Swept-Sine の畳み込みになっているため，Swept-Sine を生成する際にあらかじめ逆畳み込みのフィルタを作ることで，得られた信号からマイクのインパルス応答を得ることができる．

4.1.2 計測に用いる Swept-Sine の設計

Swept-Sine はサンプルの長さによって周波数分解能が定まるため，長いほど高い周波数分解能となる．今回はサンプリング周波数が 44.1khz の際に周波数分解能が 1Hz 以下になるよう，65536 点のサンプルを用いた．Swept-Sine の周波数の範囲は人が聞こえる 22.1kHz までとした．また上記のとおり，Swept-Sine を用いてインパルス応答を計測する際には，繰り返すほど SNR が改善されるため，複数繰り返すことが重要である．そのため 65536 点からなるからなる Swept-Sine を作成し，これを 5 回繰り返したものを一つの wav ファイルとして作成した (図 2 左)．ただし，各 Swept-Sine の開始位置を特定するため，それぞれの Swept-Sine の間に Swept-Sine と同じ長さの空白を挿入した．このように空白を含めた 5 回の Swept-Sine の入力を 10 回ずつ繰り返し，計 50 回のインパルス応答の平均を取得した．最後に，得られたインパルス応答に対して 1/3 オクターブバンドフィルタによる平滑化 [46] を行った．

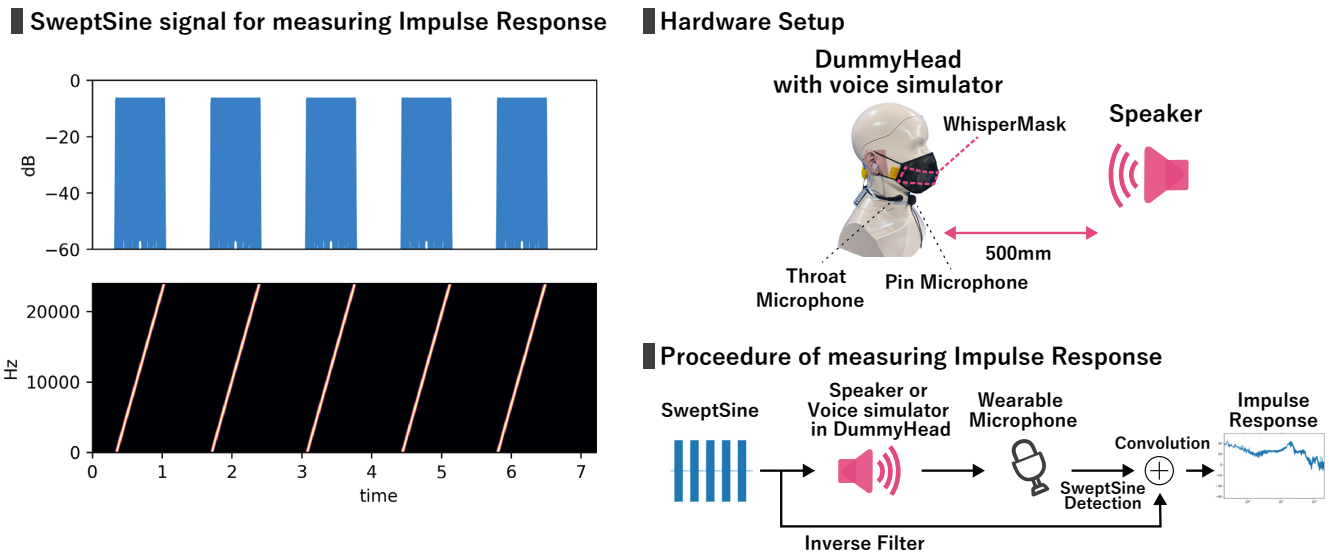

図 2 インパルス応答の計測に用いる Swept-Sine の波形．65536 点からなる Swept-Sine 波を 5 つ生成した (左). マウスシミュレーター付きダミーヘッドに身につける形でマイクを装着した．スピーカーは 500 mm 離れた位置に設置した (右上). インパルス応答は事前に インパルス応答は，前処理して得られた信号に Swept-Sine の逆フィルタを畳み込むことで計算される (右下)

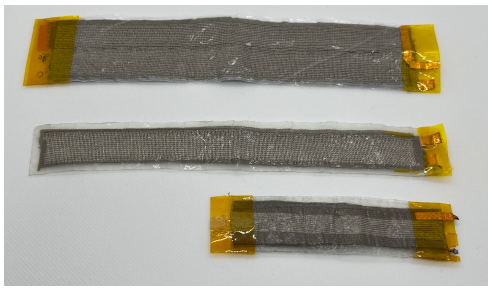

図 3 WhisperMask の振動板のパターン：
200 mm × 40 mm(上), 200 mm × 20 mm(中央),
and 100 mm × 20 mm(下)

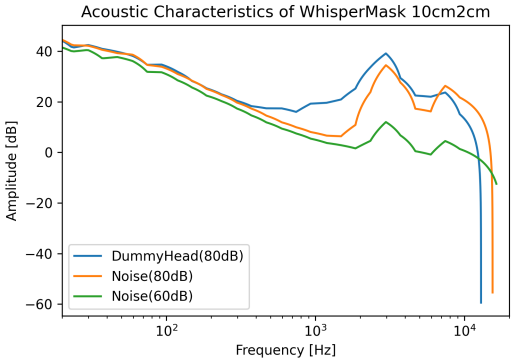

図 4 100 mm × 20 mm の振動板のインパルス応答

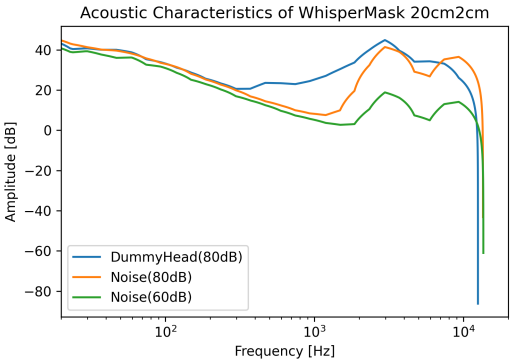

図 5 200 mm × 20 mm の振動板のインパルス応答

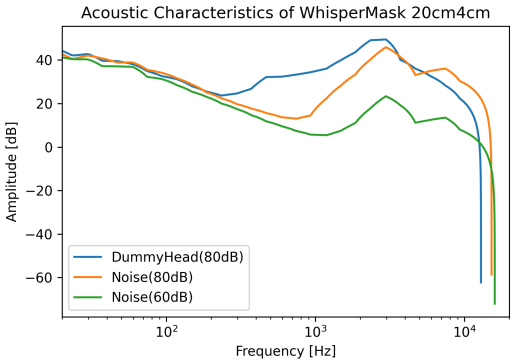

図 6 200 mm × 40 mm の振動板のインパルス応答

4.2 実験条件

4.2.1 インパルス応答計測のための実験環境

インパルス応答の計測にあたって，人の発話を模倣して出力するため，マウスシミュレーターを搭載したダミーヘッド SAMAR Type4700M (サザン音響) を利用した．SAMAR4700M は頭部形状とマウスシミュレーターが国際

規格 IEC 60318-7，ITU-T Rec.P51 にそれぞれ準拠しており，人の発話をシミュレートした計測が可能である．ダミーヘッドは床から高さ 40cm のところにあり，一脚の上にネジで固定した．また，外から発せられた音を模したスピーカーはダミーヘッドの正面前方 50cm のところに固定した．音声の計測は電波暗室で行ない，多孔質の素材で吸音されるため，電磁波や一部の音波の干渉がない．暗騒音レベルは 28.8 dB であった．

4.2.2 スピーカーから再生する信号

計測の際には，人の発話を模したスピーカー (以降，マウススピーカー) を持つダミーヘッドと，ノイズを模したスピーカー (以下，ノイズスピーカー) をダミーヘッドの前に設置した (図 2 右上)．ダミーヘッドからは，出力した Swept-Sine 波が装着するマイクに 80 dB の音圧レベルで届くように出力を設計し，その際に出力する音が歪まないようにした．一方で，外部のノイズスピーカーからは二つの音圧の Swept-Sine を出力し，マイクにそれぞれ 80 dB, 60 dB の音圧レベルで届くように出力を調節した．これらの音圧レベルは精密騒音計（RION NL-52）で計測され，誤差 0.1dB の範囲であった．

4.3 Swept-Sine の検出

出力する Swept-Sine は, 受け取るマイクの特性によって，周波数特性が異なる．すなわち，すべての帯域に渡って取得できるわけではなく, 低周波領域など部分的にしか表示されない場合がある．しかし，一方で正確なインパルス応答を計測するには，Swept-Sine の信号が始まったタイミングを正確に測定する必要がある．そこで，以下のようにして一部の周波数帯域の情報から Swept-Sine の開始と終了の時間を推定する．まず，Swept-Sine を 1000Hz の幅で周波数ごとに分解し，envelope を取得する．Swept-Sine はパルス 1 回分の間隔を開けて 5 回出力しているので，Swept-Sine がよく現れている周波数領域では 5 回の立ち上がりが現れる．これを取得することで 5 つの点が記録される．周波数を 1000Hz ごとに刻むことで，最大 20000/1000=20 個の点が記録されるので，これらの計測点から線形補完し，周波数 0Hz の点が開始時間，周波数 20kHz の点が終了時間となる．

4.4 インパルス応答の測定結果

インパルス応答の結果は図 4 から図 6 のようになった．青いラインがダミーヘッドのマウススピーカーから流した，人の発話を想定した時のインパルス応答であり，音圧レベルが 80 dB の Swept-Sine を流している．オレンジと緑はどちらもノイズを想定しており，ノイズスピーカーからそれぞれ 80 dB, 60 dB で出力している．

3 つのパターンのそれぞれにおいて，特に 200Hz 以上 5kHz 以下の周波数帯域で，ダミーヘッドのマウススピーカーからの出力がノイズスピーカーからのノイズよりも 10 dB 程度向上しており，同じ音圧の同じ波形をマイクの内側 (マウススピーカー) と外側 (ノイズスピーカー) で入力した際に内側の音をより入れやすくする性質があり，すなわちノイズを低減することが明らかになった．

また，この特性はマイクの大きさを変えても同様の傾向が見られた．

5. マイクのノイズ低減効果の評価

5.1 SNR の計測

提案するマイクは騒音環境において話者の声を拾うことができる．これを示すために，SNR を計測する．ここでの SNR はマウススピーカーからの音声なしにマイクで計測した際の純粋なノイズのみの値 (N) と，マウススピーカーから入力信号を発生させた時の出力 (S) をそれぞれ記録し，それぞれのマイクの出力に対して RMS（Root Mean Square）値を計算する．そして $20\log_{10}(S_{\mathrm{RMS}}/N_{\mathrm{RMS}})$ を計算することで行う [47]．SNR での評価はノイズ環境でのマイクの性能の一つとして用いられており，アレイマイクなどでも利用されている [48]．

5.2 SNR の測定環境の設定

SNR の計測にあたって，人の発話はインパルス応答と同様にマウススピーカーを用いた．計測に用いる信号も先ほどと同様に 20 Hz–20 kHz の SweptSine を利用した．マウススピーカーから出力する信号の音圧は人の発話に近い 60 dB を用いて行った．また，環境ノイズは，ノイズスピーカーの出力の特性から 90 dB を上限とし，30 dB から 90 dB まで 10 dB 刻みでノイズスピーカーから出力した．

計測するデバイスは, WhisperMask(200 mm × 20 mm), ピンマイク, 咽頭マイクの 3 つで行った．イヤホン型のマイクである Apple AirPods Pro (第 2 世代) は人の声に最適化されており，Swept-Sine のようなノイズのような波形を入力することができなかった．それぞれのマイクは，ダミーヘッドに理想的な位置で取り付けた．

5.3 騒音環境下における SNR の結果

30–90 dB のノイズ環境下における SNR を図 7 に示す．ピンマイクと咽頭マイクは外部のノイズが増加するにつれて SNR が減少したが，WhisperMask は外部のノイズが 30 dB(SNR:22.2) から 60 dB(SNR:21.3) までほとんど変化しなかった．SNR は信号のパワーとノイズのパワーの比率で計算されており，$\mathrm{SNR} = 20\log_{10}(S_{\mathrm{RMS}}/N_{\mathrm{RMS}})$ である．図 7 においてマウススピーカーからの信号は音圧レベルが 60 dB の Swept-Sine で一定なので，S_{RMS} は定数となる．図 7 の 30 dB から 60 dB にかけて．WhisperMask の計算された SNR がほとんど一定なので，WhisperMask へ入力される周囲のノイズである N_{RMS} は必然的にこの区間にお

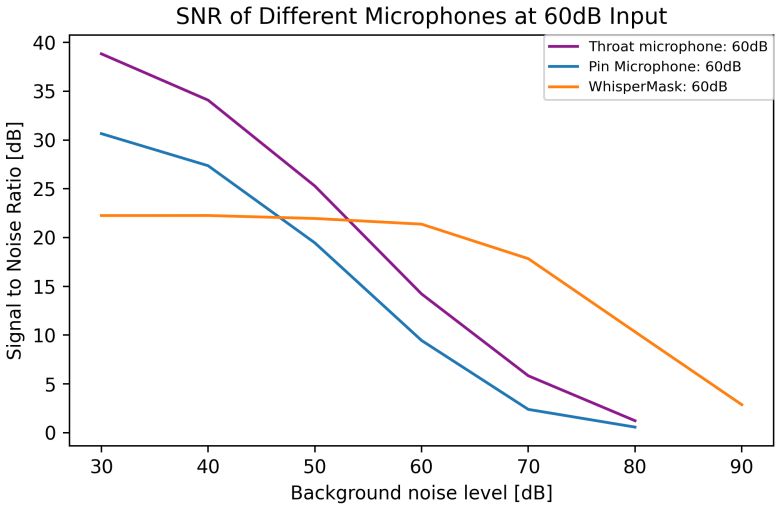

図 7 60 dB の音圧の音をマウススピーカーから出力し，騒音の音圧レベルを変化させた際の各マイクの SNR の変化

いてほとんど一定であり，周囲のノイズをほとんど拾っていないことがわかる

また，外部ノイズが 70 dB 以降は WhisperMask はノイズの影響が増えてくるが，ノイズが 70 dB において SNR が 17.83 でこの時ピンマイクは 2.3, 咽頭マイクは 5.83 と 10 dB 以上の差があることがわかった．

6. 騒音環境における音声認識の評価

音声入力は通話だけでなく，スマートアシスタントのコマンド操作や音声認識による対話的な検索など，機械との対話にも広く利用されている．今回はノイズ耐性のある音声認識手法を用いて音声認識を行う．Whisper[49] は encoder-decoder 型の Transformer モデルを用いて事前学習されている．

6.1 データ収集

データの収集は，9 名の参加者 (平均年齢 26.2, 男性 4 人，女性 5 人) を対象に実におこなった．参加者にはあらかじめ 5 段階で英語の能力を聞いており，平均は 3 となった．データの収集の際には，テキスト入力のコーパス [50] からあらかじめ用意した 20 のフレーズを各々が読み上げた．また，背景のノイズは，何もない場合に加えて，ホワイトノイズを 40 dB, 60 dB, 80 dB の 3 種類をあわせた 4 つのノイズでの測定を行なった (図 8 ではそれぞれ a30dB, w40dB, w60dB, w80dB と表記した)．評価するマイクは，提案する WhisperMask, ピンマイク, イヤホン型マイク (Apple AirPods Pro (第 2 世代)) の 3 つを用いた．データの収集の際には，テキスト入力のコーパス [50] からあらかじめ用意した 20 のフレーズを各々が読み上げた．実験中の発話には，自然な有声での発声と囁き声の 2 種類を用いた．ノイズはラップトップ PC からステレオで出力し，ユーザーの口元での音圧が求める音圧になるように騒音計で計測することで，0.5 dB 以下の差になるよう調節した．測定は防音室で行い，通常時のノイズは 30.0 dB であった．

6.2 認識結果

音声の領域では背景のノイズを低減するための音声の強調手法が提案されている [13], [14]．特に近年ではリアルタイムでのノイズ除去手法 [51] の精度が高く，ピンマイクや AirPods のようなマイクにノイズ除去を適用することで目的の音声を取得することが可能である．本研究では，ピンマイクや AirPods で得られた音声にこのリアルタイムノイズ除去を適用した場合に比べて，音声認識の精度がどれほど改善されるかを検討する．音声認識には，最新の音声認識器である Whisper[49] を用いて行った．音声認識は単語ごとの正解率と文字ごとの正解率の評価を行った．

ノイズ環境下におけるマイク種類ごとの正解率を図 8 に示す．図 8 の左は通常の発声によるもの，右は囁き声によるものである．

結果としては，提案する WhisperMask, ピンマイク, イヤホン型マイクの順番で認識精度が良くなった．特に，80 dB の騒音環境下での囁き声では 30%ポイントの差が生じた．Denoiser によるノイズ除去は，通常の発話においてはあまり変化していないが，囁き声では大幅に精度が劣化した．騒音環境での囁き声入力において，WhisperMask の精度はノイズが 60 dB と 80 dB の時に他のマイクよりも優れており，提案手法は騒音環境下でより精度の高い入力が可能なデバイスであることが言える．

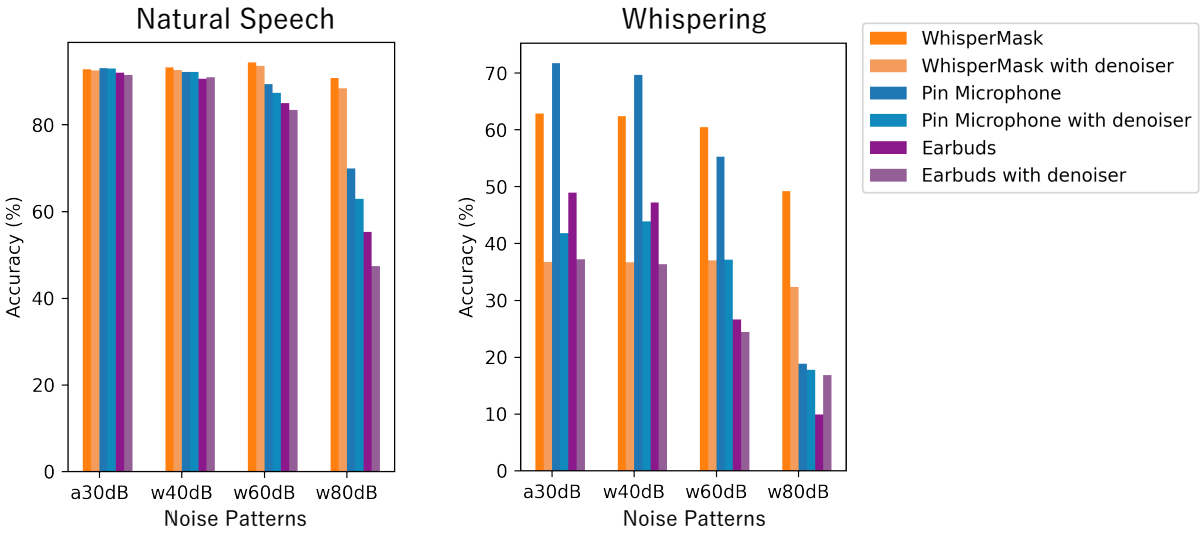

図 8 複数のノイズ環境での音声認識の結果．a30 dB は，ノイズのない環境での録音であり，w40 dB，w60 dB，w80 dB はそれぞれ，ホワイトノイズが 40 dB，60 dB，80 dB で出力された騒音環境での録音を指す．自然な発声（左）では，提案する WhisperMask，ピンマイク，Earbuds の順に認識精度が向上した．また囁き声入力では 80 dB の騒音環境において約 30 %の差がみられた（右）

7. 議論と今後の課題

7.1 騒音が減衰する機序

提案手法が騒音環境でも音声をよりよく入力できるのは，マイクの感度が低いことが一因として考えられる．図 4 から図 6 にあるように，80 dB の音をダミーヘッドのマウススピーカーやノイズスピーカーから流しているにもかかわらず，マイク側は最大 40 dB 程度の入力になっており，大きな音の入力が要求される．人の通常の有声での発話は 60 dB–80 dB であることが知られている [52] が，その計測位置は 30cm 先であるため [53]，マイクに接近した場合はより大きい音声が入力される [54]．実際に 3cm 程度の距離で計測すると 80 dB から 90 dB 程度の大きさであり，接近した位置で声を大きく拾うことが一つの要因として挙げられる．

7.2 再利用性

マイクは日常的に用いるデバイスであり，繰り返し使える必要がある．そのため，つけ外しによる影響や，繰り返し使う時の汚れに対しての有効性を検証する必要がある．今回は 9 名に利用してもらい，入力するセッションごとにマスクのつけ外しを行ったが，それによる性能上の問題は見られなかった．また実験の際には，衛生上の観点から不織布マスクの上から提案するデバイスを重ね着して計測しており，口に直接接触することなくデバイスを利用できる．

7.3 風による影響

空気を媒介して声はマイクに入力される．そのため強風の環境など口とマイクの間の空気が乱れている場合には入りにくい．これらは通常のピンマイクなどにおいても生じる現象であり，風防が必要となると考えられる．本研究では強風が当たった場合の影響を評価できていないが，不織布のマスクの中にデバイスを入れた場合でも集音できることが確認できている．

8. 結論

本研究では，騒音環境下でもユーザー自身の声を明瞭に捉えることができるマスク型マイクである，WhisperMask を提案した．具体的な音響的な特性としては，60 dB や 80 dB の発声を行う際に，10kHz 以下の周波数帯域において，WhisperMask は外からの音への感度が低下することを示した．SNR の点では 60 dB 以下においてはノイズが入力されず，70 dB においては既存のマイクに比べて 10 dB 以上良い性能を得た．

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, JST CREST グラント番号 JPMJCR17A3, 国立研究開発法人情報通信研究機構の委託研究 02901 の支援を受けたものです．

参考文献

[1] https://appleinsider.com/articles/21/03/30/apple-airpods-beats-dominated-audio-wearable-market-in-2020.
[2] Daniyal Liaqat, Salaar Liaqat, Jun Lin Chen, Tina

Sedaghat, Moshe Gabel, Frank Rudzicz, and Eyal de Lara. Coughwatch: Real-world cough detection using smartwatches. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8333–8337, 2021.

[3] Daniyal Liaqat, Robert Wu, Andrea Gershon, Hisham Alshaer, Frank Rudzicz, and Eyal de Lara. Challenges with real-world smartwatch based audio monitoring. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, page 54–59, New York, NY, USA, 2018. Association for Computing Machinery.

[4] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E. Starner, Omer T. Inan, and Gregory D. Abowd. Fingersound: Recognizing unistroke thumb gestures using a ring. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), sep 2017.

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3), sep 2018.

[6] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291, 1995.

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

[8] Robert Ingalls. Throat microphone. *The Journal of the Acoustical Society of America*, 81(3):809–809, 03 1987.

[9] 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.

[10] Tobias Röddiger, Tobias King, Dylan Ray Roodt, Christopher Clarke, and Michael Beigl. Openearable: Open hardware earable sensing platform. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 246–251, 2022.

[11] 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.

[12] 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.

[13] Asri Rizki Yuliani, M Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. Speech enhancement using deep learning methods: A review. *Jurnal Elektronika dan Telekomunikasi*, 21(1):19–26, 2021.

[14] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

[15] Ryan M Corey and Andrew C Singer. Speech separation using partially asynchronous microphone arrays without resampling. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–9. IEEE, 2018.

[16] 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.

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

[18] 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.

[19] NAKAJIMA Yoshitaka, KASHIOKA Hideki, CAMPBELL Nick, and SHIKANO Kiyohiro. Non-audible murmur (nam) recognition. *IEICE TRANSACTIONS on Information and Systems*, E89-D(1), 2005.

[20] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication*, 52(4):301–313, 2010. Silent Speech Interfaces.

[21] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. Sensing with earables: A systematic literature review and taxonomy of phenomena. 6(3), sep 2022.

[22] Hiroshi Sawada, Nobutaka Ono, Hirokazu Kameoka, Daichi Kitamura, and Hiroshi Saruwatari. A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf. *APSIPA Transactions on Signal and Information Processing*, 8:e12, 2019.

[23] 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.

[24] 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.

[25] 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.

[26] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 276–280. IEEE, 2015.

[27] Naoki Makishima, Shinichi Mogami, Norihiro Takamune, Daichi Kitamura, Hayato Sumino, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono. Independent deeply learned matrix analysis for determined audio source separation. *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing, 27(10):1601–1615, 2019.

[28] 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.

[29] 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.

[30] 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.

[31] Hyein Lee, Yoonji Kim, and Andrea Bianchi. Mascreen: Augmenting speech with visual cues of lip motions, facial expressions, and text using a wearable display. In *SIGGRAPH Asia 2020 Emerging Technologies*, SA '20, New York, NY, USA, 2020. Association for Computing Machinery.

[32] Ryoga Kumazaki and Akifumi Inoue. Development and evaluation of a mask-type display transforming the wearer's impression. In *Proceedings of 31st Australian Conference on Human-Computer-Interaction*, OzCHI '19, page 568–571, New York, NY, USA, 2020. Association for Computing Machinery.

[33] Zengrong Guo and Rong-Hao Liang. Texonmask: Facial expression recognition using textile electrodes on commodity facemasks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[34] Takahiro Kusabuka and Takuya Indo. Ibuki: Gesture input method based on breathing. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, page 102–104, New York, NY, USA, 2020. Association for Computing Machinery.

[35] Christopher Beach, Nazmul Karim, and Alexander J. Casson. A graphene-based sleep mask for comfortable wearable eye tracking. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6693–6696, 2019.

[36] Mose Sakashita, Keisuke Kawahara, Amy Koike, Kenta Suzuki, Ippei Suzuki, and Yoichi Ochiai. Yadori: Mask-type user interface for manipulation of puppets. In *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, New York, NY, USA, 2016. Association for Computing Machinery.

[37] Yutaro Suzuki, Kodai Sekimori, Yuki Yamato, Yusuke Yamasaki, Buntarou Shizuki, and Shin Takahashi. A mouth gesture interface featuring a mutual-capacitance sensor embedded in a surgical mask. In Masaaki Kurosu, editor, *Human-Computer Interaction. Multimodal and Natural Interaction*, pages 154–165, Cham, 2020. Springer International Publishing.

[38] Takumi Yamamoto, Katsutoshi Masai, Anusha Withana, and Yuta Sugiura. Masktrap: Designing and identifying gestures to transform mask strap into an input interface. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 762–775, New York, NY, USA, 2023. Association for Computing Machinery.

[39] Linda M Thibodeau, Rachel B Thibodeau-Nielsen, Chi Mai Quynh Tran, and Regina Tangerino de Souza Jacob. Communicating during covid-19: The effect of transparent masks for speech recognition in noise. *Ear and Hearing*, 42(4):772–781, 2021.

[40] Joseph C Toscano and Cheyenne M Toscano. Effects of face masks on speech recognition in multi-talker babble noise. *PloS one*, 16(2):e0246842, 2021.

[41] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. Silent speech interfaces. *Speech Communication*, 52(4):270 – 287, 2010. Silent Speech Interfaces.

[42] João Freitas, António Teixeira, Miguel Dias, and Samuel Silva. *An Introduction to Silent Speech Interfaces*. 07 2016.

[43] Hirotaka Hiraki and Jun Rekimoto. Silentmask: Mask-type silent speech interface with measurement of mouth movement. In *Augmented Humans Conference 2021*, AHs'21, page 86–90, New York, NY, USA, 2021. Association for Computing Machinery.

[44] Yusuke Kunimi, Masa Ogata, Hirotaka Hiraki, Motoshi Itagaki, Shusuke Kanazawa, and Masaaki Mochimaru. E-mask: A mask-shaped interface for silent speech interaction with flexible strain sensors. In *Augmented Humans 2022*, AHs 2022, page 26–34, New York, NY, USA, 2022. Association for Computing Machinery.

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

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

[47] Colin Breithaupt, Timo Gerkmann, and Rainer Martin. A novel a priori snr estimation approach based on selective cepstro-temporal smoothing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4897–4900, 2008.

[48] Reuven Berkun and Israel Cohen. Microphone array power ratio for quality assessment of reverberated speech. *EURASIP Journal on Advances in Signal Processing*, 2015(1):49, 2015.

[49] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[50] I. Scott MacKenzie and R. William Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, page 754–755, New York, NY, USA, 2003. Association for Computing Machinery.

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

[52] Hugo Fastl and Eberhard Zwicker. *Hearing Area*, pages 17–22. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[53] Ingo R. Titze. Acoustic interpretation of the voice range profile (phonetogram). *Journal of Speech, Language, and Hearing Research*, 35(1):21–34, 1992.

[54] Hana Šrámková, Svante Granqvist, Christian T. Herbst, and Jan G. Švec. The softest sound levels of the human voice in normal subjects. *The Journal of the Acoustical Society of America*, 137(1):407–418, 01 2015.