WESPER: 話者・言語非依存の実時間ささやき声通常音声変換によるスピーチインタラクション

暦本純一[1,a)]

概要：ささやき声の認識と通常音声への変換には音声インタラクションの多くの可能性がある．ささやき声の音圧は通常の音声よりもはるかに低いため，公共の場において他人に聞かれることなくサイレントスピーチに準ずる音声入力として利用でき，公共環境での遠隔会議も可能である．また，ささやき声やかすれ声を通常の発声に変換できれば，発声障害者や聴覚障害者の発声品質を向上させることができる．しかし，従来の音声変換技術では，ささやき声から通常声への変換には十分な変換品質が得られないか，ささやき声と通常音声のペアからなるデータセットが必要だった．本研究では，自己教師型学習に基づく実時間ささやき声音声変換機構，WESPER を提案する．WESPER は，ささやき声と通常音声の差分を吸収した潜在音声単位を生成する音声単位変換器 (Speech to Unit encoder, STU) と，音声単位から目的の音声を復元する単位音声変換器（Unit to Speech decoder, UTS）から構成される．テキストの付随しない，ペアでないささやき声と通常音声のデータのみから事前学習可能で，発話者・言語に依存しない変換を実現する．UTS は，ラベルのない対象話者の音声データのみから，対象話者の音声を復元するように学習可能である．本手法を実験参加者 50 名により評価し，ささやき声から変換された音声の品質が向上し，韻律の自然さも保持されることを確認した．また，提案手法が言語障害者や聴覚障害者の発声再構成にも有効であることも評価実験により確認した．

1. はじめに

音声対話インタフェースは広く利用されているが，他人がいるところでの使用は容易ではない．音声認識を公共の場で使用することは他者への迷惑になり，機密情報が漏洩する可能性もある．また，遠隔会議中に声を出すことも，周囲の人に不快感を与え，会話の秘匿性を阻害する可能性がある．

これらの問題を解決するために，様々な無音音声入力技術（サイレントスピーチインタフェース，以下 SSI）が研究されている [2], [22], [23], [38], [39], [40]．しかし，従来の SSI では特殊なセンサー構成により発声時の口腔情報を得るものが多く，認識のための学習用データセットを，それぞれのセンサー構成ごと，利用者ごとに採取しなければならず，利用するための準備負荷が大きい．認識精度も高くはなく，あらかじめ定義されたコマンドを認識するレベルにとどまっている．語彙や言語の制限なく，無声発話を通常の発声に変換することも，まだ実現されていない．

SSI の適用分野として，発声障害者や聴覚障害者の発話の認識や再構築も重要であるが，上記の制約により，既存の SSI の手法では実用的な音声復元は達成されていない．

そこで，本研究では，「ささやき音声」に注目する．ささやき声は音圧が低いため秘匿性が高く，SSI に準じた静寂な音声入力手段である．ささやき声は通常のマイクで採取可能で，特別なセンサー構成を必要としない．また，発声障害者で，声帯が損傷されている場合でも，ささやき声やかすれ声で話すことができるため，発話の認識や再構築の可能性がある．

本研究では，自己教師型学習により，ささやき声から通常音声への音声変換を，話者・言語非依存，実時間で行う機構，WESPER を提案する．ささやき声と通常音声の音声データのみで学習し，音声データに付随するテキストラベルや，ささやき声と通常発声の対応したパラレルデータを必要としない．

WESPER は，通常音声とささやき声音声とで事前学習し，その差分を吸収した潜在表現である音声単位を生成する音声単位変換器 (Speech-to-Unit encoder , STU) と，音声単位から目標音声波形を再構成する単位音声変換器 (Unit-to-Speech decoder, UTS) から成る (図 1)．(対になっていない) ささやき声と通常音声による自己教師型の事前学習により，STU は通常音声とささやき声との差分を吸収した共通の音声単位を出力するようになる．

[1] 東京大学, ソニーコンピュータサイエンス研究所
[a)] rekimoto@acm.org

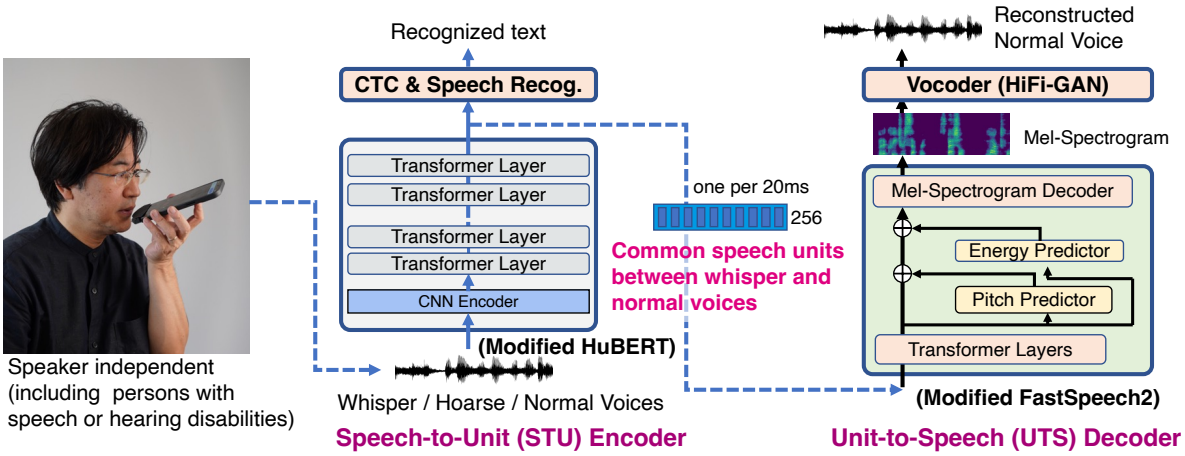

図 1 WESPER は，実時間，話者・言語非依存のささやき声→通常音声の変換機構である．自己教師型事前訓練によりささやき声と通常音声の差分を吸収した共通音声単位を生成する音声単位変換器 (Speeh-to-Unit,STU) と，音声単位から音声を復元する単位音声変換器（Unit-to-Speech,UTS）から構成される．

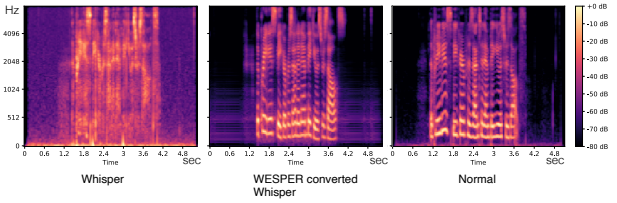

図 2 WESPER 変換結果 (左：ささやき声，中：WESPER で変換されたささやき声，右：同一発話の通常音声)

UTS は，特定の話者の音声データのみから，付随するテキストラベルを要さずに学習することができる．たとえば，声帯摘出者であれば，摘出前の本人の音声がデータとして残っていれば，摘出後のささやき声から人の摘出前の音声を再構築することができる．また，任意の他者の音声に変換することも可能である．

STU と UTS は非自己回帰的 (non-autoregressive) に動作するため，システム全体は実時間で動作する．そのため，例えば遠隔会議に応用した場合，会議参加者がささやき声で話し，それをリアルタイムで変換し，他の参加者は通常の声で再生して聞くことができる．

図 2 に，提案手法によるささやき声から通常音声への変換の例を示す．添付のビデオで，より多くの音声変換例を示している．

本研究の貢献は以下のようにまとめられる:

1. 未対応のささやき声・通常音声データのみで学習可能な，話者・語彙・言語非依存な実時間ささやき声→通常音声変換機構を提案した．
2. 任意話者のテキスト未対応の音声サンプルデータから，変換対象となるターゲット音声を学習・再構築できる．
3. 通常のささやき声，および発声障害者や聴覚障害者の発声の品質が向上することを利用者評価実験で確認した．

2. 関連研究

2.1 SSI・ささやき声インタラクションの研究

サイレントスピーチ (SSI) 研究として，読唇，筋電図，超音波などの様々なセンシング技法を用いて，利用者の無音発話や無音コマンドを認識する方式が提案されている [2], [22], [23], [38], [39], [40]．しかし，これらの手法は，特別なセンサーが必要なことに加えて，個々のセンサー構成に応じた特別なデータセットを利用者ごとに構築する必要があり，普及の大きな障壁となっている．また，有声発声と無声発声で口腔運動が異なる場合，有声発声時のセンサーデータを無声認識に用いると認識率が低減してしまう．その結果，これらのシステムは非常に限られた数十程度の語彙（コマンド）認識にとどまっている．この制限のため，発声障害者の支援に供することも実現されていない．

SilentVoice は，息を吸いながら話す “ingressive speech” の利用を提案している [12]．Ingressive speech は音圧が小さく，SSI の変形とみなすことできるが，専用のセンサーを口に近づける必要があり，ユーザが ingressive モードで正しく発話できるようになるには訓練が必要である．また，他の SSI 研究と同様，認識のために Ingressive Speech 専用かつ話者依存のデータセットを作成する必要がある．

ささやき声の音声認識に関する研究 [5], [7], [11], [14] では，ささやき声とテキストとの対応がなされたデータセットを必要とした．

ささやき声の認識を wav2vec2.0 [1] と HuBERT [16] による自己教師付き音声認識システムに基づいて end-to-end で行う方式として DualVoice [35], [36] が提案されている．DualVoice ではささやき声と通常発声を弁別するインタラクション技法も提案している．WESPER は DualVoice と組み合わせることで，ユーザのささやき声を選択的に変換

手法	学習データ	話者毎の学習	通常音声への変換	ささやき声音声認識
サイレントスピーチ [2], [22], [23], [38], [39], [40]	提案センサー毎に必要 パラレル	必要	YES/NO	YES
Parotoron [3]	W-N パラレル	必要	YES	YES(音声変換経由)
DualVoice [35], [36]	テキスト付き ノンパラレル W,N	必要	NO	YES
SilentVoice [12]	テキスト付き ingressive 音声	必要	NO	YES
CycleGAN-VC [21]	ノンパラレル，テキスト不要 W,N	不要	YES	NO
MSpeC-Net [28]	W-N パラレル	必要	YES	NO
AGAN-W2SC [13]	W-N パラレル	必要	YES	NO
WESPER (ours)	ノンパラレル，テキスト不要 W,N	不要	YES	YES (音声変換経由)

表 1 サイレントおよびささやき音声インタラクションの比較 (W:ささやき声，N:通常声)

することができる．たとえば，アバター A の声を通常音声で，アバター B の声をささやき声で発声することで，複数のアバターを演じ分けるようなシステムが構成できる．

2.2 音声変換技術

音声を別の音声に変換する通常音声変換技術が開発されているが，ささやき声では F0(基本周波数) 情報が欠損しているため、通常音声への変換に適用すると，変換品質に不満が残る．音声認識を用いてささやき声をテキスト化し，音声合成により通常音声を生成することも考えられるが，テキストが付随したささやき声専用のコーパスが必要となり，発話する言語に依存してしまう．また，認識されたテキスト表現では，原音声の非言語情報が保持されないため，変換結果が機械的になってしまうという課題がある．

Phonetic PosteriorGrams (PPGs) [41] は，自動音声認識機構 (以下 ASR) から得られる中間表現で，話者とは無関係に発話内容の調音を表現するため，多対一話者の音声変換にも利用されている． ただし，ささやき声に対する PPG の効用については，これまで研究されていない．本研究の手法は中間表現として音声単位を用いるが，ASR や PPG のようなテキストベースのコーパスを必要とせず，発話する言語にも依存しない．

Parrotron [3] は，構音障害者の発話を改善するために設計された音声変換システムである．テキスト音声合成システム Tacotron [43] に準拠したエンコーダ・デコーダモデルに基づいているが，Tacotron と異なり，入力・出力データがともにメルスペクトルグラムを利用している．ただし，音声とテキストのペアを必要とするため，利用者ごとに学習用のコーパスを構築する必要がある．

近年, 深層学習を用いたささやき声から通常音声への変換技術が研究されている [31]. これらの技術を比較するために, 変換品質だけでなく, 学習に要求されるデータセットの特性も考慮することが重要である．AGAN-W2SC (Attention-Guided Generative Adversarial Network for Whisper to Normal Speech Conversion) は，GAN を用いたささやき声から通常音声への変換である [13]. MSpeC-Net [28] は，オートエンコーダーに基づく音声変換である．これらの手法は，ささやき声と通常音声の対となったデータセットを必要とする．

これに対し，WESPER では，通常音声とささやき音声の対になっていないデータサンプルのみから学習可能である．発話者ごとのデータセットや，音声に付随するテキストの書き起こしも不要であるため，データセットの準備が容易であり，話者・対象言語に依存しないのが特徴である．関連研究との比較を表 1 に示す．既存の変換技術との変換音声例を `https://wesperproj.github.io/` に掲載する．

2.3 音声の自己教師付き表現学習

ラベルのない音声データによって事前学習 (pre-training) として自己教師付き表現学習を行い，ラベルのある音声データで微調整 (finetuning) を行う学習方式が注目されている [33], [44]．これらのシステムは主に音声認識を目的としているが，話者認識，言語認識，感情認識などにも応用されている．特に，HuBERT (hidden-unit BERT) [16] の学習方法は，自然言語処理の BERT (Bidirectional Encoder Representations from Transformers) [8] で用いられるマスク型言語モデルを適用している．すなわち，入力の一部をマスクし，残りの入力からそれに対応する表現特徴を推定することで学習する．この事前学習により，モデルは入力データの音響特性や音声の特徴を学習することができる．自己教師あり ASR は， [1], [44] で報告されているように，ラベルなし音声による事前学習後に，短時間のラベル付き音声データセットで微調整するだけで高い音声認識精度を達成することができる．したがって，この方式は，ささやき声のテキストつきコーパスが限られている場合に，ささやき声の認識に適していると期待される．

WESPER では，ささやき声と通常発声から生成される潜在ベクトルの差を小さくする目的で事前学習を用いている．ささやき声でも通常発声でも出力される音声変換結果は同等になる．

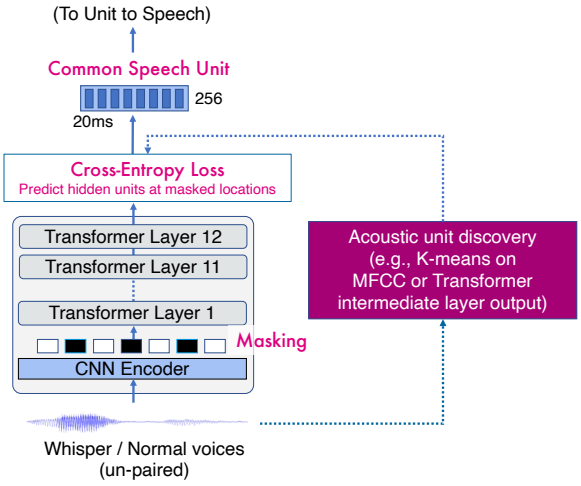

図 3 STU の事前トレーニングの概要．事前学習には，ペアリングされていないささやき声や通常の音声を使用する．マスクされた入力から当該部を推定することで学習する．変換層の後にある投影層で 256 次元のベクトル（20 ms に 1 個）を生成し，これを共通音声単位とする．

2.4 textless NLP

近年，テキストを用いない音声処理・音声変換の手法が注目されている．これらのシステムでは，自己教師付き学習によりテキストを伴わない音声データから潜在的な表現を得ることができる．Textless-NLP [25], [26] や AudioLM [4] は，音声処理システムにおいてテキストや音素記号を用いず，自己教師付き学習によって構築された離散単位を用いる．Soft Discrete Unit もテキスト表現を用いない自然言語処理へのアプローチの 1 つである [42].

本研究の手法も，自己教師付き学習から得られる潜在的な音声表現として非離散ベクトルを用い，テキストや音素記号を明示的に用いない．本研究では，自己教師付き学習により，ささやき声と通常の音声の差分を吸収した音声単位で表現できることを示した点が特徴である．また，本手法は，後段の音声生成システムとの連携が可能なように設計されている．

3. WESPER 音声変換モデル

WESPER は，STU（Speech-to-Unit）エンコーダと UTS（Unit-to-Speech）デコーダから構成される．STU はささやき声と通常音声の差分を吸収した共通音声単位 (common speech unit) を生成する．共通音声単位は 256 次元のベクトルで，20 ms ごとに生成される．UTS は音声単位をメルスペクトログラムに変換し，ボコーダでターゲット音声として再構成する．STU は，ペアになっていない，ささやき声と通常音声のデータからの事前学習のみからこの共通性を達成していることが特徴である．

3.1 STU (Speech-to-Unit) エンコーダー

STU（Speech to Unit）エンコーダは，音声波形を入力とし，音声単位を出力するエンコーダである．HuBERT (Hidden Unit BERT) [16] の，ラベルのない大量の音声を BERT [9] 的に事前学習し，部分的にマスクされた音声特徴入力から関連部分を復元する学習により，音声モデルを獲得する音声用自己教師付きニューラルネットワークをベースとしている．

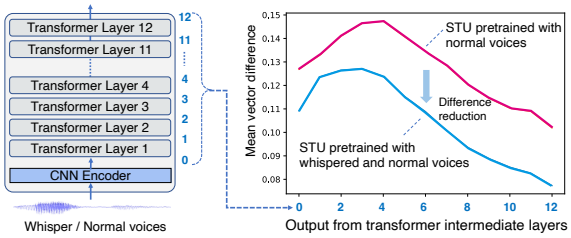

図 4 ささやき声と通常音声の違いによる比較．通常の音声とささやき声変換を施した音声で比較: 通常音声とささやき音声で事前学習した STU は，通常音声のみで事前学習した STU に比べ，差分が減少する．また，変換層が深くなるにつれても，両者の差は減少する．

本研究の目的は，ささやき声と通常音声の差分を吸収して，できるだけ同等の音声単位を生成することである．そのため，ささやき声と通常音声を混合させて事前学習する．学習には，(1) Librispeech 960h dataset から複数話者による通常の英語音声，(2) Librispeech 音声データを LPC ベースの音声変換ツールで機械的にささやき声に変換した [45] 音声，(3) 音声長が 58 時間の通常音声とささやき声の wTIMIT 音声データセット [27]（英文，複数発話者），を使用した．

図 3 に STU の事前学習方式を示す．ペアリングされていないささやき声や通常音声を入力とし，トランスフォーマー層は，マスクされた入力から当該部の離散単位を推定することにより学習する．HuBERT と同様に，まず第一段階の事前学習では，入力音声データを k-means クラスタリングした離散単位を推定するように学習する．第二段階では，トランスフォーマー中間層の出力を k-means クラスタリングして離散的な単位としたものを推定する．実験では，離散ユニット数を 100 に固定した．変換器中間層以降のプロジェクション層で，256 次元のベクトル列（20 ms ごと 1 個）を生成し，これを共通音声単位とする．

STU では，CNN 特徴抽出器の先に 12 個のトランスフォーマー層が配置されている（図 1）．事前学習後，各層の出力を比較すると，(1) ささやき声と通常音声の音声単位値の差は，層の深さが増すほど小さくなること，(2) ささやき声と通常音声の両方で事前学習を行った場合は，通常音声のみで事前学習した場合と比較して差が小さくなること，を確認した（図 4）．

このように，ささやき声と普通の声では，メルスペクトログラムなど，波形に近い音声特徴では差分があるもの，STU を解することでその差分が吸収されることがわかる．

音声認識で，複数話者の発話を事前学習することで，話者間の差分を吸収した表現を獲得できるように，WESPER でも，ささやき声と通常発話で，言語的に立場が近い発声を同一の音声単位で表現するように事前学習されていると考えることができる．

本機構はさらに，wTIMIT や Librispeech のデータセットを用いて，CTC 層 ([35]) を追加し，ささやき声と普通の声を認識する ASR（自動音声認識）としても使用できるように STU を微調整 (finetuning) することも可能でる．

3.2 UTS(Unit-to-Speech) デコーダ

UTS デコーダは，STU から生成された音声単位を入力とし，目的となる話者の（通常の）音声を再構成する．非自己回帰型テキスト音声合成システム FastSpeech2 [37] に基づいている．オリジナルの FastSpeech2 には，テキストを受け取ってベクトル列に変換するための埋め込み層 (embedding layer) を学習させる必要があるが，UTS は STU からの音声単位を直接入力とするため，この部分は不要となる．また，オリジナルの FastSpeech2 には，各音素の継続時間を推定する継続時間推定部 (duration-prediction) の学習と，内部ベクトルの数を推定継続時間に合わせて伸縮させる長さ調整部 (length regurator) が含まれている．STU の生成する音声単位の時間長は常に 20 ms としているため，これらの部分も削除できる．また，オリジナルの FastSpeech2 の学習では，各テキストに対応する継続時間を真値として与える必要があり，Montreal Forced Aligner [29] などの外部ツールによってからこの継続時間を与える必要がある．この条件のため，FastSpeech2 の学習は言語に依存していた．一方，UTS では継続時間の推定が不要であるため，WESPER は特定の言語や外部ツールに依存しない．実際，WESPER の学習は英語の音声のみで行ったが，日本語のささやき声変換にも適用可能であった．

UTS の出力はオリジナルの FastSpeech2 と同様, メルスペクトログラムである．ニューラルボコーダ（HiFi-GAN [24] 等）を介して，実際の音声波形に変換される．

通常の TTS システムでは，学習を行うために，ターゲット音声とそれに対応するテキストラベルが必要である．これに対し，提案方式では，テキストを伴わない目標音声のみで学習可能である．目標音声を STU に通すことで，音声波形に対応した音声単位列が得られ，それを用いて UTS の学習を行う（図 6）．本実験では，LJSpeech [20] の単一話者の音声データと，ナレーションデータから取得した他の話者のデータから UTS を学習させた．

4. システム構成

WESPER のモデルは Pytorch のフレームワークを用いて構築した．STU は HuBERT を，UTS は [6] による FastSpeech2 PyTorch 実装の修正版を改造することで実現した．事前学習には Librispeech と wTIMIT（いずれも通常音声とささやき声音声を収録）を使用した．NVIDIA R6000 をデュアルで使用した場合，事前学習時間は 48 時

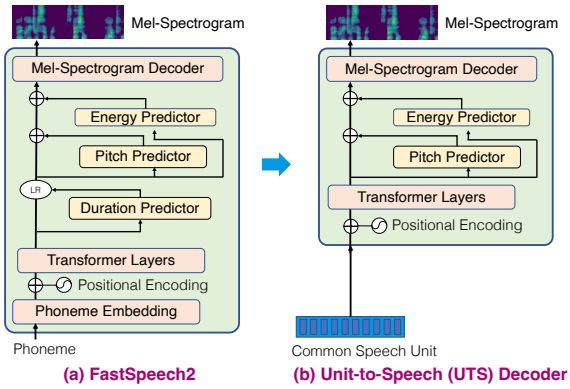

図 5 FastSpeech2 [37] と Unit-to-speech (UTS) decoder の比較. FastSpeech2 は各音素の継続時間を予測する必要があるが，UTS は同じ継続時間を持つ共通の音声単位を受け付けるため，継続時間予測器と長さ調節器（図中の LR）を省略することができる．また，共通音声単位は離散的なトークンから構成されないため，音素埋め込み (Phoneme Embedding) 層も不要になる．

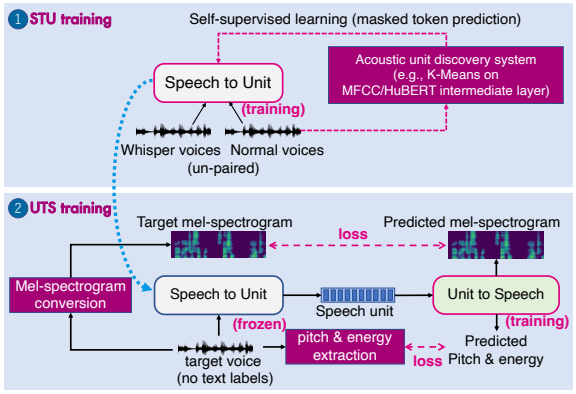

図 6 STU(1) と UTS(2) の学習：UTS (unit-to-speech) は，凍結された STU を用いて，対象音声の波形データのみから対象音声の構築を学習する．対応するテキストラベルは不要である．

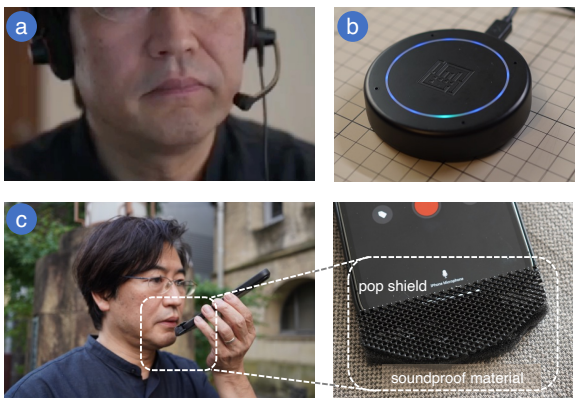

図 7 WESPER で試用した音声入力装置: (a) ヘッドセット, (b) アレイマイク, (c) ポップガードノイズを装着したスマートフォン.

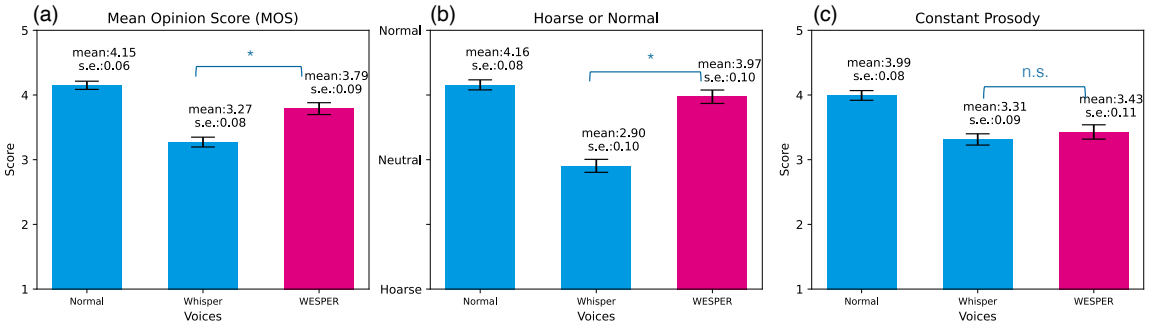

図 8 WESPER によるささやき声→通常声の変換品質評価: (a) ささやき声，通常声，WESPER 変換結果に対する Mean opinion scores (MOS). (b) かすれ声か通常声かの判定, (c) 韻律の自然さの判定 (`s.e.`:標準誤差，`*`:$p < 0.01$(t 検定), `n.s.`:有意差なし)

間程度であった．また，UTS の訓練には，各ターゲット音声に対して 26 時間を要した．

変換に必要な全体の処理時間は，NVIDIA R6000 を 1 台使用した場合は実際の音声の時間の約 1/20，Apple M1 Max の CPU を使用した場合は約 1/10 であった．実際の変換音声品質を添付のビデオで示す．

インタフェースは 2 種類構築した．1 つは，ボタンを押しながら，ささやき声や普通の声で話す，プッシュ・トゥ・トーク方式である．ボタンを離すと，その間に録音された音声波形が音声変換ニューラルネットワークに送られ，その結果がすぐに再生される．もう 1 つは，入力された音声の無音声期間を検出し，ユーザが追加入力することなく，自動的に音声セグメントを変換する．

図 7 に現在の WESPER の音声入力構成を示す．(a) ヘッドセット，(b) MEMS マイクを 4 個配列した指向性マイク，(c) 携帯電話用マイク，ささやき声のポップノイズを防ぐポップガードと環境騒音を低減する防音材を装着してテストを行っている．

5. 評価

WESPER 機構は，入力話者・言語に依存しないささやき声から通常音声への変換を可能にする．本章では，ささやき声から通常音声への変換と，言語障害者・聴覚障害者の音声再構成を考慮し，3 つの基準から変換品質を評価する．

5.1 ささやき声から通常音声への変換品質

変換された音声の品質を評価するために，クラウドソーシングシステム Prolific [19] を利用して，18 歳以上で英語が堪能な男女比を均等にした計 50 名の実験参加者をインターネット上で募集した．各参加者は，Web を利用した評価システムを介して，4 セットの通常発話，ささやき声，WESPER 変換されたささやき声 (合計 12 音声) を聴き，5 段階 MOS(平均意見スコア，Mean-Opinion Score) および二つのアンケート項目 （“Is the voice hoarse or normal?”, “Does the speech uses consistent articulation, standard intonation and prosody?”）で 5 段階評価で順位をつけた．文の内容による印象の違いを避けるため，発話例はすべて同じ書き下し文を使用した．通常音声，ささやき声音声は wTIMIT コーパスより採取したものを使用した．WESPER 音声は LJSpeech による単一話者音声で学習したものを利用し，ささやき声音声データから変換したものを用いた．結果を図 8 に示す．

図 8 (a) は MOS の評価結果である．WESPER 変換により，元のささやき声の MOS が向上することが確認された (paired t-test, $p < 0.01$)．図 8-(a) (b) は「この声はかすれ声か普通か」という質問に対する回答を示しており，WESPER で変換された音声では明らかな改善が見られる ($p < 0.01$)．図 8 (c) は，「声は一貫したアーティキュレーション，標準的なイントネーション，韻律を保持しているか」に対する回答で，WESPER と Whisper はほぼ同じスコアを示している（有意差なし）．WESPER で合成された音声は，現音声の韻律の自然さを毀損していないと解釈できる．

これらの評価から，以下の結論を得た．

- WESPER はささやき声の音声を普通の音声に変換できる．
- WESPER で変換した音声データは，元のささやき声よりも MOS が良い．
- WESPER は現音声 (ささやき声) の韻律を毀損しない．

5.2 音声認識精度

WESPER を音声認識器として利用する場合，WESPER 単体で音声認識する方法と，WESPER で変換した音声を他の音声認識器で認識する方法の 2 通りが考えられる．

前者は，ささやき声と通常音声で事前学習した WESPER モデルを，ささやき声コーパスを用いて微調整 (finetuning) し，ささやき声からテキストを推定する方法である．後者では，WESPER で変換されたささやき声は，他の音声対応機器の制御に利用することができる．ささやき声が通常の音声に変換できれば，既存の音声対応機器をささやき声用に改造することなく，すぐに利用することができる．

認識モデル	学習データ (finetuning)	評価データ	WER (%)	CER (%)	BLEU
Google	Google	wTIMIT(N)	11.55	4.66	0.76
		wTIMIT(W)	44.70	28.38	0.34
		WESPER[wTIMIT(W)]	**26.68**	**12.70**	**0.52**
HuBERT base	Librispeech	wTIMIT(N)	21.06	8.17	0.54
	Librispeech	wTIMIT(W)	33.06	15.45	0.38
	librispeech + wTIMIT(N,W)	wTIMIT(W)	**13.75**	**5.47**	**0.70**

wTIMIT(N): wTIMIT [27] normal voice　　wTIMIT(W): wTIMIT whisper voice
Google: Google Cloud Speech-to-Text [17]　　WESPER[•]: WESPER converted •
HuBERT: HuBERT base model, pretrained with Librispeech [32] + wTIMIT(N,W)

表 2 Google Cloud Speech-to-Text citegooglespeech をリファレンス ASR とした場合のささやき声音声認識精度．　通常の ASR ではささやき声の認識率は高くないが，WESPER を用いてささやき声を通常音声に変換した結果を認識すると，認識率が向上した．（WESPER はラベル付きデータで事前学習していない）

このような観点から，2 種類の評価を行った．まず，既存の音声認識装置（Google Cloud Speech-to-Text [17]) を評価手段として，通常音声，ささやき声，WESPER で変換したささやき声の音声認識精度を測定した．

測定には，TIMIT に準拠した wTIMIT コーパスを用いた．wTIMIT は，通常音声とささやき声音声をラベル付きで収録したコーパスである．このコーパスを用いて，WESPER で変換したささやき声音声の認識精度を，通常の音声とささやき声音声の認識精度を基準として評価した．

結果を WER（word error rate）と CER（character error rate），および BLEU（bilingual evaluation understudy）で表 2 に示す．表に示すように，ささやき声音声を直接認識した場合，認識精度は高くないが（WER=44.70%），WESPER で変換すると認識精度が向上する（WER=26.68%）ことがわかった．

また，wTIMIT(W) でテストした場合，Librispeech と wTIMIT(N,W) で事前学習した HuBERT ベースは Google cloud speech to text よりも良いスコアを示した（HuBERT: WER=33.06%, Google: WER=44.70% ）．これは，ささやき声と通常音声を混合した事前学習（ただし，ささやき声による微調整は行わない）の効果が，ASR の精度に寄与している可能性があると推測される．

ここで注意すべきは，WESPER はラベル付けされたコーパスを用いて学習していないことである．WESPER はラベル付けされたコーパスを用いて学習するのではなく，ささやき声と通常の音声の混合データを用いて事前学習を行っている．すなわち，ささやき声に対する教師データ（テキストラベル）を要さずに，音声認識精度が向上している点が興味深い．

5.3 発声・聴覚障害者の発話補正の評価

WESPER の重要な目的の 1 つは，言語障害者や聴覚障害者の非定型音声を再構成することであり，この性能を評価した．

発声障害とは，痙攣，声道ポリープなど，多岐にわたる要因により，不随意に声がかすれたり，息苦しくなったり，緊張したり，音量や音程が小さくなったりすることである．また，咽頭癌などで声帯を切除した場合も，声を出すことが極端に難しくなる．一方，聴覚障害の場合，発声器官そのものに異常がなくても，発話の制御が難しくなり，発声障害を起こす．声帯損傷者の発声として，喉を機械的に振動させる EL（Electrolarynx）装置が使用されることがある．しかし，ノイズが発生し，発声する音は人工的で，ピッチ変換は決定論的であり，正常な発声との乖離が大きくなってしまう．発声障害によるコミュニケーション不全は深刻な問題であり，音声変換技術による解消は大きな社会的価値を持つ．

本研究では，以下の 2 種類の音声障害を持つ人の発話を評価した:

声帯ポリープ (Vocal Fold Polyps): (以下 VFP) 声帯ポリープは，喉頭の良性病変の中で最も頻度が高く，声質に影響を与える疾患である．
痙攣性発生障害 (Spasmodic Dysphonia): (以下 SD) 喉頭ジストニアとも呼ばれ，声や会話に影響を与える代表的な神経疾患である．声を出すための筋肉が痙攣を起こすことで発声が阻害される．

言語障害者の音声を収録したコーパスとしてよく利用されている “Saarbruecken Voice Database” (SVD) [34] を使用する．このコーパスには，種々の発声障害を持つ複数の話者により，母音発話の録音と，ドイツ語の参照文である “Guten Morgen, wie geht es Ihnen?” が収録されており，この文章の発声を評価に使用した．

評価のために 50 人の実験参加者を Prolific を用いてインターネット上で募集した [19]. 募集した参加者は，男女比のバランスがとれており，全員が 18 歳以上であった．また，今回は例文がドイツ語であったため，英語に加えてド

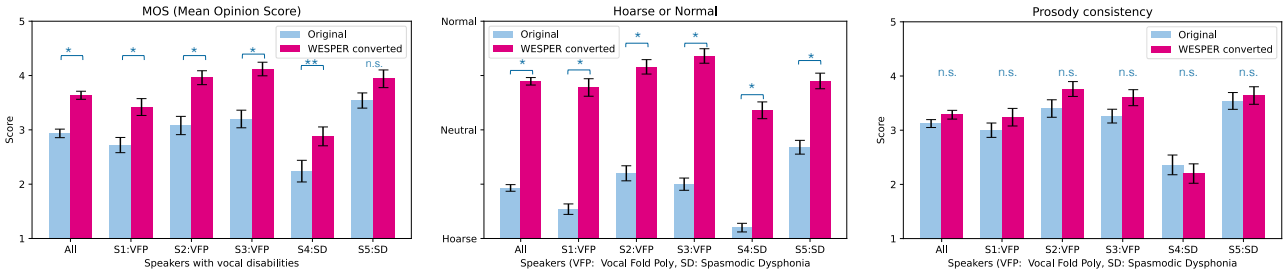

図 9 発声障害者の WESPER による声質評価（5 段階 MOS） S1-S5:話者,. *:$p < 0.01$, **:$p < 0.05$, VFP:声道ポリープ, SD:痙攣性発声障害

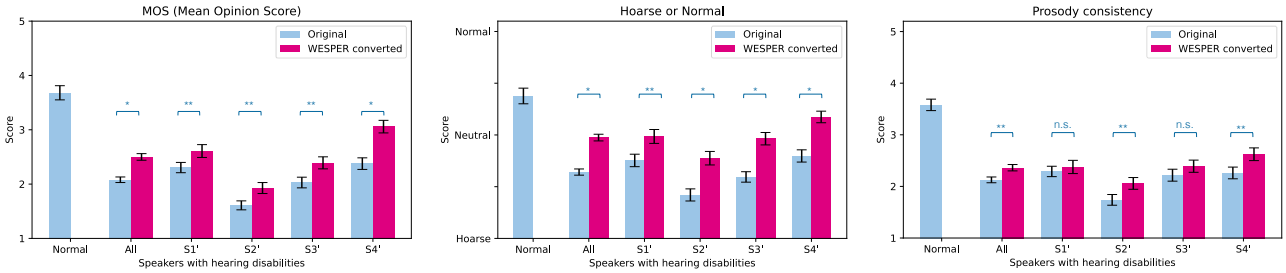

図 10 聴覚障害者の声質評価（5 段階 MOS）S1'-S4':話者,*:$p < 0.01$, **: $p < 0.05$

イツ語にも堪能なことを参加条件とした．

結果を図 9 に示す．図 9(左) は，MOS の結果を示している．SD，VFP ともに WESPER で変換した音声の方が高い MOS 値を示した（paired t-test, $p < 0.01$）．図 9(中) は「この声はかすれ声か普通か」という質問に対する回答を示しており，WESPER 変換後の音声は元の VFP と SD 音声に比べて明らかに改善している（$p < 0.01$ ）．図 9(右) は，“Is the voice using consistent articulation, standard intonation, and prosody?” という質問に対する答えを示したものである．ここでは，WESPER に変換された音声とオリジナルの音声はほぼ同じ評点を示した (有意差なし）が，WESPER に変換された音声の方が若干良い評点であった．

これらの評価から，以下のような結果を得た：

- WESPER で変換された VFP および SD の音声は MOS が向上しており，WESPER は個々の発話内容を知らない人からの理解という点で，これらの人の発話の質を向上させることができることが示された．
- WESPER は元の VFP および SD の音声を，かすれ声よりも通常音声に近くすることができた．
- WESPER は原音音声の韻律を毀損しない．

また，このテストはドイツ語の文章に対して行われたことを指摘しておく．WESPER の事前学習は英語音声のみで，ドイツ語音声での学習行っていないが，原文の韻律を保持し，その MOS を向上させることができた．この結果は，WESPER モデルが言語依存していない可能性を示している．

5.4 聴覚障害者の音声再構成評価

最後に，聴覚障害者に対する WESPER による音声再構成の効果を評価した．聴覚障害者は，健常者と比べて自分の話す声がよく聞こえないため，一般話者から理解されやすい話し方をすることが難しい傾向にある．しかし，発声器官は正常であるため，音声障害者の音声とは異なる特徴を示す．

評価には，“Corpus of deaf speech for acoustic and speech production research(音響・音声生成研究のための聴覚障害者音声コーパス)” [30] を使用した．5 人の話者（うち 1 人は健聴者，残り 4 人は聴覚障害者）の発話を使用した．実験参加者として，Prolific [19] を用いて，男女比が均等で，英語が堪能で，18 歳以上の条件で 50 名を募集した．

結果を図 10 に示す．これらの結果から，MOS などの評価結果は WESPER で変換した方が高いスコアを示す傾向があったが（$p < 0.01$ or $p < 0.05$），その改善度合いは声帯障害のある話者よりも小さかった．また，聴覚障害者の音声は，健聴者の音声と比較して，韻律評価が有意に低いことが示された．したがって，これらの結果は，聴覚障害者が発話中に韻律を制御することが困難である可能性を示している．

6. 議論

ささやき音声に適した音声入力装置

本研究で示したように，ささやき声やかすれ声も通常の音声に変換することが可能である．しかし，実際には，適切な音声入力装置を選択することが重要である．現在，通常のヘッドセットとスマートスピーカー用に開発した指向

性アレイマイクを用いて，提案手法を検証しており，これまでに良好な結果を得ている．ささやき声インタラクションが利用可能なことが立証されたので，スマートフォンのマイクロフォンをささやき声収録に適するように改善すること（アレイマイク等の導入により）が考えられる．ウェアラブル機器では，皮膚の振動を検出する NAM（Non Audible Murmur）マイクロフォンを非可聴発話に利用することが考えられる [15].

また，フィリップスやダイソンでは，大気汚染や感染症対策として，呼吸を電動で換気するマスクを開発している [10], [18]．マスクの内側にマイクを仕込んでマスクの中にマイクを入れて，ささやくような声を拾えば，無声音声に近い効果が得られる．

人間と AI の融合

本研究は，不特定多数の話者のささやき声を通常の音声に変換する機械学習技術に関するものである．しかし，実際に使ってみると，同じようなささやき声でも，簡単に通常の音声に変換できるものと，難しいものがあることにユーザが気づくことがわかった．ユーザ側で機械学習に寄り添って発話しようとしたのだ．この相互作用は，機械学習が一方的に人間の能力を拡張するだけでなく，人間側の学習によってもさらなる相乗効果が得られることを示唆している．これは，発声による人間と AI との融合の実例と考えることができる．

7. 結論

本研究では，ささやき声から通常の音声にリアルタイムに変換する仕組みとして WESPER を提案した．また，ささやき声と通常音声で音響的特徴が異なる場合でも，自己教師付き学習により共通の音声単位が得られることを確認した．音声再生は，テキストラベルを用いず，任意の対象話者の音声データのみから学習することが可能である．また，ユーザごとの学習は不要であり，ささやき声と通常音声の並列データも不要である．WESPER は，音声単位から，任意の対象話者の音声の発話を復元することができ，対象話者のラベルのない音声データのみが必要だ．変換された音声の品質が向上し，音声の韻律が保持されることを確認した．また，言語障害者や聴覚障害者の音声発話を復元した結果の評価も報告した．

謝辞

本研究は JST Moonshot R&D Grant Number JPMJMS2012, JST CREST Grant Number JPMJCR17A3, 東京大学ヒューマンオーグメンテーション社会連携講座の支援を受けて実施された．

参考文献

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for Self-Supervised learning of speech representations. arXiv [cs.CL], June 2020.

[2] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. Toward silent-speech control of consumer wearables. Computer, Vol. 48, No. 10, pp. 54–62, 2015.

[3] Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, and Ye Jia. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation, 2019.

[4] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation, 2022.

[5] Heng-Jui Chang, Alexander H Liu, Hung-Yi Lee, and Lin-Shan Lee. End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training. May 2020.

[6] Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8588–8592, 2021.

[7] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. Silent speech interfaces. Speech Commun., Vol. 52, No. 4, pp. 270–287, April 2010.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[10] Dyson. dyson zone: Air-purifying headphones with active noise cancelling. `https://www.dyson.co.uk/en`, 2022.

[11] Joo Freitas, Antnio Teixeira, Miguel Sales Dias, and Samuel Silva. An Introduction to Silent Speech Interfaces. Springer Publishing Company, Incorporated, 1st edition, 2016.

[12] Masaaki Fukumoto. Silentvoice: Unnoticeable voice input by ingressive speech. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, UIST '18, p. 237–246, New York, NY, USA, 2018. Association for Computing Machinery.

[13] Teng Gao, Jian Zhou, Huabin Wang, Liang Tao, and Hon Keung Kwan. Attention-guided generative adversarial network for whisper to normal speech conversion, 2021.

[14] Dorde T. Grozdic and Slobodan T. Jovicic. Whispered speech recognition using deep denoising autoencoder and inverse filtering. IEEE/ACM Trans. Audio, Speech and Lang. Proc., Vol. 25, No. 12, p. 2313–2322, dec 2017.

[15] Panikos Heracleous, Yoshitaka Nakajima, Hiroshi Saruwatari, and Kiyohiro Shikano. A tissue-conductive acoustic sensor applied in speech recognition for privacy. In Proceedings of the 2005 Joint Conference on

Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies, sOcEUSAI '05, p. 93–97, New York, NY, USA, 2005. Association for Computing Machinery.

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech and Lang. Proc., Vol. 29, p. 3451–3460, jan 2021.

[17] Google Inc. Google cloud speech-to-text. `https://cloud.google.com/speech-to-text`, 2020.

[18] Philips Inc. Fresh air mask series 6000. `https://www.philips.com.sg/c-p/ACM066_01/fresh-air-mask-series-6000`, 2021.

[19] Prolific inc. Prolific, 2014.

[20] Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[21] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks, 2017.

[22] Arnav Kapur, Shreyas Kapur, and Pattie Maes. Alterego: A personalized wearable silent speech interface. In 23rd International Conference on Intelligent User Interfaces, IUI '18, p. 43–53, New York, NY, USA, 2018. Association for Computing Machinery.

[23] Naoki Kimura, Michinari Kono, and Jun Rekimoto. Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, p. 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[24] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.

[25] Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. Textless speech emotion conversion using discrete and decomposed representations, 2021.

[26] Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. Generative spoken language modeling from raw audio, 2021.

[27] Boon Pang Lim. Computational differences between whispered and non-whispered speech, ph.d. thesis, university of illinois urbana-champaign, 2010.

[28] H. Malaviya, J. Shah, M. Patel, J. Munshi, and H. A. Patil. Mspec-net : Multi-domain speech conversion network. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7764–7768, 2020.

[29] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. pp. 498–502, 08 2017.

[30] Lisa Lucks Mendel, Sungmin Lee, Monique Pousson, Chhayakanta Patro, Skylar McSorley, Bonny Banerjee, Shamima Najnin, and Masoumeh Heidari Kapourchali. Corpus of deaf speech for acoustic and speech production research. The Journal of the Acoustical Society of America, Vol. 142, No. (1), p. EL102, 2017.

[31] Marco A. Oliveira. Machine learning approaches for whisper to normal speech conversion: A survey. U.Porto Journal of Engineering, Vol. 8, No. 2, pp. 202–212, 2022.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210, 2015.

[33] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. April 2021.

[34] Manfred P̈utzer and William J. Barry. Saarbruecken voice database, 2016.

[35] Jun Rekimoto. Dualvoice: A speech interaction method using whisper-voice as commands. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[36] Jun Rekimoto. Dualvoice: Speech interaction that discriminates between normal and whispered voice input. arXiv, 2022.

[37] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2020.

[38] Gabriel Reyes, Dingtian Zhang, Sarthak Ghosh, Pratik Shah, Jason Wu, Aman Parnami, Bailey Bercik, Thad Starner, Gregory D. Abowd, and W. Keith Edwards. Whoosh: Non-voice acoustics for low-cost, hands-free, and rapid input on smartwatches. In Proceedings of the 2016 ACM International Symposium on Wearable Computers, ISWC '16, p. 120–127, New York, NY, USA, 2016. Association for Computing Machinery.

[39] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. The tongue and ear interface: A wearable system for silent speech recognition. In Proceedings of the 2014 ACM International Symposium on Wearable Computers, ISWC '14, p. 47–54, New York, NY, USA, 2014. Association for Computing Machinery.

[40] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. Lip-interact: Improving mobile device interaction with silent speech commands. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, UIST '18, p. 581–593, New York, NY, USA, 2018. Association for Computing Machinery.

[41] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, 2016.

[42] Benjamin van Niekerk, Marc-Andre Carbonneau, Julian Zaidi, Matthew Baas, Hugo Seute, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, may 2022.

[43] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: A fully end-to-end text-to-speech synthesis model. CoRR, Vol. abs/1703.10135, , 2017.

[44] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. Applying wav2vec2.0 to speech recognition in various low-resource languages. December 2020.

[45] zeta chicken. towhisper, 2017.