大和大学　研究紀要　第10巻　情報学部編　2024年３月　pp.67～78

円滑なオンライン音声コミュニケーションの支援に向けた音声強調技術

Speech Enhancement Technology Toward Supporting Online Communication with Hearing-Impaired People

小林 彰夫* 安 啓一**

KOBAYASHI Akio and YASU Keiichi

要　旨

聴覚障害者と聴者の円滑なオンラインコミュニケーションに向けた支援技術について述べる．Zoom などのオンライン会議ソフトウェアを通じたオンラインの音声コミュニケーションでは，話声の明瞭度や聞き取りやすさが通信の品質に依存する．筆者らの研究の目的は，こうしたオンラインの音声コミュニケーションにおいて，音声の品質が聴者・障害者双方の話声の明瞭度や聞き取りやすさにどのような影響を及ぼすのかを明らかにすること，そして音声の品質を改善して聞き取りやすい円滑なコミュニケーションを実現するための支援技術を開発することである．なかでも聴覚障害者の聞き取り支援を目的とした音声強調技術が本研究の核である．しかし，オンラインの音声コミュニケーションにおける，聴覚障害者の立場からみたときの音声の品質の良し悪しや聞き取りやすさは明らかではない．本稿では，まずオンラインにおける音声コミュニーケーションの現実的な通信シナリオと，音声品質を阻害するいくつかの要因を定める．そして，シナリオに沿って擬似的な音声を生成し，これまでに提案されているいくつかの音声強調手法および客観的評価基準を用いて，聴覚障害者による主観評価への利用の可否について検討する．

Abstract

This paper describes technologies for accessible online communication between normal-hearing and hard-of-hearing (HoH) people. In online speech communication through online conferencing software such as Zoom, the quality of communication depends on the intelligibility and understandability of the spoken words. Our research investigates the effect of speech quality in online speech communication. This research aims to clarify how speech quality affects the intelligibility and understandability of speech for both normal-hearing and HoH in such online speech communication and to develop accessible technologies to improve speech quality and realize effortless communication. The core of this research is speech enhancement technology to help HoH people in listening comprehension. However, it needs to be made clear how good or bad the quality of speech is in online speech communication from the viewpoint of the HoH people. Thus, we first define a scenario for online speech communication and some factors degrading speech quality. Then, we generate simulated speech according to the scenario and investigate whether or not it can be used for subjective evaluation by HoH people by employing several speech enhancement methods and objective evaluation criteria.

キーワード: 音声強調，聴覚障害，主観評価，客観評価，ニューラルネットワーク

Keywords: speech enhancement, hearing inpaired, subjective evaluation, objective evaluation, neural network

1 はじめに

筆者らはこれまで，高齢者・聴覚障害者への情報保障を目的とした音声認識による情報保障・音声対話の研究に携わってきた．しかし，これまでの研究では近年増加しているオンライン環境におけるコミュニケーションが考慮されておらず，聴者と聴覚障害者との間の音声による意思疎通への対応が課題となっている．例えば [1] によるアンケートでは，聴覚障害者のおよそ半数がオンラインでのコミュニケーションに不安を抱いている．その理由は「(音声の) 品質が低いために聞き取りにくい」,「会議に (十分な品質の) 字幕がない」といったことである．

聴者と聴覚障害者との意思疎通では，筆談，手話通訳，音声認識といった手段が用いられる．オンライン環境では一般に筆談を用いるケースが多いと想定されるが，キーボードを用いたタイピングによる入力は時間がかかることから会話が中断しやすく，円滑なコミュ ニケーションが実現しにくい．一方，手話通訳は意思伝達手段として優れているが，手話を解するのは聴覚障害者の 25%程度にすぎず，音声による意思疎通を行う者は障害者の 50%以上 (補聴器・人工内耳装用と筆談の総計) と多数を占めていることから，音声によるコミュニケーションが重要であることがわかる [2].

音声認識は聴者が自身の意思を伝えることは可能だが，正しく発話する能力が不十分な聴覚障害者は，特に自由発話において音声認識の恩恵を受けることが難しい．筆者らのこれまでの研究 [3] では，聴覚障害者の発話に対する音素レベルの認識率は 20~80%程度と，話者によってきわめて認識率にばらつきが大きいことがわかっている．一方，研究 [4] によれば，文字 (字幕) と音声の双方の提示によって内容の理解が進むことが示されている．音声認識に誤りがあっても，聴覚による補償があれば頑健なコミュニケーションが可能と考えられるが，聴者は障害者の発話の曖昧性 [5] が原因で音声認識の誤りを聴覚によって補償することが困難である．

同様に，聴覚障害者オンライン会議ソフトウェアにおける音声通話の品質の低さに起因して，聴覚活用による情報の取得に困難を感じていると考えられる．例えば，無線通信を介した通話では，パケット損失により音声の一部が欠落するため，品質の劣化が明瞭度や聞き取りづらさと結びついているといえる．また，音声が電話品質程度の狭帯域である場合，高い周波成分に特徴的なパワーのある一部子音音素の聞き取りが困難になるかもしれない．

*大和大学情報学部情報学科　**筑波技術大学産業技術学部産業情報学科

令和６年２月２日受理

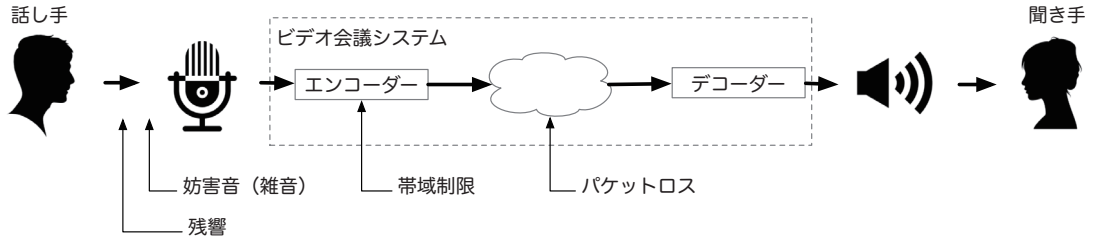

図 1: オンラインコミュニケーションにおける通信路

音声の品質は通信路だけではなく，話し手の通話環境 (収録環境) にも依存する．話し手が通話を行うマイクロフォンの周囲に音源があれば，話し手の話声はマスクされ明瞭度が著しく低下する．話声を収録する場所も品質に影響を与える．例えば，話声が室内で反射することにより残響を生じ，これが原音声に重畳することにより音声品質が低下する．

オンラインコミュニケーションにおける音声通話の品質は，上で述べた通信路や収録環境に起因するだけではなく話し手の発話スタイルにも大きく依存する．例えば，音声学的な観点からみた聴覚障害者の発話の特徴は，鼻音化や母音の中性化，二重母音化，子音の欠落や有声・無声の置換に現れるとされている [6]．したがって，聴者が聴覚障害者の発話を聞き取る際に，こうした発声の変形が聞き取りに大きな影響を及ぼすことは明らかである．

低品質の音声に対しては，音声強調による品質改善が行われることが多い．信号処理から深層学習を利用した手法まで，これまでさまざまな提案がなされている．音声強調手法により品質の改善した音声を評価する手法もまた，さまざまな提案が行われている．一方で，品質評価は聴者による主観評価であったり，聴者の聴力に基づく客観評価であるなど，聴覚障害者が聴取した際の品質について顧みられる機会があまりない．

筆者らの研究は，聴覚障害者と聴者間のオンライン環境における音声コミュニケーションを円滑に行えるような支援技術の開発を目的としている．支援技術のひとつの実現としては，聴覚障害者が残存聴力をもって聞き取りやすい音声を音声強調手法により実現することである．

この実現のためには，i) オンラインにおける通信環境の再現と実・模擬データによるデータ収集，ii) 聴者を対象とした既存の評価基準による音声強調手法の評価に基づく品質分類，iii) 聴覚障害者を対象とした主観・客観評価およびこれらの基準に基づく音声強調手法の開発，の 3 つの研究ステップが必要だと考えられる．

そこで本稿では，聴覚障害者・聴者のオンライン上の音声コミュニケーションを改善する上で，上記 i) に関してオンラインコミュニケーションにおける通信環境を再現した上で模擬データを生成し，深層学習によるアプローチに基づいて音声強調を行い，従来法として提案されている複数の評価基準により評価を行った．実験により，聴覚障害者による評価や音声強調手法の開発に向けた課題を明らかにする．

2 オンラインにおける音声コミュニケーション

本章では，オンラインの音声コミュニケーションにおける通信シナリオの概要と，通信路と収録環境の点からみた音声品質の劣化要因について述べる．

2.1 オンラインにおける音声コミュニケーション

IP 電話等に見られる通信シナリオを元に [7]，オンライン会議ソフトウェアを通じた音声コミュニケーションの通信シナリオを図 1 に示す．図 1 に示す通信路では，まず話し手（送り手）の話声がマイクロフォンで収音され，さらにオンライン会議ソフトウェアに内蔵されているエンコーダーにより符号化・圧縮されて IP ネットワークへ送出される．聞き手（受け手）側では符号化された音声を受信後，デコーダーにより復号・伸長しスピーカー/ヘッドホンを通して再生する．

図 1 に示す通信路では，音声の品質を低下させる多様な妨害要因が存在する．例えば，送り手の話声を収音する際は，周囲の妨害音 (雑音や他の人の話声) が混合する可能性がある．また，送り手の話声あるいは妨害音が室内の壁面等で反響し，残響として混合する．送り手以外の話し手がする場合，複数の話声が重なり合う．上記のいずれも，聞き取りやすさを阻害する要因となる．

VoIP による音声の伝送では通常，非可逆式のコーデックを用いて音声を符号化・圧縮する．近年広く使われる音声コーデックの一つに Opus[8] がある．Opus は電話のような狭帯域 (3.4 kHz) から AM ラジオ程度の品質のワイドバンド (7 kHz)，フルバンド (20 kHz) までを 2 つのコーデック (SILK および CELP) でカバーしていること，競合する他のコーデックと比べて遅延が小さいことが特徴であるまた，冗長化による前方誤り訂正 (Forward Error Correction) がサポートされている [9, 10]．

音声の帯域幅と音声の明瞭度には相関があるとされている [11]．例えば，4 kHz 以上の帯域にエネルギーが集中する一部の子音は狭帯域の伝送では明瞭度が低下してしまう．したがって，通信路の問題により帯域に制限がある場合，品質の低下した音声が伝送される可能性がある．

さらに，音声通話においてはパケット損失も大きな問題である．特にワイヤレス通信における遮蔽，フェージング，輻輳により音声通話の品質は著しく劣化する．こうしたパケット損失による音声品質の低下を防ぐために，パケット損失隠蔽 (Packet Loss Concealment; PLC) といった技術が提案されている．

3 音声強調による劣化音声からの原音声の復元

3.1 音声強調

現実世界の音声コミュニケーションでは，聞き取りたい音（目的音）以外にもさまざまな音が存在する．室内であればエアコンの吹き出し口やノートパソコンの回転するファンの音といった雑音や，複数の話者がいればほか

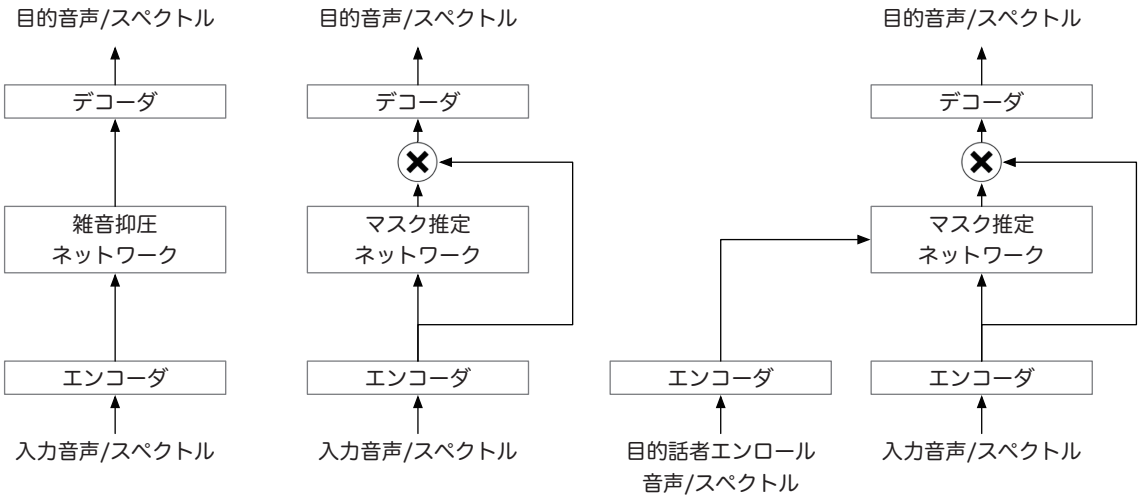

図 2: 雑音抑圧 (左)，音源分離/雑音抑圧 (中)，個人向け音声強調 (右)

の人の話声である．また，講堂のような広い室内では天井や壁面で音が反射し，響き（残響）が生じる．

しかし，私たちは選択的聴取能力（カクテルパーティー効果）によって，自身が注意を向けている音声だけを都合よく聞き取ることができる．この選択的聴取を計算機で実現すること，例えば複数の話声をそれぞれ正しく音声認識するということは，今も解決の途上にある課題である．

一般に，実環境のさまざまな聞き取りを阻害する音や要因を取り除き，目的音のみを抽出する技術を音声強調とよぶ．収録系のマイクロフォンの数が単一か複数かによって音声強調のアプローチは異なるが，本稿では主として単一マイクロフォンによる音声強調手法を取り上げる．

音の信号処理では，時間領域の信号を短時間フーリエ変換 (Short-Time Fourier Transform; STFT) を用いて時間周波数領域に変換して処理することが多い．これは音が調波構造を持っているためであり，音声であれば声帯振動とその倍音を含むような特徴を表現するのに適しているからである．一方で，最近の深層学習に基づく音声強調では，時間領域の信号をそのまま処理する end-to-end 手法が数多く提案されている．

雑音抑圧の技術に着目すると，スペクトラルサブトラクション法 [12] が古くから行われている代表的な手法である．これは，雑音の重畳した音声のパワースペクトルから雑音のパワースペクトルを減算することにより，目的音声のパワースペクトルを推定する手法である（雑音のパワースペクトルは非音声区間から推定する）．

2010 年代に入って深層学習の研究が大きく進展すると，ニューラルネットワークの一種であるオートエンコーダー（自己符号化器）を用いた雑音抑圧手法が提案され，音声認識などの分野で上に挙げた従来手法を凌駕する性能が得られることとなった．初期の研究である雑音抑圧オートエンコーダー [13, 14] では，フィードフォワード型の多層ニューラルネットワークが用いられていたが，近年は，畳み込みニューラルネットワークの一種である U-Net[15] をベースにしたモデル [16, 17] や，周波数領域におけるグローバル/ローカル特徴を用いるモデル [18] などのさまざまなアプローチが提案されている．U-Net 構造 [15] を用いた Denoiser の雑音抑圧のアプローチでは，図 2(左) のダイアグラムで示すように，入力音声もしくは STFT により得られたスペクトログラムをネットワークへの入力とする．ネットワークでは，（直観的には）多層の畳み込み層で目的音声に有用な情報を抽出する操作を繰り返す．

ニューラルネットワークのモデルパラメータは，品質低下した音声と目的音声（教師信号）のペアを用いた学習により得られる．ネットワークの出力と目的音声により定義される損失関数は，例えば Denoiser[16] では平均絶対誤差 (Mean Absolute Error; MAE) 損失 (ℓ_1損失) と多重解像度 STFT 損失 [19] の組み合わせであり，いわゆるマルチタスク学習を構成している．

3.2 帯域拡張

容量に制限のある通信路では，帯域の一部をカットして品質を犠牲としたうえ狭帯域音声として伝送を行う場合がある．かつて広く使われた固定電話では，帯域の上限がおおむね 3.4 kHz 程度であったが，これは音声が担う主要な情報が 4 kHz 以下に含まれているためである．また，AM ラジオ放送はおおむね上限が 7.5 kHz（ワイドバンド相当）である．しかし，聴者であれば帯域の欠落を品質低下として捉えることも多いため，品質改善を目的とした帯域拡張技術が開発されてきた．従来法では，低域の励起信号と高域のスペクトル包絡の推定により帯域を拡張する手法が多く [20]，推定にガウス混合分布を含む隠れマルコフモデルが利用されてきた．近年では雑音抑圧同様，深層学習を用いる方法が提案されている [21, 22]

3.3 音源分離

複数の目的話者が同時に発話する際は，それぞれの音源を分離して音声認識などの処理を行う必要がある．複数のマイクロフォン（マイクロフォンアレイ）を利用する場合，目的音声の到来方向を推定（ビームフォーミング）して音源分離する手法が使われる．

近年，マイクロフォンの本数によらず，ニューラルネットワークを用いた音源分離が多く提案されている [23, 24, 25, 26, 27] ニューラルネットワークを用いた音源分離のアプローチは，図 2(中) のダイアグラムに示すよう

に，STFT により得られたスペクトログラムから，目的音声を強調し，かつ妨害音を減衰させるゲイン調整を行うマスクを推定する構成を取ることが多い．スペクトログラムの代わりに時間領域の音声信号を入力する際は 1 次元の畳み込みニューラルネットワークを用いたエンコーダーにより，畳み込みフィルタのチャンネル次元を加えた 2 次元信号とすることが多い．

音源分離の手法うち，[25, 26, 27] は，抽出したい目的話者に関する情報がいわゆる “エンロール音声 (enrollment speech)” として得られているのであれば，これを一種のヒントとして話者情報を取り出し，目的音声を分離する，エンロール音声を用いる分離手法は “個人向け音声強調 (Personalized Speech Enhancement; PSE)” とよばれる [28]．図 2(右) のダイアグラムに示すように，ニューラルネットワークを用いて，エンロール音声から話者埋め込み (speaker embedding) と呼ばれる話者性を反映した埋め込み特徴量を抽出する [29, 30, 31]．この話者性を反映した特徴量を目的音声抽出のための話者依存マスクを推定するネットワークに入力する．目的音声は，話者依存のマスクと入力音声の埋め込みとの積を計算したうえで，デコーダーによって時間周波数特徴もしくは波形に変換することにより得る．

PSE の研究には SpeakerBeam[25, 26] がある．オリジナルの SpeakerBeam は ConvTasNet とよばれる 1 次元の畳み込みをモジュールとしたネットワークベースとしているが，近年，構造をより単純化した E3Net[27] とよばれるネットワークが提案されている．

3.4 音声強調における推定音声の品質評価

音声強調や音源分離により推定された目的音声と真の目的音声（リファレンス）との間には，干渉，雑音およびアルゴリズムによるアーティファクトに相当する誤差が加算されている [32]．それぞれの誤差に対して，SIR(Source-to-Interference Ratio)，SNR(Source-to-Noise Ratio)，SDR(Source-to-Distortion Ration)，SAR(Source-to-Artifact Ratio) が定義され，信号どうしの比較が可能となる．

一方，音声の品質評価は，平均オピニオンスコア (Mean Opinion Score; MOS) を用いた主観評価 [33] により行われるが，評価に要する時間・金銭的コストの面から客観評価手法を代替手段とすることが多い．このうち，PESQ(Perceptual Evaluation of Speech Quality)[34, 35, 36] は，代表的な客観評価基準で，国際電気通信連合 ITU-T により標準化 [34, 37, 35] されており，狭帯域/広帯域の両者で評価できる．後継の POLQA[38] に比べ，Python などのライブラリを通して利用可能であることから，現在もよく使われる指標である．また，標準ではないが ViSQOL[39, 40, 41] といった客観指標も提案されている．

PESQ，POLQA あるいは ViSQOL は，“フルリファレンスモデル”（“ダブルエンド”，“侵入型“とも）よばれる評価方法であり，評価の際に目的音声と抽出した音声の両者が必要となる．

音声強調アルゴリズムにより雑音抑圧された音声の主観品質評価方法として ITU-T P.835 が規定されている [42]．ITU-T P.835 では，目的音声の歪み，妨害音の（目的音声に対する）侵入度，総合評価をそれぞれ 5 段階で評価する．これは，音声強調アルゴリズムが雑音を抑圧すればするほど，目的音の品質にも悪影響を及ぼすトレードオフの関係にあることが理由である．しかし，コストの面で高くつくことから，それぞれの値を推定する手法が提案されている [43]．また，[44, 45] では，この ITU-T P.835 の各値をニューラルネットワークによりリファレンスなしで推定している．DNSMOS は国際会議におけるコンペティション [46, 47] で採用されている．

従来の “ノーリファレンスモデル”（“シングルエンド”）としては ITU-T P.563 が規定されている [48]．P. 563 では，音声の不自然さ，雑音，クリッピングなどの歪みを分析して MOS 値を推定する．同様のノーリファレンスモデルとして，[49] で提案されている手法では自己注意機構を使ったニューラルネットワークで推定を行っている．

PSE では目的話者の過剰抑圧問題が着目されている [28, 50]．これは，音声強調アルゴリズムにより目的音声が，目的話者以外の話声など他の妨害音や雑音と同時に抑圧されてしまう問題を指す．目的音声の過剰な抑圧は音声強調アルゴリズム適用後の下流タスクとなる音声認識性能の著しい劣化をもたらす．[28] では TSOS (Target Speaker Over-Supprestion) Rate を導入することにより，過剰抑圧の評価と削減について提案している．

パケット損失を含む通信路からパケット損失隠蔽技術により復元された音声の品質の評価方法として，PLCMOS[51] が提案されている．

雑音抑圧された音声の評価では，明瞭性の評価も重要視される [52]．音声の明瞭度に関する評価は，聴取内容の書き取りによる手法 [53] があり，正しく聞き取れた語の割合を 0 から 1 までの指標として表すことが多い．明瞭度を測定する客観評価基準としては，STOI(Short-Time Objective Intelligibility)[54] が広く使われている．STOI は歪みのないリファレンスと音声強調アルゴリズムにより抽出された目的音声との時間周波数領域における特徴量の相関として定められる指標である．

3.5 聴覚障害者を対象とした品質評価

補聴器や人工内耳を装用した聴覚障害者に対しては，これまでのところ前述の主観評価に相当する品質評価方法はあまり行われていない．これは，聴力の損失や補聴器による補償方法が障害者によって大きく異なり．同一の音声であったとしても全く異なる評価値を与える可能性があることが原因だと考えられる．

一方で，聴力損失や補聴器の増幅などを反映した客観指標が提案されており，HASPI（Hearing Aid Speech Perception Index）[55]，HASQI（Hearing Aid Speech Quality Index）[56] といった指標が，Clarity Challenge[57] のようなコンペディションで用いられている．

補聴器などによる聴力損失の補正を行った上で，ITU-T P.835 のように目的音声，妨害音，総合評価といったきめ細かい主観評価，あるいはこれを代替するような深層学習ベースの評価手法の開発も筆者らの研究課題である．

4 音声強調実験

先に述べたように，オンラインの音声コミュニケーションでは，さまざまな要因により音声の品質が劣化する．

表 1: 実験 1:（パケット損失なし，妨害音なし，帯域 8 kHz）

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	**2.76**	**3.48**	**3.91**	**3.96**	**2.51**	**3.32**	2.99	**3.56**	**3.79**	**3.95**	**4.16**	**3.57**
SpeakerBeam	2.29	3.08	3.20	3.52	2.32	2.87	2.88	2.94	3.61	3.78	4.07	3.43
E3net	2.15	3.06	3.03	3.32	2.30	2.73	**3.10**	3.20	3.59	3.70	4.12	3.36
CleanUNet	2.65	3.40	3.60	3.82	2.46	3.21	2.97	3.46	3.46	3.90	4.03	3.45

表 2: 実験 1: （パケット損失 10 %，妨害音なし，帯域 8 kHz）

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	1.17	1.26	2.52	2.58	1.70	1.80	2.80	1.44	2.04	2.69	3.24	2.33
SpeakerBeam	**1.41**	**2.19**	**3.15**	**2.81**	1.86	**2.05**	2.72	**2.00**	**3.15**	**3.45**	**3.81**	**3.03**
E3net	1.36	2.02	2.93	2.66	**1.90**	1.98	**2.99**	1.86	2.49	2.93	3.47	2.48
CleanUNet	1.33	1.96	1.97	2.71	1.89	1.98	2.72	1.94	2.14	3.09	3.42	2.56

表 3: 実験 2: （パケット損失なし，妨害音なし，帯域 4 kHz）

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	**2.34**	**3.41**	**3.08**	0.18	**2.32**	1.24	2.89	**3.23**	3.45	**3.80**	**4.26**	**3.48**
SpeakerBeam	2.08	3.10	2.66	**2.96**	2.21	**2.48**	2.92	2.73	**3.51**	3.78	4.16	3.45
E3net	1.98	2.97	2.72	2.64	2.21	2.30	**3.03**	2.93	3.35	3.51	4.11	3.20
CleanUNet	2.24	3.29	2.94	0.90	2.27	1.55	2.88	3.16	2.65	3.73	4.08	3.33

聴者と聴覚障害者との間のコミュニケーションを想定する際，i) 障害者が聴者の音声の聞き取りに困難を生じる，ケースと，ii) 聴者が障害者の音声の聞き取りに困難を生じるケースのそれぞれを考える必要がある．ここでは，i) に焦点を当て，今後聴覚障害者による主観評価を行う上で必要となる音声データの収集を目的に，オンラインコミュニケーションで想定される，さまざまな品質低下音声とその音声強調後の音声を評価する．

ただし通信路の条件のコントロールは実環境では難しいことから，擬似的に構成した通信路のもとで音声品質を低下させた音声を作成する．

本章では，既存の代表的な音声強調アルゴリズムを用いて音声強調の実験を行い，今後の聴者・聴覚障害者による主観評価への利用の可否について論じる．

4.1 学習・評価データ

学習・評価データとして日本語読み上げ音声コーパス JNAS[58] を用いた．評価データは JNAS に含まれる ATR 音素バランス文のうち，話者 32 名（男女各 16 名）による B セットの B01 から B25 までの 25 文とした（計 800 文）．また，B26 から B50 までの 25 文をエンロール音声とした．

学習データは JNAS から上記の評価データを除いた 242 名による発話 31.2k 文 (67.3 時間) とした．なお，学習データと評価データは発話者，発話内容のオーバーラップはない（話者，発話内容について評価データはオープン）．

4.2 実験条件

4.2.1 概要

図 1 の実環境を模擬するよう，品質の低下した音声を上述の学習・評価データから作成した．

4.2.2 収録環境のシミュレーション

本実験では，妨害音として千葉大学 3 人会話コーパス [59] を用いた．コーパスは男性 3 名あるいは女性 3 名による自由発話で構成されており，本実験では無音区間を除去したあと，以下に述べる室内残響を加えた上で学習・評価データに対して目的音と妨害音が同性で，SNR が 20dB から 0dB の範囲となるように混合した．

室内残響は Pyroomacoustics[60] によるシミュレーションに基づいて作成した．収音環境として会議室を想定し，室内サイズ (4.8 m × 5.8 m × 2.5 m) とした．収録位置を (0, 5 m, 2.9 m, 1.0 m) とし，目的話者を収録位置に近接した (0.3 m, 2.9 m, 1.2 m) に設定した，妨害音は (3.0 m, 2.9 m, 1.2 m) を中心として (±0.85 m, ±1.0 m, ±0.25 m) の範囲でランダムに位置を変えた．また，室内の残響時間（RT69）は，0.2 秒から 1.0 秒までの間でランダムに変えた．

4.2.3 通信環境のシミュレーション

通信路を伝送する際の音声コーデックとして Opus[8] を用いた．学習に用いた音声データのサンプリング周波数は 16 kHz であり，これをいったん 8 kHz でリサンプリングして 16 kHz に戻した（以下では帯域 4 kHz とする）．得られた 2 種類の帯域の信号をそれぞれ伝送レート 6k bps(帯域 4 kHz), 20k bps (帯域 8 kHz) として，フレームサイズ 20 msec の固定ビットレート（Constant Bit Rate）でエンコードした．

通信路におけるパケット損失のシミュレーションには Gilbert-Elliott モデル [61] を用いた．Gilbert-Elliott モデルはネットワークの状態を表す 1 次マルコフモデルであり，Good（パケット損失なし）/Bad（パケット損失あり）の 2 状態間をそれぞれの遷移確率で遷移する確率モデルである．

4.3 ニューラルネットワーク

本実験では，SpeakerBeam[25]，CleanUNet[17]，E3Net[27] の各ネットワークを用いた．CleanUNet は Denoiser[16] の LSTM（Long Short Term Memory）を自己注意機構に置き換えたネットワークで，雑音抑圧向け

表 4: 実験 3: （パケット損失なし，SNR 20 dB，帯域 8 kHz）

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	**2.13**	**3.03**	**2.89**	3.28	2.04	2.64	3.03	**3.10**	3.19	3.63	2.59	2.55
SpeakerBeam	2.09	2.90	2.75	**3.34**	**2.19**	**2.68**	2.87	2.78	**3.56**	**3.76**	**4.02**	**3.38**
E3net	1.89	2.73	2.50	3.11	2.17	2.48	**3.20**	2.92	3.26	3.38	3.87	3.01
CleanUNet	1.84	2.65	2.46	3.16	2.12	2.47	2.89	2.81	2.67	3.46	3.60	2.92

表 5: 実験 3: （パケット損失なし，SNR 0 dB，帯域 8 kHz)

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	1.15	1.76	1.56	1.89	1.21	1.36	**2.96**	1.79	2.24	**3.50**	1.60	1.76
SpeakerBeam	**1.26**	**1.89**	**1.80**	**2.43**	**1.63**	**1.76**	2.56	1.86	**2.96**	3.07	**3.46**	**2.61**
E3net	1.16	1.52	1.53	2.13	1.57	1.56	2.92	1.78	2.34	2.20	2.69	1.69
CleanUNet	1.22	1.75	1.66	2.33	1.63	1.70	2.59	**1.94**	2.01	2.74	2.60	1.99

表 6: 実験 4: （パケット損失 10%， SNR 0 dB，帯域 4 kHz）

	intrusive						non-intrusive					
	PESQ	PESQ (NB)	ViSQOL	CompM			MOSNet	NISQA	PLCMOS	DNSMOS		
				CSIG	CBAK	COVRL				SIG	BAK	OVRL
degraded	1.06	0.63	1.38	-1.96	1.23	-0.60	2.73	1.45	1.70	2.37	1.79	1.42
SpeakerBeam	**1.14**	**1.53**	1.53	**1.97**	1.52	**1.45**	2.46	**1.61**	**2.48**	**2.56**	3.29	**2.13**
E3net	1.11	1.27	1.45	1.79	1.52	1.35	**2.77**	1.37	2.00	2.43	**3.40**	2.05
CleanUNet	1.13	1.37	**1.54**	1.81	**1.57**	1.38	2.43	1.54	1.62	2.50	2.60	1.82

表 7: 実験 5: HASQI (パケット損失なし，帯域 8 kHz)

	normal hearing			mild			moderate		
	妨害音なし	20 dB	0 dB	妨害音なし	20 dB	0 dB	妨害音なし	20 dB	0 dB
degraded	**0.577**	0.412	0.135	**0.667**	**0.584**	0.211	**0.711**	**0.644**	0.244
SpeakerBeam	0.481	**0.418**	**0.203**	0.569	0.529	**0.268**	0.612	0.578	**0.302**
E3net	0.413	0.346	0.138	0.497	0.429	0.169	0.536	0.471	0.188
CleanUNet	0.557	0.350	0.169	0.647	0.461	0.215	0.688	0.514	0.246

に構成されているが，本実験では，エンコーダーブロックに SpeakerBeam 同様，エンロール音声から取得した話者埋め込みを追加するように変更している．したがって，SpeakerBeam と E3Net は目的話者マスクを用いた目的話者抽出である一方，CleanUNet はマスクなしの目的音声抽出手法となる．

4.4 実験結果

実験結果は 8 つの指標で評価した．指標のうち，PESQ, NBPESQ, ViSQOL, CompM は侵入型（intrusive）でリファレンスを必要とするもの，CompM は文献 [43] の統合指標で目的音（CSIG），背景音（CBAK），総合評価（COVRL）の各指標を表す．MOSNet, NISQA, PLCMOS, DNSMOS P.835 は非侵入型でリファレンス不要のものである．また，DNSMOS P.835 は ITU-T P.835 の模擬であり，目的音（SIG）, 背景音（BAK），総合評価（OVRL）を表す（以下では簡便のため DNSMOS と記す）．

4.4.1 実験 1: パケット損失がある場合の実験結果

通信路にパケット損失がない場合の結果を表 1 に示す．このケースでは，通信路の両端において Opus でエンコード・デコードを行ったのみであり，符号化/複合化を除く品質低下の要因を設定していない．表を見ると，MOSNet の結果を除いて，音声強調処理を行わずに通信路を経由した音声（degraded）が最も指標値が高くなっている．CleanUNet は，PESQ, VISQOL, NISQA, DNSMOS（OVRL）で degraded に続く結果となった．CleanUNet は目的音声の埋め込み特徴量に対するマスクを推定するのではなく，直接的に目的音声の埋め込みを生成することから，単純な構成のネットワークであるといえ，品質低下の少ない目的音声の抽出では他の音声強調手法よりも性能が高い．

一方，データの 10 %にパケット損失を与えた場合の結果を表 2 に示す．MOSNet を除いた指標で SpeakerBeam が最も評価値が高くなった．

4.4.2 実験 2: 帯域制限がある場合の実験結果

音声に帯域制限をかけた場合（4 kHz）の実験結果を表 3 に示す．intrusive, non-intrusive のいずれの指標も，劣化音声の方が高い評価値となった．強調音声をいくつか聴取してみたところ，聴感上は失われた高域が復元したように聞き取れたものの，調波構造の再現に到っていないため各指標の数値が軒並み劣化音声と同程度となった．今回のモデルでは，高域についてはそのパワーをおおまかに推定できる程度で，スペクトルの微細構造の頑健な推定には到っていないということである．今回利用したモデルは，雑音抑圧か目的音抽出のためのモデル構造であり，各種指標値の改善には調波構造を考慮したモデルの採用が望ましいといえる．

4.4.3 実験 3: SNR を変えた場合の実験結果

SNR を 20 dB，0 dB としたときの結果をそれぞれ表 4, 5 に示す．20 dB では，degraded 音声が intrusive な指標（PESQ, NB-PESQ，ViSQOL）で高い評価値を，SpeakerBeam が non-intrusive な指標（PLCMOS, DNSMOS）で高い評価値となった．CompM では CSIG, CBAK，COVRL

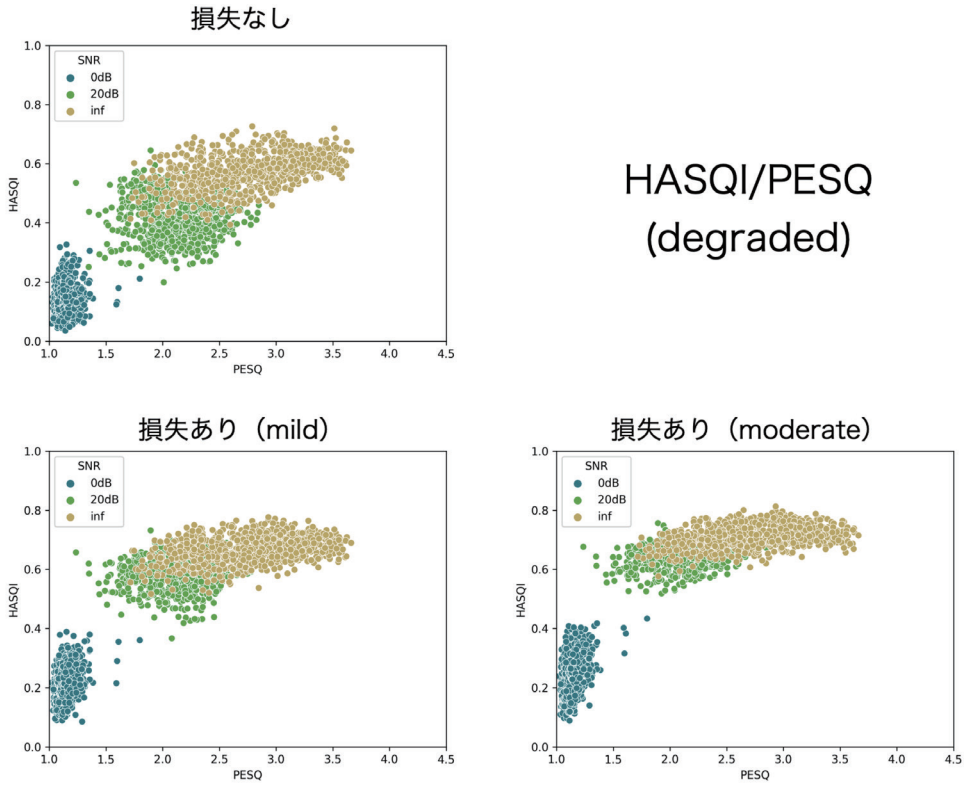

図 3: HASQI と PESQ の比較（degraded）

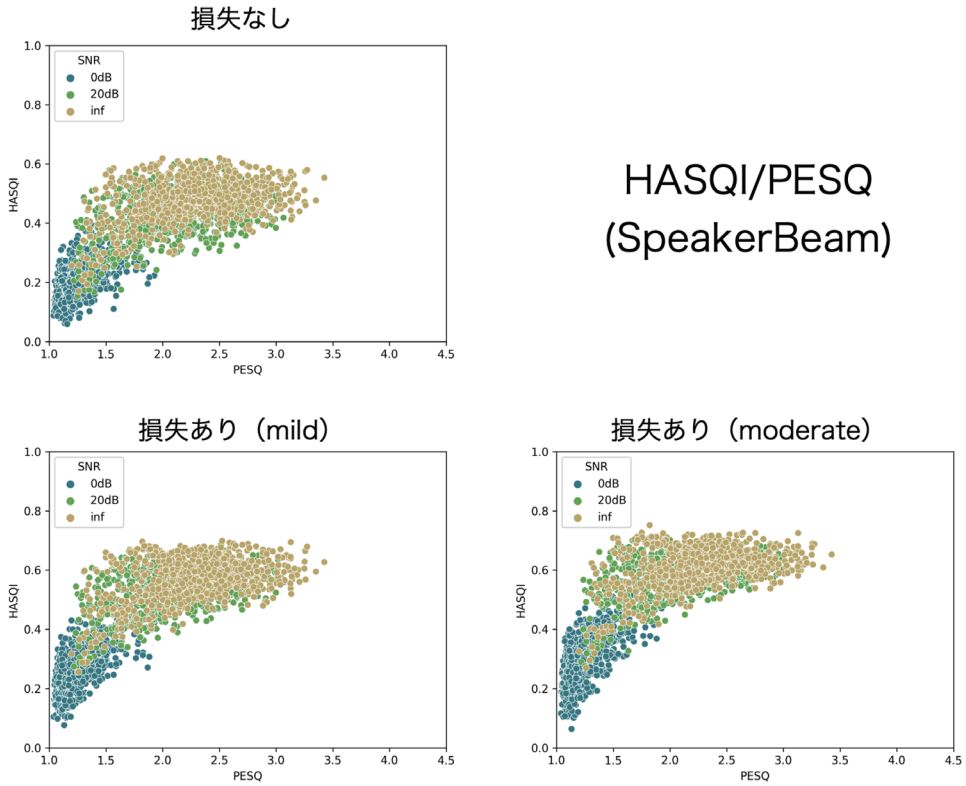

図 4: HASQI と PESQ の比較（SpeakerBeam）

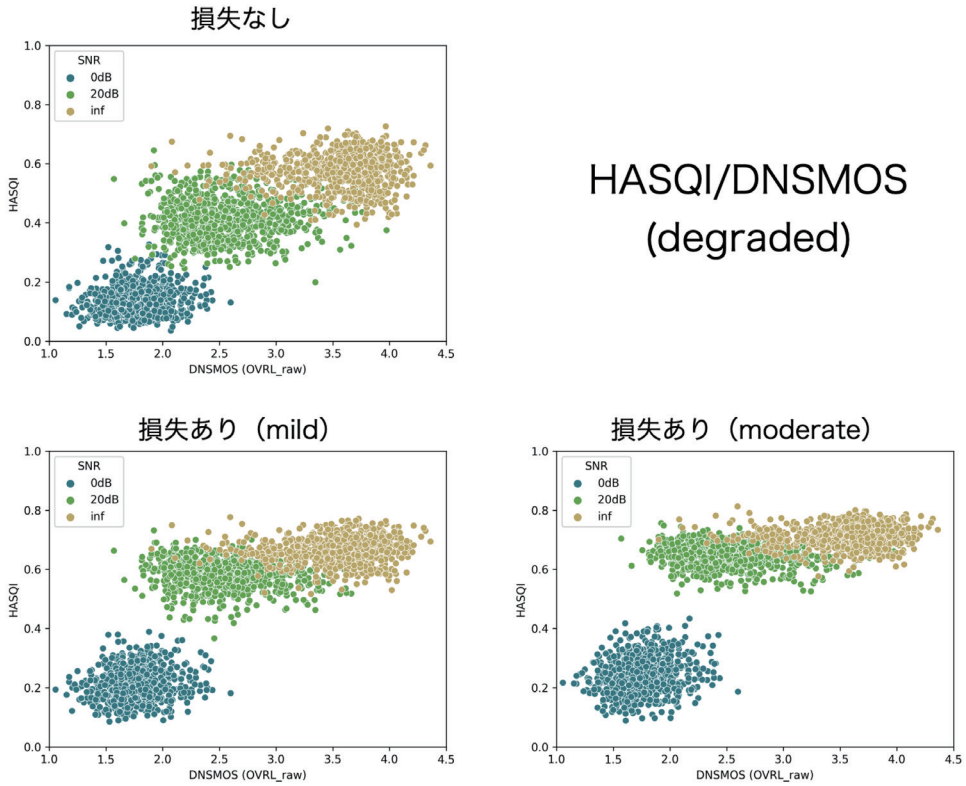

図 5: HASQI と DNSMOS の比較（degraded）

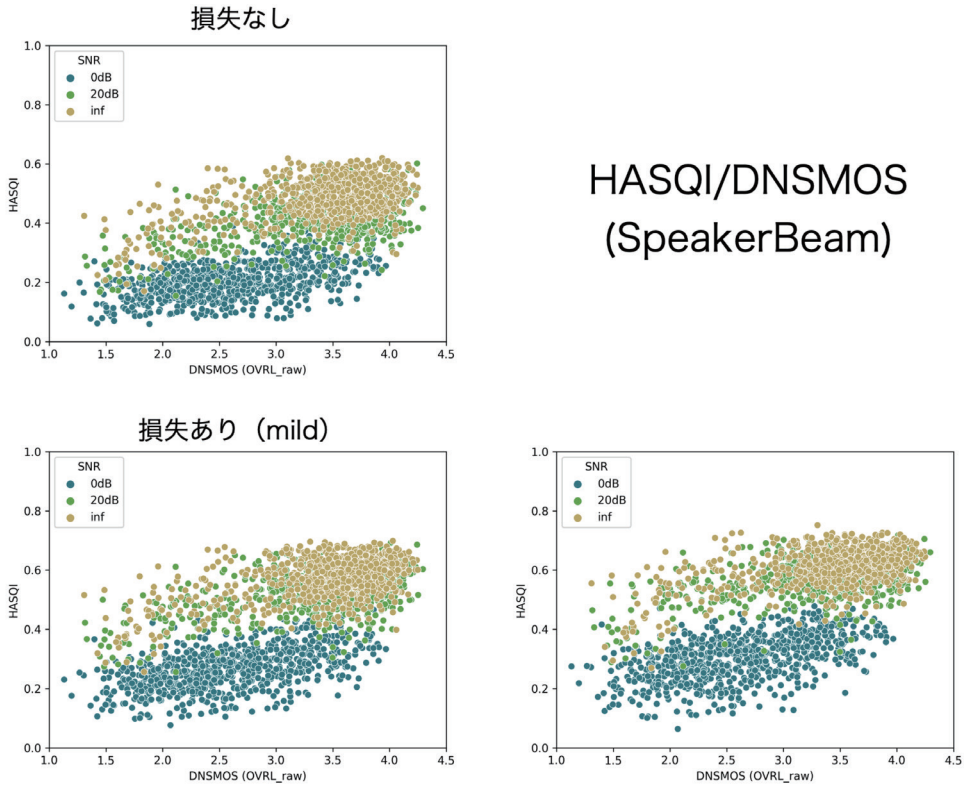

図 6: HASQI と DNSOMOS の比較（SpeakerBeam）

でわずかに SpeakerBeam の評価値が高いが，DNSMOS では BAK の評価値が degraded に対して大幅に改善し，SIG は小幅な改善にどどまった．

0 dB では，ほぼすべての指標が SpeakerBeam で最も高くなっている．DNSMOS の評価値を見ると，degraded に比べて音声強調の SIG の値は改善しておらず，BAK の改善により OVRL の値が改善している．いずれののモデルも背景となる話声の抑圧の効果が大きく，目的音声の品質の改善に到っていない．一方で CompM では CSIG，CBAK ともに改善が見られており，指標によって結果の解釈が大きく異なる可能性があるといえる．

4.4.4 実験 4: 最悪ケースの実験結果

最後に最悪ケースとしてパケット損失が 10 %，SNR が 0 dB，帯域 4 kHz としたときの実験結果を表 6 に示す．表を見ると SpeakerBeam が多くの指標値で最も高くなり，E3Net と CleanUNet が一部の指標で高い指標値となった．DNSMOS の評価値を見ると，SNR の実験結果同様，degraded に比べて SIG の改善はそれほど大きくなく，BAK の評価値の改善が総合評価に寄与している．一方，CompM では信号の著しい品質低下により degraded の CSIG が正確に推定されていない．CBAK についても強調手法との間に数値の差が見られず，筆者の聴取による印象とは異なる結果となった．本実験での妨害音は音声であるから，表にみられるような評価値の差が妥当なのかどうかは主観評価を行う必要がある．また，SpeakerBeam,E3Net のような埋め込みに対するマスクを生成する手法に比べ，直接的に目的信号を推定する CleanUNet を見ると，目的信号の品質は他の 2 つのモデルとあまり変わらないが，妨害音の侵入度の結果が他の手法よりも低くなっている．ただし，総合評価でみた場合は，intrusive な PESQ 等と同様，モデル間の評価値の差は大きくない．

4.4.5HASQI による評価

最後に実験 1 と 3 で得られた実験結果（妨害音の SNR を変えたケース）に対して，HASQI を使って評価した結果を表 7 に示す．PESQ や DNSMOS（OVRL）による結果では，妨害音なしから SNR が 20dB までは混合音そのものを評価した値が最も高くなっている（表 1，4）．一方，SNR が 0dB の場合は SpeakerBeam の値が高くなり，混合音よりも高品質であるという評価となった（表 5）．これらの傾向は，HASQI においても同様であった．

評価データの HASQI と PESQ，DNSMOS(OVRL) の値を散布図としてプロットした（図 3~6）．図をみると，いずれの SNR においても PESQ や DNSMOS の分散が大きくなっている一方で，HASQI は狭い範囲の値を取っていることがわかる．このことから，HASQI は SNR や音源の種類によらず，比較的安定した品質の推定を行なっているのではないかと考えられる．SpeakerBeam による抽出音に対する散布図からは，特に SNR が 0dB の場合，プロットが目的音のそれと比べて右側にシフトしており，HASQI 値の改善よりも PESQ や DNNMOS で評価した際の改善が大きいことがわかる．聴覚障害者を対象とした音声強調の場合，従来の指標は品質改善の目安になるといえるが，聴力損失は個人差が大きいため，HASQI（あるいは HASPI）に関連した指標に基づいた音声強調手法の開発が重要といえる．

5 おわりに

聴覚障害者との円滑な音声コミュニケーションの実現に向けた音声強調技術について述べた．本稿では，実環境を模擬して生成した低品質音声を既存のさまざまな評価基準で評価したが，評価基準によっては最良となる音声強調手法が異なるなどの結果となった．今後は，今回の実験結果に基づいてバランスのとれた音声データのセットを規定し，まずは聴者の主観評価と各種指標との相関について調べる予定である．その後，HASQI などの聴覚障害者の評価に適した客観指標との関連を調べたうえで，障害者による主観評価を行う予定である．

聴覚障害者への情報保障を目的とした音声強調技術では，例えば定期的な会合や会議のケースでは，目的話者に関する情報がまったく手に入らないという状況は考えにくく，事前に目的話者に関する音声を入手できる場合が多い．したがって，少量の目的話者音声を用いた音声強調のモデルの適応化が有効であると考えられる．また，強調された音声の品質と音声強調の処理時間（遅延）にはトレードオフがあることから，実用的な低遅延の音声強調処理の実現のためにも適応化手法の開発を進めることが肝要である．

謝辞

本研究の一部は JSPS 科研費 20H01716，21H00901，23H00493，23H00995 の助成を受けたものである．

参考文献

[1] (一財) ダイアローグ・コミュニケーション・ジャパン・ソサエティ：“コロナ禍でのオンラインコミュニケーションにおける聴覚障害者の課題・困難に関するアンケート” (2021).

[2] 厚生労働省：“平成２８年生活のしづらさなどに関する調査（全国在宅障害児・者等実態調査）結果” (2018). https://www.mhlw.go.jp/toukei/list/dl/seikatsu_chousa_c_h28.pdf.

[3] A. Kobayashi, K. Yasu, H. Nishizaki and N. Kitaoka: “Corpus design and automatic speech recognition for deaf and hard-of-hearing people”, 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), pp. 17–18 (2021).

[4] R. E. Mayer: “Multimedia Learning”, Cambridge University Press, 3rd edition (2020).

[5] 安東, 吉野, 志水, 板橋：“聴覚障害児における語音明瞭度、発音明瞭度並びに聴力レベルの相互関連性について”, 特殊教育学研究, **36**, 4, pp. 49–57 (1999).

[6] M. J. Osberger and N. S. Mcgarr: “Speech production characteristics of the hearing impaired”, Vol. 8 of Speech and Language, Elsevier, pp. 221–283 (1982).

[7] S. Möller, F. Köster, L. F. Gallardo and M. Wagner: “Comparison of transmission quality dimensions of narrowband, wideband, and super-wideband speech channels”, 2014 8th International Conference

on Signal Processing and Communication Systems (ICSPCS), pp. 1–6 (2014).

[8] J.-M. Valin, K. Vos and T. B. Terriberry: "Definition of the Opus Audio Codec", RFC 6716 (2012).

[9] K. Vos, K. V. Sørensen, S. Jensen and J.-M. Valin: "Voice coding with Opus", Audio Engineering Society Convention 135Audio Engineering Society (2013).

[10] J.-M. Valin, G. Maxwell, T. B. Terriberry and K. Vos: "High-quality, low-delay music coding in the Opus codec", arXiv preprint arXiv:1602.04845 (2016).

[11] J. Rodman: "The effect of bandwidth on speech intelligibility", Polycom inc., White paper (2003).

[12] S. Boll: "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, **27**, 2, pp. 113–120 (1979).

[13] X. Lu, Y. Tsao, S. Matsuda and C. Hori: "Speech enhancement based on deep denoising autoencoder.", Interspeech, Vol. 2013, pp. 436–440 (2013).

[14] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda and T. Nakatani: "Exploring multichannel features for denoising-autoencoder-based speech enhancement", 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 116–120 (2015).

[15] O. Ronneberger, P. Fischer and T. Brox: "U-Net: Convolutional networks for biomedical image segmentation", Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, pp. 234–241 (2015).

[16] A. Defossez, G. Synnaeve and Y. Adi: "Real time speech enhancement in the waveform domain", Interspeech 2020 (2020).

[17] Z. Kong, W. Ping, A. Dantrey and B. Catanzaro: "Speech denoising in the waveform domain with self-attention", 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7867–7871 (2022).

[18] X. Hao, X. Su, R. Horaud and X. Li: "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement", 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6633–6637 (2021).

[19] R. Yamamoto, E. Song and J.-M. Kim: "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199–6203 (2020).

[20] H. Yu and W.-P. Zhu: "Deep neural network based complex spectrogram reconstruction for speech bandwidth expansion", 2020 18th IEEE International New Circuits and Systems Conference (NEWCAS), pp. 110–113 (2020).

[21] L. Wen, L. Wang, Y. Zhang and K. P. Choi: "Multi-stage progressive audio bandwidth extension", 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 422–427 (2023).

[22] M. Mandel, O. Tal and Y. Adi: "AERO: audio super resolution in the spectral domain", 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023).

[23] Y. Luo and N. Mesgarani: "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation", IEEE/ACM Transactions on Audio, Speech, and Language Processing, **27**, 8, pp. 1256–1266 (2019).

[24] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong: "Attention is all you need in speech separation", 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–25 (2021).

[25] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget and J. Černocký: "SpeakerBeam: speaker aware neural network for target speaker extraction in speech mixtures", IEEE Journal of Selected Topics in Signal Processing, **13**, 4, pp. 800–814 (2019).

[26] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani and S. Araki: "Improving speaker discrimination of target speech extraction with time-domain speakerbeam", 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 691–695 (2020).

[27] M. Thakker, S. E. Eskimez, T. Yoshioka and H. Wang: "Fast Real-time Personalized Speech Enhancement: End-to-End Enhancement Network (E3Net) and Knowledge Distillation", Interspeech 2022, pp. 991–995 (2022).

[28] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen and X. Huang: "Personalized speech enhancement: New models and comprehensive evaluation", 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 356–360 (2022).

[29] E. Variani, X. Lei, E. McDermott, I. L. Moreno and J. Gonzalez-Dominguez: "Deep neural networks for small footprint text-dependent speaker verification", 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056 (2014).

[30] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur: "X-Vectors: Robust DNN embeddings for speaker recognition", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333 (2018).

[31] N. R. Koluguri, T. Park and B. Ginsburg: "TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context", 2022 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), pp. 8102–8106 (2022).

[32] E. Vincent, R. Gribonval and C. Fevotte: “Performance measurement in blind audio source separation”, IEEE Transactions on Audio, Speech, and Language Processing, **14**, 4, pp. 1462–1469 (2006).

[33] “ITU-T Rec. P. 800 methos for subjective determination of transmission quality” (1996).

[34] “ITU-T rec. P. 862 perceptual evaluation of speech quality (PESQ)” (2001).

[35] “ITU-T Rec. P. 862.2 wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs” (2008).

[36] A. Rix, J. Beerends, M. Hollier and A. Hekstra: “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”, 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 749–752 vol.2 (2001).

[37] “ITU-T Rec. P. 862.1 mapping function for transforming P.862 raw result scores to MOS-LQO” (2003).

[38] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy and M. Keyhl: “Perceptual objective listening quality assessment (POLQA)”, Journal of The Audio Engineering Society, **61**, 6, pp. 366–384 (2013).

[39] A. Hines, J. Skoglund, A. C. Kokaram and N. Harte: “ViSQOL: An objective speech quality model”, **2015**, 13 (2015).

[40] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O’Gorman and A. Hines: “ViSQOL v3: An open source production ready objective speech and audio metric”, 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2020).

[41] A. Hines, J. Skoglund, A. Kokaram and N. Harte: “Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA”, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3697–3701 (2013).

[42] “ITU-T Rec. P. 835 subjective test methodology for evaluating speech communication systems that include noise suppression algorithm” (2003).

[43] Y. Hu and P. C. Loizou: “Evaluation of objective quality measures for speech enhancement”, IEEE Transactions on Audio, Speech, and Language Processing, **16**, 1, pp. 229–238 (2008).

[44] C. K. Reddy, V. Gopal and R. Cutler: “DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors”, ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6493–6497 (2021).

[45] C. K. Reddy, V. Gopal and R. Cutler: “DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors”, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 886–890 (2022).

[46] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper and R. Aichner: “ICASSP 2022 deep noise suppression challenge”, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9271–9275 (2022).

[47] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh and R. Aichner: “Deep speech enhancement challenge at ICASSP 2023”, ICASSP (2023).

[48] “ITU-T Rec. P. 563 single-ended method for objective speech quality assessment in narrow-band telephony applications” (2004).

[49] G. Mittag, B. Naderi, A. Chehadi and S. Möller: “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets”, Interspeech 2021, pp. 2127–2131 (2021).

[50] Q. Wang, I. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika and A. Gruenstein: “VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition”, Interspeech 2020, pp. 2677–2681 (2020).

[51] L. Diener, M. Purin, S. Sootla, A. Saabas, R. Aichner and R. Cutler: “PLCMOS – A Data-Driven Non-Intrusive Metric for The Evaluation of Packet Loss Concealment Algorithms”, Interspeech 2023, pp. 2533–2537 (2023).

[52] 山田, 牧野, 北脇：“雑音抑圧音声の主観・客観品質評価法”, 日本音響学会誌, 第 67 巻, pp. 476–481 (2011).

[53] 坂本, 鈴木, 天野, 小澤, 近藤, 曽根：“親密度と音韻バランスを考慮した単語了解度試験用リストの構築”, 日本音響学会誌, **54**, 12, pp. 842–849 (1998).

[54] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen: “A short-time objective intelligibility measure for time-frequency weighted noisy speech”, 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4214–4217 (2010).

[55] J. M. Kates and K. H. Arehart: “The hearing-aid speech perception index (HASPI) version 2”, Speech Communication, **131**, pp. 35–46 (2021).

[56] J. M. Kates and K. H. Arehart: “The hearing-aid speech quality index (HASQI) version 2”, Journal of The Audio Engineering Society, **62**, pp. 99–117 (2014).

[57] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska and Z. Tu: “The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement:

Overview and outcomes", 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023).

[58] T. A. S. of Japan: "ASJ Japanese newspaper article sentences read speech corpus (JNAS)". https://doi.org/10.32130/src.JNAS.

[59] Y. Den and M. Enomoto: "Chiba three-party conversation corpus (Chiba3Party)". https://doi.org/10.32130/src.Chiba3Party.

[60] R. Scheibler, E. Bezzam and I. Dokmanić: "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 351–355 (2018).

[61] G. Hasslinger and O. Hohlfeld: "The Gilbert-Elliott model for packet loss in real time services on the internet", 14th GI/ITG Conference - Measurement, Modelling and Evalutation of Computer and Communication Systems, pp. 1–15 (2008).