社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声品質と音響環境の潜在変数で条件付けた Denoising Trainingによるノイズロバスト音声変換

五十嵐琢斗† 齋藤 佑樹† 関 健太郎† 高道慎之介† 山本 龍一††
橘 健太郎†† 猿渡 洋†

† 東京大学　〒113–8656 東京都文京区本郷 7-3-1
†† LINE ヤフー株式会社　〒102–8282 東京都千代田区紀尾井町 1-3

あらまし　本稿では，ノイジーな入力音声に対し，その音声品質と音響環境を表現する潜在変数の条件付けを行うノイズロバストな音声変換を提案する．先行研究では，クリーン音声に雑音や残響を人工的に付加することで得た疑似ノイジー音声のデータから noisy-to-clean の音声変換を学習する denoising training と呼ばれる手法により，既存のモデル構造に変更を加えることなく，ノイズロバストな音声変換を提案した．しかし，この手法は音声変換モデルが入力音声の多様な雑音や品質劣化を十分に学習できないため，推論時に未知ノイズで劣化した入力音声に対して，変換された音声の音韻や韻律が乱れる傾向にある．本研究では，入力音声の品質・雑音の多様性を解釈する機構を取り入れた音声変換を行うことを目的とし，denoising training の際に入力音声の音声品質および音響環境の潜在変数で条件付けたノイズロバストな音声変換の学習法を提案する．客観および主観評価により，提案手法により変換された音声の品質が従来手法と比較して向上することを示す．

キーワード　DNN 音声変換，ノイズロバスト，denoising training，any-to-any VC

Takuto IGARASHI†, Yuki SAITO†, Kentaro SEKI†, Shinnosuke TAKAMICHI†, Ryuichi YAMAMOTO††, Kentaro TACHIBANA††, and Hiroshi SARUWATARI†

† The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan
†† LY Corporation, 1–3 Kioicho, Chiyoda–ku, Tokyo, 102–8282, Japan

1. はじめに

音声変換 (Voice Conversion: VC) は，言語内容を保持したまま，ソース話者の音声特性をターゲット話者の音声特性に変換する技術である．VC は，映画の吹き替え [1]，パーソナライズされたテキスト音声合成 [2]，会話支援 [3] など，実世界において広く応用可能な技術である．近年，ディープラーニングの登場により，ニューラルネットワークベースの手法が VC 研究の中心を占めるようになり，変換音声の自然性と話者類似性の点で多大な改善がもたらされている [4].

現在の VC 研究では，雑音や残響のないクリーンな音声データの入力を前提とすることがほとんどだが，実世界で収録される音声には様々な環境ノイズが含まれる．このような劣化した音声が入力された場合，クリーン音声を前提とする従来の VC の性能は自然性と話者類似性の点で顕著な劣化が見られる [5].

この問題に対処するため，先行研究 [6] では，クリーン音声に雑音や残響を人工的に付加することで得た疑似ノイジー音声のデータから noisy-to-clean の VC を学習する denoising training と呼ばれる手法により，既存のモデル構造に変更を加えることなく，入力のノイジー音声に起因する変換音声の劣化を緩和可能なノイズロバスト VC を提案した．この先行研究は未知のソース話者の音声特性を任意の未知話者に変更する any-to-any VC の設定で調査を行い，複数の any-to-any VC モデルで denoising training を行い，既存の VC モデルのノイズロバスト性を比較検証した．その結果，Contrastive Predictive Coding (CPC) [7] という自己教師あり学習 (Self-Supervised Learning: SSL) 表現を用いた VC モデルである S2VC が話者類似性の観点で他モデルを卓越した．一方で，先行研究で提案された denoising training では，VC モデルは入力音声がどのようなノイズによりどの程度劣化を受けたかという情報を明示的に学習していないため，多様な品質劣化や雑音への汎化が困難である．これにより，推論時に未知ノイズで劣化した入力音声に対し，変換された音声の音韻や韻律が乱れる傾向にある．

本研究では，実世界においてユーザが VC モデルを利用する

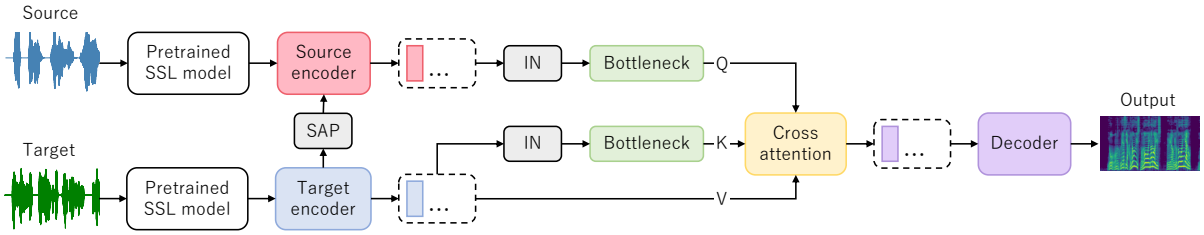

図 1: S2VC モデルの概略図．“SAP”は Self-attention pooling，“IN”は Instance Normalization の意．

上でクリーンなターゲット音声の確保が比較的容易である一方，ユーザ自身から発せられるソース音声はノイジーであるという前提に立ち，入力のノイジーなソース音声に対し，その音声品質と音響環境を表現する潜在変数で条件付けを行った S2VC における denoising training を提案する．本研究は，これらの潜在変数の条件付けによりノイジーなソース音声の品質・雑音の多様性を解釈する機構を取り入れた VC モデルを実現し，先行研究で提案された denoising training のノイズロバスト性をさらに向上させることを目的とする．具体的には，音声品質の潜在変数として，音声の自然性に関する Mean Opinion Score (MOS) 値を予測する UTMOS [8] の中間特徴量を用いる．この特徴量は，入力音声のノイズによる劣化程度を反映する傾向にある．また，音響環境の潜在変数として，音響シーン分類モデルの一種である PaSST [9] の中間特徴量を用いる．この特徴量は，入力音声に含まれる様々な環境で収録されたノイズの多様性を反映する傾向にある．したがって，学習中にこれら 2 つの潜在変数で条件付けることで，VC モデルはソース音声がどのようなノイズによりどの程度劣化したかを考慮して VC を行うように暗黙的に学習される．客観および主観評価により，提案手法により変換された音声の品質が従来手法と比較して向上することを示す．

2. 先 行 研 究

Huang らの先行研究 [6] では，既存の any-to-any VC モデルにノイズロバスト性を付与するために 2 つのアプローチを提案した．一つ目は，事前学習済み音声強調モデルと既存の any-to-any VC モデルの連結である．二つ目は，先述した End-to-end denoising training である．一般に前者のような sequential アプローチは，未知ノイズにより劣化された音声に対して音声強調を行う際に顕著に生じるアーティファクトが後段のタスクに影響を及ぼすため，後者のような End-to-end アプローチと比較して性能に限界がある [10]．そこで，本研究は End-to-end denoising training を拡張し，ノイズロバスト性の向上を目指す．本節では，先行研究の提案法の一つである S2VC [11] を用いた End-to-end denoising training の手法を説明する．

2.1 S2VC

S2VC は，FragmentVC [12] を改良した最新の any-to-any VC モデルである．S2VC の全体的なフレームワークを図 1 に示す．図に示すように，S2VC は，ソースとターゲットの SSL 特徴抽出器，ソースエンコーダ，ターゲットエンコーダ，Self-attention pooling，Instance Normalization，クロスアテンション機構，デコーダから構成される exemplar-based モデルである．推論時には，ソースエンコーダ由来の特徴量 Q とターゲットエンコーダ由来の特徴量 (K, V) がまず計算され，K と Q により計算されたアテンション重みと V の内積がデコーダに入力される．デコーダはこの情報をもとにメルスペクトログラムを生成する．学習時は，両エンコーダの入力とデコーダの再構成ターゲットに，共通の発話が使用される．このアーキテクチャでは，エンコーダは明示的な制約なしに，内容情報と話者情報を分離することを自動的に学習し，クロスアテンション機構は，類似した音素情報を持つターゲット特徴量にソース特徴量を合わせるように学習する．

S2VC の先行研究 [12] との差分は，ソース音声だけでなくターゲット音声の表現にも SSL 特徴量を採用したことと，Self-attention pooling および Instance Normalization を導入したことである．これにより先行研究である FragmentVC [12] の性能を卓越した．

2.2 End-to-end denoising training

先行研究 [6] では，クリーン音声とクリーン音声に雑音や残響を人工的に付加することで得た疑似ノイジー音声のデータから noisy-to-clean の VC を学習する denoising training と呼ばれる手法により，S2VC にノイズロバスト性を付与することに成功した．具体的には，VC モデルは，学習中にクリーンな音声または，ランダムな信号対雑音比 (Signal-to-Noise Ratio: SNR) で劣化させた疑似ノイジー音声のいずれかを一定確率で受け取り，Ground-Truth のクリーン音声と変換音声のメルスペクトログラムを比較して損失を計算する．この denoising training を S2VC に適用することで，S2VC はノイズ除去オートエンコーダのように学習され，ノイジー音声に対する変換音声の劣化を緩和可能なノイズロバスト VC が実現される．

2.3 従来法の問題点

先行研究で提案された denoising training では，疑似ノイジー音声が直接 VC モデルに入力されクリーン音声に復元されるように学習する過程において，VC モデルは入力音声がどのようなノイズによりどの程度劣化を受けたかという情報を明示的に学習していないため，多様な品質劣化や雑音への汎化が困難である．その結果，推論時に，様々な SNR や環境で収録された未知ノイズで劣化した入力音声に対し，変換された音声の音韻や韻律が乱れる傾向にある．

3. 提案法：条件付け denoising training

提案法では，any-to-any VC の一種である S2VC [11] をベースラインとする．図 1 に示すように，S2VC は未知のソース音声

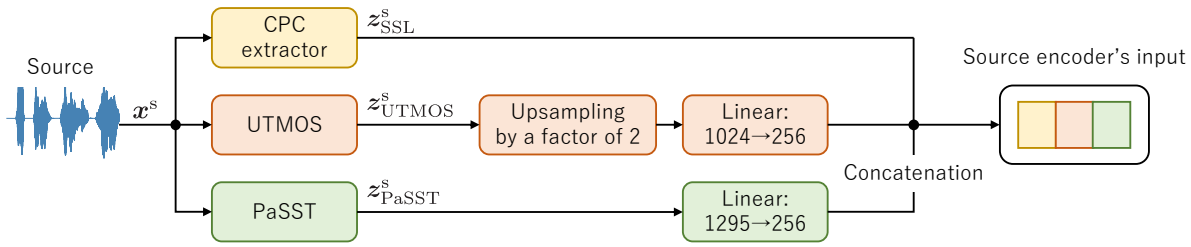

図 2: UTMOS と PaSST の中間表現の条件付けの概念図.

とターゲット音声の 2 つを入力にとり，ソース音声の発話内容を保持したまま，ソース話者の音声特性をターゲット話者の音声特性に変換する．本研究では，実用性を考慮し，ソース音声のみノイズにより劣化を受けていると想定する．この時，先行研究で提案された denoising training の問題点に対し，入力音声の品質・雑音の多様性を解釈する機構を取り入れた音声変換を行うことを目的とし，denoising training の際に，ソース音声の音声品質および音響環境を表現する潜在変数で条件付けた VC モデルを提案する．音声品質の潜在変数として音声の自然性に関する Mean Opinion Score (MOS) 値を予測する UTMOS [8] の中間特徴量を使用し，音響環境の潜在変数として音響シーン分類モデルの一種である PaSST [9] の中間特徴量を使用した．従って，学習用データセット内の任意のクリーンな音声を $\boldsymbol{x}^c$ とすると，提案手法の損失関数 $\mathcal{L}$ は以下のように定義される．

$$\boldsymbol{x}^s = \boldsymbol{x}^c + \boldsymbol{n} \tag{1}$$

$$\boldsymbol{x}^t = \boldsymbol{x}^c \tag{2}$$

$$\boldsymbol{z}^s_{SSL} = f_{SSL}\left(\boldsymbol{x}^s\right) \tag{3}$$

$$\boldsymbol{z}^s_{UTMOS} = f_{UTMOS}\left(\boldsymbol{x}^s\right) \tag{4}$$

$$\boldsymbol{z}^s_{PaSST} = f_{PaSST}\left(\boldsymbol{x}^s\right) \tag{5}$$

$$\boldsymbol{z}^t_{SSL} = f_{SSL}\left(\boldsymbol{x}^t\right) \tag{6}$$

$$\mathcal{L} = \left\| f_\theta\left(\boldsymbol{z}^s_{SSL}, \boldsymbol{z}^s_{UTMOS}, \boldsymbol{z}^s_{PaSST}, \boldsymbol{z}^t_{SSL}\right) - g_{mel}(\boldsymbol{x}^c) \right\| \tag{7}$$

ここで，$\boldsymbol{x}^s$，$\boldsymbol{x}^t$ はそれぞれソース，ターゲット音声で，$\boldsymbol{n}$ はソース音声に対して人工的に加えられるノイズである．$g_{mel}(\cdot)$ は入力音声からメルスペクトログラムを計算する関数である．また，$f_{SSL}(\cdot)$，$f_{UTMOS}(\cdot)$，$f_{PaSST}(\cdot)$ はそれぞれ入力音声に対し，SSL 特徴量 $\boldsymbol{z}_{SSL}$，UTMOS 中間表現 $\boldsymbol{z}_{UTMOS}$，PaSST 中間表現 $\boldsymbol{z}_{PaSST}$ を抽出するニューラルネットワークであり，$f_\theta(\cdot)$ はこれらの音声表現を入力とし，変換音声のメルスペクトログラムを出力する，学習可能パラメータ θ でパラメタライズされたニューラルネットワークである．

提案法では，$f_{UTMOS}(\cdot)$，$f_{PaSST}(\cdot)$ により抽出される UTMOS，PaSST の中間特徴量として，フレーム依存 (frame-wise) 特徴量とフレーム非依存 (average) 特徴量の 2 通りを使用する．前者は，フレーム単位で入力の音声表現を与えるので，ノイジーなソース音声を入力した場合，それに含まれるノイズの時間変化を反映することが期待される．後者は，フレーム依存 (frame-wise) 特徴量をフレーム方向に平均化することで，フレームごとに統一された音声表現を与える．したがって，フレーム非依存 (average) 特徴量は，ノイジーなソース音声を入力した場合，入力音声の全体的な品質を安定的に反映することが期待される．

4. 実験的評価

4.1 実 験 条 件

提案する条件付け denoising training の際には，Japanese Versatile Speech コーパス (JVS) [13] のパラレル発話データである voiceactress100 を 16 kHz にダウンサンプリングして使用した．voiceactress100 は日本人話者 100 名（男性話者 49 名，女性話者 51 名）の計 22 時間の発話データ（話者ごとに 100 発話）を含む．学習には話者 90 名分の発話データを使い，うち 4 話者 (“jvs087”から “jvs090”) を検証用とした．学習データに含まれていない 10 話者 (“jvs091”から “jvs100”) の発話を，提案法の設定である any-to-any VC の評価用の未知話者データセットとした．

学習・検証・評価用のノイジーなソース音声のデータセット作成のために，式 (2) にある通り，クリーン音声 $\boldsymbol{x}^c$ に対して，人工的なノイズ $\boldsymbol{n}$ が付加された．ここで，学習・検証用のノイジー音声の作成の際は，0 dB から 20 dB までの一様分布から乱択された SNR に基づき，様々な環境雑音が収録されている DEMAND [14] ノイズがクリーン音声に付加された．また，評価用のノイジー音声の作成の際は，5 dB および 15 dB の SNR に基づき，DEMAND とは異なる様々な環境雑音が収録されている WHAM!48kHz [15] を 16 kHz にダウンサンプリングしたノイズがクリーン音声に付加された．

UTMOS と PaSST の中間表現の条件付けは図 2 のように行われた．図 2 に示すように，入力音声の SSL 特徴量，UTMOS の中間表現，PaSST の中間表現のサイズを統一した後にそれらが特徴量次元で結合され，ソースエンコーダに入力された．ここで，UTMOS の中間表現は，SSL 特徴量と比較してフレーム次元数が半分だったため，1 フレーム分の表現を 2 フレーム分に拡張するアップサンプリングを行い，UTMOS 中間表現と SSL 特徴量のフレーム次元数を統一した．さらに特徴量次元を揃えるために，1024 次元から 256 次元に圧縮する線形射影層を挿入した．また，PaSST の中間表現は，入力音声を適当なフレーム区間でセグメント化し，それぞれに対して PaSST を適用することで，SSL 特徴量とフレーム次元数を揃えた．さらに特徴量次元を揃えるために，1295 次元から 256 次元に圧縮する線形射影層を挿入した．

提案法の学習は，式 (7) で示す損失関数に基づき行われた．なお，$f_{SSL}(\cdot)$，$f_{UTMOS}(\cdot)$，$f_{PaSST}(\cdot)$ は事前学習済みモデルを

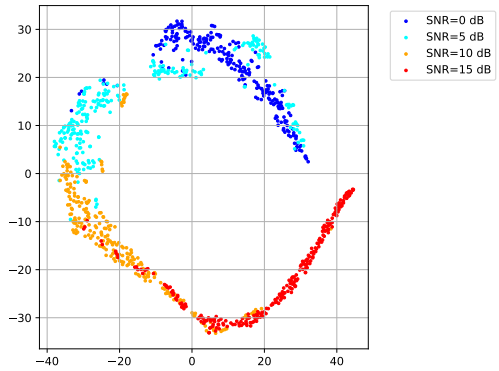

図 3: SNR= $\{0, 5, 10, 15\}$ dB の計 1000 個のノイジー音声の avUTMOS で用いた中間特徴量を t-SNE で 2 次元に射影し可視化した散布図．ノイズは DEMAND の “PRESTO”を使用した．

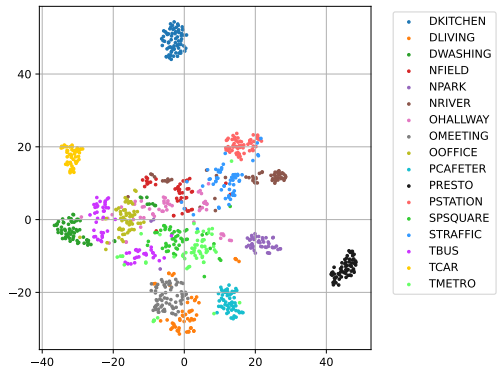

図 4: 17 種類の DEMAND ノイズにより劣化された計 1000 個のノイジー音声の avPaSST で用いた中間特徴量を t-SNE で 2 次元に射影し可視化した散布図．SNR は 5 dB とした．

使用しており，学習時は固定された．ここで，$f_{\mathrm{SSL}}(\cdot)$ は，先行研究 [6] と同様に CPC [7] による特徴抽出器を使用した．また，$f_{\mathrm{UTMOS}}(\cdot)$ は UTMOS [8] の公式実装[1]を使用し，$f_{\mathrm{PaSST}}(\cdot)$ は PaSST [9] の公式実装[2]を使用した．学習時の最適化の実装は，先行研究 [6] の公式実装[3]を参考に行い，最適化パラメータおよび最適化スケジューラはこの実装に準拠した．学習は，validation loss が完全に収束した時点で停止した．従来法と提案法それぞれでモデルの学習，評価を行った．

4.2 客観評価

評価話者 10 名の計 1000 発話から，話者の異なるソース発話とターゲット発話 250 対をランダムに選択し，従来法および提案法で VC を行った．ここで，提案法を条件付けの仕方に応じてそれぞれ “avUTMOS-avPaSST”，“avUTMOS-fwPaSST”，“fwUTMOS-avPaSST”，“fwUTMOS-fwPaSST”と命名した．ここで，“av”は “average”の略称でフレーム非依存特徴量を示し，“fw”は “frame-wise”の略称でフレーム依存特徴量を示す．

変換された音声の自然性を評価するために，自動音声認識 (Automatic Speech Recognition: ASR) システムによる文字誤り率 (Character Error Rate: CER) を使用した．CER 計算時の ASR モデルは HuggingFace 上で公開されている ReazonSpeech 事前学習済みモデル[4]を用いた．CER が小さいほど，変換音声がソース音声の言語内容をよく保っており，CER が大きいほど，変換音声が認識できないほどひどく歪んでいると考えられるため，CER は変換音声の自然性を反映した指標だと考えられる．

また，変換音声とターゲット話者の話者性類似度を測定するために，話者埋め込み cos 類似度を使用した．この指標は，変換音声とターゲット話者による音声を入力とし，2 つの固定次元埋め込みベクトルを生成した後，それらの cos 類似度を取ることにより計算された．埋め込みベクトルの生成には，HuggingFace 上で公開されている WavLM [16] ベースの x-vector [17] の事前学習済みモデルを[5]使用した．このベクトルは，入力音声の話者性を反映するため，話者埋め込み cos 類似度が大きいほど，変換音声がターゲット話者に類似していると考えられる．

表 1 (a)， (b) に従来法および提案法の VC モデルによる 250 対の評価用音声に対する変換音声の CER および話者埋め込み cos 類似度の平均を示す．VC モデルを UTMOS と PaSST 由来の中間特徴量で条件付けることで，条件付けを行わない従来法より CER，話者埋め込み cos 類似度の観点で VC 性能が向上したことがわかる．図 3， 4 に avUTMOS で用いた中間特徴量の SNR 依存性，avPaSST で用いた中間特徴量の DEMAND ノイズの種類の依存性を t-SNE [18] で可視化した散布図を示す．図 3， 4 に示すように，UTMOS 中間特徴量は SNR に関してク

表 1: 従来法および提案法の VC モデルによる 250 個の変換音声の CER および話者埋め込み cos 類似度の平均．条件付けを行わない従来法の結果を最上段に示す．“cos.” は話者埋め込み cos 類似度の意．

(a) ソース音声の SNR が 5 dB の場合

Method		CER	cos.
UTMOS	PaSST		
-	-	43.5%	0.925
av	av	38.0%	0.928
av	fw	32.1%	0.931
fw	av	35.0%	0.935
fw	fw	**30.2%**	**0.936**

(b) ソース音声の SNR が 15 dB の場合

Method		CER	cos.
UTMOS	PaSST		
-	-	17.1%	0.919
av	av	13.3%	0.923
av	fw	11.5%	0.929
fw	av	10.4%	0.931
fw	fw	**9.9%**	**0.932**

(注1)：https://github.com/sarulab-speech/UTMOS22
(注2)：https://github.com/kkoutini/passt_hear21
(注3)：https://github.com/cyhuang-tw/robust-vc
(注4)：https://huggingface.co/reazon-research/reazonspeech-espnet-next
(注5)：https://huggingface.co/microsoft/wavlm-base-sv

ラスターを形成しており，PaSST 中間特徴量は収録されたノイズのドメインに関してクラスターを形成している．これらの結果は，UTMOS 中間特徴量と PaSST 中間特徴量がそれぞれノイジー音声の音声品質と音響環境を反映した潜在変数であることを示唆する．したがって，これらの特徴量を条件付けることにより，VC モデルはソース音声がどのようなノイズによりどの程度劣化したかを考慮して VC を行うことができ，結果としてノイズロバスト性の向上に繋がったと考えられる．また，フレーム依存 (frame-wise) の特徴量を条件付ける手法が，フレーム非依存 (average) の特徴量を条件付ける手法より CER，話者埋め込み cos 類似度の観点で VC 性能が卓越した．この結果は，フレーム依存 (frame-wise) 特徴量の条件付けにより，VC モデルが入力音声内のノイズの非定常性をうまく解釈して VC を行ったことを示唆する．

4.3 主観評価

Lancers[6] でのクラウドソーシングにより，提案法および従来法で変換された音声に対し，自然性・話者類似性の主観評価を行った．主観評価に用いた変換音声は 4.2 節で評価の対象としたものと同一であり，ソース音声の SNR が 5 dB と 15 dB の 2 種類の評価用データそれぞれに対して，条件付けを行わない従来法と 4 種類の提案法により変換された各 250 個の変換音声を 1 つのデータセットにまとめた．したがって，評価セットはソース音声の SNR に応じて 2 種類あり，各セットは $250 \times 5 = 1250$ 個の変換音声を含む．

自然性および話者類似性の主観評価の際，SNR に応じた 2 種類の評価セットそれぞれにおいて，評価者はそれぞれ 100 名であった．自然性を評価する際は，各評価者はランダムに 20 個の変換音声サンプルを聴き，それらの自然性を 1（非常に悪い）から 5（非常に良い）の 5 段階の MOS 値で評価した．話者類似性を評価する際は，各評価者はランダムに 20 個の変換音声とそれに対応する Ground-truth のターゲット音声のサンプル対を聴き，それらの話者類似性を 1（非常に悪い）から 5（非常に良い）の 5 段階の MOS 値で評価した．

表 2 (a)， (b) に自然性・話者類似性の主観評価結果を示す．この結果は，4.2 節と同様，フレーム依存 (frame-wise) の特徴量の条件付けが変換音声の自然性および話者類似性の観点で VC 性能の向上に有効であることを示す．一方で，フレーム非依存 (average) の特徴量の条件付けによる性能の向上は認められなかった．これは，4.2 節の結果に反する．音声の自然性における人間の評価は，発話内容の歪みやネイティブらしさなど様々な要因により総合的に決定され，話者類似性における人間の評価は，発話リズムやイントネーションなど様々な要因により総合的に決定されるため，主観評価結果と客観評価結果は必ずしも一致しない．フレームごとに統一された音声表現を与えるフレーム非依存 (average) の特徴量の条件付けは，CER や話者埋め込み cos 類似度と異なり定量化が困難な人間の評価を改善するには情報量が不足していたと考えられる．

(注6)：https://www.lancers.jp/

表 2: 従来法および提案法の VC モデルによる変換音声の自然性および話者類似性の主観評価結果．条件付けを行わない従来法の結果を最上段に示す．“Nat.” は自然性，“Sim.” は話者類似性の意． 数値の後の ( ) は 95% 信頼区間を示す．

(a) ソース音声の SNR が 5 dB の場合

Method		Nat.	Sim.
UTMOS	PaSST		
-	-	2.17 (±0.11)	2.33 (±0.11)
av	av	2.17 (±0.10)	2.31 (±0.11)
av	fw	2.38 (±0.11)	2.42 (±0.11)
fw	av	2.27 (±0.11)	**2.49** (±0.11)
fw	fw	**2.48** (±0.11)	2.46 (±0.11)

(b) ソース音声の SNR が 15 dB の場合

Method		Nat.	Sim.
UTMOS	PaSST		
-	-	2.94 (±0.12)	2.31 (±0.11)
av	av	2.92 (±0.12)	2.36 (±0.11)
av	fw	3.09 (±0.11)	2.41 (±0.11)
fw	av	3.15 (±0.11)	2.44 (±0.10)
fw	fw	**3.23** (±0.12)	**2.48** (±0.10)

5. おわりに

本稿では，入力のノイジーなソース音声に対し，その音声品質と音響環境の潜在変数により条件付けを行う VC モデルを提案し，客観および主観的評価により提案法の有効性を検証した．評価の結果，フレーム依存 (frame-wise) 特徴量を条件付けることで変換音声の自然性および話者類似性が大きく向上することが明らかになった．今後は，本研究で考慮した加法性ノイズ以外にも残響，帯域制限等を含む様々な音質劣化要因を考慮した VC モデルを検討する．また，スマートフォン等の実機で収録されたノイジー音声に対する提案法のノイズロバスト性を検証し，本研究の実世界への応用を目指す．

謝辞: 本研究は，LINE ヤフー株式会社と東京大学 猿渡・高道研究室の共同プロジェクトとして実施したものです．

文　献

[1] F. M. Mukhneri, I.Wijayanto and S. Hadiyoso: “Voice conversion for dubbing using linear predictive coding and hidden markov model”, Journal of Southwest Jiaotong University, **55**, 4 (2020).

[2] A. Kain and M. W. Macon: “Spectral voice conversion for text-to-speech synthesis”, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98 (Cat. No. 98CH36181), Vol. 1IEEE, pp. 285–288 (1998).

[3] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano: “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech”, Speech communication, **54**, 1, pp. 134–146 (2012).

[4] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling and T. Toda: “Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion”, arXiv preprint arXiv:2008.12527 (2020).

[5] T.-h. Huang, J.-h. Lin and H.-y. Lee: “How far are we from robust voice conversion: a survey”, 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 514–521 (2021).

[6] C.-Y. Huang, K.-W. Chang and H.-Y. Lee: "Toward degradation-robust voice conversion", ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 6777–6781 (2022).

[7] A. v. d. Oord, Y. Li and O. Vinyals: "Representation learning with contrastive predictive coding", arXiv preprint arXiv:1807.03748 (2018).

[8] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi and H. Saruwatari: "Utmos: Utokyo-sarulab system for voicemos challenge 2022", arXiv preprint arXiv:2204.02152 (2022).

[9] K. Koutini, J. Schlüter, H. Eghbal-Zadeh and G. Widmer: "Efficient training of audio transformers with patchout", arXiv preprint arXiv:2110.05069 (2021).

[10] C. Zhang, Y. Ren, X. Tan, J. Liu, K. Zhang, T. Qin, S. Zhao and T.-Y. Liu: "Denoispeech: Denoising text to speech with frame-level noise modeling", ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 7063–7067 (2021).

[11] J.-h. Lin, Y. Y. Lin, C.-M. Chien and H.-y. Lee: "S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations", arXiv preprint arXiv:2104.02901 (2021).

[12] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee and L.-s. Lee: "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention", ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 5939–5943 (2021).

[13] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari: "Jsut and jvs: Free japanese voice corpora for accelerating speech synthesis research", Acoustical Science and Technology, **41**, 5, pp. 761–768 (2020).

[14] J. Thiemann, N. Ito and E. Vincent: "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments", Proc. Meetings Acoust, pp. 1–6 (2013).

[15] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow and J. Le Roux: "Wham!: Extending speech separation to noisy environments", Proc. Interspeech (2019).

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al.: "Wavlm: Large-scale self-supervised pre-training for full stack speech processing", IEEE Journal of Selected Topics in Signal Processing, **16**, 6, pp. 1505–1518 (2022).

[17] 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)IEEE, pp. 5329–5333 (2018).

[18] L. Van der Maaten and G. Hinton: "Visualizing data using t-sne.", Journal of machine learning research, **9**, 11 (2008).