音環境に適応するテキスト音声合成のための一人称視点コーパス構築

武 伯寒[1]　高道 慎之介[1]　関 健太郎[1]　坂東 宜昭[2]　猿渡 洋[1]

概要：テキスト音声合成 (text-to-speech, TTS) モデルが環境雑音や対話相手の発話といった聴取者の音環境を考慮することは, 多様な音環境に適応して自然な対話を行う対話ロボットなどへの応用に必要である. 本研究では音環境に適応できる TTS モデルを構築を目指す. まず, 話者の発話音声・受聴音・一人称視点映像を時刻同期して収録する自発対話コーパスを構築した. 次に構築したコーパスを用いて学習した TTS モデルと比較実験により, TTS モデルにおける音環境を考慮する機構の必要性について検討した.

1. はじめに

対話ロボットといった音声対話システムの運用シーンとして, 周囲に環境雑音が存在する実環境下での人間との会話が想定される. 実環境における人間同士の音声コミュニケーションでは, その場に応じた自然で聞き取りやすい発話を生成するために, 人間自身が見聞きして得た視聴覚情報をもとに, 周囲の環境雑音や会話相手同士の物理的な関係といった環境要素を考慮しそれらに適応する [1], [2]. 本稿ではこれらの環境要素をまとめて音環境と呼称する. 対話ロボットについても, 異なる環境に応じて人間にとって自然な発話の仕方が存在することが報告されており [3], 音声対話システムが会話において音環境へ適応することは, 自然で円滑な音声コミュニケーションが実現するために必要であると考えられる.

音環境を反映した人間同士の音声コミュニケーションで現れる現象として, 人間は環境雑音の中で自身の発話を自ら耳で聴き, その結果をフィードバックして以後の発話の仕方を脳で調整していることが知られており, これは speech chain という枠組みの中で説明される [4]. 例えば, 雑音環境の中で, 人間は聞き手にとって明瞭で円滑に発話を伝達するために, 雑音環境の中で無意識に声を張り上げるなどして発話音声の特徴量を変化させる Lombard 効果が現れることが知られている [5]. Lombard 効果が現れた発話は

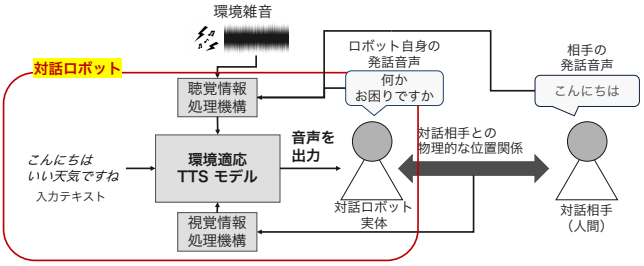

図 1　環境適応 TTS モデルを用いて発話を生成する対話ロボットの例を示すブロック図.

Lombard 発話と呼ばれ, そのパワーや基本周波数, スペクトル特性といった特徴量が変化することが報告されている [6]. また, 会話相手との物理的な距離関係 [7] や, 会話相手の発話 [8] に応じて発話音声の特徴量が変じることも知られている. このような人間の音環境への適応能力を模倣する機構を対話ロボットに実装できれば, 実環境において聞き手にとって自然で明瞭な発話を生成することが期待される.

対話ロボットにおける発話音声の生成方法として, 本研究ではテキスト音声合成 (text-to-speech, TTS) を用いる方法を想定する [9]. 深層ニューラルネットワーク (deep neural network, DNN) を用いた TTS モデル [10] とコーパスの拡充により, 静音環境での再生を想定した朗読音声については人間と遜色ない自然な音声の生成が可能となっている. また, 音環境を入力として音声生成に反映する TTS モデルの検討も行われている. このような TTS モデルを本稿では 環境適応 TTS モデルと呼称する. 環境適応 TTS モデルは**図 1** のように, 対話においてテキストのみではなく, 周囲の環境雑音や対話相手の発話音声, そして物理的な関係といった音環境を, その場に応じた自然な発話音声の

[1] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology （AIST）, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.

生成に利用する．この環境適応 TTS モデルに関連して, 環境雑音 [11] や, 対話相手の顔を捉えた視覚情報 [12] を考慮して適応する DNN に基づくモデルが研究されている．しかし, こうした環境適応 TTS モデルを学習するために必要な, 一人称視点の視聴覚情報から得られる音環境を発話と同時に収録した TTS コーパスが不足しており, 環境雑音に関する環境適応 TTS モデルについては擬似データを使用した学習も行われているのが現状である．また, これまでの環境適応 TTS モデルでは朗読音声の生成に目的を限定している．人間同士の音声コミュニケーションを模倣するためには, 環境適応 TTS モデルの学習において自発的な会話を考慮することも重要である．

本研究では, 環境適応 TTS モデルに必要な, 一人称視点での視聴覚情報を発話と同時に収録した自発対話コーパスの構築を行う．また, 構築したコーパスを用いて DNN に基づく環境適応 TTS モデルを作成し, その比較評価を行う．

2. 関連研究

Lombard 効果を模擬して実環境下で聞き手にとって明瞭な発話を生成する手法として, スペクトルの整形と Dynamic range compression による発話音声の加工で実現する手法 [13] や, 発話音声の音量と基本周波数といった音声特徴量を加工して実現する手法 [14] が提案されている．これらの手法は信号処理に基づく加工を利用した手法であり, 加工後音声の自然性を損なう．そのため, 自然な発話音声が合成可能な DNN に基づく TTS モデルに Lombard 効果を模擬する機能を搭載する手法が提案されている．例えば, 既に学習された DNN に基づく TTS モデルを Lombard 発話のデータでファインチューニングする手法 [15] や, 多話者 TTS モデルを応用した手法 [16] が提案されている．また, TTS モデルを発話音声の平均基本周波数といった大域的な音声・韻律特徴量で制御する手法 [17] が提案されており, この手法を用いて信号処理に基づく加工と比較して自然な音声が生成可能であることが示されている．さらに, 音声認識技術と TTS を組み合わせて動的な雑音環境に適応した発話を生成する手法 [11] が提案されている．以上の手法では評価時の音声は実環境での評価者の聴取音と乖離しており, 空間の残響特性や環境雑音の耳への伝達特性を考慮できていない．また, これらの手法は研究対象を朗読音声に限定しており, 自発対話についての研究は行われていない．これについては, 自発的な発話や会話においても環境雑音下では Lombard 効果が現れることが指摘されている [18] ことから, 対話ロボットによる自然な音声コミュニケーションを実現するために, 自発会話シーンにおける環境雑音に適応する TTS モデルの構築が必要である．

周囲の環境雑音を考慮して明瞭な Lombard 朗読発話を生成するための TTS コーパスは, Hurricane Challenge というコンペティションにおいて提供されている [19] が, このコーパスにおいては収録時の雑音信号として他者の発話音声や擬似雑音が用いられている．また発話音声の他に収録時に再生した雑音信号とそのパワーが与えられているのみである．Lombard 発話が収録されたコーパスはその他に Lombard Grid コーパス [20] が存在し, Lombard 発話の生成に用いられている．しかし, いずれも朗読音声の収録であり, 自発的な会話の環境雑音への適応を捉える TTS コーパスの構築は進められていない．

自発的な会話を収録したコーパスについては, 自然な会話を実現する音声対話システムの作成のために進められており, DNN に基づく TTS モデルの学習に利用されている [21]．実環境下で自発的な会話を音環境と同時に収録したコーパスとしては, 日本語日常会話コーパス (CEJC) [22] が TTS モデルの学習に用いられている．しかし, CEJC にて収録された発話時の環境雑音や視覚映像といった音環境は三人称視点で収録されている．音環境への適応を捉えるためには、 一人称視点で見聞きしている情報を収録するコーパスの構築を進める必要がある．実環境下の音声コミュニケーションにおける一人称視点で見聞きした情報を収録したデータセットとして, EgoCom [23] が整備されている．このデータセットでは外部雑音が極力含まれないような発話音声の収録が行われておらず, TTS モデルの学習に直接用いることは難しい。 また, EasyCom [24] というデータセットでは一部の話者に対して外部雑音が含まれる発話音声の収録が行われている．しかし, 被験者の置かれた雑音環境は多様ではないため, このデータセットを通して音環境のうち環境雑音への発話の適応をモデリングすることは難しい．

3. 自発対話コーパスの構築

本節では環境適応 TTS モデルに用いる, 一人称視点情報付きの自発対話コーパス SaSLaW (**S**o, what **a**re you **S**peaking, **L**istening, **a**nd **W**atching?) の構築について述べる．

3.1 自発対話コーパス SaSLaW の概要

SaSLaW では, 人間の自発的な音声コミュニケーションにおける音環境への適応を捉える．そのために, 実環境を模擬した雑音環境下において, 被験者である 2 人が向かい合って自発的な対話を行うシーンを収録する．被験者 2 人の会話の際に収録する情報を**図 2** に示す．SaSLaW における会話の収録では, TTS モデルの構築に必要な被験者の発話音声の収録を行う．その収録と時刻を同期して, 一人称視点での受聴音と映像を収録する．これらの情報を用いることにより, TTS モデルが発話を生成する際に, 人間のように聴覚・視覚情報から得られる音環境を考慮できるようになることが期待される．

SaSLaW と既存の TTS コーパスやデータセットと比較

図 2 自発対話コーパス SaSLaW において収録する情報を示した図. 特に収録する情報を**青の太字**で示した. これらの情報は会話時に時刻を同期して収録される.

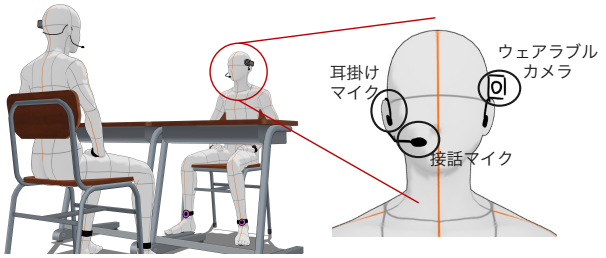

図 3 会話収録における被験者 2 人の会話時の様子と, 頭部に装着する収録機器のイメージ図.

した表を**表 1** に示す. このように, SaSLaW は TTS のための自発対話コーパスとして, 雑音が少ないクリーンな発話音声の他に, 一人称視点での聴覚・視覚情報にアクセスできるため, 人間を模倣して環境に適応できる機構の構築に利用できることが利点である.

3.2 会話収録環境のセットアップとその手順

被験者 2 人の会話の収録は全て単一の室内（以下, 収録室と呼ぶ）で行う. 収録室の室内において, 被験者 2 人は机を挟んで向かい合って座った状態で会話を行い, その様子を収録される. 被験者 2 人の間の距離は 1.5 m から 3 m の範囲で厳密に制御せず設定する. 会話の収録時に, 被験者はそれぞれ接話マイク (SHURE PGA31-TQG ワイヤレス用ヘッドセットマイク[*1]), 耳掛けバイノーラルマイク (Sound Professional MS-EHB-2[*2]), ウェアラブルカメラ (Ordro EP8[*3]) を**図 3** に示すように装着する. 接話マイクは環境雑音が極力混合しない発話音声の収録に, 耳掛けバイノーラルマイクは人間が耳で聴いている音である一人称視点の受聴音の収録に, そしてウェアラブルカメラは一人称視点での視覚映像情報の収録に用いる. 接話マイクと耳掛けバイノーラルマイクのサンプリング周波数は 44.1 kHz, ウェアラブルカメラのフレームレートは 30 fps で収録する. 収録の際, 環境雑音ごとの発話音量の違いが想定されるため, 話者ごとに全ての収録を通して接話マイクと耳掛けバイノーラルマイクのゲインを固定する. なお, 被験者によっては体格の問題から 3 つの装備の装着に支障がある

[*1] https://shure.com/ja-JP/products/microphones/pga31
[*2] https://soundprofessionals.com/product/MS-EHB-2/
[*3] https://ordro.online/en-jp/products/camcamcorder-ep8

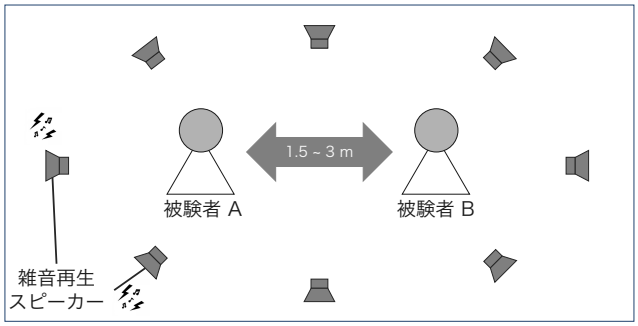

図 4 会話収録における収録室内の被験者 2 人と, 雑音を再生するスピーカーの配置イメージ.

場合がある. その際は聴覚情報を重視し, 接話マイクと耳掛けバイノーラルマイクを用いた発話音声と受聴音の収録を行う.

環境雑音が存在する様々な実環境の模擬のために, **図 4** に示されるように話者の周囲を覆うように 8 台のスピーカーを配置し, それぞれのスピーカーから同一実環境雑音の別区間音源を再生して実環境の雑音を模擬する. このスピーカーの配置により, 被験者は実際の環境雑音に近い, 特定方位からではない四方八方から到来する拡散性雑音が模擬できる. 実環境雑音データは DEMAND [27] の一部の種類の環境音を使用する. 一定回数の会話の収録ごとに, スピーカーから再生される雑音の種類とパワーを変更する. 会話を収録する前に, 被験者 2 人の中央において騒音計 (サンワサプライ デジタル騒音計 CHE-SD1[*4]) を用いて環境雑音の騒音値を dB 単位で測定する.

被験者 2 人は 1 つの会話を収録する前に, 収録する会話のテーマと, それに応じた 2 人それぞれの役と, その役が会話を通して達成するタスクを指示される. その例を**表 2** に示す. 収録の際には, 被験者 2 人は指示された内容に沿って, 即興で 5 から 8 ターンの会話を行う.

収録した発話音声を TTS コーパスとして整備するために, 自動での発話区間検出 [28] と whisper による自動書き起こし [29], そして手動での修正により収録音声の各発話への分割と発話テキストのアノテーションを行う.

3.3 評価用のデータ収録・評価用音声の作成手順

本研究で構築した SaSLaW を用いた環境適応 TTS モデルの理想的な評価では, まず収録室内で被験者 2 人と同様に TTS 合成音声を再生するスピーカーと評価者が向かい合う. 周囲のスピーカーから環境雑音を再生した中で TTS モデルが生成した音声を再生し, 評価者がそれを受聴してその自然性と聞き取りやすさを評価する. しかし, この評価方法は高コストであるほか, 他の研究にとって SaSLaW を用いた手法の評価が困難であるという欠点がある. これについて, 実環境下での音声加工に基づく音声強調手法の

[*4] https://www.sanwa.co.jp/product/syohin?code=CHE-SD1

表 1 SaSLaW とその他の TTS コーパス・データセットの比較表. なお, 発話音声における “あり (※)” は発話音声に雑音が混合していることを示している. また, 評価用 IR の IR はインパルス応答を表す.

コーパス名	発話様態	雑音環境	受聴音	映像	発話音声	評価用 IR
TTS コーパス						
SaSLaW(ours)	自発的	実環境を模擬	一人称視点あり	一人称視点あり	あり	あり
Hurricane [19]	朗読 (独話)	実環境を模擬	なし	なし	あり	なし
CEJC [22]	自発的	実環境	三人称視点あり	三人称視点あり	あり (※)	なし
Guo et al. [16]	演じるような	雑音なし	なし	なし	あり	なし
TTS が目的ではないデータセット						
EgoCom [23]	自発的	実環境	一人称視点あり	一人称視点あり	なし	なし
EasyCom [24]	自発的	実環境	一人称視点あり	一人称視点あり	一部あり (※)	なし
noisy-CSJ [25]	自発的&朗読	実環境	三人称視点あり	なし	なし	なし
Hurricane 2.0 [26]	朗読 (独話)	実環境を模擬	三人称視点あり	なし	あり	あり

表 2 1 つの会話の収録において被験者 2 人に指示する内容の例.

テーマ	美術館での絵画の解釈	
被験者 A	役	観光客
	タスク	絵画の意味や背景を理解する
被験者 B	役	ガイド
	タスク	絵画の解釈や背景を分かりやすく説明する

性能を競うコンペティションである Hurricane Challenge 2.0 [26] では, 実環境での聴取を模擬した評価用音声作成に必要な収録室のインパルス応答と, 収録室における環境雑音のみの受聴音を収録し, 手法の性能評価に用いている例がある. 本研究における SaSLaW の構築においてもこの例を参考にし, インパルス応答と環境雑音のみの受聴音の収録を行う. インパルス応答は被験者間の距離関係が複数の場合について収録する. 評価用音声の作成に際し, インパルス応答を収録していない被験者間の距離関係の評価用音声の作成には, 距離関係が最も近いインパルス応答を代用する. また, コーパスに収録された発話音声と, その発話を聴く被験者の受聴音声に含まれる発話音声の音量は異なる. そのため, 評価用音声の作成においては, TTS モデルが生成した音声のパワーが聞き手側の受聴音声に含まれる発話音声のパワーと同じになるように, 手動で指定した調整量に基づきゲインを調整してからインパルス応答を畳み込み, 雑音のみの受聴音を重畳する.

4. SaSLaW コーパスの分析

これまでに, 男性同士と女性同士の被験者ペア 1 組に付いて 3 章で述べた自発対話の収録を行い, 男女 2 名ずつの自発対話時発話音声, 一人称受聴音, 一人称視点映像を収録した. 本節では SaSLaW コーパスの分析として, そのうち男女 1 名（男性：spk01, 女性：spk06）ずつの収録結果について分析した結果について述べる. 収録により得られた発話音声長の合計は spk01 が 24 分, spk06 が 41 分程度であった.

4.1 分析対象と手順

発話音声が環境雑音の音圧レベルによってどのように変化するかを分析するために, 各会話収録前に計測した環境雑音の音圧レベルに基づき, 各会話に対して環境雑音の大きさに関するラベル “noisy”（音圧レベル大）, “moderate”（音圧レベル中）, “silent”（音圧レベル小）を付与した. これにより, 同時に分割した発話音声にも環境雑音の大きさに関するラベルを付与した. なお, この分析では被験者 2 人の距離関係による発話の特性変化について調査しなかった. また, 環境雑音の音圧レベルによる発話音声の特性変化を調べる際, 被験者 2 人の距離関係による要因を考慮せずに分析した.

発話音声の分析として, 各発話ごとに特徴量としてフレームごとに RMS 値とスペクトル傾斜, そして第 1 フォルマント周波数を計算してその有声フレーム間平均を取り, 発話ごとの各特徴量の分布を調べた. まず各発話に対して, 窓長 5 ミリ秒での harvest [30] により有声フレームを取得した. 次に窓長 20 ミリ秒, ホップサイズ 5 ミリ秒で フレームごとの RMS 値とスペクトル傾斜, 第 1 フォルマント周波数を計算した. スペクトル傾斜の単位は dB/oct であり, その値が大きいほどスペクトルが平坦であり, 高域が相対的に強調されていることを示す. スペクトル傾斜の計算方法は関連研究 [31] を参考に, 0.25 kHz, 8 kHz を起点, 終点とした 1/3 オクターブ間隔のフィルタバンクを用いて各バンドパスフィルタの出力ごとの RMS 値を計算し, その対数周波数軸上の線形回帰直線の傾きを出力量とした. 第 1 フォルマント周波数は Praat を用いて計算した. 有声区間内フレーム平均をとる際に, スペクトル傾斜と第 1 フォルマント周波数については, 中央値から大きく離れた値のフレームは外れ値として計算時に除外した.

計算した発話ごとの特徴量は, 発話に付与した環境雑音の音圧レベルのラベルで分けて集計した. ラベルごとに特徴量に差があるかを評価するために, 特徴量の平均がラベル間で差があるかを有意水準 $\alpha = 0.05$ の Welch の両側 t

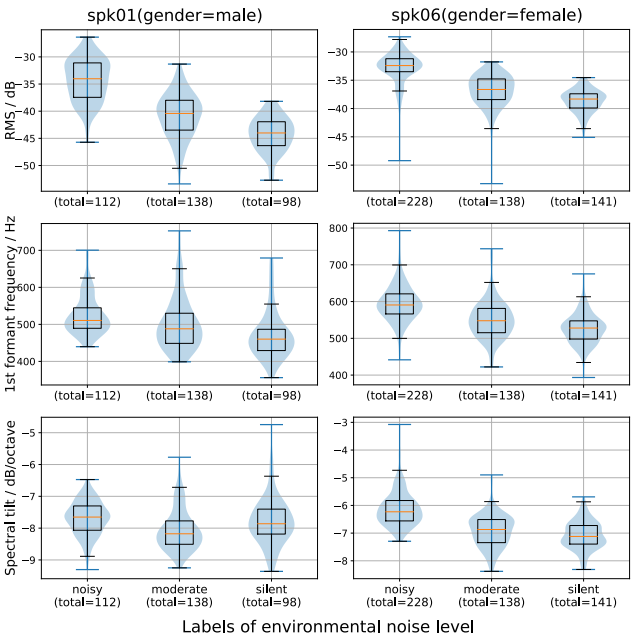

図 5 収録した発話音声の特徴量分布の箱ひげ図と violin-plot を重ねたものを, 男性・女性話者ごとに環境雑音の音圧レベルで分けて表示したもの.

検定により評価した.

4.2 分析結果と考察

男性・女性話者ごとに計算した発話特徴量の分布について, その箱ひげ図と violin-plot を重ねたものを**図 5** に示す. 各特徴量の分布を観察すると, 男性, 女性のいずれの話者においても, RMS と第 1 フォルマント周波数は環境雑音の音圧レベルが silent から noisy になるにつれ中央値が大きくなり、 分布全体が値が増大する方向に平行移動していることがわかった. また 2 人の話者について, RMS と第 1 フォルマント周波数の分布の平均は silent から noisy になるにつれ有意に大きかった. これらの結果は, 既存のコーパスである Lombard Grid コーパス [20] に関する分析から得られた, Lombard 朗読発話は静音環境における発話に比べて発話のエネルギー (RMS) と第 1 フォルマント周波数が上昇したという結果と一致した. そのため, SaSLaW コーパスに収録された発話音声において, 環境雑音の音圧レベルが大きいことにより自発的な発話音声に Lombard 効果が確かに現れたと考えられる.

スペクトル傾斜については, 男性話者では環境雑音が moderate から noisy になることで中央値が大きくなり, 分布も値が大きくなる方向に平行移動しているが, silent と noisy の間で中央値に差はなかった. 分布についても, moderate よりも noisy の方が有意に平均が大きく, silent と noisy の間で有意差はなかった. 女性話者では RMS や第 1 フォルマントと同様にして, silent から noisy になるにつれ中央値が大きくなっており, 分布の平均も有意に大きくなっていくことがわかった. この結果から, 環境雑音の音圧レベルへの適応において, 話者の間で共通する特徴量もあるが, そうではないものもあることがわかる. そのため, 先行研究で行われているような信号処理に基づく Lombard 発話作成だけでは, 音環境に適応した実 Lombard 発話は模擬するのは困難であり, DNN に基づく音環境へのルールベースでない柔軟な適応能力の獲得が必要であると考えられる. 今後は収録を継続して話者数を増やすことで, 人間の音環境への適応についての一般的な傾向や話者依存性について詳しく分析していきたい.

5. SaSLaW を用いた環境適応 TTS モデルの実験的評価

本節では, 3 節で収録した SaSLaW コーパスを用いて環境適応 TTS モデルについて行った比較評価実験について述べる. 男性・女性話者 1 名ずつ (spk01, spk06) それぞれについて複数の環境適応 TTS モデルを作成し, 合成した音声が環境雑音に適応した自然な発話音声であるかについて比較評価を行った.

5.1 作成した環境適応 TTS モデルと学習条件

本研究では環境適応 TTS モデルの作成にあたり, 音声のメルスペクトログラムを出力する DNN モデルとして FastSpeech 2 [10] をベースとしたモデルを, 音声波形を生成するボコーダーは事前に学習した HiFi-GAN [32] を用いた. これらの実装は公開されている PyTorch で記述したもの[*5]を用い, モデル内の FastSpeech 2 部分と HiFi-GAN のハイパーパラメーターは全てこの実装と同一のものを採用した. 本実験では以下の 3 つの環境適応 TTS モデルを学習した:

FS2:テキストのみを入力とする FastSpeech 2. JSUT [33] を用いて 90 万ステップの事前学習を行い, その後 SaSLaW の単話者の発話データで 10 万ステップのファインチューニングを行う. 環境適応 TTS モデルのベースラインとして用いる.

FS2-predsty:FastSpeech 2 を global style token (GST) [34] で条件付ける構造を採用したモデル. その詳細は**図 6** を参照されたい. 推論時には, FastSpeech 2 と同時に学習した Env-to-style predictor を用いて音環境から GST の出力するトークンを推定し, FastSpeech 2 の Encoder 出力後に条件付ける. 学習時はまず JSUT を用いて Style token layer と Env-to-style predictor 以外の部分を JSUT を用いて 90 万ステップ事前学習し, その後モデル全体について SaSLaW の単話者発話データと音環境データで 10 万ステップのファインチューニングを行う.

FS2-predsty-ptrn:モデルの構造は FS2-predsty と同じ. 学習時に FS2-predsty のモデル全体を, 後述する JSUT を

[*5] https://github.com/Wataru-Nakata/FastSpeech2-JSUT

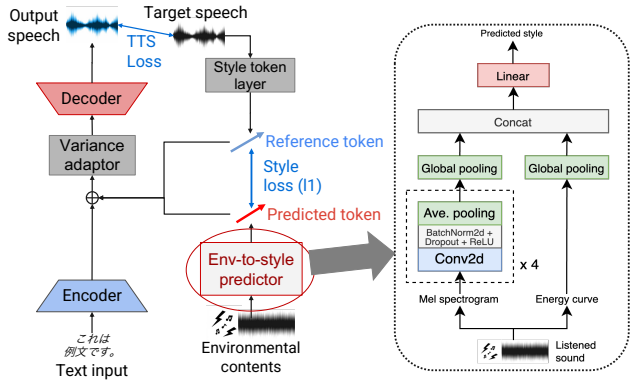

図 6 提案する global style token [34] を用いた FastSpeech 2 に基づく環境適応 TTS モデルのブロック図. 右側に Env-to-style predictor のモデル図も示した.

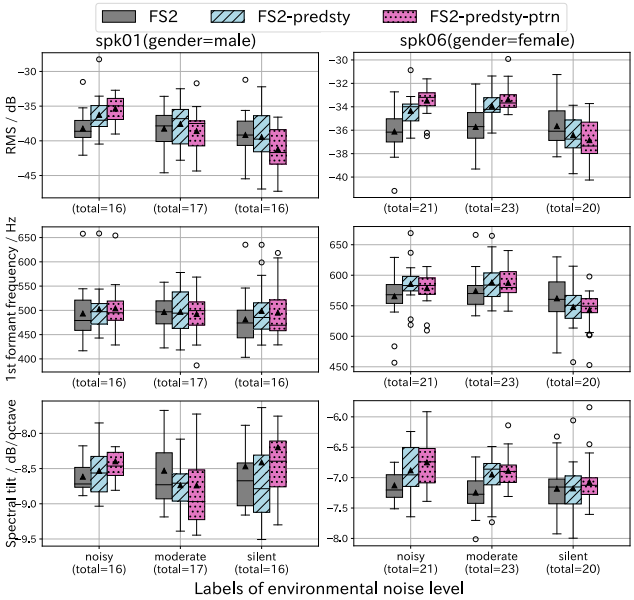

図 7 学習した環境適応 TTS モデルによる合成した発話音声から得られた特徴量分布の箱ひげ図を, 男性・女性話者ごとに環境雑音の音圧レベルごとに分けて表示したもの. 箱ひげ図中の三角点は特徴量の平均値を表す. 男性話者のスペクトル傾斜についてのみ, 表示上の都合から外れ値をプロットしなかった.

ベースとした擬似データで 90 万ステップ事前学習 (“pseudo data training”) し, その後 SaSLaW の単話者発話データと音環境データで 10 万ステップのファインチューニングを行う.

本実験では, FS2-predsty と FS2-predsty-ptrn の Env-to-style predictor に入力する音環境データとして, 各発話が存在する会話における直前ターンの受聴音を用いた. この受聴音には, 環境雑音と当該話者の会話相手の発話を耳で聴いた音が含まれる. なお, FS2-predsty-ptrn の事前学習において入力する環境雑音には会話相手の発話が含まれないことを注意されたい. 本実験で用いた Env-to-style predictor のモデル構造についても図 6 に示す.

SaSLaW を用いたファインチューニングにおいて, 学習/検証用にデータを分割する際, 学習/検証データ間で受聴音に含まれる環境雑音にオーバーラップが生じないように検証データを作成した. さらに, 検証データ内で発話音声に付与された環境雑音の音圧レベルのラベルが網羅されるように検証データを作成した. 本実験で学習したモデルはいずれも検証データを用いたモデル選択を行わないことを踏まえ, 実験の評価用データは検証データと同一とした.

FS2-predsty-ptrn の事前学習に用いた JSUT をベースとした擬似データは次のように作成した. まずは JSUT の各発話音声が DEMAND に収録された cafe, station, square 雑音のいずれかの中で, 発話と雑音の SN 比が $-10\,\mathrm{dB}, 0\,\mathrm{dB}, 10\,\mathrm{dB}$ のいずれかになるように発されたものとして, 対応するモノラルの環境雑音を割り当てた. 次に, 割り当てた環境雑音の中で各発話が明瞭で自然となるように, 発話の 1 次の線形予測分析係数を事前に決定した量に基づき変更することでスペクトル傾斜の絶対値を減じて高域を強調する処理を施し, テキスト–環境雑音–発話音声の組で構成される擬似データを作成した.

5.2 客観評価実験

5.1 節で学習した 3 つの環境適応 TTS モデルが合成した音声についての客観評価として, 4.1 節で挙げた発話の特徴量である RMS, 第 1 フォルマント周波数, スペクトル傾斜を評価用データ内の全ての発話について計算し, 得られた分布が環境雑音の音圧レベルのラベルや手法によってどのように異なるかについての比較評価を行った. 環境雑音のラベルやモデルによる特徴量の分布に差があるかを調べるために, 4 節と同様にして特徴量の平均について有意水準 $\alpha = 0.05$ の Welch の両側 t 検定により差分を評価した.

男性, 女性話者 (spk01, 06) それぞれについて, 得られた分布の箱ひげ図を**図 7** に示す.

男性, 女性話者のいずれにおいても, ベースラインモデルの FS2 が合成した音声の特徴量分布を観察すると, いずれの特徴量においても, 環境雑音の音圧レベルによる分布の増減方向での平行移動が見られなかった. また, 異なる音圧レベルのラベルの間で, いずれの特徴量の平均にも有意差がなかった.

男性話者について, FS2-predsty と FS2-predsty-ptrn が合成した音声の特徴量分布を観察すると, まず RMS は環境雑音が silent, moderate, noisy の順で大きくなる方向に平行移動していた. RMS の平均については, FS2-predsty-ptrn では同様の順で有意に大きくなっていたが, FS2-predsty では silent と比較して noisy は有意に大きく, それ以外のモデルの組み合わせでは有意差がなかった. 次に第 1 フォルマント周波数は, いずれのモデルにおいても環境雑音の音圧レベルによる分布の増減方向での平行移動が見られず, 平均の有意差も見られなかった. 最後にスペクトル傾斜では, FS2-predsty と FS2-predsty-ptrn が合成し

た音声の分布は silent と noisy の場合に比べて moderate が小さくなる方向に平行移動していたが, いずれのラベル間でも有意差は見られなかった. この分布の平行移動の関係は, 図 5 で得られたコーパスの分析結果と整合している.

女性話者について FS2-predsty と FS2-predsty-ptrn が合成した音声の特徴量分布を観察すると, いずれの特徴量においても, 環境雑音が silent の場合に比べて moderate, noisy の特徴量分布は大きくなる方向に平行移動していた. また, RMS と 第 1 フォルマント周波数では silent に比べて moderate, noisy の平均が有意に大きかった. 一方で, スペクトル傾斜については silent に比べて noisy の平均は有意に大きかったが, moderate の平均との有意差はなかった.noisy と moderate の間では, いずれのモデルでも平均に有意差はなかった.

以上の結果から女性話者については音環境を考慮する機構を付加したモデルを用いることで初めて, 直前の受聴音に適応して発話に現れた Lombard 効果がある程度制御できるようになることがわかった. 一方で男性話者が RMS 以外の特徴量について, 音環境を考慮する機構を付加したモデルでも環境雑音のラベル間で RMS 以外に有意差が見られなかった. その理由として, 女性話者に比べて発話音声長が 6 割程度しかなく, 音量を司る RMS 以外の音声特徴量と環境雑音の関係を学習するのに十分な音声データを収集できていなかった可能性と, 本稿で実験やコーパスの分析を通して調べた話者の特性による可能性の 2 つが考えられる. いずれの可能性についても, 今後話者数を増やして同様の分析・実験を行っていくことで検証していきたい.

5.3 主観評価実験

学習したモデルの主観評価として, まずはモデルが評価用データから合成した音声から 3.3 節に基づき実環境での聴取を模擬した評価用音声を作成した. その後, 3 つのモデルから 2 つ一対を評価対として取り出し, どちらのモデルの評価用音声が雑音環境の中で自然で聞き取りやすいかについての AB テストを各話者計 3 対それぞれに対して行った. AB テストの評価結果はランサーズ[*6]を介したクラウドソーシングにより収集した. 話者ごと・モデルの評価対ごとにクラウドワーカーを重複を許容して 72 名, 合計 720 件の回答を収集した. 話者間・評価対間でのクラウドワーカーの重複は許容した. 評価結果の収集過程で, 評価対によっては 72 名以上（最大 74 名）からの回答が得られたが, これらの除外は行わなかった. 評価結果から, 検証データに付与された環境雑音の音圧レベルのラベルごとに評価対間のプリファレンススコアを算出した. プリファレンススコアの差に関する評価は有意水準 $\alpha = 0.05$ の Student の両側 t 検定により行った. 男性・女性話者 (spk01, 06) ごとに算出したプリファレンススコアについて, **図 8** にて棒グラフでプロットした.

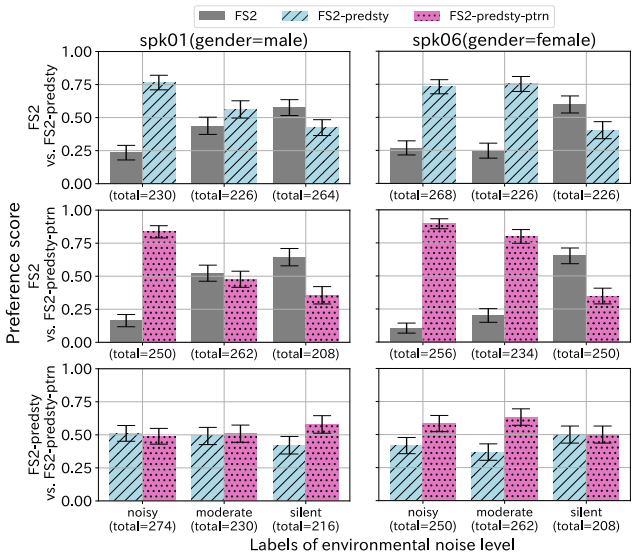

図 8 環境適応 TTS モデルが合成する音声の各評価対における AB テストの結果を, 男性・女性話者ごとに環境雑音の音圧レベルで分けて表示したもの. エラーバーは 95%信頼区間を示す.

まずベースラインモデルの FS2 に比べて, 男性話者については FS2-predsty が合成した音声は環境雑音が noisy と moderate であるときはスコアが有意に高かった. FS2-predsty-ptrn は環境雑音が noisy であるときはスコアが有意に高く, moderate ではスコアに有意差がなかった. 環境雑音が silent であるときは, いずれのモデルもスコアは有意に低かった. 女性話者については, いずれのモデルにおいても環境雑音が moderate と noisy である場合はスコアが有意に高く, silent である場合は有意に低いことがわかった.

次に FS2-predsty と FS2-predsty-ptrn を比較すると, 男性話者では環境雑音が silent の場合のみ FS2-predsty-ptrn の方がスコアが有意に高く, それ以外では有意差がなかった. 女性話者では moderate, noisy の場合で FS2-predsty-ptrn の方がスコアが有意に高く, silent の場合では有意差がなかった.

音環境を考慮する機構が環境に応じて自然で聞き取りやすい発話を生成するのに必要かという点で, ベースラインモデルと他の 2 つのモデルを比較したプリファレンススコアは 5.2 節における客観評価結果と moderate, noisy の場合は整合するが, silent の場合は整合しない. この結果が現れた理由として, 主観評価実験のデザインによるものが考えられる. silent の場合にモデルが合成した音声は, 客観評価結果からいずれの特徴量においても FS2 の方が平均が大きいか, 有意に差がなかったことがわかる. よって FS2 は環境雑音が silent の場合には, 音環境の考慮をしたモデルより明瞭で聞き取りやすい発話音声が平均的に生成される. AB テストの際, クラウドワーカーには自然で聞き取

[*6] https://www.lancers.jp/

りやすい音声を選択するよう指示した. そのためクラウドワーカーは環境雑音における自然さと聞き取りやすさを切り分けられず, 音声を選択するタスクにおいて聞き取りやすさに偏重してしまい, silent の場合において客観評価と整合しない結果が得られた可能性がある. 今後この考察の検証として, 収録した話者数の増加に伴って評価項目を切り分けた主観評価実験デザインを再検討していきたい.

FS2-predsty と FS2-predsty-ptrn を比較した結果, FS2-predsty-ptrn の方が音環境に応じてより自然で聞き取りやすい発話音声を出力できるモデルであることがわかった. この結果は事前学習の段階から図 6 で提案したモデル全体を, 音環境を入力するコーパスで学習することでより自然で聞き取りやすい音声を出力できることを示唆している.

6. おわりに

本研究では対話ロボットといった音声対話システムの実環境運用への応用を志向し, 音環境に応じた自然な発話を生成できる環境適応 TTS モデルのためのコーパスである SaSLaW を構築した. SaSLaW では話者の一人称視点での音環境情報を捉えるため, 実環境を模擬した様々な雑音環境における 2 人の自発対話から各話者の発話音声, 一人称受聴音声と一人称視覚映像を同時に収録する. SaSLaW を用いた環境適応 TTS モデルの主観・客観評価実験を通して, 音環境を考慮する機構が音環境に応じてより自然な発話を生成する能力に寄与したが, その寄与の程度は話者によって変化することがわかった. 今後はより多くの話者のデータを収集し, 話者ごとの環境への適応も含めた分析・評価を行っていきたい. また, 本稿の実験では収録した視覚映像を含めた評価を行わなかったため, 発話の視覚映像への適応の程度もコーパスの分析や実験を通して評価していきたい.

謝辞 本研究の一部は，科研費 22H03639，23K18474，JST 創発的研究支援事業 JP23KJ0828，及び JST ムーンショット型研究開発事業 JPMJMS2011 の助成を受け実施しました. また, 原稿の作成に際して, 渡邊 亞椰さんには図の作成でご協力頂きました. この場を借りて感謝申し上げます.

参考文献

[1] Cooke, M., King, S., Garnier, M. and Aubanel, V.: The listening talker: A review of human and algorithmic context-induced modifications of speech, *Computer Speech & Language*, Vol. 28, No. 2, pp. 543–571 (2014).

[2] Hazan, V. and Baker, R.: Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions., *The Journal of the Acoustical Society of America*, Vol. 130 4, pp. 2139–52 (2011).

[3] Tuttosi, P., Hughson, E., Matsufuji, A., Zhang, C. and Lim, A.: Read the Room: Adapting a Robot's Voice to Ambient and Social Contexts, *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3998–4005 (2023).

[4] Denes, P. B. and Pinson, E.: *The speech chain*, Macmillan (1993).

[5] Lombard, E.: Le signe de l'élévation de la voix, *Annales des Maladies de L'Oreille et du larynx*, Vol. 37, pp. 101–119 (1911).

[6] Junqua, J.-C.: The Lombard reflex and its role on human listeners and automatic speech recognizers, *The Journal of the Acoustical Society of America*, Vol. 93, No. 1, pp. 510–524 (1993).

[7] Pelegrín-García, D., Smits, B., Brunskog, J. and Jeong, C.-H.: Vocal effort with changing talker-to-listener distance in different acoustic environments, *The Journal of the Acoustical Society of America*, Vol. 129, No. 4, pp. 1981–1990 (2011).

[8] 西村良太，北岡教英，中川聖一：音声対話における韻律変化をもたらす要因分析 (特集 リズムとタイミング)，音声研究， Vol. 13, No. 3, pp. 66–84 (2009).

[9] Kawahara, T.: Spoken dialogue system for a human-like conversational robot ERICA, *9th International Workshop on Spoken Dialogue System Technology(IWSDS)*, Springer, pp. 65–75 (2019).

[10] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *Proceedings of the 9th International Conference on Learning Representations (ICLR)* (2021).

[11] Novitasari, S., Sakti, S. and Nakamura, S.: A Machine Speech Chain Approach for Dynamically Adaptive Lombard TTS in Static and Dynamic Noise Environments, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2673–2688 (2022).

[12] Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T. and Mei, T.: Visual-Aware Text-to-Speech*, *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023).

[13] Zorila, T.-C., Kandia, V. and Stylianou, Y.: Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression, *Proc. INTERSPEECH*, pp. 635–638 (2012).

[14] T. V. Ngo, R. Kubo, M. A.: Mimicking Lombard Effect: An Analysis and Reconstruction, *IEICE Transactions on Information and Systems*, Vol. E103.D, No. 5, pp. 1108–1117 (2020).

[15] Bollepalli, B., Juvela, L. and Alku, P.: Lombard Speech Synthesis Using Transfer Learning in a Tacotron Text-to-Speech System, *Proc. INTERSPEECH*, pp. 2833–2837 (2019).

[16] Hu, Q., Bleisch, T., Petkov, P., Raitio, T., Marchi, E. and Lakshminarasimhan, V.: Whispered and Lombard Neural Speech Synthesis, *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 454–461 (2021).

[17] Raitio, T., Petkov, P., Li, J., Shifas, M., Davis, A. and Stylianou, Y.: Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise, *Proc. INTERSPEECH*, pp. 1936–1940 (2022).

[18] Laura Folk, F. S.: The Lombard Effect in Spontaneous Dialog Speech, *Proc. INTERSPEECH*, pp. 2701–2704 (2011).

[19] Cooke, M., Mayo, C. and Valentini-Botinhao, C.: Intelligibility-enhancing speech modifications: the Hurricane Challenge, *Proc. INTERSPEECH*, pp. 1341–1345 (2013).

[20] Alghamdi, N., Maddock, S., Marxer, R., Barker, J. and Brown, G. J.: A corpus of audio-visual Lombard speech with frontal and profile views, *The Journal of the Acoustical Society of America*, Vol. 143, No. 6, pp. EL523–EL529 (2018).
[21] Guo, H., Zhang, S., Soong, F. K., He, L. and Xie, L.: Conversational End-to-End TTS for Voice Agents, *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 403–409 (2021).
[22] 小磯花絵, 臼田泰如, 川端良子:『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation, CEJC), 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 96, p. 31 (2022).
[23] Northcutt, C. G., Zha, S., Lovegrove, S. and Newcombe, R.: EgoCom: A Multi-Person Multi-Modal Egocentric Communications Dataset, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 6, pp. 6783–6793 (2023).
[24] Donley, J., Tourbabin, V., Lee, J.-S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V. K. and Mehra, R.: EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments, *arXiv preprint arXiv:2107.04174* (2021).
[25] 三村正人, 井上昂治, 河原達也, 中村友彦, 猿渡洋: 実環境下日本語話し言葉音声コーパスの構築と音声認識ベンチマーク, 音声言語情報処理研究報告 (SLP), Vol. 2023, No. 12, pp. 1–6 (2023).
[26] Rennies, J., Schepker, H., Valentini-Botinhao, C. and Cooke, M.: Intelligibility-Enhancing Speech Modifications — The Hurricane Challenge 2.0, *Proc. INTERSPEECH*, pp. 1341–1345 (2020).
[27] Thiemann, J., Ito, N. and Vincent, E.: DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, Zenodo, (online), DOI: 10.5281/zenodo.1227121 (2018).
[28] Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, *Proc. INTERSPEECH*, pp. 1983–1987 (2023).
[29] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I.: Robust speech recognition via large-scale weak supervision, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, JMLR.org, pp. 28492–28518 (2023).
[30] Morise, M.: Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals, *Proc. INTERSPEECH*, pp. 2321–2325 (2017).
[31] Sato, Y. and Villegas, J.: Spectral Tilt May Have a Smaller Impact on the Intelligibility of Speech in Noise, *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–5 (2023).
[32] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., pp. 17022–17033 (2020).
[33] Sonobe, R., Takamichi, S. and Saruwatari, H.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, *arXiv preprint arXiv:1711.00354* (2017).
[34] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F. and Saurous, R. A.: Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Vol. 80, PMLR, pp. 5167–5176 (2018).