Style-Bert-VITS2 を用いた災害時避難を促す緊迫感のある音声合成

Urgent voice synthesis to encourage evacuation during disasters using Style-Bert-VITS2

横田 小次郎
Kojiro Yokota
法政大学情報科学部ディジタルメディア学科
kojiro.yokota.4s@stu.hosei.ac.jp

Abstract

Lessons from the Great East Japan Earthquake have emphasized the importance of a "urgency" in evacuation calls to prompt immediate action. However, it is difficult for non-professionals to maintain a stable, strong tone during emergencies. This study proposes a method to generate evacuation guidance speech with multiple levels of urgency using the speech synthesis model "Style-Bert-VITS2," trained on speech mimicking actual broadcasts. An objective evaluation using a classification model trained in actual broadcast audio showed that 84 % of high urgency synthetic speech was classified as "high urgency," approaching the 88 % accuracy of recorded human speech. For both low urgency speeches, they were classified as 100 % "low urgency." These results demonstrated the effectiveness of the proposed method in generating multiple levels of urgency.

1 はじめに

平成 23 年 3 月 11 日に起こった東日本大震災では, 多くの人が地震だけでなく, 津波の被害に遭い, 犠牲となった. その中には避難をしなかった人や逃げ遅れた人も多くいた. その原因の一つとして, 防災無線やテレビなどの避難アナウンスが平時と同様に丁寧で穏やかであったことがあげられた. 「ここまで津波は来ないだろう」「周りが逃げていないから大丈夫だろう」などの正常性・同調バイアスが働き, それを覆すほどの避難を呼びかけることができなかった. 一方で, 茨城県大洗町などの一部の自治体では, 防災無線の放送で, 緊迫感のある声で呼びかけを行った結果, 多くの住民が避難をし, 人的被害が抑えられた事例も報告されている.

NHK は東日本大震災の教訓を踏まえて, 避難の呼びかけを見直した. 放送による情報伝達はできても, 住民の避難行動を促進するのに十分な効果がなかったという指摘があった [1]. 震災の 3 か月後には全国のアナウンサーを招集して呼びかけに関する検討会が開かれた. その後, 呼びかけのポイントは「確実に伝わること」「行動を促すこと」「予断を与えないこと」の 3 点に集約された. これに基づき, 普段とは異なる強い口調で呼びかけを行う方針が決定された.

しかしながら, 公共放送においては本来, 情報は冷静に伝えられることが求められる. 実際に「NHK 放送ガイドライン 2025[2]」の国内番組基準では, 表現の原則として「人心に恐怖や不安または不快の念を起こさせるような表現はしない」と定められている. これは災害報道においても適用される基本姿勢であり, 不用意に視聴者の不安を煽ることは避けなければならない. その一方で, 一刻を争う津波災害においては, この原則よりも人命救助が最優先される. 先ほどのガイドラインの「災害・非常事態」の項目では, 津波のおそれがある場合, 「アナウンサーは強い口調で避難を呼びかけ」るように明記されており, 状況に応じた発話スタイルの明確な使い分けが規定されている.

この方針が実際に適用され, その効果が示されたのが令和 6 年 1 月 1 日に起こった能登半島地震である. 東日本大震災以来の大津波警報が発令された際, NHK のアナウンサーは絶叫に近い口調で避難を呼びかけた [3]. 特筆すべきは, 1 回目の緊急地震速報が発表されてから津波警報が出るまでの約 5 分間は, ガイドラインの原則通り落ち着いた口調で状況を伝え, 津波警報の発令とともに規定の「強い口調」へと変化させた点である. 「今すぐ逃げること！」といった強い命令形や「東日本大震災を思い出してください」という言葉を用いた呼びかけは, 視聴者に「ただごとではない」という危機感を抱かせ, 避難行動を強く喚起したと評価されている.

避難を呼びかける際は, NHK のアナウンサーのように緊迫感のある発話を行うことが望ましい. しかし, 彼らは訓練を受けたプロであり, 一般の自治体職員が緊急時に意識して発話スタイルをコントロールすることは困難である. 加えて, 津波が迫る極限状態において, 冷静かつ適切な緊迫感を持ってアナウンスを行うことはさらに難易度が高い. そこで, 音声合成を活用する. 音声合成であれば, 発話の内容を事前に作成するため, 発話者の心理状態に依存せず, 安定して生成することができる.

一方で, 避難の呼びかけにおいては「誰が呼びかけているか」という信頼性も重要な要素である. 茨城県大洗町の事例では, 町長が呼びかけていたから避難をしようと思ったという意見もあった. しかしながら, 避難を呼びかける人が逃げ遅れる場合も考えられる. 実際に, 東日本大震災では, 南三陸町の職員の方が防災無線での避難を呼びかけ続けた結果, 津波の犠牲となる痛ましい事例もあった. 呼びかける側の安全を確保しつつ, 住民の避難行動を最大限に促すためには, 音声合成の活用が有効である.

そこで本研究では, 話者性を保持したまま発話の緊迫度といったパラ言語的属性を付与および制御可能な音声合成手法を提案する. 提案手法では, 緊迫感を伴う発話音声と緊迫感を伴わない発話音声の双方を学習データとして用い, 表現強度の差異を潜在表現としてモデル化することで, 同一話者音声に対する緊迫度の操作を可能にする特徴がある. これにより, 緊迫感がある避難呼びかけ音声を作成することだけではなく, 町長などの「信頼できる人物」の声で音声合成を作成することができ

Supervisor: Prof. Katunobu Itou

るため, アナウンス担当者の安全確保と住民への訴求力の向上を両立できると考える.

2 従来研究

2.1 緊迫感のある音声の特徴

小林らは, 緊迫感のある音声の特徴量を調査するために, プロのアナウンサーが発した避難音声の特徴量を変換した [4]. 変換した特徴量は, 音の高さを表す fo（基本周波数）と音の高さを表すスペクトル, 話す速さを表す持続時間の 3 つである. 評価方法は, マグニチュード推定法と SD 法と因子分析を用いた. マグニチュード推定法は元の刺激を基準に比較刺激がどのくらいの大きさかを推定する評価方法である.

この実験は, 元の音声を基準に変換した音声がどのくらいの緊急性があるのかを調査したものである. SD 法は刺激に対して, 対になる様々な形容詞のどちらが近いのかを比べ, その刺激がどの要素かを調べる評価方法である. この実験では,「落ち着いている-緊張している」「ゆっくり-速い」などの形容詞対を用いている. 因子分析は, SD 法で得た数値を相対的に比較することでその刺激がどの要素に近いのかを推定する手法である.

実験の結果, fo が緊急性に影響があり, スペクトルと持続時間は緊急性に影響がないことが示されている. fo の時間平均と時間変動（抑揚）についても調べたが, どちらも影響があることが示されている.

2.2 音声合成モデルを用いた避難呼びかけ音声

2.2.1 基本周波数と話速を操作した緊迫感の付与

原田らは, ニューラル TTS（Text-to-Speech）モデルを用いて音声合成を行い, 緊迫感のある音声の印象評価実験を行った [5]. ニューラル TTS モデルは FastSpeech2[6] という End-to-End の TTS モデルを使用した. FastSpeech2 は, 非自己回帰型モデルであり, Tacotron などの自己回帰型モデルよりも推論が速く, 声の高さや強さ, 話す速さなどの韻律特徴量を操作できることが特徴である. 学習データとして, JSUT 音声コーパス [7] と ITA コーパス [8] を用いた. JSUT 音声コーパスは単一の日本語女性話者による 10 時間の音声データである. 日本語の音声合成に使われるコーパスとして有名なもののひとつである. ITA コーパスは著作権の消滅した文献やオリジナルの文章・単語から文セットを構築することで, パブリックドメインで公開される文章コーパスである. 日本語の単語では出現しにくいモーラも一定量カバーしつつも読みやすさを考慮している.

実験では, MOS 評価という 5 段階評価での主観評価実験を行った. 被験者に対象の音声を聞かせ, その音声の評価を緊急性, 聞き取りやすさ, 信頼性の 3 つの項目で評価をしてもらった.

まず, 実際のアナウンス音声の特徴量を変換し, 変換する前の音声と比べたところ, 声の高さを上げた場合と話す速さを速くした場合が緊急性が上がることがわかった. その結果をもとに, FastSpeech2 で特徴量の操作を行った. FastSpeech2 での操作を行った場合でも同様に声の高さを上げた場合と話す速さを速くした場合に緊急性が上がった. よって, 音声合成でも緊迫性の操作が有効であることが示された. 一方で, 音声合成での声の高さを上げたときに音声不良が見られた. これにより, 聞き取りやすさが低下したという結果になった.

2.2.2 公開されている避難用音声合成

NHK は災害用の音声合成を「命を守る”防災の呼びかけ”」として公開（2025 年 10 月 1 日まで）していた. 大雨や大雪のときの場合を想定した避難や防災を呼びかける音声である. しかし, この音声には緊迫感が不足しており, 危険性を伝えるには不十分であると考える. NHK のアナウンサーが避難の呼びかけを見直したように, 多くの人の避難を促すような緊迫感のある音声が必要だと考える.

2.3 緊迫感がもつ話し方の特徴

能登半島地震で NHK のアナウンサーが地震速報と津波警報で緊迫感を強弱で制御をしていた. その緊迫感の特徴について調査を行った. 緊迫感を表現するための一つの要素としてアクセントのエネルギーの違いがある. そのアクセントを分析するために藤崎モデルを用いる. 藤崎モデルとは, 人間の声の基本周波数の変化を数式化したモデルである. そして、その基本周波数を推定する際にフレーズ成分とアクセント成分の 2 つの要素が関係する. フレーズ成分は, 最初が高く, 終わりにかけてゆっくり下がる部分である. アクセント成分は, 特定の単語や強調したい部分で一時的に上がる部分である. 基本周波数の底（最小値）, フレーズ成分, アクセント成分を足すことでモデル化が可能となっている.

$$\begin{aligned}\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{p,i} G_p(t - T_{0,i}) \\ + \sum_{j=1}^{J} A_{a,j}\{G_a(t - T_{1,j}) - G_a(t - T_{2,j})\}\end{aligned} \tag{1}$$

フレーズ制御機構のインパルス応答は以下のようになる.

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & (t \geq 0) \\ 0, & (t < 0) \end{cases} \tag{2}$$

アクセント制御機構のステップ応答は以下のようになる.

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & (t \geq 0) \\ 0, & (t < 0) \end{cases} \tag{3}$$

また, 初期値として $\alpha = 3.0\text{rad/s}$, $\beta = 20.0\text{rad/s}$, $\gamma = 0.9$ が与えられる.

成澤らは藤崎モデルの問題点であるフレーズ指令やアクセント指令の初期値を手動で行う問題に着目し, それらを自動で推定するシステムを提案した [9]. このシステムを参考に, 音声からアクセント成分を推定するプログラムを実装した. 今回は, 推定したアクセントのエネルギーの総和を求める. これを比較して, アクセントが緊迫感に影響を与えていることを示す.

津波警報の音声と地震速報の音声をそれぞれ約 1 分入力した. 結果, 津波警報が 13.90 で地震速報が 7.83 となった. 津波警報は地震速報の 1.78 倍であることがわかった. また, この藤崎モデルを用いて基本周波数を推定したものとアクセントのエネルギーを視覚的に表したグラフを Fig. 1 に示す. 赤色で表したグラフが津波警報で, 青色で表したグラフが地震速報を表している. 下のアクセントエネルギーの情報では, 津波警報は山が

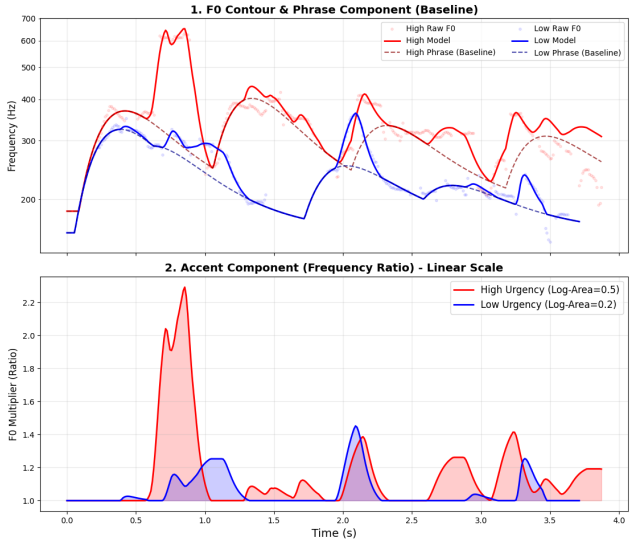

Fig.1: アナウンサーの津波警報と地震速報のアクセントのエネルギーの比較

10 個ほどあり, 地震速報の 6 個と比べると多いことがわかる. また, 津波警報の 1, 2 番目の山は高くなっており, 強いアクセントであることがわかる. このように, アクセントが頻繁に強く含まれている津波警報の音声は, 地震速報とは異なることがわかる. このような話し方の特徴を制御することで緊迫感の付与を行う.

3 提案手法

本研究では, 話者性を保持したまま発話の緊迫度といったパラ言語的属性を付与および制御可能な音声合成手法を提案する. 本手法では, 発話スタイルの制御が可能な音声合成モデルをもとに, 災害時の報道におけるプロのアナウンサーの韻律特徴を再現した音声を学習させる. また, 強度の異なる緊迫感を含む音声データを用いてモデル化を行うことで, 生成音声における緊迫感の段階的な操作を実現する.

従来の手法（FastSpeech2 等）において, 基本周波数や話速を一律に上昇させる操作では, 単に「音が高く速い」だけの音声となり, 避難を呼びかける際に発声する特有の緊迫感を再現することが困難であった. これに対し, 提案手法の音声合成モデルでは, テキストの文脈情報とスタイル設定に基づき, 文末までピッチを下げずに維持するといった, プロのアナウンサーに近い韻律特徴を学習・生成することが可能である.

3.1 音声合成モデル

本研究では, 発話スタイルの柔軟な制御が可能な音声合成モデルとして, Style-Bert-VITS2[10] を用いる. 本モデルは, 高精度な波形生成が可能な VITS[11] をベースに, VITS2[12] による品質改善と, Bert-VITS2[13] による意味理解, そしてスタイル制御機能を統合したアーキテクチャである（Fig. 2）. 本研究の目的である「緊迫感のある避難呼びかけ」の生成において, 本モデルの各構成要素は以下の重要な役割を果たしている.

3.1.1 VITS2 ベースの高品質な波形生成

基本構造には, 条件付き変分オートエンコーダ（CVAE）と敵対生成ネットワーク（GAN）を組み合わせた VITS が採用されている (Fig. 2 Training procedure). また, テキストと音声の時間的な対応（アライメント）を推定するために MAS（Monotonic Alignment Search）を用いており, 事前の人手によるアライメント作業を不要にしている. これにより, 従来の自己回帰型モデルよりも高速かつ高音質な波形生成が可能である. さらに, 「正規化フローへの Transformer ブロックの導入」により, 長文の発話であっても自然な抑揚とリズムを維持できる点は, 避難放送の生成において大きな利点である.

3.1.2 BERT による文脈理解

TextEncoder には音素情報に加え, 大規模言語モデル BERT から抽出された「BERT ベクトル」が入力される（Fig. 2 Inference procedure）. 「BERT ベクトル」とは, 文章の内容から得られる情報をもとに, その文章が持つ意味情報を数値化したものである. これにより, モデルは単なる文字の並びだけでなく, テキストが持つ「意味」や「文脈」を考慮することが可能となった. これは, テキストの内容（例：「津波が来ます」といった災害の情報）に応じた適切なニュアンスを生成することに寄与する.

3.1.3 スタイルベクトルによる音色の制御

本モデルの最大の特徴は, 「スタイルベクトル」の入力である. 学習した音声から抽出された「スタイルベクトル」が, TextEncoder と Decoder の主要部分に注入される構造となっている（Fig. 2 Inference procedure）. 「スタイルベクトル」とは, 発話者の声色や韻律特徴を数値化したものである. これにより, 基本周波数や話速といった単純なパラメータだけでなく, 声の張りや息遣いといった「声色」そのものを直接制御することが可能となる. 本研究では, この機能を用いることで, 緊迫度を制御する避難呼びかけ音声を生成する.

3.2 スタイルの生成と強度制御

本手法では, 学習データをスタイルごとにサブディレクトリに分割して配置することで, 各クラスのスタイルを定義する手法を採用した. 具体的なスタイルベクトルの抽出には, 話者照合タスク等で用いられる事前学習済みモデル「pyannote/wespeaker-voxceleb-resnet34-LM」を利用する. 本モデルは, 音声データをメルスペクトログラムに変換し, そのスペクトログラム画像を ResNet-34 に通すことでベクトルを作成している. この処理を各音声ファイルに対して行い, 個別のスタイルベクトルを取得する. また, 学習データに含まれる全音声から抽出したベクトルの平均値を算出し, これをニュートラルスタイル $\boldsymbol{\mu}$ と定義する.

推論時におけるスタイルの強度調整は以下の式で（1）に基づいて行われる.

$$\boldsymbol{v}_{\text{new}} = \boldsymbol{\mu} + (\boldsymbol{v}_{\text{org}} - \boldsymbol{\mu}) \times w \tag{4}$$

ここで, $\boldsymbol{v}_{\text{new}}$ は調整後のスタイルベクトル, $\boldsymbol{v}_{\text{org}}$ は元のスタイルベクトル, $\boldsymbol{\mu}$ はニュートラルスタイル, w は強度を制御する重みパラメータである. この式は, 平均からの「変化の方向」と「距離」を取り出し, その変化量だけを重み w 倍して増幅・減衰させることを意味する. $w > 1$ と設定すれば特定の特徴が強調され, $0 < w < 1$ であれば平均的な発話スタイルに近づく.

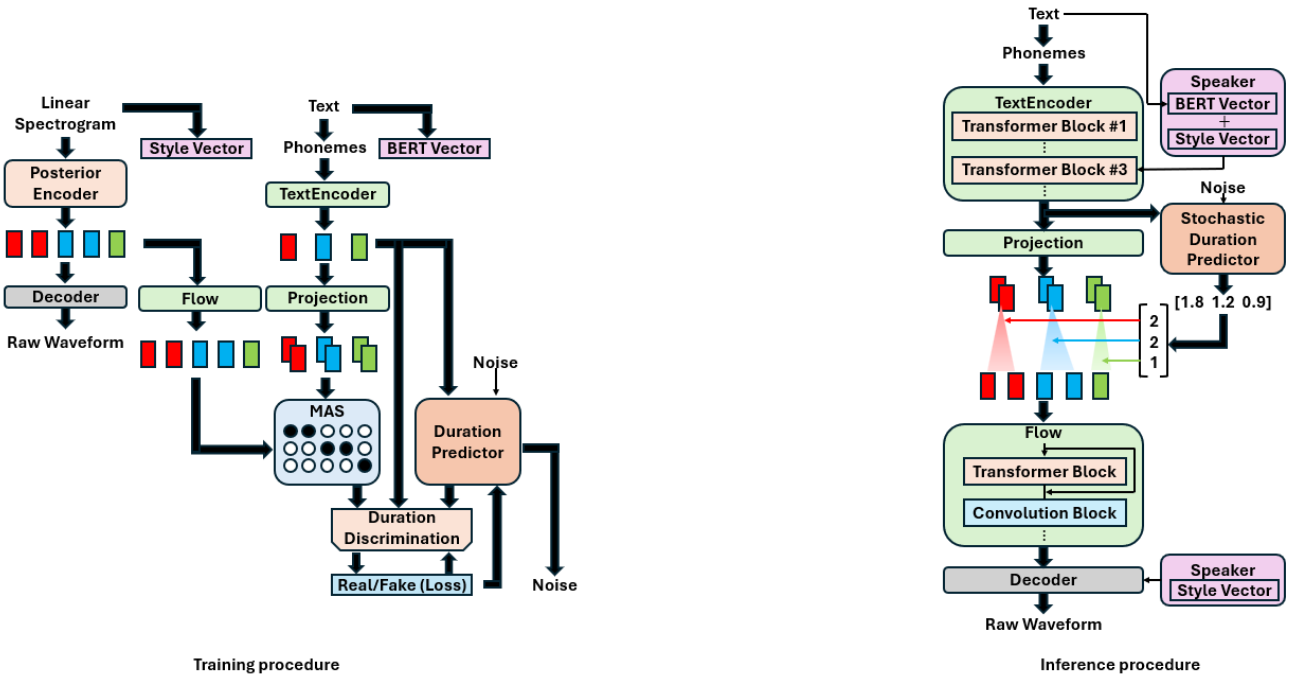

Fig.2: Style-Bert-VITS2 の学習・推論プロセス

3.3 音声の録音

緊迫感のある音声を得るために, 男子大学生 1 名に発話者として録音の協力を依頼した. 録音環境は研究室にある防音室を利用した. 防音室内には, 録音ソフトを操作する筆者と, 発話を行う発話者の 2 人のみである. 防音室は換気扇がついているが, 録音中はノイズが入らないようにするために, 停止しなければならない. そのため, 換気をするために 20 分の録音と 10 分の休憩という流れを繰り返して録音を行った.

録音ソフトは Audacity を用いた. サンプリング周波数は 48000Hz, サンプル形式は 32 ビット浮動小数点数, 録音レベルは 80 %に下げて録音し, ノーマライズは −1.0db に設定した.

録音方法として, PC にマイクを接続し, 発話者はマイクを口に近づけて持つようにした. 録音ボタンを押してから 1〜2 秒で発話を行い, 発話終了から 1〜2 秒で録音停止ボタンを押した. 録音前には, 腹式呼吸を行い発声をしやすいようにした.

録音する音声は, 「高緊迫」と「低緊迫」のレベルで分ける. 発話する内容は Google Gemini に「緊迫感が出やすいアナウンスの文章」というプロンプトを入力し, 出力を得た. 例は, 「有毒ガスが検知されました. 指定された避難所へ向かってください.」「堤防が決壊しました. 呼吸を確保してください.」「大津波警報が発表されました. 決して外に出ないでください.」のような文章である. 「高緊迫」は津波警報時, 「低緊迫」は地震速報時のアナウンスを真似してもらった.

まず, 高緊迫音声を録音した音声はかなり早口でありながらも, 聞き取りやすい音声である. このような音声を 169 個録音した. 1 つの音声が約 5 秒なので, 全部で約 15 分の長さの緊迫感のある音声を入手した.

低緊迫音声を録音した音声はゆったりとしていながらも聞き取りやすい音声である. このような音声を 200 個録音した. 全部で約 20 分の長さの緊迫感のある音声を入手した.

3.4 録音音声の学習

録音した音声を学習させることで緊迫感のある音声を出力することができる. Style-Bert-VITS2 の学習は, GUI での操作が基本なため, 初心者の人でも扱いやすいのが特徴である. まず, 音声とその書き起こしデータが必要だが, 書き起こしを行うツールも用意されている. また, 長い音声の場合でも, スライスできるため, 1 つの音声があれば学習が可能である.

本研究の学習では, 録音した音声である「高緊迫」音声と「低緊迫」音声を学習に使用する. 合計で約 35 分の音声データである.

学習の設定として, バッチサイズは 1, エポック数は 100, ステップごとの保存は 1000 ごと, 学習のために音声を切り取る長さである segment size は VRAM 不足により, 初期設定 16384 から半分の 8192 とした.

3.5 Style-Bert-VITS2 の操作

Style-Bert-VITS2 の操作画面では, テキストの入力欄とテキストの削除ボタン, 追加ボタンがある. テキストごとにモデルの変更, 話者の変更, スタイルの変更が可能である. スタイルの強さや話速, 基本周波数の変更, イントネーションの変更などを操作できる. 音声合成ボタンを押すと左下にプレビュー操作（再生や音量調整, ダウンロード）が可能となる.

4 評価

まず, 従来の音声合成モデルと Style-Bert-VITS2 による音声合成で同じ文章を用いて緊迫感を付与したモデルと付与していないモデルの比較を行う.

さらに, 高緊迫の合成音声と低緊迫の合成音声のアクセントエネルギーを比較する.

次に, 「高緊迫」と「低緊迫」を付与した音声が正しくその緊迫度を表しているのかを評価する. 先行研究では, 音声から

緊急度を評価し, 優先順位を付ける手法が提案されている [14]. この論文では, MFCC を用いて音声の特徴量を抽出し, 複数の機械学習の手法を用いてどの手法が性能が良いのかを比較した実験である. 本実験では, この先行研究を参考に音声の特徴量を用いたランダムフォレスト識別器を作成し, 客観評価指標とする.

4.1 評価用音声と使用テキスト

比較検証および客観評価のために以下の音声データを作成した.

従来の音声合成モデルの音声は, JSUT 音声コーパスを学習させた音声を用いた. Style-Bert-VITS2 を用いた音声は「高緊迫」スタイルを付与した音声である. 文章はどちらも「津波警報が発表されました. 今すぐ逃げてください.」である.

緊迫度のある音声として, Style-Bert-VITS2 を用いて 2 種類のスタイルを持つ音声を作成した. それぞれ, 「高緊迫」スタイルを付与した音声と「低緊迫」スタイルを付与した音声である. アクセントエネルギーの比較を行う音声には, 「ただちに高台へ避難してください.」という文章を入力した.

判定システムに入力する音声に使用したテキストは, 避難を促す際に呼びかけられる文章を 100 文を用意した.

4.2 評価方法

従来の音声合成モデルの音声と Style-Bert-VITS2 で作成した音声の基本周波数のグラフを比較する. このグラフは縦軸が基本周波数, 横軸が時間を表しているため, 基本周波数と話速の差を比較することができる.

アクセントエネルギーの比較には, 「高緊迫」を付与した合成音声と「低緊迫」を付与した合成音声のそれぞれ 30 発話を入力とした. すべての音声から算出されたアクセントエネルギーの総数を求め, その差を比較する.

緊迫度の評価に使用するシステムにおける音声判定処理は, 特徴抽出フェーズと識別フェーズの 2 段階で構成される. まず入力された音声信号から, 声色や発話の強弱などの特徴を捉えるために MFCC (40 次元) を抽出する. 可変長の音声入力に対応するため, 抽出された MFCC の時間平均を算出し, 固定長のベクトルへと変換する. 次にこの特徴ベクトルをランダムフォレスト識別器に入力する. ランダムフォレストは 100 本の決定木に縁る多数決を行うことで, ノイズや外れ値の影響を低減しつつ, 「高緊迫」「低緊迫」「それ以外」かを判断する. その後, まず, 「高緊迫」と「低緊迫」の確率を足したものと「それ以外」の確率を比較する. 「それ以外」の確率が高ければ, 終了する. 「高緊迫」と「低緊迫」を足したものの確率が高ければ, 「高緊迫」と「低緊迫」で判断を行う. そのうち, 高い方がその判定となる.

この評価システムの正解データには, 実際に放送で使用された NHK のアナウンサーの音声を使用した. ここで, 避難を強く呼びかける「津波警報時」の音声を「高緊迫」, 「地震速報時」の音声を「低緊迫」, 「JSUT コーパス」の音声を「それ以外」と定義した.

評価システムの精度検証として, 録音した音声を入力した. その結果, 「高緊迫」を意図して録音した音声の判定は, 89 件が「高緊迫」と判定され, 10 件が「低緊迫」と判定された. 「それ以外」と判定された音声は 1 件であった. 「低緊迫」を意図して録音した音声の判定は, 3 件が低緊迫と判定され, 97 件が「それ以外」と判定された. 「高緊迫」と判定された音声は 0 件であった.

Table 1: 学習データの音声を入力した際の判定結果の内訳 (単位：件)

	高緊迫と判定	低緊迫と判定	それ以外と判定
高緊迫録音	89	10	1
低緊迫録音	0	3	97

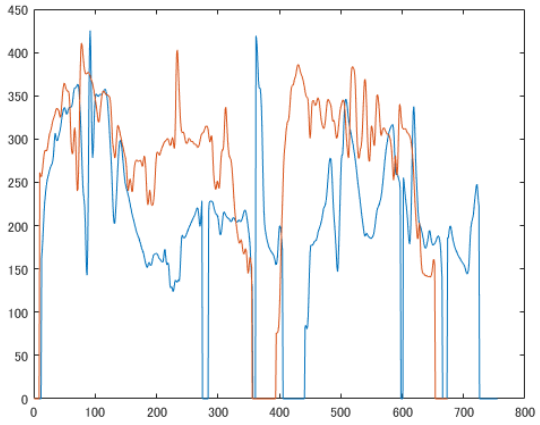

Fig.3: 緊迫感を付与したモデルと付与していないモデルの基本周波数の推移

Style-Bert-VITS2 に学習させた元の録音音声と, 生成した音声の識別結果を比較する. 「高緊迫 (または低緊迫)」として学習させた元音声 100 件と, 同様のスタイルを付与して生成した音声 100 件を評価システムに入力し, その分類結果の一致率を調査する. 元音声と合成音声の判定結果に大きな差がなければ, 提案手法により意図した緊迫度が付与され, 段階的な操作が可能であると判断する.

4.3 結果

従来の音声合成モデルの音声と Style-Bert-VITS2 で作成した音声の比較実験の結果を Fig. 3 に示す. 青い線が従来の音声合成モデルの音声である. オレンジ色の線が Style-Bert-VITS2 で作成した音声である. 緊迫感を付与したモデルの生成音声は, 青い線よりもオレンジの線が短いため, 全体的な話速が速いことがわかる. また, オレンジの線が最初の高さから大幅に下がっていないことから, 基本周波数の高推移がわかる. 日本語の標準的なイントネーション特徴である文末のピッチ降下が抑制され, 語尾まで高音を維持するという特徴が確認された. これは, 単なるパラメータ操作では表現できない緊迫感を聴取者に与えることが有効であることを示している.

Fig. 4 にアクセントエネルギーの比較の一例を示す. 赤い線が高緊迫を付与した合成音声で青い線が低緊迫を付与した合成音声である. 図の下は, アクセントの強さと時間を表したグラフである. 赤い方が面積が大きく, アクセントエネルギーが大きいことがわかる. また, 30 発話のアクセントエネルギーの総

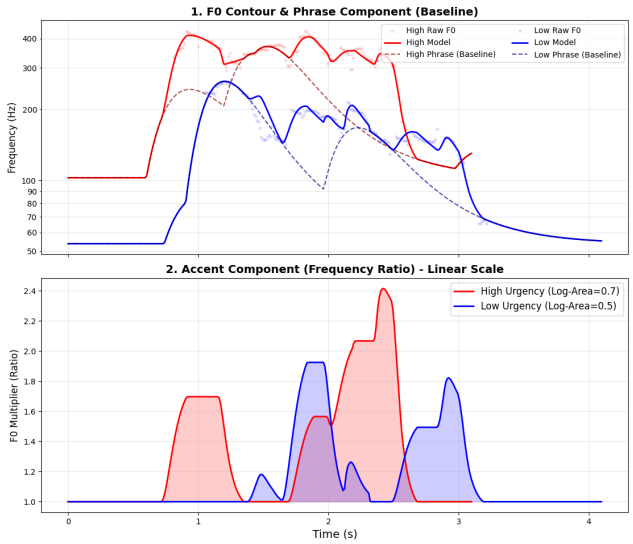

Fig.4: 高緊迫と低緊迫を付与した合成音声のアクセントエネルギーの例

Table 2: 識別器による「高緊迫」条件の録音音声と合成音声の判定結果の内訳（単位：件）

	高緊迫と判定	低緊迫と判定	それ以外と判定
録音音声	89	10	1
出力音声	97	3	0

Table 3: 識別器による「低緊迫」条件の録音音声と合成音声の判定結果の内訳（単位：件）

	高緊迫と判定	低緊迫と判定	それ以外と判定
録音音声	0	3	97
出力音声	1	6	93

量は, 高緊迫を付与した合成音声は 39.35, 低緊迫を付与した合成音声は 24.67 となった. 高緊迫を付与した合成音声は, 低緊迫を付与した合成音声よりも 1.59 倍アクセントエネルギーが大きいことがわかった.

Table 2, Table 3 に評価結果を示す. 「高緊迫」の元音声（録音音声）は 89 件が高緊迫と判定されたのに対し, 出力された「高緊迫」スタイルの合成音声は 97 件が高緊迫と判定された. 「低緊迫」の元音声（録音音声）は 97 件が「それ以外」として判定されたのに対し, 出力された「低緊迫」スタイルの合成音声は 93 件が「それ以外」と判定された.

Style-Bert-VITS2 を用いたスタイル変換は, 元の音声が持つ緊迫感の特徴を十分に学習し, 合成音声においてもその特徴を再現できているといえる.

4.4 考察

4.4.1 「低緊迫」音声が「それ以外」と判定されたこと

「低緊迫」を想定して録音した音声と出力した合成音声のどちらも 9 割以上が「それ以外」と判定されてしまった. これは, 録音音声がアナウンサーの地震速報の音声の特徴を真似できなかったといえる. 録音時, 「低緊迫」の特徴として, 「ニュースを読むような感じで落ち着いて話すように」と指示をした. ニュースを読むような感じとは異なるが, 緊迫感を少し感じるような音声の特徴を理解する必要があり, それを反映する必要があると考える.

一方で, 「低緊迫」を付与した合成音声も 9 割以上が「それ以外」と判定された. これは, 録音音声の特徴をそのまま付与したことを証明している. したがって, Style-Bert-VITS2 を用いたスタイルの付与, および操作が可能であると考える.

4.4.2 録音音声よりも合成音声の精度が高い

評価システムの結果, 「高緊迫」を付与した合成音声はその学習データである録音音声の評価よりも高い結果が得られた. これは, Style-Bert-VITS2 のスタイルベクトルの平均値の抽出が関係すると考える. 録音音声はノイズが含まれたり, 文章による緊迫感を強く付与できないことがある場合がある. しかし, スタイルベクトルではそれらを含めた特徴の平均値を出すため, 安定して緊迫感を付与することができたと考える.

4.4.3 合成音声で意図した緊迫感を付与できていないもの

「高緊迫」を付与した合成音声では 3 %, これは, 前節で述べたスタイルベクトルの平均値の抽出が関係すると考える. スタイルの平均値を抽出して, 音声を出力するため, 文章ごとにアクセントの強さが小さくなることがあるとすると, 緊迫感を十分に付与できない可能性があると考える.

4.4.4 理想的な緊迫音声と本手法の現時点

災害時における避難呼びかけ音声の理想は, 単に基本周波数が高いことや話速が速いことだけでなく, 発話者の「必死さ」や「危機感」といったパラ言語情報が聴取者に伝わり, 即座に避難行動を促すものである. 従来手法のような基本周波数や話速の操作ではなく, 人間が叫んでいるような声の震えや張りを再現することが求められる.

この理想に対し, 本手法は一定の到達水準に達したと考える. この成果は, Style-Bert-VITS2 の特徴であるスタイルベクトルが, アクセントエネルギーを維持したままの発話を生成でき, 話し方の特徴を再現することが可能であることを示しているからである.

Fig. 5 にスタイルベクトルの分布図を示す. 従来の音声合成モデルは学習データの平均的な特徴を出力する. よって, Fig. 5 の「Neutral」の特徴のみの音声が合成できるようになっている. Style-Bert-VITS2 では, スタイルベクトルを利用したスタイルごとの平均的な特徴を抽出することができる. これにより, Fig. 5 の Style__1 や Style__2 のような特徴を抽出する. この特徴というのは, アクセントの頻度や強度, テンポや間の取り方などである. これにより, 緊迫感のある音声を複数生成できると考えている.

4.4.5 学習データに依存することによる欠点

一方で, 本手法の制約として, モデルの表現力が学習データの質と量に強く依存する点が挙げられる. 基本周波数や話速の操作では学習データの範囲外も操作が可能であるのに対し, 本手法は学習データに含まれない未知のスタイルや感情表現を生成することは困難である. そのため, 高い緊迫感を生成するには, 相応の演技を行った教師データの収録が必要となり, データセット構築のコストが増大することが課題である.

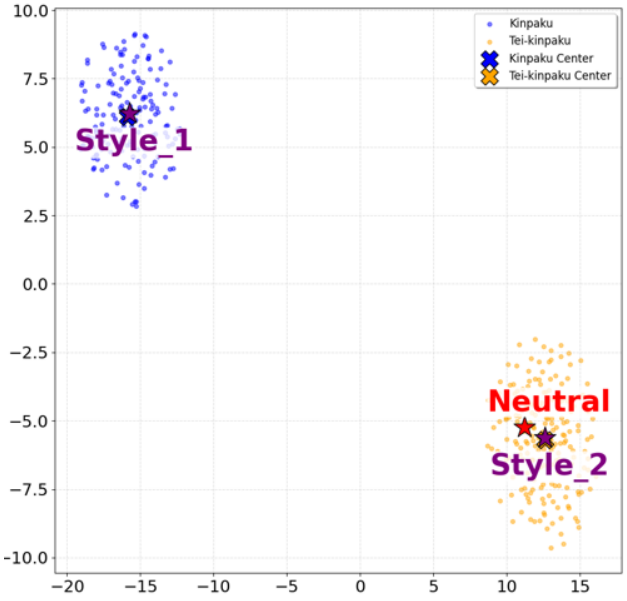

Fig.5: スタイルベクトルの分布図

4.4.6 「信頼できる人物」の声の音声合成

前節で述べたように, Style-Bert-VITS2 は学習データの声質を再現するという特徴を持つ. この「学習データの依存性」を肯定的に捉えれば, 特定の人物の声を模倣する能力が高いといえる.

そこで, 町長や地域のリーダーなど, 住民にとって「信頼できる人物」の音声を学習データとして用いることで, あたかもその本人が呼びかけているかのような音声を生成することが可能となる. これにより, その訴求力を利用し, 住民の避難を強く促すことができると考えられる.

また同時に, アナウンス担当者の安全確保を行うことができる. 事前にモデルを構築しておくことで, 担当者は安全を確保した状態で, 状況に応じた適切な緊迫度の音声を生成し, 発信することが可能となる.

以上のことから, 本研究で提案した手法は, 単に緊迫感のある音声を合成することだけでなく, 避難呼びかけにおける「アナウンス担当者の安全確保」と「住民への避難の訴求力の向上」を両立できる有効な手段である.

しかし, 音声の録音方法でも述べたように, 学習データは倍音成分の高域が大きい人が良いとした. よって, 「信頼できる人物」の音声を学習データとして用いる場合, 倍音成分の高域が大きい人であるかを確認する必要がある.

5 おわりに

本研究では, Style-Bert-VITS2 を用いて異なる緊迫感を複数生成可能な避難呼びかけ音声合成の手法の検討を行った. プロのアナウンサーの韻律特徴量を再現した音声として, 「高緊迫」と「低緊迫」の 2 種類の音声を録音し, そのデータをもとに学習を行い, 2 つのスタイルを生成し, 付与した音声を出力することを可能にした.

実際の放送音声で学習した判定モデルによる客観評価を行った. 「高緊迫」を付与した合成音声, 「低緊迫」を付与した合成音声の 2 種類を 100 件ずつ出力し, 入力とした. その結果, 高緊迫を意図して生成した合成音声は 97 %が「高緊迫」と判定され, 録音音声の 89 %を超える精度を確認した. また, 低緊迫を意図して生成した合成音声は 93 %, 録音音声の 97 %が「それ以外」と判定された. 「高緊迫」を付与した合成音声は意図した緊迫感の音声を生成することができたが, 「低緊迫」を付与した合成音声はそれができなかったと結論づける. しかし, これは Style-Bert-VITS2 のスタイル設定による音声の特徴の付与が可能であることを示唆している.

本手法の有用性は, 単なる緊迫感の制御に留まらない. 「信頼できる人物」を学習データとして事前にモデルを構築しておけば, 「アナウンス担当者の安全確保」と「住民への避難の訴求力の向上」を両立できる有効な手段となる.

参考文献

[1] 中島沙織. 「今すぐ逃げること!」という呼びかけ表現「能登半島地震における津波からの避難呼びかけ全国調査」から. 放送研究と調査, Vol. 74, No. 6, pp. 30–41, 2024.

[2] 日本放送協会. NHK 放送ガイドライン 2025 インターネットサービス必須業務化版. 日本放送協会, 10 2025.

[3] 中丸憲一, 中山準之助. 能登半島地震 緊急論考 「命を守る呼びかけ」「災害関連死」 過去の災害の教訓は生かされたのか. 放送研究と調査, Vol. 74, No. 4, pp. 2–31, 2024.

[4] Maori Kobayashi, Yasuhiro Hamada, and Masato Akagi. Acoustic features correlated to perceived urgency in evacuation announcements. *Speech Communication*, Vol. 139, pp. 22–34, 2022.

[5] 原田そら, more. 避難呼びかけ音声の持つ緊急性の分析と音声合成への適用の検討. 日本音響学会研究発表会講演論文集, pp. ROMBUNNO.2–Q–41, 2022.

[6] 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. *arXiv preprint arXiv:2006.04558*, 2020.

[7] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[8] 小口純矢, 金井郁也, 小田恭央, 齊藤剛史, 森勢将雅. Ita コーパス: パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価. 情報処理学会研究報告, Vol. 2021, pp. 1–6, 2021.

[9] 成澤修一, 峯松信明, 広瀬啓吉, 藤崎博也. 音声の基本周波数パターン生成過程モデルのパラメータ自動抽出法. 情報処理学会論文誌, Vol. 43, No. 7, pp. 2155–2168, 2002.

[10] litagin02. Style-bert-vits2. `https://github.com/litagin02/Style-Bert-VITS2`, 2024. GitHub repository.

[11] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

[12] Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin

Kim, Dohee Kong, and Sangjin Kim. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*, 2023.

[13] fishaudio. Bert-vits2. `https://github.com/fishaudio/Bert-VITS2`, 2023. Accessed: 2025-12-22.

[14] Naman Kumar Sinha. Voice stress analysis for emergency dispatchsystems: Enhancing real-time decision-makingthrough ai-driven speech processing. *IJCRT*, Vol. 13, No. 6, pp. 628–634, 2025.