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 *

☆原田 そら (木更津高専)，中田 亘 (東大・工)，高道 慎之介，
齋藤 佑樹 (東大院・情報理工)，齋藤 康之 (木更津高専)，猿渡 洋 (東大院・情報理工)

1 はじめに

洪水や津波などの災害時には，早期の避難行動が重要となる．意識モデルの事例研究などから，避難行動の喚起には「避難呼びかけ」の有効性が示唆されている．柿本らは，令和 2 年 7 月豪雨における避難行動意思決定モデルの推定から，避難情報や災害への備えより，避難呼びかけや置かれた状況を認知する刺激が有意としている [1]．この「呼びかけ」の効果を高める手法の一つに，避難情報を伝達する「避難呼びかけ音声」の適切な設計がある．

この設計指針の一つが「緊急性をいかに伝えるか」である．松本は，平成 30 年 7 月豪雨における事例から，(音声に限らず) 危機感が充分に伝わる伝達手法を検討する必要があるとしている [2]．この指針に関して Ofuji らと小林らはそれぞれ，音声の与える緊迫感が基本周波数 (F0) と話速に関連すること [3]，緊迫感の高い音声が避難行動を促すこと [4] を明らかにしている．音声の危機感については，具体的なパラメータ操作による変化を検討する研究もある [5,6]．また，これらに関係し伊藤らは，屋外の行政防災無線における合成音声の利便さに言及し，ピッチの加工は可聴性に影響を及ぼさないことを示唆している [7]．

以上を受け本稿では，音声特徴量加工による緊急性の操作を学習済みニューラル TTS (text-to-speech) に適用し，避難呼びかけ音声を合成可能な TTS システムを構築する．本稿ではまず，プロのアナウンサーによる緊急性の異なる音声を用いて，緊急性操作に寄与する音声特徴量を検討する．その後，その特徴量に対する加工を，緊急性の低い音声で学習されたニューラル TTS に施した結果を報告する．

2 緊急性に寄与する音声特徴量の解明とニューラル TTS への適用

2.1 緊急性に寄与する音声特徴量の解明

緊急性に寄与する特徴量を見つけるには，同一話者による緊急性の異なる音声群を利用することが好ましい．また，その音声は，所望の緊急性を適切に有さなければならない．これらを踏まえ本稿では，(1) 当該話者が発話訓練を十分に受けていること (例えば，アナウンサー [4])，(2) 緊急性の高い音声は，当該話者による実際の避難呼びかけ音声 (例えば，実際の自然災害において緊急避難を呼びかけた音声) であることを条件とし，音声群を収集する．収集する音声群は以下の 3 種類である．

平静時音声：緊急性が低く，災害と無関係の言語内容を発話している平静音声．

Table 1 各音声群の緊急性

音声群	内容の緊急性	音声の緊急性
平静時音声	平静	平静
加工対象音声	平静	緊急
緊急時音声	緊急	緊急

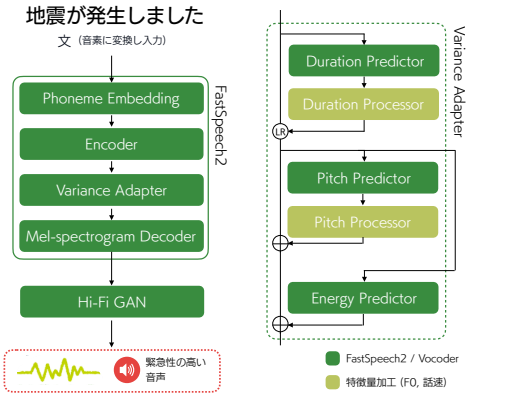

Fig. 1 FastSpeech2 [8] に対する緊急性操作

加工対象音声：緊急性は低いが，災害を避けるための後程の行動 (例えば，備え) を聴衆に求める音声．

緊急時音声：緊急性が高く，災害を避けるための即座の行動を聴衆に求める音声．

Table 1 は，これらの音声群の特徴である．平静時音声と緊急時音声の音声特徴量を比較し，その変換規則を加工対象音声に付与することで，緊急性操作に寄与する音声特徴量を検討する．

音声特徴量の候補として，本研究では F0 と話速 [3, 5] に着目する．変換規則を検討するため，各特徴量の分散，歪度，尖度を計算する．なお，本研究で扱う尖度は，正規分布の尖度を 0 とする値である．また，音声特徴量の変換として，2 群間の平均を揃える平均シフトと，2 群間の分布を揃えるヒストグラム等化を検討する．

2.2 学習済みニューラル TTS に対する緊急性操作

特徴量加工による緊急性操作を，学習済みニューラル TTS に適用する．利用するニューラル TTS モデルは FastSpeech2 [8] である．

FastSpeech2 における緊急性操作の手法を，Fig. 1 に示す．当該モデルは，推論時の中間表現として韻律特徴量を有するため，外部からの特徴量加工を反映した音声を合成できる．学習済みニューラル TTS を利用するのは，既存の音声合成システムに対し緊急性操作を追加できる可能性を模索するためである．ここで，他話者での適応可能性を探るために，このニューラル TTS は，2.1 節で用いる音声群の話者と異なる話者の読み上げ音声で学習されているものとする．

---

* Analysis of urgency of evacuation announcement speech and its application to text-to-speech, by HARADA, Sora (NIT, Kisarazu College), NAKATA, Wataru, TAKAMICHI, Shinnosuke, SAITO,Yuki (UTokyo), SAITO, Yasuyuki (NIT, Kisarazu College), and SARUWATARI, Hiroshi (UTokyo).

Table 2　各音声群の F0 統計量

音声群	平均 (Hz)	分散	歪度	尖度
平静時音声	128.69	1207.56	0.245	−0.842
緊急時音声	189.59	1572.80	−0.271	−0.337
加工対象音声	127.01	1076.41	0.221	−0.892

Table 3　各音声群のモーラ長統計量

音声群	平均 (s)	分散	歪度	尖度
平静時音声	0.119	3.084×10^{-3}	1.710	0.119
緊急時音声	0.106	1.940×10^{-3}	0.627	1.011
加工対象音声	0.112	2.548×10^{-3}	0.460	0.350

3　実験 1：各群音声における避難呼びかけ音声としての適性の評価

本節では，2.1 節にて定義した 3 種類の自然音声（各群音声）について，聴取実験によって各群の印象差異を検討し，2.1 節における定義と合致するか検討する．また平静時音声と F0 の近い加工対象音声について，避難呼びかけ音声としての適性を評価する．

3.1　実験条件

縁故法で集めた 18 名に協力を得て，音声の評価データを収集した．音声群は 2.1 節の定義に沿って，NHK の同一男性アナウンサーによる読み上げ音声を集めた．音声サンプルは．平静時音声を 32 個（計 7 分 52 秒），加工対象音声を 13 個（計 1 分 57 秒），緊急時音声を 97 個（計 5 分 9 秒）集めた．Table 2, 3 は，各群音声の F0，モーラ長の統計量について，Julius [9] と，WORLD [10] を Python 用に拡張した PyWorld [11] を用いて分析したものである．

評価では Ofuji らと坂本らの研究 [3, 6] を参考に，避難呼びかけ音声の適性を図るパラメータとして「聞き取りやすさ」「緊急性」「信頼性」を定義する．「緊急性」は，音声の与える緊急性の度合いを測る．「聞き取りやすさ」は，実験参加者がどの程度発話内容を聞き取れたかを測る．「信頼性」は，音声が伝達する情報の信頼性に関わる印象を測る．Table 4 は，評価に用いるスケールである．実験参加者は各項目ごとに 15 個の音声を聞き，適切と思う選択肢を選択する．

評価はオンラインで実施した．音声サンプルは，実験参加者毎に各群音声からランダムで計 15 個を提示した．受聴できる回数は制限せず，後戻りでの再評価は禁止した．評価値は実験参加者の評価した値を各群ごとで平均し，さらに全回答を平均して検討した．各項目は $P < 0.05$ を基準に有意差を判定した．

3.2　実験結果・考察

評価の平均値を Table 5 に示す．聞き取りやすさの緊急時音声・加工対象音声間を除き，各群音声間の評価評価に有意差が認められた．音声は，Table 1 に示した緊急性の性質により，聞き取りやすさ，緊急性，信頼性が変化した．項目別では，緊急性について，平静時音声，加工対象音声では知覚された緊急度が相対的に低く，2.1 節の定義と一致する結果を得た．知覚した緊急性の差は，緊急性のある項目数と増減関係が一致した．信頼性，聞き取りやすさは，緊急性と比べ音声群間の差が小さい．これは，収集音声は訓練された話者が発話していることに起因する可能性がある．

Table 4　参加者に提示した評価スケール

評価項目	質問文	スケール（5 段階）
緊急性	この音声を聴いて，どの程度緊急な状況だと思いましたか？	緊急性の低い状況 (1)〜緊急性の高い状況 (5)
聞き取りやすさ	この音声は，どの程度聞き取りやすかったですか？	聞き取りにくい (1)〜聞き取りやすい (5)
信頼性	この音声から伝達された情報や呼びかけは，どの程度信頼できますか？	全く信頼できない (1)〜とても信頼できる (5)

Table 5　実験 1 における評価結果

	平静時	緊急時	加工対象
信頼性	3.260	4.010	3.620
聞き取りやすさ	3.770	4.070	4.130
緊急性	1.690	4.460	2.220

4　実験 2：音声特徴量の加工による緊急性操作と実験的評価

3.2 節では，緊急時音声は緊急性が高く，加工対象音声と平静時音声は緊急性が相対的に低いことが示された．本節では，加工対象音声に平静時音声と緊急時音声における音声特徴量差を反映させ，緊急性，聞き取りやすさ，信頼性の変化を検討する．生じた緊急性の変化から，その操作が可能か考察する．予備実験として，2.1 節で候補とした F0，話速操作についての最適な手法を検討する．

4.1　加工手法

本節では，音声特徴量を反映するための加工手法について述べる．F0 は「平均シフト」と「ヒストグラム等化」の 2 手法により加工する．話速は「平均シフト」により加工する．

4.1.1　平均シフト (F0，話速)

F0 操作：平静時音声と，緊急時音声での F0 平均の倍率を用いて，加工対象音声の F0 平均値を緊急時音声に揃える．F0 操作は，PyWorld [11] により音声特徴量を抽出し，F0 のみを加工し，再合成する．

話速操作：平静時音声のモーラ長と，緊急時音声でのモーラ長の倍率を用いて，加工対象音声の話速を緊急時音声に揃える．話速操作は，PyWorld [11] により音声特徴量を抽出し，その特徴量に対しリサンプリング処理を施して再合成する．

Table 6，7 をみると，モーラ長の統計量は，平均シフトによって緊急時音声に近づいている．F0 は平均シフトによって平均値は近づいたが，高次統計量（尖度や歪度）までは反映できていない．

4.1.2　ヒストグラム等化 (F0)

F0 操作：緊急時音声の音声群（3.1 節にて収集）をもとに，緊急性の高い発話における，F0 についての累積分布関数を求める．このとき他話者への適用を考慮し（2.2 節），累積分布関数は，各フレームの F0 と緊急時音声群の F0 平均値との差を用いて求める．ヒストグラム等化では，操作したい音声の F0 平均値と累積密度関数を求め，フレームごとに加工元音声

Table 6　加工前後の F0 統計量

音声群	平均 (Hz)	分散	歪度	尖度
加工対象音声	127.01	1076.41	0.221	−0.892
平均シフト後（1.47 倍)	186.05	1572.80	0.166	−0.919
ヒストグラム等化後	188.76	1802.82	−0.109	−0.421

Table 7　加工前後のモーラ長統計量

音声群	平均 (s)	分散	歪度	尖度
加工対象音声	0.112	2.548×10^{-3}	0.460	0.350
平均シフト後 (0.89 倍)	0.101	2.245×10^{-3}	0.684	0.950

の F0 と，累積密度に対応する累積密度関数の F0 を入れ替え，操作前の F0 平均値を足す操作によって，F0 の分布を緊急時音声に揃える．

Table 6 をみると，ヒストグラム等化によって，F0 の高次統計量は緊急時音声に近づいている．すなわち F0 の変動パターンを近づけられている．

4.2　実験条件

縁故法で集めた 16 名に協力を得て，音声の評価データを収集した．評価は 3.1 節と同様の評価基準，評価手続，有意水準にて行った．実験は 3.1 節にて収集した加工対象音声を用いた．実験参加者には，加工対象音声について,「無操作」「F0 操作」「話速操作」のいずれかから，計 12 個の音声サンプルをランダムに提示した．定倍操作の倍率は，F0 は 4.1.1 節の定義を，話速は小笠原らの先行研究 [3] をもとに，以下の通り定めた．

平均シフト（F0）: 1.47 倍．

平均シフト（話速）:無加工（等速），1.125 倍，1.25 倍．

4.3　実験結果・考察

評価の平均値を Table 8，9 に示す．話速は，信頼性の無操作と 1.125 倍速，聞き取りやすさの各操作後音声間，F0 は全指標の操作後音声間で有意差が認められなかった．

4.3.1　F0 の操作に関する考察

F0 操作により緊急性が向上し，信頼性が低下した．この傾向は先行研究と一致する [3]．各操作手法間ではいずれの項目も有意差が示されず，緊急性操作に F0 変動パタン操作が F0 平均値操作より有意とした先行研究 [5] とは異なる結果を得た．先行研究では操作効果の比較に緊急時音声に相当する音声を用いている．操作元音声における F0 の分布傾向や平均値が異なることで，有意差が検出されなかった可能性がある．

4.3.2　話速の操作に関する考察

話速操作において，倍率の増加につれて緊急性の評価は高まった．1.125 倍速における聞き取りやすさの維持，信頼性の向上は，先行研究では見られない [3]．先行研究では話速を等速，高速，低速（±20%）にわけ操作した．各操作での倍率検討は行われておらず，本結果からは，話速の操作により，必ずしも信頼性は低下しないこと，適切な操作値を設定すれば，聞き取りやすさや信頼性を損なわずに緊急性を付与できる可能性も示唆された．

Table 8　実験 2 における評価結果 (F0 操作)

	無操作	F0 平均シフト	F0 ヒストグラム等化
信頼性	3.778	3.389	3.200
聞き取りやすさ	4.283	3.500	3.423
緊急性	2.578	3.339	3.500

Table 9　実験 2 における評価結果 (話速操作)

	無操作 (等速)	1.125 倍速	1.25 倍速
信頼性	3.433	3.750	3.367
聞き取りやすさ	3.833	3.967	3.267
緊急性	2.156	2.719	3.078

5　実験 3: 音声特徴量操作の組合せによる緊急性操作と実験的評価

本節では，加工対象音声に，最良の話速操作，F0 操作，また両方を組み合わせた操作手法（以下，F0-話速操作）を音声に適用し，緊急性付与のための最適な操作手法を検討する．F0 操作は先行研究を参考に [5] ヒストグラム等化を選択し，話速操作は 4 節の結果より緊急性が最も高い 1.25 倍速を適用する．

5.1　実験条件

縁故法で集めた 17 名に協力を得て，音声の評価データを収集した．評価は 3.1 節と同じ評価基準，評価手順をとり，3.1 節で収集した加工対象音声を用いた．実験参加者には，無操作，F0 操作，話速操作，F0-話速操作のいずれかから，計 12 個の音声サンプルをランダムに提示した．各項目は $P < 0.05$ を基準に有意差を判定した．$0.05 < P < 0.1$ をとる項目は，有意傾向ありと判定した[1]．

5.2　実験結果・考察

評価の平均値を Table 10 に示す．信頼性の話速操作と無操作，話速操作と F0 操作，聞き取りやすさの F0 操作と F0-話速操作，緊急性の F0 操作と無操作間で有意傾向があった．信頼性の F0 操作と F0-話速操作，聞き取りやすさの F0 操作と話速操作，緊急性の F0 操作と話速操作間で有意差が認められなかった．F0-話速操作は，知覚された緊急性が最も高く，3 手法の中では緊急性付与に最も適すると考えられる．

6　実験 4：学習済みニューラル TTS に対する緊急性操作と実験的評価

本節では，実験 3 までに検討した緊急性操作を学習済みニューラル TTS に適用し，既存の音声合成システムへの適用可能性を検討する．印象の比較には，無操作の推論音声と，F0-話速操作により加工した音声特徴量に基づく推論音声を用いる．F0-話速操作の加工手法は，5 節で用いたヒストグラム等化による操作（以下，ヒスト等化操作）に話者依存性がみられたため，4 節で同等効果が示された F0 平均シフトによ

[1] 4.3 節において有意差が検出された項目について，実験 4 では 0.05 をわずかに上回った．それらの項目も議論するため，有意差とは異なる有意傾向の区分をここでは設ける．

Table 10　実験 3 における評価結果

	無操作	F0 操作	話速 操作	F0-話速 操作
信頼性	4.157	3.294	3.725	3.275
聞き取りやすさ	4.216	3.176	3.412	2.765
緊急性	2.588	2.980	3.333	3.863

Table 11　実験 4 における評価結果

操作手法 (うち F0 操作)	無操作	F0-話速 操作 (平均 シフト 1.20 倍)	F0-話速 操作 (平均 シフト 1.47 倍)	F0-話速 操作 (ヒスト 等化)
信頼性	3.985	3.494	2.930	3.037
聞き取り やすさ	4.216	3.286	2.388	2.688
緊急性	3.280	3.565	3.612	3.698

る操作（以下，平均シフト操作）を加える．平均シフト操作は，4 節で用いた倍率での操作では，音声に不良が見られた[2]．そのため，より低い 1.20 倍による操作も加え，計 4 種類の操作による音声を比較する．

6.1　ニューラル TTS の実装

ニューラル TTS は 2.2 節（Fig. 1）に示した形で実装する．緊急性付与では，2.1 節で述べた特徴量の F0，音素継続長を推論過程で取り出す．これらを F0-話速操作により加工し，メルスペクトログラムを予測する．音声は HiFi-GAN [12] を用いて合成する．実装には第二著者による FastSpeech2 の公開実装[3]を利用した．ニューラル TTS モデルは，JSUT [13] で学習したものを，第一著者による ITA コーパス [14] 読み上げ音声で転移学習したモデルを用いた．

6.2　実験条件

クラウドソーシングサービス Lancers を使用し，100 名から音声の評価データを収集した．評価は 3.1 節と同じ評価基準，評価手順，有意水準にて行った．実験参加者には本節冒頭に述べた 4 種類の音声から，計 24 個の音声サンプルをランダムで提示した．

6.3　実験結果・考察

評価の平均値を Table 11 に示す．信頼性はヒスト等化操作と平均シフト（1.47 倍）操作間，緊急性は各平均シフト操作間，ヒスト等化操作と F0 平均シフト操作（1.47 倍）間で有意差が認められなかった．無操作と比較し，緊急性について，操作後の各音声で緊急性が向上している．F0-話速操作による緊急性操作は，ニューラル TTS においても有効であることが示唆された．聞き取りやすさと信頼性は，操作後の音声において低下している．特に聞き取りやすさでは，平均シフト操作（1.47 倍）とヒスト等化操作において，評価スケールの中間値を下回る結果を示した．要因には，(1) 平均シフト (1.47 倍) について，本節冒頭で述べたとおり音声不良があった，(2) ヒスト等化操作に，話者依存性がみられた，の 2 点が推測できる．これらを改善すれば，評価は向上する可能性がある．

[2]具体的には，推論結果の音声にかすれが生じた．平静な音声読み上げデータにより学習したことで，高い F0 をとる音声の学習データが不足していたことに起因する可能性がある．

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

7　まとめ

本研究では，音声特徴量加工による緊急性操作について検討し，効果が特に認められた操作手法を緊急性の低い音声で学習されたニューラル TTS に実装した．これらに対し，実験的評価結果から，(1) 緊急性の低い音声，緊急性の高い音声には，音声特徴量の差がある，(2) F0 操作と話速操作は緊急性の付与に有効である，(3) 各々の単独操作より，両手法を組合せた操作が緊急性付与において有効である，(4) 学習済みニューラル TTS においても，緊急性操作は有効である，の 4 点が示唆された．今後の課題として，6.3 節に述べた，ニューラル TTS における緊急性操作手法の改善があるほか，本稿では音声聴取による避難行動への具体的な影響まで検討がなされていない．今後はこの 2 点について，研究を進める．

謝辞: 本研究の一部は，科研費 21H04900 の助成を受け実施した．

参考文献

[1] 柿本　竜治，吉田　護, “状況認識を考慮した令和 2 年 7 月豪雨時の避難行動意思決定モデルの推定,” **土木学会論文集** D3（土木計画学）, vol. 78, no. 2, pp. 45–57, 2022.

[2] 松本　浩司, “『危機感』は伝わったのか～豪雨のダム大量放流」（時論公論）,” https://www.nhk.or.jp/kaisetsu-blog/100/301682.html（参照：2022-07-21）.

[3] Ofuji Kenta and Ogasawara Naomi, “Verbal disaster warnings and perceived intelligibility, reliability, and urgency: The effects of voice gender, fundamental frequency, and speaking rate,” *Acoustical Science and Technology*, vol. 39, no. 2, pp. 56–65, 2018.

[4] 小林　まおり，赤木　正人, “避難呼びかけ音声の心理的評価,” **日本音響学会誌**, vol. 74, no. 12, pp. 633–640, 2018.

[5] 小林　まおり 他, “音声の緊迫感に影響する音響特徴の検討,” **信学技報**, vol. 118, no. 149, pp. 79–84, 2018.

[6] 坂本　湧暉, “避難誘導音声における緊迫感の操作を行うための音響特徴に関する研究,” **北陸先端科学技術大学院大学**（JAIST）2020 **年度修士論文**, 2021.

[7] 伊藤　憲三，田村　幸子, “防災行政無線に音声合成を用いるための最適制御法に関する検討,” **公立大学法人岩手県立大学地域政策研究センター 地域協働研究研究成果報告集** 1【平成 24 **年度教員提案型/地域提案型・前期**】, pp. 32–33, 2013.

[8] Yi Ren *et al.*, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” vol. arXiv 2006.04558, 2020.

[9] A. Lee *et al.*, “Julius – an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.

[10] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[11] “Pyworld - a python wrapper of world vocoder,” https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder（参照：2022-07-21）.

[12] Jungil Kong *et al.*, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” vol. arXiv 2010.05646, 2020.

[13] R. Sonobe *et al.*, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” vol. arXiv 1711.00354, 2017.

[14] 小口　純矢 他, “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-MUS-131, no. 31, pp. 1–6, 2021.