第 24 回日本バーチャルリアリティ学会大会論文集（2019 年 9 月）

自信声フィードバックによる緊張緩和手法の提案：クラウドソーシングを利用した自信声加工パラメータの推定

A Proposal of Alleviating Tension by Confidence Voice Feedback:
Estimating Confidence Voice Conversion Parameters Using Crowdsourcing

成瀬加菜 [1)]，吉田成朗 [1),2)]，高道慎之介 [1)]，鳴海拓志 [1),2)]，谷川智洋 [1)]，廣瀬通孝 [1)]
Kana NARUSE, Shigeo YOSHIDA, Shinnosuke TAKAMICHI, Takuji NARUMI, Tomohiro TANIKAWA and Michitaka HIROSE

1) 東京大学大学院情報理工学系研究科（〒 113-0033 東京都文京区本郷 7-3-1）
2) 国立研究開発法人科学技術振興機構さきがけ（〒 332-0012 埼玉県川口市本町 4-1-8）

概要: 口頭発表におけるパフォーマンス低下の一因として緊張感が挙げられる．本研究では発話音声を加工し，自信を持ったように聞こえる声をフィードバックすることで，使用者の自信を生起し緊張感を和らげるシステムを提案する．システム開発にあたり，声の自信知覚に関するデータをクラウドソーシングで収集し，音声加工パラメータから声の自信度スコアを推定するモデルを構築した．

キーワード: 変換聴覚フィードバック，緊張感，クラウドソーシング

1. 研究の背景と目的

円滑な発話はコミュニケーションに不可欠であり，人同士の理解を深める助けとなる．しかし，聴衆を前にして話す場面は緊張感を喚起し，スピーチパフォーマンスを低下させることが知られている [1]．

本研究では，変換聴覚フィードバック (Altered Auditory Feedback, AAF) を利用し，話者の緊張感を緩和することで発話を円滑化するシステムを提案する (図 1)．AAF とは，マイクへの入力音声を加工し，ヘッドホンなどを介して話者自身へフィードバックする工学的手法であり，話者の感情に影響を与えることが知られている [2]．本研究では入力音声を自信のある声に加工し，話者へリアルタイムにフィードバックすることで緊張感を緩和することを考える．しかし，声の加工パラメータの種類は多岐に渡るため，最適な値の特定は困難である．

そこで本研究では，加工音声の印象に関する評価データをクラウドソーシングで収集し，加工パラメータから声の自信度スコアを推定するモデルを構築する．これは複雑な人間の知覚のモデル化に利用される手法である [3]．本稿ではこの手法を声の知覚に適用し，自信のある声への加工に最適なパラメータの特定が可能であるか検証する．

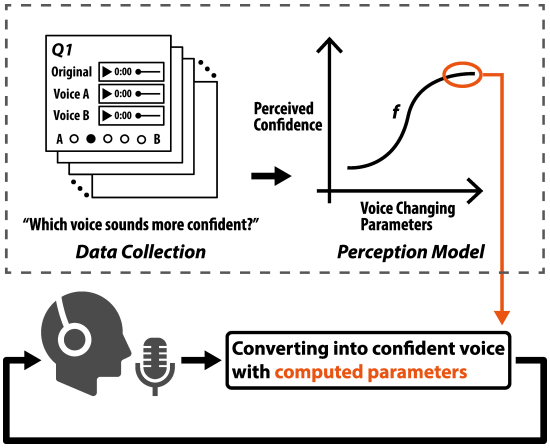

図 1: 提案手法の概念図

2. 関連研究

2.1 変換聴覚フィードバックと感情

AAF は話者の感情に作用することが知られている．Aucouturier et al. [2] は，入力音声を喜び，悲しみ，恐れの各感情を表現した声へと加工するプラットフォームを開発し，文章朗読中の実験参加者の声を徐々に加工してフィードバックすることで効果を検証した．結果，参加者にはフィードバック音声に表現された感情が実際に喚起されることが確認された．本研究では AAF の感情への作用を利用し，発話時の緊張緩和とスピーチパフォーマンスの向上を目的としたシステムの開発を目指す．

2.2 工学的手法による感情制御

ユーザの感情の鎮静化を目的としたシステムの例を紹介する．EmotionCheck[4] は，ユーザの心拍と同期したリズムの振動を手首に伝える腕時計型のデバイスである．検証実験では，実際の心拍よりも遅い振動をフィードバックすることで，参加者の緊張感が緩和されることが示された．

EmotionCheck が触覚刺激を用いているのに対し，聴覚刺激を用いたシステムも存在する．Costa et al. [5] は口論中の実験参加者の一方に対し，落ち着いたように加工した声をフィードバックする実験を行った．結果，加工音声をフィードバックされた参加者だけでなく，口論相手の感情も鎮静化されたことが示された．このことから，声の知覚が話者の緊張感の緩和に有効であることが示唆される．

Costa et al. の実験では，音声加工パラメータの値は声の音響特性と感情に関する先行研究に基づいて決定されている．しかし，感情に関係する音響パラメータには声の高さ，抑揚，周波数分布など多くの種類があり [6]，更にそれらと感情との関係は一対一ではない [7] ため，最適な加工パラメータの決定は難しい．そこで本研究では，クラウドソーシングで収集した音声評価データに基づき，AAF を用いた緊張緩和システムの構築に最適な加工パラメータを算出することを目指す．

3. 声の自信度スコア推定モデル

3.1 概要

入力音声をより自信を持ったように聞こえる声へと加工するためには，最適な加工パラメータの推定が必要となる．本研究では，入力された加工パラメータから加工音声の自信度スコアを推定するモデルを構築する．モデルの構築にあたっては，クラウドソーシングによって収集された，音声の自信度知覚に関する評価データを利用する．声の音響パラメータと声から受ける印象との関係は言語や国，文化圏などに依存するため [8]，本研究では音声評価の実験参加者を日本人に限定し，日本語音声の知覚に関する評価を行った．データの収集には国内クラウドソーシングサイト Lancers[1] を利用し，各加工音声がどの程度自信を持ったように聞こえるか実験参加者に評価してもらった．

3.2 クラウドソーシング実験

3.2.1 実験参加者

実験参加者は日本人成人 300 名 (男性 173 名，女性 127 名，平均年齢 41.29 ± 9.85 歳) で，クラウドソーシングサイト Lancers 上で募集した．実験中，参加者はヘッドホンやイヤホンを装着するものとした．実験参加の謝礼として，各参加者には Lancers を介して 100 円を支払った．

3.2.2 音声評価タスク

実験参加者はランダムなパラメータ値で加工された音声を Web サイト (図 2) 上で試聴し，それらの音声がどの程度自信を持って話しているように聞こえるか評価した．加工前のオリジナル音声には，20 代から 40 代の日本人話者 6 名 (男性 3 名，女性 3 名) が日本語の短文を読み上げたものを使用した．このオリジナル音声は国立情報学研究所 (NII) が配布している日本語の音声コーパスである JNAS[2] に収録されているものからの抜粋であり，サンプリング周波数は 16kHz であった．本実験におけるコーパスの使用に際しては，事前に NII からの認可を受けた．

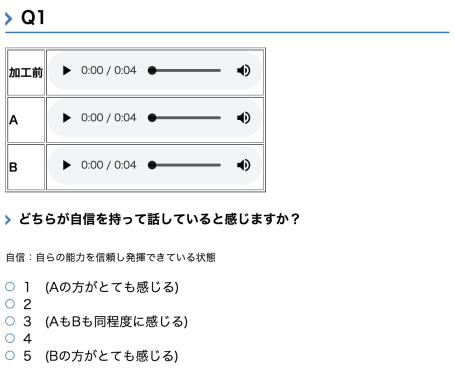

図 2: 評価タスクに用いた Web サイト（抜粋）

オリジナル音声はピッチシフト，フォルマントシフト，ピッチ範囲，スペクトル傾斜の各パラメータについて加工された．スペクトル傾斜の操作には以下の式を用いた．

$$A_{new} = \frac{A}{(1 + f/100)^x} \tag{1}$$

ここで，A と A_{new} はそれぞれ操作前と後の音声信号の振幅を，f は周波数を表す．指数 x が正の値であるときは周波数が大きくなるほど振幅は減衰していき，反対に x が負の値であるときは増幅していく．また，x の絶対値を大きくするほどスペクトル傾斜は急になる．今回は指数 x にランダムな数値を代入することでオリジナル音声のスペクトル傾斜を操作した．

パラメータの値はピッチシフト -200〜200 セント，フォルマントシフト 0.9〜1.1，ピッチ範囲 0.5〜1.5，スペクトル傾斜の指数 -0.2〜0.2 の範囲でランダムに決定した．音声加工には音声分析・合成ソフトウェアである Praat[9] を利用した．

参加者はオリジナル音声と 2 種類の加工音声を試聴し，「どちらが自信を持って話していると感じますか？」という質問に基づき，加工音声の相対評価を対比較により行った．加工音声は 2 種類の音声をそれぞれ A，B として，1(A の方がとても感じる) から 5(B の方がとても感じる) までの 5 段階リッカート尺度によって評価された．

参加者は 1 セットの評価タスクの中で話者 6 名分の音声を 1 対ずつ評価した．評価タスクは参加者 1 名につき 2 セットずつ課された．また，不正な回答データを解析に含めるのを防ぐため，2 種類の加工音声の順番を入れ替えた評価も 1 セットあたり 6 問ずつ含めた．したがって，1 セットの評価タスクは 12 問の音声評価から構成され，参加者は 1 名あたり 12 問のダミー質問を含む 24 問に回答した．音声の再生順序は特に指定せず，同一の音声を繰り返し試聴することも可能とした．

3.2.3 音声評価データの解析

クラウドソーシングで収集した 2 種類の加工音声の相対評価データから，各変換音声の自信度スコアの絶対値を推定する．スコアの推定には，以下の数式における $E(\boldsymbol{y})$ の最

[1] https://www.lancers.jp/
[2] http://research.nii.ac.jp/src/en/JNAS.html

小化問題として定式化する Koyama et al. [10] の手法を用いる．

$$E(\boldsymbol{y}) = E_{relative}(\boldsymbol{y}) + \omega E_{continuous}(\boldsymbol{y}) \quad (2)$$

$$E_{relative}(\boldsymbol{y}) = \sum_{(i,j)\in \boldsymbol{P}} \|y_i - y_j - d_{i,j}\|^2 \quad (3)$$

$$E_{continuous}(\boldsymbol{y}) = \sum_{i\in \boldsymbol{Q}} \left\| y_i - \sum_{j\in \boldsymbol{Q}(j\neq i)} \frac{y_j}{\alpha(i) D(y_i, y_j)} \right\|^2 \quad (4)$$

$$\alpha(i) = \sum_{j\in \boldsymbol{Q}(j\neq i)} \frac{1}{D(y_i, y_j)} \quad (5)$$

$$D(y_i, y_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \quad (6)$$

ここで $E_{relative}(\boldsymbol{y})$ はクラウドソーシングで得られた相対評価スコアに基づく項である．P はタスク中で評価された加工音声の対の集合を表す．$d_{i,j}$ はリッカート尺度による評価に対応する項であり，リッカート尺度における (1, 2, 3, 4, 5) はそれぞれ (1, 0.5, 0, -0.5, -1) となる．$E_{continuous}(\boldsymbol{y})$ は 4 種類の加工パラメータ $\boldsymbol{x}$ の値が近い加工音声同士の自信度スコアを近いものに補正する項である．ω は重み付けであり，本実験では $\omega = 5$ とした．また，Q は評価に用いた加工音声全体の集合である．

推定された各加工音声の自信度スコアの絶対値 $\boldsymbol{y}$ に対して RBF 補間を行い，変換パラメータを入力として自信度スコアを推定する関数を求めた．

$$f(\boldsymbol{x}) = \sum_{i\in \boldsymbol{Q}} \omega_i \phi(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \quad (7)$$

$$\min_{\boldsymbol{\omega}} \sum_{i\in \boldsymbol{Q}} \left\| \sum_{j\in \boldsymbol{Q}} \omega_j \phi(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) - y_i \right\|^2 + \lambda \|\boldsymbol{\omega}\|^2 \quad (8)$$

ここで定数 λ は $\lambda = 0.5$ とし，放射基底関数 ϕ は $\phi(\boldsymbol{r}) = \boldsymbol{r}$ とした．

3.3 モデルの評価

実験参加者 300 名分の評価データのうち，最後まで回答されていないもの，12 個のダミー質問との比較により評価の正当性が不十分と判断されたもの，回答時間が著しく短かったものを除いた 190 名分のデータを解析に用いた．

3.3.1 重回帰分析による評価

構築された音声の自信度スコア推定モデルを重回帰分析により評価した．重回帰分析にあたっては，4 種類の音声加工パラメータ (ピッチシフト，フォルマントシフト，ピッチ範囲，スペクトル傾斜) を独立変数，モデルによって推定された自信度スコアを従属変数とした．

表 1 は 6 名の話者における重回帰分析の結果である．決定係数は Adjusted R^2=0.396 で，$F(4, 4555) = 748.3$，$p < .01$ であった．各パラメータにおける t 値の比較より，ピッチ範囲が声の自信度知覚に最も大きく影響したと考える．この結果から，抑揚の大きい声が自信を持ったように知覚されるということが示唆された．

表 1: モデルの重回帰分析の結果

	偏回帰係数	t	p
(定数)	0.4993	2154.377	<.001
フォルマントシフト	0.0080	10.138	<.001
ピッチシフト	0.0029	3.616	<.001
ピッチ範囲	0.0374	46.957	<.001
スペクトル傾斜	-0.0224	-27.403	<.001

最も影響の大きかったピッチ範囲については，基本周波数の標準偏差が小さく単調な声ほど不安な声として知覚される [11] ということが知られている．したがって，声から知覚される不安の少なさが自信として知覚されたと考える．

3.3.2 モデルに基づく加工音声の印象評価

音声の自信度スコア推定モデルが実際の人間の知覚と合致するか検証するため，加工音声の評価実験を行った．

実験参加者は 20 代から 30 代の 4 名 (うち女性 2 名) だった．参加者はモデルによる自信度スコア推定の結果に基づいて加工された音声について，それぞれがどの程度自信を持って話しているように聞こえるか評価した．加工前のオリジナル音声には，3.2 節で用いた JNAS 収録の読み上げ音声 6 名分に，JNAS 収録の音声と同じ短文を参加者自身が読み上げた音声 6 種類を追加した計 12 種類の音声を使用した．

参加者自身による読み上げ音声はサンプリング周波数 44.1kHz で録音された後，発話者の自己聴取音再現のための処理がなされた．これは本研究での音声加工を AAF システムに適用した際，マイクによって観測された音声と話者の自己認知音との差異に由来する違和感が生じるのを防ぐためである．今回は落合ら [12] の手法を参考に，録音された音声にカットオフ周波数 800Hz の 4 次バターワースフィルタを用いることで擬似的な骨導音を生成し，元の音声と 1 対 1 の割合でミキシングすることで自己聴取音を再現した．参加者自身による読み上げ音声についてはこの擬似自己聴取音をオリジナル音声とし，ピッチやフォルマントなどに関する音声加工を後から行った．

加工パラメータの値は，ランダムに 50000 組生成したパラメータから自信度スコアを推定した際，スコアが最大値，中央値，最小値となった 3 組を用いた．

参加者は 3 種類の加工音声にオリジナル音声を加えた 4 種類の音声を試聴し，各音声から知覚される自信を 1(自信を持って話しているとは言えない) から 5(とても自信を持って話している) までの 5 段階リッカート尺度で評価した．参加者は 1 名あたり，4 種類の音声を 1 組とした評価タスクを 12 組 (うち自身の声 6 組) 分，計 48 種類の音声を評価した．

音声は「他者声」と「自分声」の 2 群に分け，それらを加工パラメータに基づき「加工前」「自信度高 (推定スコア

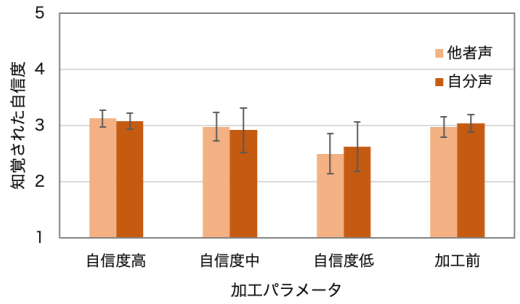

図 3: 音声から知覚された自信度の評価値

最大値)」「自信度中 (推定スコア中央値)」「自信度低 (推定スコア最小値)」の 4 群へと更に分類し，全 8 群それぞれについて参加者ごとにリッカート尺度による評価値の平均を求めた．その後，算出した参加者ごとの平均について，参加者 4 名分の平均と標準誤差を求めた．

結果を図 3 に示す．各群の平均値を比較すると，評価者自身の声であるかどうかに拘らず，推定スコアの低い加工音声はオリジナル音声よりも自信のないように知覚される傾向が見られた．一方，推定スコアの高い加工音声は，実際の評価値においてオリジナル音声との間に大きな差は見られなかった．入力音声を自信のある声へと加工するシステムの実現には，より実際に知覚される自信度の高くなる加工パラメータの特定が必要だと考える．

しかし全体として，構築されたモデルによる推定スコアが高いほど実際に自信のある声として知覚される傾向があり，モデルと人間の知覚が概ね合致している可能性が示唆された．

4. 結論と今後の展望

本研究では，口頭発表における緊張感の緩和を目的とし，発話音声を自信を持ったように聞こえる声へと加工した上で話者にフィードバックするシステムを提案した．最適な音声加工パラメータを特定するため，クラウドソーシングで収集した評価データに基づき，パラメータから音声の自信度スコアを推定するモデルを構築した．

音声評価実験の結果，モデルによる自信度推定が人間の知覚と合致している可能性は示唆された．しかし，今回推定された最適なパラメータによる加工音声は，実際に知覚される自信度においてオリジナル音声と変わらない結果であった．そのため，今後はクラウドソーシングで評価する音声の加工パラメータの種類や範囲を再検討するほか，話者の性別や発話音声の特性などを考慮しつつ最適な加工パラメータを探索し，加工音声から知覚される自信度の向上を目指す．

また，今回構築したモデルを AAF システムに適用し，自信のある声をユーザへフィードバックするシステムを実装する．音声フィードバックがユーザの緊張感やスピーチパフォーマンスに及ぼす効果を検証しつつ，コミュニケーションの円滑化を可能にするシステムの開発を目指す．

参考文献

[1] 岩田彩香, 川井智理, 齋藤順一, 嶋大樹, 熊野宏昭：社交不安傾向によるスピーチ場面でのパフォーマンス低下に関する検討，早稲田大学臨床心理学研究，Vol. 15, No. 1, pp. 53-63 (2015).

[2] Aucouturier, J. J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. : Covert digital manipulation of vocal emotion alter speakers ’ emotional states in a congruent direction, Proceedings of the National Academy of Sciences (2016).

[3] 川瀬佑司, 吉田成朗, 鳴海拓志, 上田祥代, 池田まさみ, 渡邊淳司, 谷川智洋, 川本哲也, 廣瀬通孝: Mob Scene Filter : 顔部位の形状・位置変形を利用した他人顔変換手法, 日本バーチャルリアリティ学会論文誌, Vol.21 No.3 (2016)

[4] Costa, J., Adams, A., Jung , M., Guimbretière, F. and Choudhury, T. : EmotionCheck: Leveraging Bodily Signals and False Feedback to Regulate Our Emotions, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ’16), 758-769 (2016).

[5] Costa, J., Jung, M., Czerwinski, M., Guimbretière, F., Le, T. and Choudhury, T. : Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-perception, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18) Article 631, 13 pages (2018).

[6] Juslin, P. and Scherer, K. : Vocal expression of affect, The new handbook of methods in nonverbal behavior research, 65-135 (2005).

[7] Gobl, C. and Chasaide, A. : The role of voice quality in communicating emotion, mood and attitude, Speech communication 40, 1-2, 189-212 (2003).

[8] Xu, A., Leung, S. and Lee, A. : Universal vs. language-specific aspects in human vocal attractiveness: An investigation towards Japanese native listeners ’ perceptual pattern, Proceedings of Meetings on Acoustics 172ASA, Vol. 29. ASA, 060001 (2016).

[9] Boersma, P., et al. : Praat, a system for doing phonetics by computer, Glot international, Vol. 5 (2002).

[10] Koyama, Y., Sakamoto, D. and Igarashi, T. : Crowd-powered Parameter Analysis for Visual Design Exploration, Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, 65-74 (2014).

[11] Banse, R. and Scherer, K. R. : Acoustic profiles in vocal emotion expression，Journal of personality and social psychology，70(3), 614 (1996).

[12] 落合裕貴, 丹治寛樹, 村上隆啓, 松本直樹 : 低域通過フィルタを用いた気導音からの自己聴取音の再現，研究報告音楽情報科学 (MUS)，2019(66), 1-5 (2019).