細粒度非言語的表現制御による自然な感情テキスト音声合成の実現 *

☆周王子茜, アトマジャバグストリ, サクティサクリアニ
奈良先端科学技術大学院大学

1 はじめに

現在の感情的なテキスト音声合成モデルは、言語的なプロソディの制御には成功していますが、人間らしい本物の感情表現に不可欠な非言語的発声(NVs) [6]を見落としがちです。近年、いくつかのNVデータセット [5]が登場していますが、それらは往々にして高品質で細粒度なアノテーションを欠いており、その結果、モデルがNV生成を正確に制御する能力が制限されています。この制約に対処するため、我々は細粒度な非言語表現合成のための新規なアプローチを提案します。具体的には、我々は既存のコーパスから女性のNV発話を選別し再処理するとともに、NVのタイプ、頻度、持続時間を符号化するための新しいタグ付けアノテーションスキームを開発しました。これに基づき、我々は感情TTSのベンチマークを構築し、その有効性を実証します。このNVアプローチは、自然さとのわずかなトレードオフと引き換えに、表現力と感情認識の精度を著しく向上させます。感情分析の結果、happyやfearに特に有効であり、sadの伝達においてはほぼ完璧に機能することが確認されました。

2 提案手法

2.1 細粒度非言語表現データの構築

Table 1: Differences between the two designs.

Coarse-grained	Fine-grained
<crying>	<(crying) wuuuuu whep>
style-level tag	Style-level tag
-	Vocalization type and control
-	Frequency and duration control

提案のデータセットは、EARSコーパス [7] にアノテーションとフィルタリングを施して構築されました。本節では、元のデータソースと、最終データセットにおける非言語音 (NVs) および感情タグの統計について詳述します。我々は、高品質な音声コーパスであるEARSデータセットを利用しました。本研究では、女性話者のNVsを抽出し、**laughter-open**、**laughter-closed**、**cheering**、**yelling**、**crying**、**screaming**の6つの明確なカテゴリに焦点を当てました。

Table 2: The specific nonverbal transcript table.

Category	Transcript
Cheering	'Wo ho', 'Yo'
Yelling	'Hey'
Laughter-open	'Ha'
Laughter-closed	'Ha'
Crying	'Whep', 'Wuu', 'Sneeze'
Screaming	'Ah'

元の連続的な音声ファイルは、無音の閾値に基づいてpydubライブラリを用いて360の録音から739個の個別発話（約2〜6秒）にセグメント化されました。セグメント化後、初期の書き起こしにはWhisperモデルを使用し、その後に手動検証を行い、細粒度な非言語書き起こしを確立しました。Table 1に詳述するように,従来の粗粒度タグ付け（例: <laugh>）とは異なり、我々のシステムは精密な制御のために特定のタグ構造を使用します。

離散的発声: 音節の繰り返し（例: <(Laughter-open) ha ha ha>）により、離散的な音（'ha'、'whep'）の頻度を制御し、非言語のタイプを区別します。

連続的発声: 最後の文字を繰り返すこと（例: 'wuu' の各追加 'u' が約 0.2 秒を追加）により、連続的な音（'Wo ho'、'wuu'）の持続時間を制御し、精密な時間制御を可能にします（詳細は表 2を参照）。

2.2 非言語感情TTS

我々は、バックボーンモデルとして Grad-TTS [2] を採用し、連続的な覚醒度と価 (A-V) ラベル [1] を活用する感情エンコーダによって機能を拡張しました。非言語的発声 (NV) アノテーションを効果的に処理するため、我々は、スタイル、離散ユニット、持続時間の各パーサーから構成される NV プロセッサ（図 1）をテキストパイプラインに導入しました。本プロセッサは、<(crying) wuuuuu whep> のような転写入力から、NVs のタイプ、頻度、および持続時間を構造的に符号化します。この構造化アプローチにより、感情 Grad-TTS は、言語音声と、極めて表現豊かで感情的

*Granular Control of Nonverbal Expressions for Achieving Natural Emotional Text-to-Speech System

に整合性の高い NV の両方を合成可能となりました。

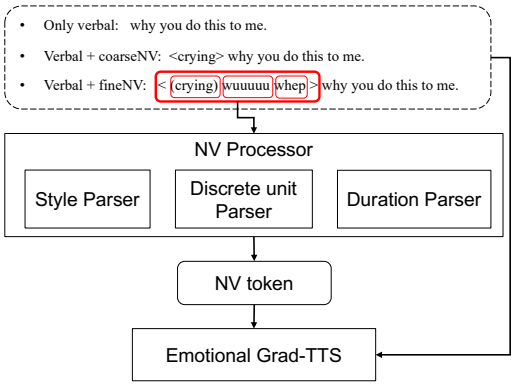


Fig. 1: The pipeline of non-verbal processing

3 実験

本研究では、感情豊かな言語音声合成を実現するため、混合英語女性話者データセット [3, 4]と連続的な感情ラベルを用いてモデルを学習させました。NV合成の比較対象として、提案手法である細粒度NVデータと、既存の粗粒度NVTTSコーパス [5]を使用しました。

評価設定: 15名の参加者による主観評価を実施し、(1) 言語のみ、(2) 言語＋粗粒度NV、(3) 言語＋細粒度NV（提案手法）の3設計を比較しました。評価指標には、自然さ (nMOS)、表現力 (eMOS)、および感情認識精度を用いました。テキスト入力には「What did you do」のような言語的に曖昧な20文章を選定し、参加者は計60サンプルを評価しました。さらに、幸福 (Happy) と悲しみ (Sad) におけるNV選好テストも実施しました。

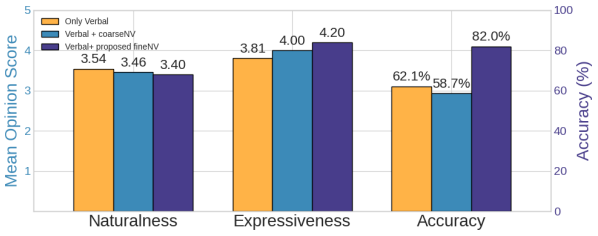
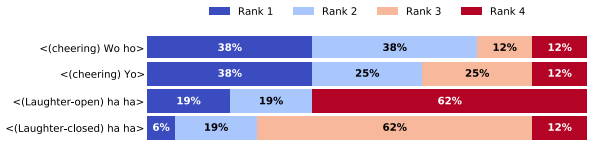
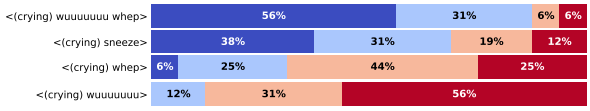


Fig. 2: The results of nMOS, eMOS and emotion recognition accuracy.

結果: 主観評価の結果、提案する細粒度NV手法は自然さにおいて僅かな低下が見られたものの、それを補って余りある優れた表現力 (eMOS = 4.20) と最高の感情認識精度 (78.8%) を達成しました。感情別分析では、細粒度NVが高覚醒感情（happy 82.5%、fear 82.7%）およびsad (98.3%) の伝達に決定的な役割を果たすことが確認されました。選好テストでは、幸福に対する「Cheering」、および悲しみに対する「複雑な多部構成のCrying」への強い支持が示され、特定の構造化されたNV発声の重要性が裏付けられました。



(a) The results for the Happy emotion.



(b) The results for the Sad emotion.

Fig. 3: Preference evaluation results for different non-verbal expressions.

4 おわりに

本論文は細粒度非言語表現データを構築し、ベンチマークを設定しました。この設計が高覚醒感情において特に、感情表現力と認識精度（平均 82.0%）を大幅に改善することを実証しました。我々の知見は、真に表現力豊かな感情TTSには細粒度な非言語キューが不可欠であることを検証し、リスナーの選好が、happyとsadを伝えるための特定のNVタイプの重要性を浮き彫りにしました。

■**謝辞** 本研究の一部は、JSPS科研費JP21H05054、JP23K21681、JP25H01139、およびJST NEXUS（JPMJNX25C1）の助成を受けて実施したものである。

参考文献

[1] J. A. Russell, “A circumplex model of affect,” in *Journal of personality and social psychology*, Vol. 39, No. 6, pp. 1161, 1980.

[2] V. Popov, et al., “Grad-tts: A diffusion probabilistic model for text-to-speech,” in *International conference on machine learning*, pp. 8599–8608, 2021.

[3] T. A. Nguyen, et al., “Expresso: A benchmark and analysis of discrete expressive speech resynthesis,” in *arXiv preprint arXiv:2308.05725*, 2023.

[4] G. McKeown, et al., “The SEMAINE corpus of emotionally coloured character interactions,” in *2010 IEEE international conference on multimedia and expo*, pp. 1079–1084, 2010.

[5] M. Borisov, et al., “NonverbalTTS: A Public English Corpus of Text-Aligned Nonverbal Vocalizations with Emotion Annotations for Text-to-Speech,” in *arXiv preprint arXiv:2507.13155*, 2025.

[6] A. Mehrabian, “Nonverbal communication,” in *Routledge*, 2017.

[7] J. Richter, et al., “EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation,” in *ISCA Interspeech*, pp. 4873–4877, 2024.