氏名	KHALIFAH MUJAHID JAMAL A
授与学位	博士（工学）
学位記番号	博甲第220号
学位授与年月日	令和7年3月21日
学位授与の要件	学位規則第4条第1項
学位論文題目	Emotional Text-To-Speech in Japanese Using Artificially Augmented Dataset （人工的に拡張されたデータセットを用いた日本語感情音声合成）
論文審査委員	主査　准教授　ﾌﾟﾀｼﾝｽｷ ﾐﾊｳ ｴﾄﾞﾑﾝﾄﾞ 教　授　桝井　文人 教　授　前田　康成 教　授　升井　洋志 教　授　奥村　貴史

学位論文内容の要旨

In my dissertation, I investigate the potential of using artificially generated emotional speech datasets as substitutes for human-recorded data in Japanese text-to-speech (TTS) synthesis. Given the limited availability of emotional datasets in Japanese, collecting high-quality human-recorded emotional speech is both costly and time-consuming. My objective was to determine whether a synthetic dataset could effectively support TTS models in generating speech that convincingly conveys emotions without relying on extensive human-recorded data.

To start, I discuss the significance of emotional speech synthesis. Integrating emotions into synthetic voices enhances the quality of human-computer interaction, making applications like virtual assistants and customer service more engaging and relatable. However, due to the scarcity of available emotional data, developing expressive TTS in Japanese has been challenging. This led me to test synthetic data as a potential solution to bridge this data gap.

For this study, I chose Tacotron 2, an end-to-end TTS model that converts text to mel-spectrograms, which represent sound frequencies over time. These spectrograms are then converted to audio using WaveGlow as the vocoder. Tacotron 2's accessibility and effectiveness in other TTS tasks made it an ideal choice for examining emotional synthesis in Japanese. I used several objective metrics, including Mel-Cepstrum Distortion (MCD) and Signal-to-Distortion Ratio (SDR), to evaluate the quality of the synthesized speech and to quantify how closely it approximated human speech in clarity and accuracy.

To establish a baseline, I conducted a preliminary test with Tacotron 2 using a standard Japanese dataset from Mozilla's Common Voice. This test demonstrated that while Tacotron 2 could manage basic Japanese synthesis, it faced consistency issues, particularly in pitch and naturalness. These results highlighted areas where Tacotron 2 would require refinement for more effective emotional synthesis.

In the main experiment, I created an artificial emotional dataset using Voiceroid 2, a Japanese TTS system capable of generating speech in various emotional tones—happiness, anger, sadness, and neutrality. By synthesizing speech samples across these emotional categories, I built a dataset large enough to train Tacotron 2. This allowed me to examine whether Tacotron 2 could learn to express different emotions in Japanese using a synthetic dataset in place of human-recorded data.

I employed two training approaches for Tacotron 2: one model was trained from scratch, while the other used a "warm-start" with pre-trained parameters from the initial experiment. My findings indicated that the warm-start model generally performed better, achieving higher spectral accuracy and producing more authentic emotional expressions, which suggested that pre-training enhanced Tacotron 2's ability to convey emotions.

To obtain subjective feedback on the synthesized speech, I developed a simplified Mean Opinion Score (MOS) survey. The dataset was divided into "unique" and "shared" sentences, both of which were synthesized by Tacotron 2 and Voiceroid 2. Unique sentences contained only one emotion type per batch, while shared sentences were distributed across all batches to enable comparative analysis. This setup allowed participants to assess each system's performance in terms of naturalness,

intelligibility, and emotional congruence—the alignment between the voice tone and the emotional content of each sentence.

The survey results indicated that Voiceroid 2 generally received higher ratings than Tacotron 2, particularly in emotional fidelity and naturalness. Emotionally congruent sentences, where the emotional tone matched the sentence's content, received higher scores, demonstrating the importance of alignment between emotional tone and meaning. While Tacotron 2 was able to replicate some basic emotions, it encountered challenges with more nuanced or complex expressions, such as mixed or subtle emotional cues. Accurate pitch, especially for emotions like anger or sadness, emerged as a critical area for improvement in achieving convincing emotional synthesis.

Overall, I found that while Tacotron 2 could capture basic emotional expressions with the synthetic dataset, replicating the full range of human emotional nuances remained a challenge. The predefined emotional styles in Voiceroid 2 lacked the depth and complexity typically found in human speech, which limited Tacotron 2's ability to model more subtle emotional variations. Nonetheless, this study shows that synthetic emotional datasets offer a viable solution for overcoming data scarcity in Japanese emotional TTS.

In conclusion, my research suggests that while synthetic datasets are a promising resource for developing emotional TTS models in under-resourced languages, achieving high-quality emotional synthesis will require further advancements in both dataset quality and model training methods. This work contributes to the broader effort to create expressive and engaging TTS systems for languages with limited emotional data, while also identifying future directions for improving emotional fidelity and naturalness in synthetic speech.

審査結果の要旨

この論文では、日本語の感情音声合成について調査する。まず、感情音声合成における課題を明確にし、感情音声合成において少資源の感情音声データに関する従来研究を徹底的にレビュー調査を行なった。次に、人工的データ拡張を活用した感情音声合成手法を提案する。Voiceroid 技術を用いて日本語の感情音声データセットを作成し、その内容には音声学的特徴量の変調を用い多様性を保ったデータセットを作成した。さらに、徹底的な実験過程を通じて、性能の高い音声合成モデルを特定している。そして、感情音声合成に最適な設定を実験的に特定し、自然な感情表現を実現するためのモデルの微調整（ファインチューニング）を行う。さらに、元の Voiceroid で作成されたデータと自動生成されたデータを多面的ユーザ調査を用いて比較した。結果、音声合成モデルを用いて自動生成された音声の音質は，元の音質を下回るものの，実用性のある十分に高い音質を維持していることを確認した。したがって、人工的に拡張されたデータが感情音声合成というタスクにおいて訓練データとして使用できることを確認した。本研究において明らかにしたことを、日本語以外の言語にも拡張することを期待でき、感情音声合成の理解を深めることにさらに貢献できると考えられる。

結論として、著者は、今後の感情音声合成とその合成手法開発への根本的な貢献をした他、人工知能、自然言語処理分野、信号処理分野、そして特に感情的音声合成の分野において課題となっている少資源の訓練データの場合でも人工的データによるデータ拡張により自然な音声合成を出力するという課題について示唆を与える新知見を得たと判断される。なお、自然言語処理、信号処理、人工知能など複数の分野に跨る深淵な課題に対して貢献するところ大なるものがある。

よって著者は、博士(工学)の学位を授与される資格があるものと認める。