2023「中技社科技獎學金」
2023 CTCI Foundation Science and Technology Scholarship
境外生生活助學金
Bursary Award for Overseas Students

Empowering Speech Recognition: Deep Generative AI for Limited Datasets, Voice Spoofing Defense, and Human Rights Violation Detection in Social Media Voice Reports

Yeshanew Ale Wubet (4th year Ph.D. Candidate)[1], Kuang-Yow Lian (Professor)[2]
[1]Department of International Program in Electrical Engineering and Computer Science, National Taipei University of Technology
[2]Department of Electrical Engineering, National Taipei University of Technology

Abstract

A deep generative artificial intelligence (AI) technique creates artificial voices efficiently. In our study, this method is employed to address the issue of limited speech datasets in speech recognition and protect the speaker's identity from potential attackers using zero-shot voice conversion. The study also focuses on the intricate task of accent detection and evaluating accent similarities. Furthermore, the study demonstrates the detection of human rights violations and safety crises in voice-based news reports in digital media using deep learning algorithms.

Research Focus

I. Improving Speaker-Independent Keyword Recognition

Main Contribution

- ✓ Applying the advanced parallel voice conversion (VC) techniques to real applications, specifically to increase the training data size of speaker-independent voice recognition.
- ✓ We realized that exact VC is not useful for voice-based augmentation.
- ✓ The LSTM model improved the inconsistent performance of the CNN model when CNN and LSTM were hybridized together.
- ✓ A delicate CNN-LSTM framework is also designed carefully for feature extraction and classification.

Proposed Model

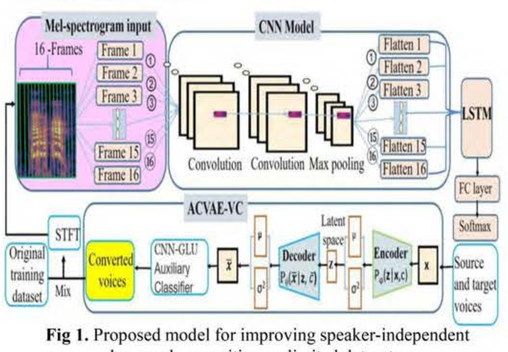

Fig 1. Proposed model for improving speaker-independent keyword recognition on limited datasets.

Experimental Results

Table I. Model performance comparison on dataset-II

Training data	CNN	LSTM	CNN-LSTM
Original	92	86	93
Original +VC	94	94	94
Baseline augmentation	95	95	96
VC + baseline augmentation	**96**	**94**	**97**

- ✓ The overall results showed that the pure CNN and pure LSTM models were highly affected by the limited training dataset size when compared to our proposed CNN-LSTM model.
- ✓ Finally, we conclude that applying augmentation and regularization techniques to a mixture of both original and converted data enhanced the deep learning performance.

- ✓ The generated sample spectrograms are shown in Fig. 2

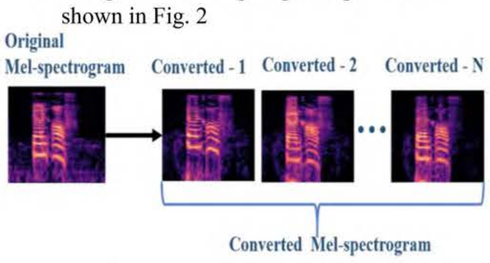

Fig. 2. VC augmentation samples

- We used two dataset for experiment.
- ✓ **Dataset I:** Voice keywords ("open", "close", "down", "up", "turn on", "turn off", "bed", "bad", "computer", "hello", "welcome", and "university").
- ✓ **Dataset II**: English keywords (0-9 digit voices).

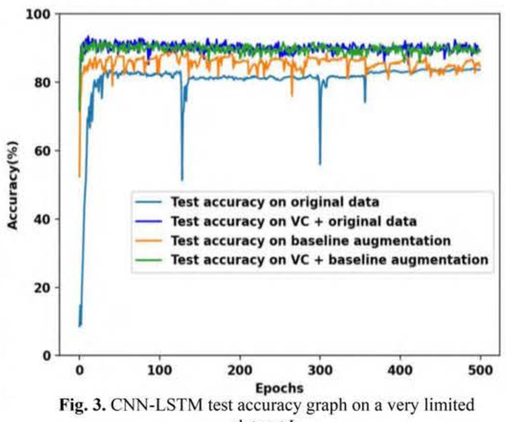

Fig. 3. CNN-LSTM test accuracy graph on a very limited dataset I.

II. Accent Classification and Accent Similarity Evaluation[2]

Main Contribution

- ✓ This paper proposed a novel intra-native accent shared features-based native accent identification (NAI) framework. To our knowledge, the proposed feature extraction approach has never been applied in the existing works.
- ✓ We conducted a deep investigation for obtaining a spectrogram frame size value that accomplishes an exemplary neural network model performance with an optimized computational time.
- ✓ We introduced and investigated the non-existing work which is the native and non-native English accent proximity evaluation model using the NAI pre-trained model.

Proposed Model

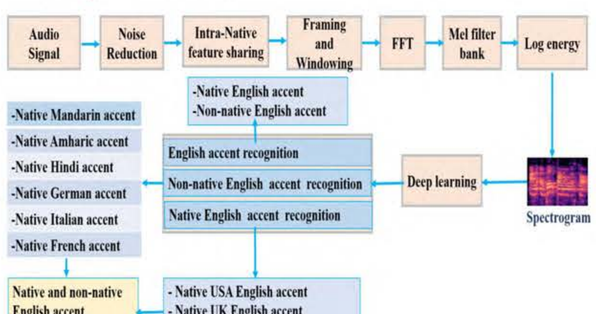

Fig 4. The overall intra-native shared features-based NAI framework.

Accent detection tasks

- The NAI framework is employed for:

1) Non-native English accent classification (Mandarin native, Hindi native, Amharic native, French native, German native, and Italian native)
2) Native English accent classification (USA, Canadian, and UK)
3) Identification of native and non-native English accents
4) Accent similarity evaluation

Experimental Results

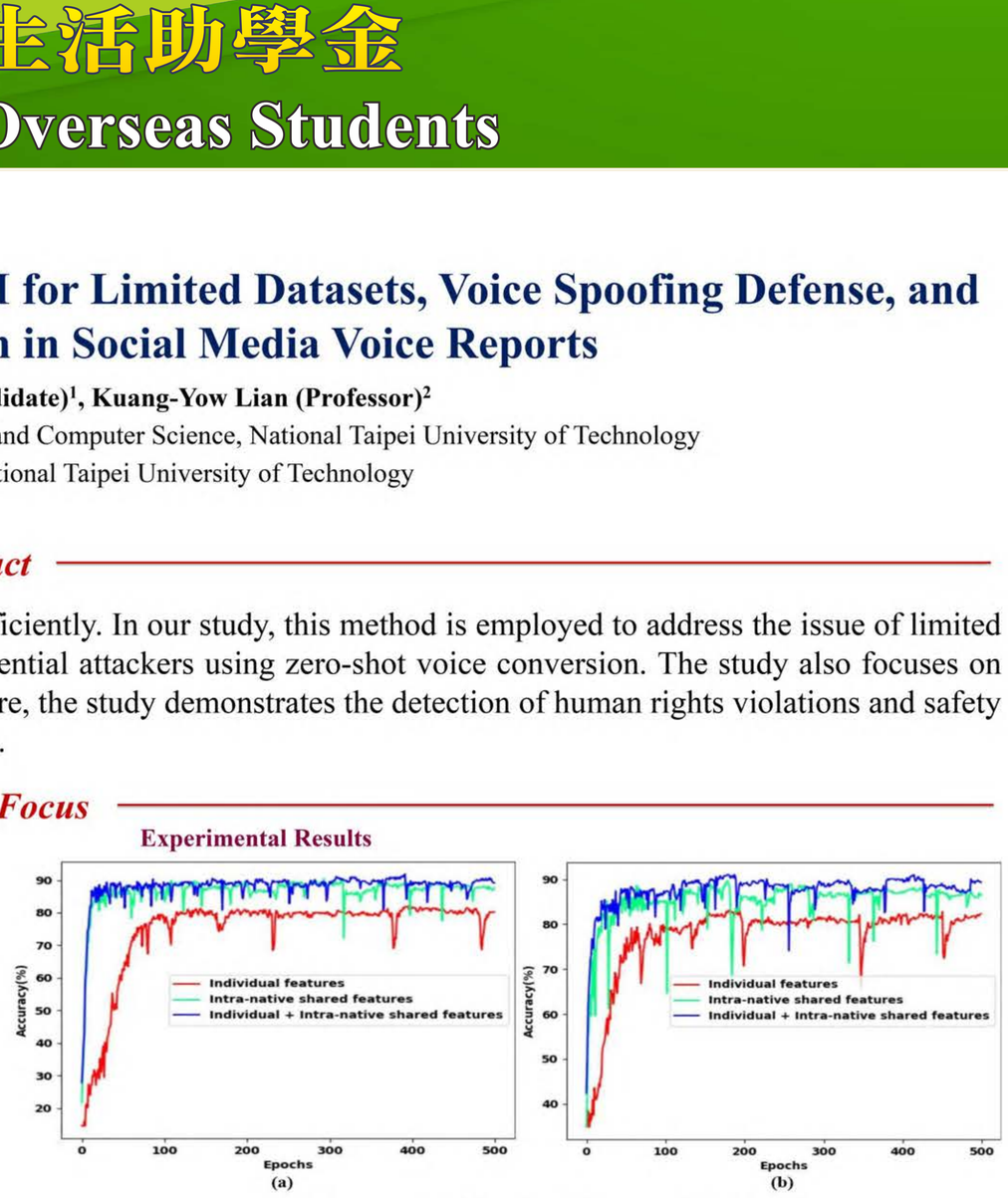

Fig 5. Accuracy graph: (a) Non-native English accents; (b) Native English accents.

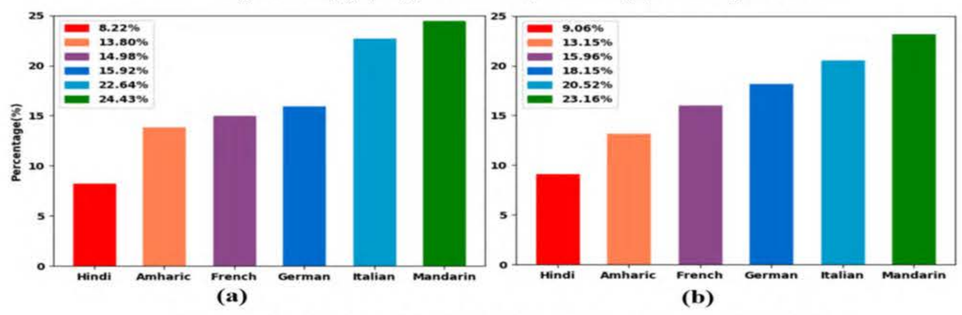

Fig 6. Accent similarity evaluation: (a) using CNN-LSTM model (b) using Bi-LSTM model.

- ➢ The proposed approach boosted the accuracy of the baseline method with an average accuracy value of 3.7% - 7.5% on different vigorous deep learning algorithms.
- ➢ The model makes the rank for non-native English accents based on their similarity to native English accents and the proximity rank is **Mandarin, Italian, German, French, Amharic, and Hindi**.

III. Detection of Human Rights Violations (HRV) Voice Reports

Main Contribution

- ✓ It provides an opportunity for human rights advocacy groups to quickly detect and access HRV reports from online news media platforms. Furthermore, they can investigate the reports with evidence, which can save the lives of victimized individuals and communities, ensure justice, and provide welfare support.
- ✓ We consider both text and voice-based Amharic HRV news detection.
- ✓ We compare the performance of attention (att)-based rigorous deep learning models and classical models.

Proposed Model

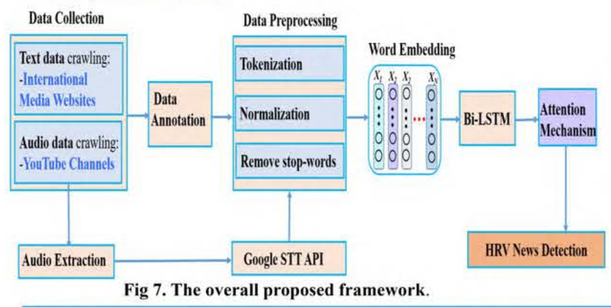

Fig 7. The overall proposed framework.

Experimental Results

- ✓ The att-Bi-LSTM, Bi-LSTM, att-GRU, att-Bi-GRU, and att-LSTM achieved accuracies of **91.38**%, 91.34%, 91.33%, 91.29%, and 91.06%, respectively, for detection of text-based HRV.
- ✓ The att-Bi-LSTM, att-GRU, att-Bi-GRU, and att-LSTM models achieve accuracy rates of **81.81**%, 80.58%, 80.21%, and 79.93%, respectively, for detection of audio-based HRV.

IV. Voice Conversion (VC) for Anti-Voice Spoofing

Main Contribution

- ✓ We employed the VC and multi-target speaker voice fusion method as an effective countermeasure against the speaker's identity attackers.

Experimental Results: In recent investigations, zero-shot VC has also been used to anonymize speakers; however, confidentiality for the target speaker remains compromised. Our model secured both source and target speakers for end-to-end speakers' privacy.

Publications

1. **Y. A. Wubet** and K. -Y. Lian, "Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets," in IEEE Access, vol. 10, pp. 89170-89180, 2022, doi: 10.1109/ACCESS.2022.3200479.
2. **Y. A. Wubet,** D. Balram and K. -Y. Lian, "Intra-Native Accent Shared Features for Improving Neural Network-Based Accent Classification and Accent Similarity Evaluation," in IEEE Access, vol. 11, pp. 32176-32186, 2023.