

# System Description for First VoicePrivacy Attacker Challenge

Oscal Tzyh-Chiang Chen, and Yi-Chen Tsai

Department of Electrical Engineering, National Chung Cheng University, Chiayi, 62102, Taiwan

**Abstract**— This study proposes a comprehensive anonymized speaker recognition system aimed at achieving efficient and privacy-preserving speech recognition. The system employs Constant-Q Transform (CQT) for spectral analysis. CQT offers higher frequency resolution in the low-frequency range and lower resolution in the high-frequency range, enabling more effective representation of fundamental frequency and formant details, which are crucial for speaker recognition. These features are closely tied to the low-frequency components of a speaker's identity. The spectral signals are then processed through the MCG-Res2Net-RH model, which integrates multi-scale branches and a Refined Highway mechanism to enhance speech feature representation. The system uses GE2E-Loss for speaker clustering, optimizing cosine similarity to minimize intra-speaker distances while maximizing inter-speaker distances, thus achieving the core goal of anonymization. The system is trained on the train-clean-360 subset of the LibriSpeech dataset. Experimental results reveal that the proposed system achieves a 2.47% reduction in Equal Error Rate (EER) compared to the baseline architecture, demonstrating high efficiency and robustness while balancing privacy protection and performance.

**Keywords**— *Speech recognition, speaker anonymization, Constant-Q Transform, GE2E-Loss*

## I. INTRODUCTION

Speech data contains a wealth of personal information, such as age, gender, emotional state, health condition, ethnicity, geographic background, and socioeconomic status [1]. This information can potentially be improperly extracted and used to infer personal privacy through speech recognition technologies. To safeguard the privacy of speech data, the VoicePrivacy initiative [2], launched in 2020, has been dedicated to advancing privacy-preserving speech technologies. It aims to provide privacy solutions through standardized datasets, protocols, and evaluation metrics.

One of the core objectives of the initiative is to develop speech anonymization techniques that effectively conceal speaker identities while preserving the accuracy of speech content. Participants are required to design robust speech anonymization systems that can mask the original speaker's identity while ensuring that linguistic and emotional content remains unaffected. The competition aims to test the capabilities of Automatic Speaker Verification (ASV) systems in re-identifying anonymized speakers, thereby driving the development of more advanced ASV systems.

## II. METHODOLOGY

Constant-Q Transform (CQT) is employed as a speech pre-processing method to extract fine-grained low-frequency

features, including the fundamental frequency and formants of speech. This enhances the stability and generalization capability of speech features. The core architecture adopts the MCG-Res2Net-1RH-E model proposed in [3] which integrates multi-level feature extraction capabilities to effectively address variability in speech. Furthermore, the system utilizes GE2E-Loss [3] to compute the similarity between speech embeddings, enabling the determination of whether anonymized speech has been previously registered in the system, thereby achieving identity verification functionality.

### A. Speech Signal Preprocessing

During the speech preprocessing stage, CQT is used to capture detailed features such as the fundamental frequency and formants in speech, with the CQT parameters listed in Table 1. These features are closely related to speaker identification. Time-frequency analysis of fixed length is performed based on an average speech duration of 13 seconds, resulting in fixed-length time-frequency features. Since CQT employs an exponential scale for spectral analysis, its spectral characteristics are closer to the natural spectral distribution of human speech compared to FFT. Additionally, CQT has a smaller bandwidth for low-frequency waves, providing higher frequency resolution, which enhances speaker identification accuracy.

Table 1 CQT parameter setup.

Parameters	Our work
Sample rates (kHz)	16
Hop length (frame step)	128
$n_{bin}$	84
$n_{fft}$	256
Frame sizes (ms)	16
Bins Per Octave (BPO)	12
$f_{min}$ (Hz)	63
Window function	Hanning
Duration (sec)	13
Feature size (H, W)	(84, 500)

### B. Method

The architecture proposed in [3], MCG-Res2Net-1RH-E, as shown in Fig. 1, is adopted as the core framework of this method. The referenced work explores the impact of varying the number of multi-scale branches  $X_m$  within residual blocks and introduces multi-layer refined highway (RH) connections [5] to enhance global feature learning while preserving critical speech feature information. By adjusting the  $X_m$  values in the residual blocks, receptive fields of different sizes can be obtained, enabling the capture of subtle features in speech and

improving the distinction between registered and unregistered speakers.

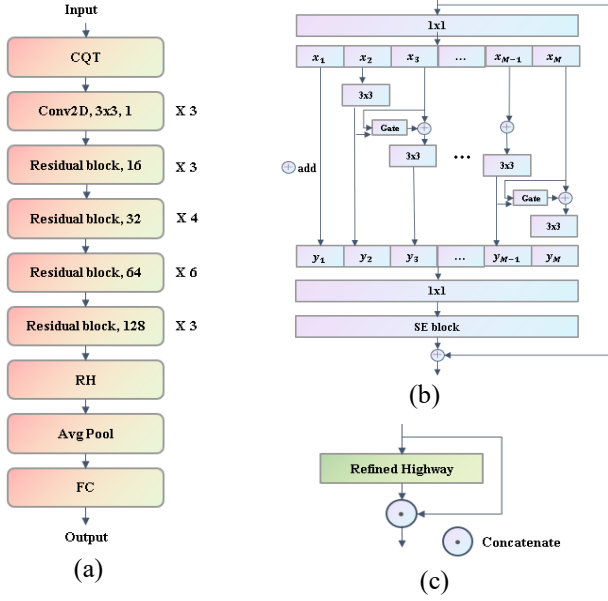


Figure 1. MCG-Res2Net-1RH-E. (a) Model architecture (b) Residual block (c) Refined Highway (RH).

### C. Cost function

Supervised contrastive learning uses the concept of clustering to bring samples with the same label closer in the feature space while pushing samples with different labels farther apart. The core idea of this approach is to learn the intrinsic features of samples with the same label in a high-dimensional feature space, rather than focusing on the fine-grained details of each individual sample.

This cost function learns from a fixed number of speech samples per batch, with the sample size determined by the number of speakers and the number of utterances each speaker provides. This cost is illustrated as follows.

$$e_{ji} = \frac{f(x_{ji}; w)}{\|f(x_{ji}; w)\|_2} \quad (1)$$

The feature vector  $x_{ji}$  represents the features extracted from the  $i$ -th utterance of speaker  $j$ , where  $e$  in Eq. (1) denotes the embedding vector of the  $i$ -th utterance from the  $j$ -th speaker.

$$S_{ji,k} = w \cdot \cos(e_{ji}, c_k) + b \quad (2)$$

$$c_k = \mathbb{E}[e_{km}] = \frac{1}{M} \sum_{m=1}^M e_{km} \quad (3)$$

The matrix similarity  $S_{ji,k}$  defines the cosine similarity between the embedding vector  $e_{ji}$  of each speaker's utterance and the centroid  $c_k$ . The centroid  $c_k$  is defined as the mean of all embedding vectors for speaker  $k$ . Both the embedding vector  $e_{ji}$  and the centroid  $c_k$  are  $L_2$ -normalized, while  $w$  and  $b$  are learnable parameters.

$$\mathcal{L}(e_{ji}) = -S_{ji,k} + \log \sum_{k=1}^N \exp(S_{ji,k}) \quad (4)$$

$$\mathcal{L}_G(x; w) = \sum_{j,i} \mathcal{L}(e_{ji}) \quad (5)$$

After applying Softmax to  $S_{ji,k}$ , feature vectors of similar utterances are assigned higher probabilities, pushing them closer to their corresponding centroid while reducing the likelihood of their proximity to other centroids. The final total loss  $\mathcal{L}_G(x; w)$  is obtained by summing the losses calculated for all similarity matrices. Thus, through GE2E-Loss, the extracted features are clustered to optimize the speaker learning process and achieve the core objective of identifying whether an anonymous speaker has been registered.

### D. Experimental Setup and Dataset

The experiments adopt a computer platform with an NVIDIA GeForce GTX 2080 Ti GPU, at the setup of Adam as the optimizer, a learning rate of  $1e-4$ , and parameter configurations of  $\beta_1 = 0.9$ ,  $\beta_2 = 0.998$ , and  $\epsilon = 1e-9$  to assist in convergence. Additionally, the batch size is 24. The dataset is derived from the train-clean-360 subset of the LibriSpeech [6] corpus, processed by the T25-1 anonymization system [7], which removes speaker identity features for subsequent training and evaluation. Table 2 lists LibriSpeech T25-1 dataset.

Table 2: LibriSpeech T25-1 dataset.

Subset (LibriSpeech)		Speaker	Utterances
Train	train-clean-360	921	104,014
Develop	dev-clean	Enroll	29
		Trial	40
Test	test-clean	Enroll	29
		Trial	40
			438
			1,496

The training set contains 921 speakers with a total of 104,014 speech samples. The development set is divided into a registration speech set and a validation speech set, consisting of 2,321 speech samples in total. The test set is similarly divided into a registration speech set and a validation speech set, totaling 1,934 speech samples. Each dataset ensures a balanced gender distribution to guarantee fairness in model training and evaluation.

### E. Evaluation Indicators

Equal Error Rate (EER) is one of the key metrics used to evaluate binary classification systems. It is determined by adjusting the threshold  $\theta$  to make the miss rate  $P_{miss}$  and false alarm rate  $P_{fa}$  equal, with this value representing the EER.  $P_{miss}$  refers to the probability that the registered speaker's voice is incorrectly identified as an imposter's voice, while  $P_{fa}$  represents the probability that the model misidentifies an imposter's voice as that of a registered speaker. The equations of  $P_{miss}$ ,  $P_{fa}$ , and EER are defined as follows.

$$P_{miss}(\theta) = \frac{\#\{ASV \text{ score} \leq \theta\}}{\#\{\text{total target trials}\}} \quad (6)$$

$$P_{fa}(\theta) = \frac{\#\{ASV \text{ score} > \theta\}}{\#\{\text{total impostor trials}\}} \quad (7)$$

$$EER = P_{fa}(\theta) = P_{miss}(\theta) \quad (8)$$

### III. EXPERIMENTAL RESULTS AND DISCUSSION

This system first experiments with MCG-Res2Net-1RH-E using one-layer and two-layer RH, and compares the results with the baseline system in the competition. The comparison results are listed in Tables 3 and 4. "Average dev" refers to the average values of male and female samples in the LibriSpeech-dev dataset, while "Average eval" refers to the average values of male and female samples in the LibriSpeech-test dataset. To compare the model's performance on different datasets, the last column, "Average," represents the combined average of Average dev and Average eval.

The results indicate that due to the ability of CQT to effectively capture critical low-frequency features, formants, and fundamental frequencies in speech, performance on the male dataset showed significant improvement. However, since CQT has a larger bandwidth at high frequencies, despite its higher temporal resolution, it may miss important spectral features in speech, leading to more limited improvement on the female dataset. Additionally, experiments tested different numbers of RH layers, including 1 layer, 2 layers, and 3 layers. It was found that the 1-layer structure performed the best, while the 3-layer structure, possibly due to the introduction of excessive features, failed to focus on the critical features in speech, resulting in overfitting during training. Therefore, the corresponding table is not included.

Table 3: EER (%) Comparison between MCG-Res2Net-1RH-E and the baseline architecture.

Dataset (LibriSpeech)	Gender	EER(%) Baseline	EER(%) Our
dev	female	42.65	<b>41.27</b>
dev	male	40.06	<b>39.91</b>
Average dev		41.36	<b>40.59</b>
test	female	<b>42.34</b>	43.23
test	male	41.92	<b>40.47</b>
Average eval		42.13	<b>41.85</b>
Average		42.24	<b>41.22</b>

Table 4: EER (%) Comparison between MCG-Res2Net-2RH-E and the original architecture.

Dataset (LibriSpeech)	Gender	EER(%) Baseline	EER(%) Our
dev	female	<b>42.65</b>	44.62
dev	male	40.06	<b>39.32</b>
Average dev		<b>41.36</b>	41.97
test	female	42.34	<b>42.33</b>
test	male	41.92	<b>41.87</b>
Average eval		42.13	<b>42.10</b>
Average		42.24	<b>42.04</b>

### IV. CONCLUSION

This system was trained on 104,140 speech samples from 921 speakers, utilizing CQT as a speech pre-processing method to extract key spectral features, formants, and fundamental frequency information from the speech. The core architecture employed is MCG-Res2Net-RH-E. To assess the effectiveness of the experiments, the impact of 1, 2, and 3-layer architectures

on performance was tested. Finally, the voice signals were grouped using the GE2E-Loss function, which helps to better identify whether the voice is from a registered speaker. The experimental results demonstrate that MCG-Res2Net-1RH-E outperforms the original ASV architecture, with the EER reduced by approximately 2.47%, highlighting the proposed system's excellence and stability in speech recognition.

### REFERENCES

- [1] Nautsch *et al.* "Privacy in speaker and speech characterization," *Computer Speech and Language*, vol. 58, pp. 441–480, 2019.
- [2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. of Interspeech 2020*, 2020, pp. 1693–1697.
- [3] Oscar T.-C. Chen, Yun-Chia Hsu, and Tun-Sheng Yang, "Detecting speech deepfakes through improved speech features and cost functions," in *Proc. of IEEE Asia Pacific Conference on Circuits and Systems*, Taipei, Taiwan, Nov. 7<sup>th</sup>-9<sup>th</sup>, 2024.
- [4] Li *et al.* "Generalized end-to-end loss for speaker verification." In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [5] M.-H. Ha, and O. T.-C. Chen, "Deep neural networks using residual fast-slow refined highway and global atomic spatial attention for action recognition and detection," *IEEE Access*, vol. 9, pp. 164887-164902, 2021.
- [6] Vassil *et al.* "Librispeech: an asr corpus based on public domain audio books." In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
- [7] Gu *et al.* "A voice anonymization method based on content and non-content disentanglement for emotion preservation." In *Proc. of SPSC 2024*.