Source Speaker Tracing Challenge (SSTC) 2024: Challenge Evaluation Plan

Ze Li[1], *Ming Li*[1]

[1] Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan, China

ming.li369@dukekunshan.edu.cn

1. Introduction

Speaker verification (SV) stands as a pivotal biometric authentication technology in the real world, exerting widespread influence on our daily lives. Particularly in recent years, with the advancement of deep neural networks, SV has witnessed extensive application across various domains, encompassing mobile devices, smart homes, smart cities, and the financial sector. Given the intrinsic significance of security to these applications, the resilience of speaker verification systems against spoofing attacks (e.g., speech synthesis, voice conversion (VC), adversarial sampling, etc.) is paramount.

Countermeasures have been developed in recent years to defend SV systems from spoofing attacks. The Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenges [1, 2, 3, 4] and Audio Deepfake Detection (ADD) Challenges [5, 6] are held to facilitate independent assessments of spoofing vulnerabilities and assess the performance of countermeasures against spoofing. However, these countermeasures typically focus on discriminating between bona fide speech and spoofed speech for SV systems, and there is no challenge to address the source speaker tracing problem – identify the information of the source speaker or eventually reconstruct the speech of the source speaker from the manipulated speech signals. Source speaker identification has potential applications in crime investigation and judicial procedures. For example, source speaker identification can help identify a suspect involved in financial fraud with voice conversion-based impersonation spoofing.

The Source Speaker Tracing Challenge (SSTC) is designed to identify the information of the source speaker in manipulated speech signals. This year's challenge focuses on the spoofing attack of VC on SV systems. The objectives of this challenge are to: 1) benchmark the current speech verification technology under this challenge condition, 2) promote the development of new ideas and technologies in speaker verification, and 3) provide an open, accessible and large-scale converted speech database for source speaker verification tasks.

2. Task Setting

The challenge comprises two tasks:

- **Task I Source speaker verification against voice conversion:** As shown in Fig 1, given a source speaker speech and a target speaker speech, VC manipulates the speech signal of the source speaker to make it sound like the target speaker while preserving the linguistic content. Participants will be asked to develop models to extract information about the source speaker from the converted speech for source speaker verification.
- **Task II Research Paper Track:** Participants are invited to contribute research papers at the SLT2024 conference. This is an opportunity to explore and document innovative approaches and findings related to source speaker identification technologies.

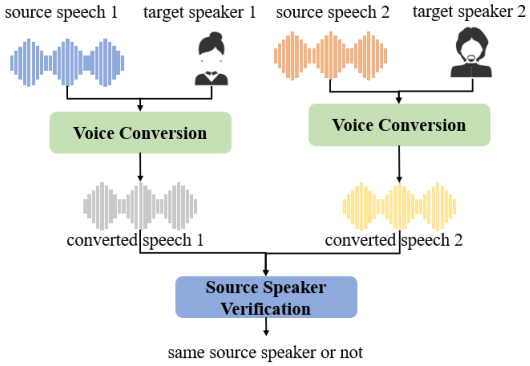

Figure 1: *Source speaker verification against voice conversion.*

3. Database

Table 1: *Source speaker dataset and target speaker dataset.*

Dataset	Subset	Speakers	#Utterances
Source Speaker	Train	1,172	132,553
	Dev	40	2,703
	Eval	40	2,602
Target Speaker	Train	5,994	1,092,009
	Dev	40	4,847
	Eval	40	4510

3.1. Source and target speaker datasets

As shown in Table 1. We utilize Librispeech [19] as the source speaker dataset and VoxCeleb [20, 21] as the target speaker dataset. Within Librispeech, to ensure the quality of the converted speech, the train-clean section (train-clean-100 and train-clean-360, comprising 132,553 utterances from 1,172 speakers), dev-clean subset (consisting of 2,703 utterances from 40 speakers) and test-clean subset (composed of 2,620 utterances from 40 speakers) are chosen as training set, development set and evaluation set, respectively. In VoxCeleb, our training data is sourced from the VoxCeleb2 development set (encompassing 1,092,009 utterances from 5,994 speakers), the VoxCeleb1 test set (comprising 4,847 utterances from 40 speakers) is chosen as the development set, and a subset of VoxCeleb1 development set (consisting of 4,510 utterances from 40 speakers) is utilized

Table 2: *Train, dev and eval sets and repositories for each VC method.*

Method	Train set		Dev set		Eval set		Repository
	ID	# Utterances	ID	# Utterances	ID	# Utterances	
AGAIN-VC [7]	Train-1	327,600	Dev-1	14,622	Eval-1	13,530	KimythAnly/AGAIN-VC
FreeVC [8]	Train-2	327,561	Dev-2	14,622	Eval-2	13,530	OlaWod/FreeVC
MediumVC [9]	Train-3	327,609	Dev-3	14,622	Eval-3	13,530	BrightGu/MediumVC
StyleTTS [10]	Train-4	327,546	Dev-4	14,622	Eval-4	13,530	yl4579/StyleTTS-VC
TriAAN-VC [11]	Train-5	327,609	Dev-5	14,622	Eval-5	13,530	winddori2002/TriAAN-VC
VQMIVC [12]	Train-6	327,498	Dev-6	14,622	Eval-6	13,530	Wendison/VQMIVC
SigVC [13]	Train-7	327,603	Dev-7	14,622	Eval-7	13,530	-
KNN-VC [14]	Train-8	327,765	Dev-8	14,622	Eval-8	13,530	bshall/knn-vc
BNE-PPG-VC [15]	-	-	Dev-9	14,622	Eval-9	13,530	liusongxiang/ppg-vc
DiffVC [16]	-	-	Dev-10	14,622	Eval-10	13,530	huawei-noah/Speech-Backbones
S2VC [17]	-	-	Dev-11	14,622	Eval-11	13,530	howard1337/S2VC
YourTTS [18]	-	-	Dev-12	14,622	Eval-12	13,530	Edresson/YourTTS

* All the repositories can be retrieved in the https://github.com.

for evaluation.

3.2. Converted Speech dataset

We adopt the method proposed by Cai [22] to generate converted speech. For each target speech, three source speech samples are randomly selected for VC, simulating attacks from three distinct attackers on the same target speech. To optimize storage efficiency without compromising training data diversity, we partition the VoxCeleb training set into ten subsets of equal size while preserving the number of speakers. Each VC method utilizes one of these subsets as the target speech set for generating converted speech.

We introduce 12 any-to-any VC methods to generate the large-scale converted speech dataset, as shown in Table 2. For SigVC, details of the SigVC model can be found in [13], and we use the in-house implementation.

In addition, we will explore N additional VC methods in the future to expand the eval set.

3.3. Named Structure

The naming format of the converted speech in the training set is <target speech utterance id>$-$<source speech utterance id> (such as id00012-21$Uxsk$56VDQ-00005-688-1070-0022.wav). You can split the audio's name with a $-$ character, where the first and third from last are the target speaker ID and source speaker ID, respectively.

4. Evaluation Protocal

4.1. The trials

The trial file consists of three segments: label (which denotes whether the trial is target or non-target), enrolment utterance ID, and test utterance ID.

For evaluation, We divide the enrollment and test utterance into four scenarios: (1) the same source speaker and the same target speaker, (2) the different source speakers and the same target speaker, (3) the same source speaker and the different target speakers, and (4) the different source speakers and the different target speakers. We randomly generate enrollment and test pairs according to these four scenarios and ensure that the number of pairs for each scenario is the same to create a balanced set of trials.

4.2. Evaluation metrics

We will use the Equal Error Rate (EER) metric in this challenge to evaluate the system's performance. For each pair of enrollment and test utterance vectors in the development and evaluation sets, the cosine similarity is computed, and the decision on same-speaker vs. different-speaker is made by threshold. Denoting by $P_{fa}(\theta)$ and $P_{miss}(\theta)$ the false alarm and miss rates at threshold , the EER metric corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., EER = $P_{fa}(\theta_{EER})$ = $P_{miss}(\theta_{EER})$. The lower the EER, the greater the performance. We will compute the average EER for all eval sets as the final evaluation criteria.

4.3. Leaderboard platform

We utilize the Codalab competition platform as a challenge-scoring platform. An online leaderboard for each task will be provided, and participants submit results up to 2 times daily. There were a total of 30 submissions during the evaluation phase. The leaderboard shows the performance of the systems on the full evaluation set. The challenge leaderboard platforms are available at:

- Task I

4.4. Evaluation Rules

Only VoxCeleb2 development, the train-clean section of Librispeech, and the training set of converted speech can be used for model training. The MUSAN [23] and RIR Noise [24] datasets can be used for data augmentation.

Suppose participants want to use another training data (e.g., voxblink, etc.) and pretrained models (e.g., wav2vec, wavLM, etc.). In that case, each registered participant can submit a list of proposed data and models to the organizers at $sstc2024.challenge@gmail.com$ before 21st April. The organizers will verify these requests and publish the list of training data and pretrained models allowed on https://sstc-challenge.github.io.

5. Registration and Submission

5.1. Registration

Since the challenge will be held on the Codalab platform, please create a Codalab account if you do not have one. We kindly request that you associate your account with an institutional email (e.g., edu.cn). Make sure to set your team's name in the user's profile, or it will not be visible on the leaderboard.

Please note that any deliberate attempts to bypass the submission limit (for instance, by creating multiple accounts and using them to submit) will lead to automatic disqualification.

5.2. Submission

5.3. Score submission

Participants are required to submit at least one valid score file for each participating task to the Codalab platform. The score files must be named as scores_n.txt .n denotes the sequence number of the test set, for example, the score file corresponding to the Eval-3 set should be named scores_3.txt .

The score files should be in UTF-8 format with one line per trial. Participants must zip the score files and submit the zipped archive on the Codalab platform. We will provide a sample on the challenge website.

5.4. System description submission

Each registered team is required to submit a technical system description report before the end of the challenge. Please submit this report using the SLT 2024 paper template. All reports must be a minimum of 2 pages (including references). Reports must be written in English. The system description does not need to repeat the content of the evaluation plan, such as the introduction of the database, evaluation metric, etc. The system description must include the following items:

- A complete description of the system components, including front-end (e.g., speech activity detection, features, normalization, front-end speech enhancement) and back-end (e.g., background models, i-vector/embedding extractor, speaker features fusion) modules along with their configurations (i.e., filterbank configuration, dimensionality and type of the acoustic feature parameters, as well as the acoustic model and the backend model configurations).
- A complete description of the data partitions used to train the various models.
- Performance of the submission systems on the development dataset (calculated based on the provided tools) and the evaluation dataset (calculated by the challenge platform).
- Novel ideas, strategies and methods are strongly recommended to be shared.
- A report of the model size and usage of GPUs or CPUs.

The reports should be sent to us as a link to an arXiv document or PDF file. In both cases, we will place links to the reports from the challenge website. The report may be used to form all or part of a submission to another conference or workshop. We recommend you send the report as a link to the arXiv document if you intend to do so. The links and PDF files should be sent to $sstc2024.challenge@gmail.com$.

6. Schedule

- April 7th, 2024: Registration opens.
- April 15th, 2024 : Training set, Development set and baseline system release.
- May 24th, 2024 : Test sets release and leaderboard for Task I open.
- June 7th, 2024 : Leaderboard freeze for Test sets.
- June 17th, 2024 : System report submission.

7. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. of Interspeech*, 2015.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. of Interspeech*, 2017.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. of Interspeech*, 2019.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *Proc. of ICASSP*, 2022, pp. 9216–9220.

[6] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

[7] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *Proc. of ICASSP*, 2021, pp. 5954–5958.

[8] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. of ICASSP*, 2023, pp. 1–5.

[9] Y. Gu, Z. Zhang, X. Yi, and X. Zhao, "Mediumvc: Any-to-any voice conversion using synthetic specific-speaker speeches as intermedium features," *arXiv preprint arXiv:2110.02500*, 2021.

[10] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[11] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion," in *Proc. of ICASSP*. IEEE, 2023, pp. 1–5.

[12] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. of Interspeech*, 2021, pp. 1344–1348.

[13] H. Zhang, Z. Cai, X. Qin, and M. Li, "Sig-vc: A speaker information guided zero-shot voice conversion system for both human beings and machines," in *Proc. of ICASSP*, 2022, pp. 6567–65 571.

[14] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc.Interspeech*, 2023.

[15] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[16] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Proc. of ICLR*, 2022.

[17] J. hao Lin, Y. Y. Lin, C.-M. Chien, and H. yi Lee, "S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations," in *Proc. of Interspeech*, 2021, pp. 836–840.

[18] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. of ICML*. PMLR, 2022, pp. 2709–2720.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. of Interspeech*, 2017.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of Interspeech*, 2018, pp. 1086–1090.

[22] D. Cai, Z. Cai, and M. Li, "Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems," in *Proc. of ICASSP*, 2023, pp. 1–5.

[23] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484*.

[24] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of ICASSP*, 2017, pp. 5220–5224.