2022 IEEE Signal Processing Cup

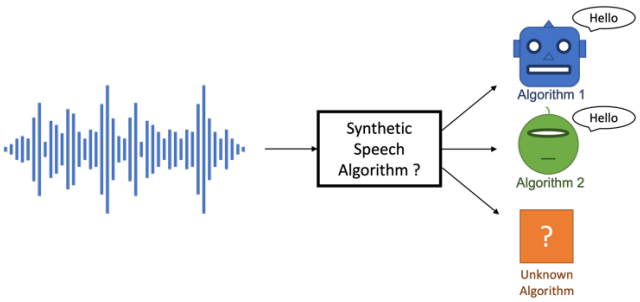
Synthetic Speech Attribution

January 6, 2022

This competition is sponsored by the IEEE Signal Processing Society and MathWorks

Introduction

The IEEE Signal Processing Society's 2022 Signal Processing Cup (SP Cup) will be a synthetic speech attribution challenge. Teams will be requested to design and develop a system for synthetic speech attribution. This means, given an audio recording representing a synthetically generated speech track, to detect which method among a list of candidate ones has been used to synthesize the speech. The detector must rely on the analysis of the speech signal through signal processing and machine learning techniques.



The IEEE Signal Processing Cup is designed to provide teams of undergraduate students with an opportunity to solve a challenging real-world problem. Students will be given an opportunity to test their signal processing skills against a real-world forensics challenge. They will be exposed to the synergy between signal processing and data science. They will need to exploit their signal processing knowledge to extract meaningful and characteristic information from audio signals. Then they will need to familiarize with the world of multi-class classification.

Finally, the challenge will help raising concerns about the forensic issues linked to the malicious use of synthetic speech (e.g., identity theft, spread of misinformation, etc.).

Background

The possibility of manipulating digital multimedia objects is within everyone's reach. Since a few years ago, this was possible thanks to several user-friendly software suites enabling audio, image, and video editing. Nowadays, media manipulation has become even easier thanks to the use of mobile apps that perform automatic operations such as face-swaps, lip-syncing, photo retouching, and audio auto-tune. Moreover, the huge technological advances determined by deep learning has delivered a series of data-driven tools that make manipulations extremely realistic and convincing for each kind of media.

All these tools are surely a great asset in the arsenal of a digital artist. However, if used maliciously to generate fake media, they can have a negative social impact. A recent example of synthetically manipulated media that raised a lot of concern is that of deepfakes, i.e., videos in which the identity of a person is replaced with the identity of another one through face manipulation [1]. These have been used to disseminate fake news through politician impersonation as well as for revenge porn.

Video deepfakes are not the only threat when it comes to people impersonation. As a matter of fact, both frauds and misinformation have taken place due to the use of voice manipulation techniques. Fake synthetic speech audio tracks can be nowadays generated through a wide variety of available methods. Indeed, synthetic speech can be obtained by simple cut-and-paste techniques performing waveform concatenation [2]. Alternatively, it can be obtained by vocoders exploiting the source-filter model of speech signal [3]. More recently, even multiple methods based on Convolutional Neural Networks (CNNs) for synthetic audio generation have been proposed [4]. These produce extremely realistic results that are hard to disambiguate from real speech also from human listeners.

In the literature, huge effort has been devoted to the development of forensic detectors capable of distinguishing original speech recordings from synthetically generated ones [5]. However, the problem of attributing a synthetic speech track to the generator used to synthesize it, has been less explored. Yet, being able to tell which algorithm has been used to generate a synthetic speech track can be of paramount importance to pinpoint the author of some illicit material [6].

Challenge Organization

The goal of this challenge is to develop a method for synthetic speech attribution in open set starting from an audio recording. This means, given a synthetic speech track under analysis:

- To detect if the track has been generated with a *known* algorithm (i.e., an algorithm within a set of methods available to the analyst).
- If yes, to detect which algorithm has been used.

In other words, the challenge consists in developing a multi-class audio detector, where the number of classes is the cardinality of the algorithms considered as *known*, plus an extra class to accommodate for the *unknown* algorithms.

The challenge will consist of two stages: an open competition (divided into two parts) that any eligible team can participate in, and an invitation-only final competition. Eligible teams

must submit their entries by **March 31, 2022**. The three teams with the highest performance will be selected by **April 7, 2022** and invited to join the final competition. The final competition will be judged at ICASSP 2022, which will be held on **May 22-27, 2022**.

Open Competition - Part 1

Part 1 of the open competition is designed to give teams a simplified version of the problem at hand to become familiar with the task. Participants will be provided with a labeled training dataset of audio recording. This dataset will consist of 1000 synthetic audio recordings generated from 5 different algorithms considered as the *known* algorithms, for a total amount of 5000 recordings of a few seconds each. These recordings are noiseless and saved as high-quality audio files.

Approximately one month prior to the competition deadline, the participants will be provided with the evaluation dataset. This consists of an additional set of unlabeled synthetic speech tracks generated with both known and unknown algorithms. Teams will be requested to associate a label (from 0 to 5, where 0-4 are associated to known algorithms and 5 to unknown) to each one of the evaluation tracks. This data will be used for testing each team method.

Open Competition - Part 2

Part 2 of the competition is designed to address a more challenging task: synthetic speech attribution in presence of noise. The task remains the same as for Part 1, but audio tracks will be edited by means of operations like noise addition, reverberation, filtering, and lossy compression.

Teams will be provided with MATLAB scripts to apply these operations to the dataset they already own.

Approximately one month prior to the competition deadline, Part 2 evaluation dataset will be released. Teams will be requested to provide a label (from 0 to 5) to each track in the evaluation dataset.

Final Competition

The three highest scoring teams from the open competition will be selected, and an additional training and evaluation dataset will be provided. The goal of this task is to test how the three finalists methods scale when more methods are available, and edited and unedited tracks are mixed.

Challenge Evaluation Criteria

Open Competition

The finalist teams will be selected based on the results achieved during the open competition. Results will be judged for Part 1 and Part 2 by means of accuracy defined as

$$Accuracy = \frac{Number\ of\ correctly\ attributed\ audio\ tracks}{Number\ of\ evaluation\ audio\ tracks}$$

For audio tracks belonging to the known class, a correct attribution is considered when the correct label is guessed. For audio tracks belonging to the unknown classes, a correct attribution is considered if the track is detected as unknown.

The final competition score will be the weighted average between the accuracy obtained in Part 1 and Part 2 computed as

$$Score = (0.7 \times Accuracy\ Part\ 1) + (0.3 \times Accuracy\ Part\ 2)$$

Final Competition

A judging panel will assess the three finalist teams' performance based on three criteria

- The score achieved on the final competition dataset.
- The quality of their report.
- The quality of their presentation (and possibly demo) at ICASSP 2022.

Open Competition Submission

Teams that wish to participate in the open competition should submit the following material by March 31, 2022 in order to be considered for the final competition:

1. A report in the form of an IEEE conference paper describing the technical details of their system. This should include a description of the signal processing techniques used to extract forensic as well as a description of how their classifier was designed and trained.
2. Evaluation dataset results in the form that will be specified on the Piazza forum.
3. A Matlab implementation of their attribution system. This should be able to accept an input in the form of a directory of "wav" audio files, and produce a text file containing the estimated labels associated to each input file.

Team Composition and Requirements

The purpose of this competition is that teams of 3-10 undergraduate students (enrolled as Bachelor or Master students) compete and develop a solution, under the supervision of a faculty member (or someone else with a PhD degree) and (optionally) the tutorship of a PhD student or postdoc. It is the undergraduate students that are responsible for presenting their work on ICASSP 2022, if the team makes it to the final competition.

Each team should contain 3-10 undergraduate students, who carry out the majority of the work. An undergraduate student is a person without a Master degree (or equivalent) at the time of submission. The students can be enrolled to a Bachelor or Master program, or equivalent. At least three of the undergraduate team members must hold either regular or student memberships of the IEEE Signal Processing Society at the time of submission. It is not mandatory to be enrolled at a university at the period of the cup. Undergraduate students who are in the first two years of their college studies as well as high-school students who are capable to contribute are welcome to participate in a team. An

undergraduate student can only be the member of one team. The students are to be supervised by a faculty member or someone else with a PhD degree that is employed by the university. The supervisor can be assisted by a tutor who has earned at least a Master degree at the time of submission; for example, a PhD student, postdoc, or equivalent. Each person can only be a member of one team

Prize for Finalists

According to the guideline from the IEEE Signal Processing Society, a maximum of three members from each of the three finalist teams will receive travel support to attend ICASSP2022 for the final competition (up to $1,200 for continental travel or $1,700 for intercontinental travel per member, and at most three people from each team will be supported). The participants can claim their travel expenses on a reimbursement basis.

More team members are also welcome to attend. Although these additional members will not be provided with any travel grant, those members who will not be presenting a paper at the conference will be offered a complimentary ICASSP registration. The finalist teams will also be invited to join the Conference Banquet as well as the Student Career Luncheon, so that they can meet and talk to SPS leaders and global experts.

A Judging Panel will be set up to select the ultimate winners at the conference. The teams will present the technical details of their approach to solve the challenge, demonstrate their results at a scheduled session, and answer questions raised at the session. The winner will be selected on the basis of the obtained results, the quality of their report, the quality of the final presentation and the capability to address questions.

The champion team will receive a grand prize of $5,000. The first and the second runner-up will receive a prize of $2,500 and $1,500, respectively, in addition to the above mentioned travel grants and complimentary conference registrations.

Important Dates

January 10, 2022	Competition webpage, Piazza forum and info
January 15, 2022	Dataset available
March 15, 2022	Team registration
March 31, 2022	Team final submission
April 7, 2022	Finalists announced
May 22-27, 2022	Final competition at ICASSP 2022

Online Resources

Main page of SP Cup on the SPS Website:
http://signalprocessingsociety.org/get-involved/signal-processing-cup

General information and resources are available on Piazza:

https://piazza.com/ieee_sps/spring2022/spcup2022
To set up a free account, use the access code “spcup2022” to join as a student the “SPCUP 2022: IEEE Signal Processing Cup 2022” class.

Organizers

The challenge is organized as a joint effort between the Image and Sound Processing Lab (ISPL) of the Politecnico di Milano (Milan, Italy) and the Multimedia and Information Security Lab (MISL) of the Drexel University (Philadelphia, USA).

The ISPL team is represented by Dr. Paolo Bestagini (Assistant Professor), Dr. Fabio Antonacci (Assistant Professor), Clara Borrelli (Ph.D. Student) and Davide Salvi (Ph.D. Student).

The MISL lab is represented by its founder Dr. Matthew C. Stamm (Associate Professor) and Brian Hosler (Ph.D. student).

Bibliography

[1] Verdoliva, L. “Media forensics and deepfakes: an overview”, CoRR abs/2001.06564, 2020

[2] Schroder, M., Charfuelan, M., Pammi, S., Steiner, I. “Open source voice creation toolkit for the MARY TTS platform”, Conference of the International Speech Communication Association (INTERSPEECH), 2011

[3] Morise, M., Yokomori, F., Ozawa, K. “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications”, IEICE Transactions on Information and Systems, 2016

[4] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. “Wavenet: A generative model for raw audio”, CoRR abs/1609.03499, 2016

[5] Todisco, M., Wang, X., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., Lee, K.A. “ASVspoof 2019: Future horizons in spoofed and fake audio detection”, Conference of the International Speech Communication Association (INTERSPEECH), 2019

[6] Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., and Tubaro, S. “Synthetic speech detection through short-term and long-term prediction traces”, EURASIP Journal on Information Security, 2021