MUDEEFA - MUltimodal DeEEp Fake detection using Text-To-Speech Synthesis, Voice Conversion and Lips Reading

Team IRISA/EXPRESSION

March 2022

Mots-clés

Artificial intelligence, automatic classification, speech synthesis, voice conversion, lip reading, heterogeneous information.

Start date : 01/10/2022

Summary

In recent years, Automatic Speaker Verification (ASV) has been increasingly used for voice biometrics. Securing these voice biometric systems for real-world applications is therefore becoming a major issue. The problem that we pose here is that of identity theft through an attack on a voice-based biometric identification system, and the countermeasures that could be implemented to respond to these attacks. Recent scientific works [KSPL20] show the diversity of possible attacks. Among them, we can count on attacks by high quality voice synthesis systems or high performance voice conversion systems. Recent advances in these two synthetic speech technologies are due to the use of deep learning techniques (DNN) and newly available massive data [SPW+18]. The security of voice biometric systems against impersonation attacks remains a difficult and unsolved topic. This is a major issue in these times of presidential elections, widerange geopolitical and territorial conflicts and massive dissemination of erroneous, altered or falsified information.

Thanks to deep learning and the availability of quality data in large quantities, the quality of text-to-speech systems and voice conversion (from a source voice to a target voice) has therefore seen unprecedented progress. For some systems, the quality is such that a naive ear cannot distinguish generated (synthetic) speech from natural speech. Automatic speaker recognition systems (verification of the speaker's voice identity) are therefore vulnerable due to the lack of countermeasures allowing the systems to verify, in addition to the speaker's voice identity, the nature of the speech (artificial or natural). Numerous studies demonstrate the effectiveness of combining modalities for speech and speaker recognition [FB07] and for building robust authentication systems [CWL18, IKF+18]. The implementation of systems allowing to synchronize audio recordings and lip expressions present in a video [PMNJ20] and the availability of corpora including both modalities [RSZ+21, FDLM18] will allow us to explore the link between these two modalities.

Thesis program

The proposed research topic is at the border of several domains and requires the acquisition of skills in speech processing (voice synthesis and conversion), in facial expression analysis, in particular lip expressions, and will have to take into account the works mixing the two modalities. A consequent work of bibliography and acquisition of technologies in these fields will thus be necessary at the beginning of this thesis.

This work will make it possible to implement these two skills with the aim of proposing an automatic system for detecting voice fraud (speaker impersonation), based on the state of the art in this field [DEK+21] and on the expertise developed by the EXPRESSION team (speaker identity and anomaly detection). The results of the thesis will be compared to those presented in the ASVspoof workshop (satellite of Interspeech), which focuses on the challenge of automatic speech verification and identity theft countermeasures.

Contexte scientifique

The EXPRESSION team of IRISA's Media and Interactions department focuses on the study of human-generated data (especially language) conveyed by different media : gesture or movement, speech and text. Two of its research axes concern the synthesis and recognition of expressive gesture and expressive speech. The EXPRESSION team has a wealth of experience in the field of automatic speech processing (speech synthesis and voice conversion) as well as in the field of anomaly detection on voice and facial expressions.

The team also has the technical means of recording that will facilitate the generation of data useful for speech synthesis, conversion and analysis. In this context, the team has developed a corpus dedicated to multimodal analysis, named EMO&LY (for EMOtion and anomaLY) to deepen and validate its research on anomaly detection conducted during the previous theses of Cédric Fayet [Fay18] and Valentin Durand de Gevigney [DdG21], supervised within the team.

Supervision team :

— Damien Lolive, Maître de conférences HDR (director),
— Pierre-François Marteau, Professeur des Universités (co-director),
— Arnaud Delhay, Maître de conférences (co-supervisor).

Candidate profile

The candidate will be expected to conduct cutting-edge applied research in one or more of the following areas : signal processing, statistical machine learning, speech and gesture recognition. He/she should have excellent computer programming skills (e.g. C/C++, Python/Perl, etc.), and knowledge of machine learning, signal processing or human-computer interaction.

The candidate must hold a master's degree in computer science or an engineering degree giving the title of master in computer science.

Please send a CV, a cover letter, one or more letters of reference and the academic results of the previous degree (Master's degree or Engineering degree giving the title of Master) **to all contacts (firtsname.lastname@irisa.fr) before Friday, April 8, 2022, strict deadline.**

Références

[CWL18] Feng Cheng, Shi-Lin Wang, and Alan Wee-Chung Liew. Visual speaker authentication with random prompt texts by a dual-task cnn framework. *Pattern Recognition*, 83 :340–352, 2018.

[DdG21] Valentin Durand de Gevigney. *Machine Learning Models for Multimodal Detection of Anomalous Behaviors.* PhD thesis, Université de Bretagne Sud, 2021. to appear.

[DEK$^+$21] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. Asvspoof 2021 challenge - speech deepfake database, May 2021.

[Fay18] Cédric Fayet. *Multimodal anomaly detection in discourse using speech and facial expressions.* PhD thesis, Rennes 1, 2018.

[FB07] Maycel-Isaac Faraj and Josef Bigun. Synergy of lip-motion and acoustic features in biometric speech and speaker recognition. *IEEE Transactions on Computers*, 56(9) :1169–1175, 2007.

[FDLM18] Cédric Fayet, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. Emo&ly (emotion and anomaly) : A new corpus for anomaly detection in an audiovisual stream with emotional context. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[IKF$^+$18] Denis Ivanko, Alexey Karpov, Dmitrii Fedotov, Irina Kipyatkova, Dmitry Ryumin, Dmitriy Ivanko, Wolfgang Minker, and Milos Zelezny. Multimodal speech recognition : increasing accuracy using high speed video data. *Journal on Multimodal User Interfaces*, 12(4) :319–328, 2018.

[KSPL20] Madhu R. Kamble, Hardik B. Sailor, Hemant A. Patil, and Haizhou Li. Advances in anti-spoofing : from the perspective of asvspoof challenges. *APSIPA Transactions on Signal and Information Processing*, 9 :e2, 2020.

[PMNJ20] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery.

[RSZ$^+$21] Manuel Sam Ribeiro, Jennifer Sanger, Jingxuan Zhang, Aciel Eshky, Alan Wrench, Korin Richmond, and Steve Renals. Tal : A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1109–1116, 2021.

[SPW$^+$18] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.