Speaker Authorization for Air Traffic Control Security

Marian Trnka[1], Sakhia Darjaa[1], Milan Rusko[1(✉)], Meilin Schaper[2], and Tim H. Stelkens-Kobsch[2]

[1] Institute of Informatics of the Slovak Academy of Sciences (Ústav Informatiky Slovenskej Akadémie Vied, UI SAV), Bratislava, Slovakia
milan.rusko@savba.sk

[2] Institute of Flight Guidance, German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR), Braunschweig, Germany

Abstract. The number of incidents in which unauthorized persons break into frequencies used by Air Traffic Controllers (ATCOs) and give false instructions to pilots, or transmit fake emergency calls, is a permanent and apparently growing threat. One of the measures against such attacks could be to use automatic speaker recognition on the voice radio channel to disclose the potential unauthorized speaker. This work describes the solution for a speaker authorization system in the Security of Air Transport Infrastructures of Europe (SATIE) project, presents the architecture of the system, gives details on training and testing procedures, analyses the influence of the number of authorized persons on the system's performance and describes how the system was adapted to work on the radio channel.

Keywords: ATC security · Radio channel speaker verification · Speaker authorization

1 Introduction

An unauthorized station can make malicious transmissions on an aeronautical frequency with the intention of misleading pilots. Such transmissions made at critical stages, e.g. during the take-off run or landing, can have potentially very dangerous consequences [1]. For example, in 2005, the pilot of an USAir flight approaching Washington's Reagan National Airport was instructed to divert his landing by a voice breaking into his frequency, which caused confusion for himself and for two other planes in position to land [2]. Or, in October 2016 a Virgin Australia flight was about 80 m away from the runway at Melbourne's Tullamarine Airport when it received instruction from an anonymous unauthorized person transmitting from an unknown location causing the pilot to pull up and change course [3]. Advanced automatic speaker verification technologies provide the ability to detect such attacks. It is possible to a) continuously monitor voice radio communication and b) verify the authorization of each speaker online. We discuss both points in 1.1 and 1.2 respectively.

A. Karpov and R. Potapova (Eds.): SPECOM 2021, LNAI 12997, pp. 716–725, 2021.
https://doi.org/10.1007/978-3-030-87802-3_64

1.1 Voice Radio Communication in Air Traffic Control

The analogue radio is still in use in the Air Traffic Control (ATC) voice communication. The radio-voice communication takes place between 118 and 137 MHz. The Very High Frequency (VHF) radio channel spacing is 25 kHz (or the reduced channel spacing 8.33 kHz) and double-sideband amplitude modulation (AM) is utilized. These frequencies are used for the voice communication of Tower (ground movements), Center (air movements) Automatic Terminal Information Service (ATIS). Audio frequency bandwidth for the voice-radio is limited to 300–2700 Hz. For the 8.33-kHz channel spacing, the speech frequency bandwidth was reduced to the 350 Hz to 2.5 kHz range [4].

In contrast to the advantage of relative robustness, this simple system has several disadvantages. One of them is the vulnerability to signals sent by attackers, such as the transmission of deceptive voice commands or interference by audio signals. In the past there were several attempts to increase the security of radio-voice communication using various approaches, such as watermarking [5] or voice biometrics [6] which are being implemented in one of the tools focusing on the security of airports [7].

1.2 Speaker Authorization

Language Dependence. English is used by default in civil international air traffic. Of course, pilots and ATCOs, whose mother tongue is not English, may speak with a strong foreign accent. This can affect automatic speaker verification and, for example, treat voices with the same foreign accent as to be more similar. In the current state of research, this aspect is not addressed. We took advantage of the fact that speaker verification can be considered a language independent task in the first approximation.

Speaker Verification. In the Speaker Verification a Binary Decision is Made and the Claimed Identity of a Speaker is Confirmed or Refused. There Are Two Types of Speaker Verification: Text-Independent Speaker Verification Verifies the Identity Without Constraints on Speech Content, and Text-Dependent Speaker Verification that Requires the Speaker Uttering Exactly the Given Password. The Approach Used in This Work is Text Independent.

Speaker Authorization/Speaker-Group Verification. Speaker authorization (SA) is the ability of a system to identify whether a speaker belongs to those having the permission to access the voice communication channel. The speakers trying to take part in the communication without the permission are designated as intruders. There are dozens of speakers who are authorized to communicate in a certain flight sector in any particular time and the number of potential intruders is practically unlimited.

To address this, a model of the incoming voice is created and compared to the group of models belonging to the authorized persons. The list of authorized persons is called the “whitelist” and the group of persons actually listed is called the “whitelist cohort”.

A speaker recognition can generally be done on a closed set of speakers, in which all the possible speakers are known, or on an open set, where the test sample may belong to a speaker that is unknown to the system. The speaker authorization is an open-set task that can be considered as a group-verification problem, as the specific identity of the speaker in the group is not important and only the affiliation to the group is verified.

Consequently, a binary decision is done in the speaker-group verification, by which the affiliation of a speaker to the "authorized" group is confirmed or refused. However, to achieve this goal multiple binary comparisons (speaker verifications) have to be done between the incoming sample and all the enrolled voices from the actual whitelist cohort. If the maximum score of all these comparisons is lower than a pre-defined threshold, the tested speaker is considered an unauthorized person. From a theoretical point of view, this is a special case of the multi-target cohort detection task [8] or open-set text-independent speaker identification [9].

Approach and Challenges. In this paper we present the solution of a speaker authorization module applied in the SATIE project's tool TraMICS (Traffic Management Intrusion and Compliance System), as well as the method and the results of the tests showing its feasibility in the context of the Air Traffic Management (ATM) security. Our approach uses automatic speaker verification to check whether the current speaker in the radio voice communication between the ATCOs and pilots belongs to the group of persons authorized to communicate in the particular time, channel, or sector. Additional challenges arise from the real-world implementation of such a speaker authorization system. First, the nature of a real-time application dictates constraints on robustness or speed. Second, the utterances submitted to the system are typically very short, usually ranging from 2 to 5 s. This is of course challenging since the reliability of the system increases with the amount of speech data under consideration. Additionally, we are facing communication on various channels. For the simulation environment VoIP speech in ATC is used, but the VHF radio channel is used in real operation.

2 Proposed Approach

2.1 Architecture of the Speaker Authorization (SA) Module

The illustrative schematic diagram of the architecture of the SA module is presented in Fig. 1. Technically, the SA module is based on the X-vector approach [10]. The Deep Neural Network (DNN), which was trained to discriminate between speakers, maps variable-length utterances to fixed-dimensional embeddings that are called X-vectors. Simply put, the X vectors serve as speaker models. Reverberation and noising were used for data augmentation. The module was created in the Kaldi environment [11]. In the verification phase an X-vector is extracted from the tested utterance and the Probabilistic Linear Discriminant Analysis (PLDA) is used to calculate a similarity score against the X-vectors of the whitelist cohort [12]. Decision on the affiliation of the speaker to the whitelist cohort is made by comparing the maximum similarity score with a threshold.

2.2 Training Data

The training data must represent a sufficient number of speakers, equipment, transmission channels, acoustic environments and background noises; we chose two large freely available databases VoxCeleb [13] and VoxCeleb2 [14].

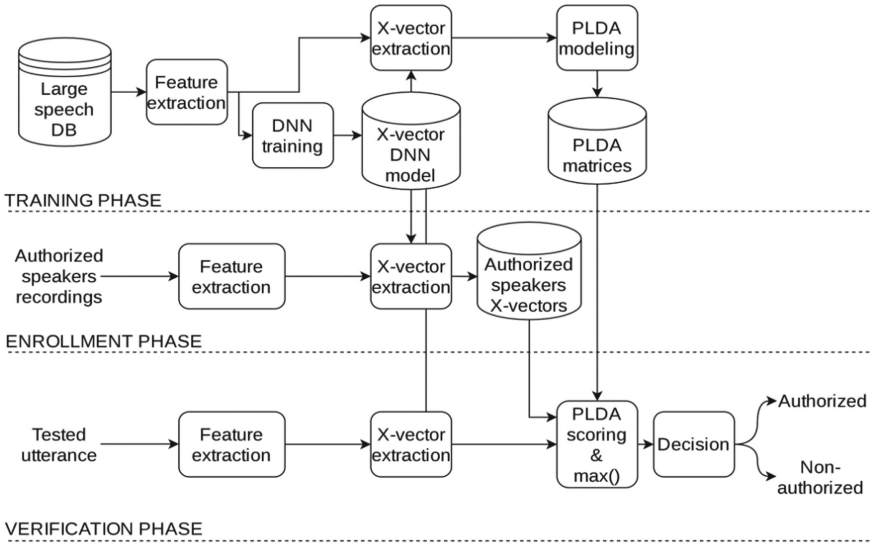

Fig. 1. Schematic diagram of the SA module.

In VoxCeleb the number of speakers in the training set is 1.211 and 40 in the test set. The number of utterances in the training set is 148.642 and 4.874 in the test-set.

VoxCeleb2 contains over 1 million utterances from over 6.000 speakers. The datasets are fairly balanced in terms of gender. The speakers span a wide range of different ethnicities, accents, professions and ages. Records are shot in a large number of challenging auditory environments. They are degraded with real-world noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and there is a broad range in the quality of recording equipment and channel noise [14].

2.3 Test Data

The results of speaker verification tests are highly dependent on the choice of speakers and test utterances. Experiments with users in real operation are not statistically representative, because any time-constrained testing in real operation can provide only a limited number of speakers and acoustic conditions while the variability of speakers and conditions the system has to deal with is enormous. They do not give enough information on the overall reliability of the system. Rather they can confirm usability of SA as an add-on tool in the system of ATC security.

Relevant reliability tests must therefore be performed on speech databases representing a wide variety of speakers and acoustic conditions. It was decided to use the following publicly available speech databases for testing: LibriSpeech, SpeechDatE Sk, and VoxForge.

LibriSpeech [15] offers 2.444 speakers and is large enough for testing the SA module. SpeechDat-E Sk [16] contains telephone speech (1.000 Slovak speakers), and is therefore suitable for some experiments evaluating the channel mismatch. VoxForge

[17] is a medium-size database which allows to conduct tests that do not require high computational power and time.

2.4 Tests

The off-line tests of the SA module were focused on three test cases that will be discussed in turn:

- Single-target speaker verification test;
- Group-verification test;
- Radio channel speaker verification test.

2.5 Single Target Speaker Verification

As the speaker authorization involves multiple repetitions of the speaker verification operation, speaker verification is the basic function that has to be reliably performed with the lowest possible error rate. Here the question is if the claimed identity of a speaker is confirmed or refused. To test this function, each particular utterance from the test database is fed into the speaker verification module and the true and false decisions are counted. The specific threshold value at which False Rejection Rate (FRR) is equal to False Acceptance Rate (FAR) is found and the Equal Error Rate (EER) is determined [18]. The EER gives information on the quality of the verification system when tested on the test-set of the VoxForge database.

Detailed results of the off-line tests on the three databases are presented in Table 1.

Table 1. Detailed results of the off-line single-target speaker verification tests.

Database name	No. of target speakers	No. of test files	EER [%]
LibriSpeech	2.444	24.440	0.86
SpeechDat-E Sk	888	888	0.90
VoxForge	579	7.215	1.63

2.6 Speaker Authorization/Speaker-Group Verification

The VoxForge database contains approximately 500 speakers. All possible groups that can be assembled from speakers in the test database should be considered, but as it will be shown, even when choosing a medium-size DB, the computational and time requirements for such a complete test can quickly exceed the capabilities of the available computational resources.

From a mathematical point of view, creating groups is making combinations without repetition. The number of k-element combinations of n objects, without repetition can be calculated as in (1):

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

As it can be seen in Table 2, the number of possible groups of speakers that can be created from the whole test database heavily depends on the number of persons in the group (i.e. currently authorized persons, or the whitelist cohort). It was decided to limit the number of the test-groups to 500.000 for each group size. Although such random selection does not cover all the possibilities of dividing speakers into groups, for the purposes of this paper this is considered a reasonable trade-off between time and computational requirements on the one hand and statistical representativeness of the test on the other hand. All speakers in the database are used for the random selection. All test samples are used in the test.

After this consideration, we are in position to formulate the question for group-verification tests as follows: "Does the test sample belong to some of the authorized speakers?". To answer the question, the test sample has to be compared to each of the members of the whitelist cohort. Naturally, a single comparison may be erroneous with a certain probability. As the size of the whitelist cohort increases, the number of needed comparisons increases, and the overall probability of error is rising.

The overlapping between the distributions of the non-target and the target scores in an open-set identification is greater than the overlapping of impostor scores and target scores in speaker verification. The bigger the target-set size, the greater is this overlapping [19].

For each size of the whitelist cohort the EER (i.e. equality of FRR and FAR) is reached at a different threshold. Therefore, an adaptive-threshold test needs to be performed, in which the threshold is changed so that a respective EER can be computed for each particular size of the authorized group.

Discussions with ATC practitioners have indicated that the number of people authorized to communicate over the voice channel on the given frequency, in the given moment, and in the given sector is typically up to 20. It was therefore decided to choose 30 as the maximum size for the group during the group verification tests. The VoxForge database was used as the test database for this evaluation. The results of speaker authorization tests are presented in Fig. 2.

Table 2. Number of groups that can be created from a test database of 500 speakers in relation to the number of members in the white list cohort.

Whitelist cohort	1	10	20	30
No. of possible groups	5×10^2	2.5×10^{20}	2.7×10^{35}	1.5×10^{48}

Histograms of score distributions of the target speaker and non-target speakers at various whitelist cohort sizes S (S = 1, 10, 20 and 30) are shown. The histogram of target score hardly changes depending on the whitelist cohort size, so only TAR_1 is shown.

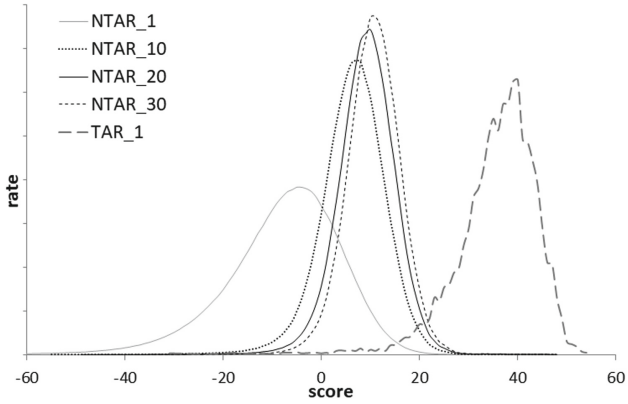

Fig. 2. Histograms of score distributions of the target speaker (TAR_1 - rightmost) and non-target speakers at various whitelist cohort sizes S (S = 1, 10, 20 and 30 – NTAR_1, 10, 20 and 30).

The score distribution of the target speakers' group is reasonably well separated from that of non-target groups even for the size of the whitelist cohort S = 30, indicating that the speaker authorization system is relatively reliable also for larger numbers of the whitelist cohort for the VoxForge database. Detailed results of the tests are presented in Table 3.

Table 3. Detailed results of the off-line multitarget group speaker verification tests - dependence of the EER on the number of currently authorized speakers.

Group size	1	5	10	15	20	25	30
EER [%]	1.6	2.5	2.9	3.2	3.5	3.7	3.8

2.7 Radio Channel Speaker Verification

For the purposes of the project SATIE the SA module was first designed for broadband clean speech signals, and was meant for the DLR's Tower-simulator in Braunschweig that uses VoIP channels for simulating ATC voice communication [20]. (DLR stands for German Aerospace Center, Deutsches Zentrum für Luft- und Raumfahrt e. V., literally German Center for Air- and Space-flight) However, in the real-life operation the system will be monitoring the voice-radio traffic with narrow frequency range, noises, distortions and other effects caused by the transmission via the radio channel.

To ensure that single speaker verification will work well on the radio channel, which is an inevitable condition for speaker authorization to work well, the speaker verification had to be tested on radio speech. There is a difference in the signal quality between the clean training data the original system was trained on and radio-quality test data, which is called channel mismatch. Hence, our task was to determine the influence of channel mismatch on the reliability of speaker verification and eliminate this influence.

No real-life radio communication database was available to the authors that was large enough and appropriately annotated for speaker verification testing. Therefore, it was decided to obtain radio recordings by transmitting an existing VoxCeleb [13] recording via a radio channel. Baofeng UV5R hand-held radios were used as transmitter and receiver. Due to local limitations, the transmission was performed on private mobile radio frequency 446.0 MHz. These frequencies belong to the UHF band, but the distortion and noise caused by the channel are similar to those on the VHF frequencies used in ATC.

Both, the original VoxCeleb database and the newly created radio channel "RadioVoxCeleb" database were split in corresponding non-overlapping training (VoxCeleb_train, RadioVoxCeleb_train) and test sets (VoxCeleb_test, RadioVoxCeleb_test). Detailed results of the off-line tests with various combinations of training and testing data are presented in Table 4.

Table 4. Detailed results of the channel mismatch.

Training data	Test data	EER
VoxCeleb_train	VoxCeleb_test	1.50%
RadioVoxCeleb_train	RadioVoxCeleb_test	2.80%
VoxCeleb_train	RadioVoxCeleb_test	5.90%
RadioVoxCeleb_train	VoxCeleb_test	3.00%
VoxCeleb_train + RadioVoxCeleb_train	VoxCeleb_test	1.20%
VoxCeleb_train + RadioVoxCeleb_train	RadioVoxCeleb_test	**2.60%**

The baseline system reached EER = 1.5% on the "clean" (i.e. original) signal and 5.9% on the radio signal. The best results were achieved by a system with multi-condition-trained models, which achieved an EER of 2.6% on the radio signal. Interestingly, the error in recognizing the clean signal has also been reduced (EER = 1.2%), which is likely a consequence of data augmentation with radio-channel signal.

The influence of the whitelist cohort size on the speaker authorization was tested using a system trained on VoxCeleb_train + RadioVoxCeleb_train set, and tested on RadioVoxCeleb_test set. The results are presented in Table 5.

Table 5. Results of the SA on radio channel with whitelist cohort size from S = 1 to S = 30.

Group size	1	5	10	15	20	25	30
EER [%]	2.6	4.9	6.5	7.2	7.6	8.2	8.6

The EER is relatively high, because the volume of the training data is small. The authors plan to create the radio version of the VoxCeleb2 database and use it for training. It can be assumed that the results will be significantly better.

3 Discussion and Conclusion

We proposed an architecture for a speaker verification system and tested its feasibility for the radio channel that corresponds to real-world use of such a system. In the VoIP mode we showed that the system reaches an EER below 4% even for the whitelist cohort size that exceeds normal application conditions. In the radio-channel mode, a sizeable decrease of the error (2.6% EER) compared to the original channel mismatched data (5.9% EER) was shown when the system was trained on the combination of augmented radio and original training data. The achieved performance meets the current expectations for a real-world air-traffic management application.

One such deployment of the speaker-authorization module is in the Traffic Management Intrusion and Compliance System (TraMICS) that serves as a detector for potential security incidents [21]. TraMICS analyses different indicators of the traffic situation combined with analyzing voices participating in radio-communication. This multimodal system can generate different kinds of alerts that TraMICS aggregates to a security situation indicator.

Due to the COVID 19 pandemic physical access to airports, air traffic control simulator facilities, and other areas where system validation would normally also take place, was drastically limited. Therefore, an on-site validation will be carried out in the near future, which will provide a scientifically based evaluation of personal opinions of future users of the system.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 832969. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein. For more information on the project see: http://satie-h2020.eu/. The results presented in this work were reported in the deliverable D4.2 of the above mentioned project SATIE. The work was also partly funded from the Slovak Scientific Grant Agency VEGA project No 2/0161/18.

References

1. EUROCONTROL: Radio interference (2020). https://www.skybrary.aero/index.php/Radio_Interference
2. Morgan, D.: Hackers attack air traffic control (2006). https://abcnews.go.com/US/story?id=95993&page=1
3. Morris, H.: Hoax caller impersonating air traffic control forces pilot to abandon landing (2016). https://www.telegraph.co.uk/travel/destinations/oceania/australia/articles/hoax-caller-impersonating-air-traffic-control-forces-pilot-to-abort-landing/
4. Eurocontrol: "Implications of end-to-end communication for air traffic control", EUROCONTROL (2009)
5. Hagmüller, M., Kubin, G.: "Speech watermarking for air traffic control", EUROCONTROL (2005)
6. Rusko, M., Trnka, M., Darjaa, S., Rajčáni, J., Finke, M., Stelkens-Kobsch, T.: Enhancing air traffic management security by means of conformance monitoring and speech analysis. In: Klempous, R., Nikodem, J., Baranyi, P. (eds.) Cognitive Infocommunications, Theory and Applications. Topics in Intelligent Engineering and Informatics, vol. 13. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-95996-2_9

7. SATIE project: Security of air transport infrastructure of Europe (2020). http://satie-h2020.eu/
8. Shon, S., et al.: The 1st multi-target speaker detection and identification challenge evaluation. In: Proceedings Interspeech, Graz (2019)
9. Reynolds, D., Singer, E., Douglas, A.: Analysis of multitarget detection for speaker and language recognition. In: ODYSSEY The Speaker and Language Recognition Workshop, number 4 (2004)
10. Snyder, D., Garcia-Romero, G., Sell, D., Povey D., Khudanpur, S.: X-Vectors: Robust DNN embeddings for speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary (2018)
11. Povey, D., et.al: The Kaldi speech recognition toolkit. In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011)
12. Kenny, P., et al.: PLDA for speaker verification with utterances of arbitrary duration. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada (2013)
13. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: Proceedings of INTERSPEECH (2017)
14. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: Deep speaker recognition. In: Proceedings INTERSPEECH (2018)
15. Panayotov, V., Chen, G., Povey, D.K.S.: Librispeech: An ASR corpus based on public domain audio books. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane (2015)
16. Pollak, P., et al.: SpeechDat(E) – Eastern European telephone speech databases. In: Proceedings of XLDB Workshop on Very Large Telephone Speech Databases, Athens (2000)
17. VoxForge: VoxForge (2006). www.voxforge.org
18. CISSP: The CISSP open study guide web site. https://web.archive.org/web/20081017165633/http://www.ccert.edu.cn/education/cissp/hism/039-041.html. Accessed 2021
19. Zigel, Y., Wasserblat, M.: How to deal with multiple-targets in speaker identification systems? In: Proceedings of IEEE Odyssey - The Speaker and Language Recognition Workshop, San Juan (2006)
20. Institute of Flight Guidance, German Aerospace Center: Apron and Tower Simulator (ATS). https://www.dlr.de/fl/en/desktopdefault.aspx/tabid-1964/1601_read-3011/ Accessed 2021
21. SATIE project D4.2 - Traffic Management Intrusion and Compliance System. SATIE project, 2021