DOI: 10.17469/O2111AISV000026

FRANCESCO SIGONA, GIUSEPPE VITOLO, MIRKO GRIMALDI

Forensic Automatic Speaker Recognition based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network

In the field of automatic forensic voice comparison (FVC), the use of neural networks models in the processing chain is increasingly frequent. Recently, Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) demonstrated high discrimination and accuracy in the non-forensic speaker verification task. In this contribution, after illustrating the fundamental differences between the forensic and non-forensic tasks of speaker verification – from a linguistic, logical and methodological perspective – the performances of a software implementation of automatic FVC based on ECAPA-TDNN, at different simulated operating conditions (noise level, net speech duration), are verified. Preliminary results confirm excellent performance in non-critical operating conditions. The hypotheses on the performance trend are also preliminarily confirmed: as the duration of the speech samples increases, and the noise level decreases, the evaluation metrics improve at a rate that depends on the combination of these two factors.

Keywords: Forensic linguistics, voice comparison, speaker verification, speaker recognition.

1. *Introduction*

Automatic Speaker Recognition (ASR), i.e. the task to verify if the characteristics of the voice sample of a test speaker match the enrolled model of another speaker, finds numerous applications, mainly where some kind of user authentication is required. Therefore, ASR is employed to grant authorized access to info-telematic services, or personal devices (for example, smartphones), or physical environments with restricted access, and so on, in the most varied application sectors (such as banking-financial and healthcare: Sigona, 2018).

On the other hand, speaker recognition also plays an extremely important role in the forensic field: Forensic Speaker Recognition (FSR, or Forensic Voice Comparison, FVC) makes it possible to establish from a probabilistic, qualitative and quantitative point of view, to what extent a recorded voice sample of unknown identity (the questioned-speaker recording) can be attributed to a particular speaker of known identity (e.g. a suspect speaker), according to some methodological requirements.

It is not surprising that ASR technologies are often used to build FASR (Forensic Automatic Speaker Recognition) systems (Slavíček, Swart, Klčo & Brümmer *et al.*, 2021, 2021; van der Vloed, Kelly & Alexander, 2020; Jessen, Meir & Solewicz,

2019): however, including such technologies in the forensic application is not straightforward, due to the delicacy of the matter and the consequences that the outcome of a vocal comparison can have on the fate of a judicial proceeding.

In this work, the fundamental differences between the forensic and non-forensic tasks of speaker verification, from a linguistic, logical and methodological perspective, are highlighted (cf. § 2.)

This contribution also presents one of the first implementations, with relative experimentation, of a FASR system based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN). This is a very recent neural network model which, for non-forensic applications, has already shown better performance in terms of recognition error rates, compared to previous models (Desplanques, Thienpondt & Demuynck, 2020). In § 3 the FVC based on ECAPA-TDNN is described, from a typical FVC processing chain to the actual implementation. In § 4, the experimental design used to achieve the system performance in different case-work operating conditions is illustrated, and in § 5 the corresponding results are presented. Finally, in § 6 the results are discussed, while conclusions are drawn in § 7.

2. *Forensic and non-forensic speaker verification tasks.*

Multiple ASR techniques have been developed over the years: first they model appropriate features of the voice sample of the user to be authenticated, and then compare this model with one or more pre-enrolled speaker models, usually stored in the recognition system. These techniques include the Joint Factor Analysis (Kenny, 2005; Kenny, Boulianne, Ouellet & Dumouchel *et al.*, 2007; N. Dehak, Kenny, R. Dehak, Glembek, Dumouchel, Burget, Hubeika & Castaldof, 2009), the i-vector-based paradigms (N. Dehak, Kenny, R. Dehak, Dumouchel & Ouellet, 2011), up to the more recent introduction of Deep Neural Networks (DNN) in some place in the processing chain (Lei, Scheffer, Ferrer, & Mclaren, 2014, among the first attempts) or as an end-to-end approach (Snyder, Ghahremani, Povey, Garcia-Romero, Carmiel & Khudanpur, 2016). The result of an ASR is generally a numerical value (score) which in most cases expresses little more than a similarity between the input and the enrolled model(s). The outcome of the procedure is binary (authenticated/rejected) and is determined by comparing the score against a threshold value, suitably designed to meet certain requirements in terms of probability of decision error (false authentication of a foreign speech sample, false rejection of a legitimate speech sample). These requirements are in turn chosen based on the best acceptable compromise between the two error rates, according to the addressed application.

The voice comparison in the forensic application is a conceptually very different thing, in which the "general-purpose" ASR technologies can certainly be exploited, but this constitutes only one of the building blocks of the system. The FVC is a particularly dedicated field of application: in this case the need for the comparison methodology to be based on quantitative measurements lead to acceptable error

rates, and above all to be reproducible and validated in the scientific community has been repeatedly highlighted (Enzinger, Morrison & Ochoa, 2016; Morrison, 2014). The most established methodology in the literature is based on a Bayesian approach recently also recommended by the European Network of Forensic Science Institutes (Drygajlo, Jessen, Gfroerer, Wagner, Vermeulen & Niemi, 2016). In this setting, the FVC is not configured as a decision test and therefore the outcome is not binary (identified/rejected). Rather, it is expressed in the form of a Likelihood Ratio (LR; Morrison, 2009; Morrison, 2011) between the prosecution hypothesis (of identification, in which the degree of similarity between the two test samples prevails) and the defence hypothesis (non-identification, in which the degree of typicality of the items prevails, i.e. the similarity with a set of samples taken from a suitable "reference population").

There are many factors that can influence the performance of an ASR system. With the same choice of algorithms and related parameters, a decisive role is played by the operating conditions of the comparison, case by case. These conditions include:

(1) the effect of the transmission/recording channel of the vocal sample, and therefore of alteration of the frequency spectrum of the voice.
(2) the level and type of noise present in the recordings.
(3) the net duration of the speech inside the recorded samples.

In the case of FVC, an additional condition must be taken in account:

(4) socio-linguistic factors such as gender and the linguistic variety of the speakers, which also concern the logical-methodological correctness of the comparison (in light of which the results of the comparison and the system performance metrics are to be interpreted.).

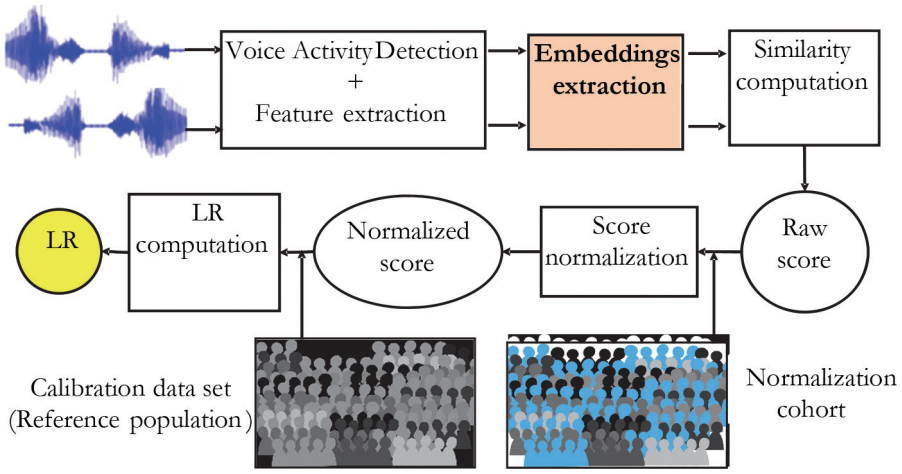
The linguistic variety of the spoken samples must be assessed to correctly formulate the defence hypothesis: accordingly, the questioned voice sample has been produced by a different speaker than the one investigated. It is immediate to understand that this "generic" different speaker cannot be one speaking a different language or variety but must be able to speak the same variety of the anonymous speaker as fluently as he does. Therefore, unlike the non-forensic ASR, in a FVC it is essential to have a so-called "reference population", which reflects not only the recording conditions of the case at hand, but also the linguistic variety. The selection of a suitable "reference population", in terms of recording conditions, linguistic variety, speaker gender and, where possible, also speaking style, is a key factor that can dramatically influence the outcome of FVCs.

3. *Typical processing chain for the FASR task.*

Figure 1 shows a typical sequence of steps for a FASR/FVC. The input of the system usually consists of vocal samples in which only the voice of the speaker of interest is present. Therefore, in the case of an interception of a conversation between two (or more) speakers, the forensic expert usually takes care of segmenting and

concatenating only the parts of the sample containing the requested voice, before introducing this data into the system. Each of the two resulting vocal samples to be compared are subsequently converted into a sequence of a variable number of fixed-length feature vectors, relating only to the spoken part (excluding any relevant pauses still present). These sequences constitute the input for the extraction of the vectors of embeddings, one vector for each of the two voice samples. The two vectors of embeddings are input into a similarity function to get a raw similarity score. A score normalization technique can also be used to mitigate the effect of mismatch between the recording conditions of the speech samples to be compared. This usually requires a normalization cohort, i.e. a dataset of speech recordings in various recording conditions, including those of the actual comparison. The last step is a score-to-forensic likelihood ratio conversion, also known as "calibration".

Figure 1 - *Schematic of a typical processing chain of a Forensic Voice Comparison system, from the input voice samples to the finale likelihood Ratio calculation*



3.1 Implementation of the processing chain

We have implemented the reference scheme illustrated in Fig. 1, as follows:

a. Input audio samples: each voice sample only contains voice segments, with pauses no longer than 100 ms.
b. Feature extraction and embeddings calculations: these steps are in charge to the pre-trained ECAPA-TDNN model, that will be presented in § 3.2.
c. Similarity computation: a raw comparison score is calculated by means of the well-known cosine similarity scoring function between the two input vectors of embeddings w_1 and w_2:

$$(1) \qquad score(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|}$$

d. Score normalization: Symmetric Normalization (S-norm) is used (Shum, N. Dehak, R. Dehak & Glass, 2010). The questioned-condition vector of embeddings

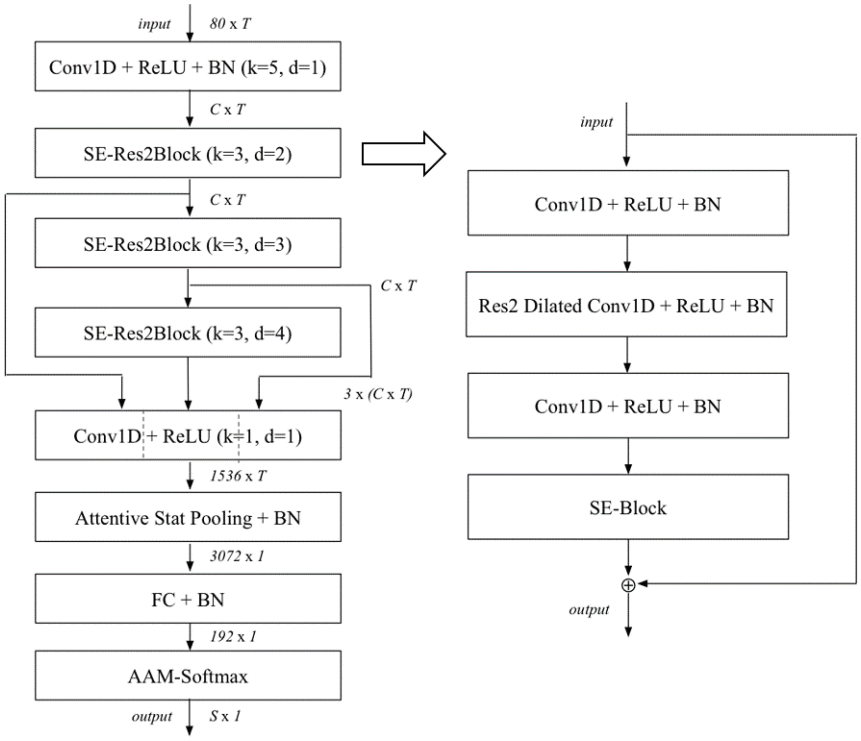
is compared with each vector of embeddings of the normalization cohort, producing a set of scores, from which the normalization statistics (mean and standard deviation) are calculated. The suspect-condition vector of embeddings is also compared with each vector embeddings of the cohort, producing a second set of scores, and a second pair of normalization statistics (a second mean and a second standard deviation). Finally, the final normalized score is obtained by normalizing the score with both the statistics sets (subtracting the mean and dividing by the standard deviation), and then taking the average of the two resulting values. In the current work, the normalization cohort is selected as a portion of whole dataset, while the remaining portion is used as reference population.

e. Score-to-forensic-likelihood-ratio conversion (calibration). For each speaker in the reference population, at least one speech sample in the questioned-speaker recording condition and one sample in the known-speaker recording condition must be available. By this way, it is possible to compute the set of the same-speaker scores (i.e. the scores between the two samples of each speaker), and the set of the different-speakers scores (i.e. the scores between the actual questioned-speaker sample and each sample of the reference population in the known-speaker recording conditions). The two set of scores are used to calculate respectively the same-speaker scores distribution and the different-speakers scores distribution, by means of a Kernel Density Estimate (Parzen, 1962). The same-speaker density evaluated at the actual score value has the meaning of a likelihood of the prosecution hypothesis, while the different-speaker density evaluated at the actual score value has the meaning of a likelihood estimate of the defense hypothesis. The LR is achieved as the ratio between these two likelihood values.

3.2 The ECAPA-TDNN model: a short overview.

According to Desplanques *et al.* (2020), ECAPA-TDNN includes multiple enhancements to the baseline TDNN-based x-vector architecture which applies statistics pooling to project variable-length utterances into fixed-length speaker characterizing embeddings. Such enhancements include restructuring the frame layers into 1-dimensional Res2Net modules with impactful skip connections, also introducing Squeeze-and-Excitation (SE) blocks in these modules to explicitly model channel interdependencies. Moreover, hierarchical features from different network layers are aggregated and propagated, and the statistics pooling module has been improved with channel-dependent frame attention, to make able the network to focus on different subsets of frames during each channel's statistics estimation.

Figure 2 - *Left diagram: network topology of the ECAPA-TDNN. k is the kernel size and d is the dilation spacing of the Conv1D layers or SE-Res2Blocks. C (=1024) and T are the channel and temporal dimension of the intermediate feature-maps respectively. S is the number of training speakers. BN stands for Batch Normalization (Ioffe and Szegedy, 2015). Right diagram: the SE-Res2Block of the ECAPA-TDNN architecture. The standard Conv1D layers have a kernel size of 1. The central Res2Net Conv1D with scale dimension s = 8 expands the temporal context through kernel size k and dilation spacing d. Adapted from Desplanques et al. (2020)*



In the present work, we used the ECAPA-TDNN model available on the Hugging Face platform[1], which was pre-trained using Voxceleb1 + Voxceleb2 data sets (Nagrani, Chung, Xie & Zisserman, 2020) and Additive Margin Softmax Loss. The network model has been implemented within the SpeechBrain toolkit (Ravanelli, Parcollet, Plantinga, Rouhe, Cornell, Lugosch, Subakan, Dawalatabad, Heba, Zhong, Chou, Yeh, Fu, Liao, Rastorgueva, Grondin, Aris, Na, Gao, De Mori, & Bengio, 2021), a well-known Open-Source Conversational AI Toolkit. The model is configured to work with 80 log Mel filter banks energies as frame level input features, 1024 channels in the convolutional layers, dimension of the bottleneck in the SE-Block and attention module set to 128, scale dimension *s* in the Res2Block

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb.

(Gao, Cheng, Zhao, Zhang, Yang & Torr, 2019) set to 8, and 192 nodes in the final fully-connected layer. The model has more than 20 million trainable parameters.

In the context of the present work, the model is essentially used to transform a vocal sample to its fixed-length vector of embeddings, for further processing.

3.3 Evaluation metrics

The performances of a FSR system can be represented by means of graphics (such as Tippet plots: Morrison, 2010), and numerical metrics, such as the Equal Error Rate (EER) and, even better, the Log-Likelihood Ratio Cost (Cllr).

The C_{llr} summarizes the overall system quality and have the following formulation (Brümmer and du Preez, 2006; van Leeuwen and Brümmer, 2007; Morrison, 2011):

$$(2) \qquad C_{llr} = \frac{1}{2}\left[\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} \log_2\left(1+\frac{1}{LR_{ss_i}}\right) + \frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}} \log_2\left(1+LR_{ds_j}\right)\right]$$

In the above formulation, LR_{ss} are the likelihood ratio values in the case of same-speaker comparison, while LR_{ds} are likelihood ratio values in the case of different-speakers comparisons. Since LR_{ss} values much greater than 1 better support the same-speaker hypothesis, and LR_{ds} values much less than 1 better support the different-speakers hypothesis, smaller values of C_{llr} indicate better performances. On the contrary, a system which provides no useful information and always responds with a likelihood ratio of 1 will result in a C_{llr} value of 1.

The EER is another largely used metric to evaluate the discriminating power of the system. Likelihood ratios test values can be combined with some prior odds to achieve posterior odds; in turn, posterior odd can be matched against a threshold to classify a test comparison as same-speaker (prosecutor hypothesis) or different-speakers (defense hypothesis). By this way, false "identification" and false "rejection" error rate can be computed as proportion of wrong classifications. Adjusting the priors and the threshold in such a way that the two error rates are equals, the resulting error rate is called the EER. This metric can be calculated using the Receiver Operator Characteristic Convex Hull method (Brümmer and de Villiers, 2013).

In this work, EER and C_{llr} are used to assess the performance of the tested FVC system.

4. *Experimental design*

In this pilot experiment, we wanted to test the pre-trained ECAPA-TDNN model as previously described, carrying out batteries of comparison tests between speech recordings in Italian. As far as we know, structured databases of Italian voice recordings with an adequate number of speakers, different sociolinguistic profiles and different speaking styles are not available for the validation of FVC systems. For this reason, a test dataset was built by the authors, collecting extracts of audiobook narratives. A dataset was created with 127 audiobook male narrators, collecting two recording excerpts for each speaker, lasting approximately three minutes each,

without any particular selection criteria, with the exception of the total absence of background melodies. Considering the goal of our work, this dataset can still be useful for the purposes of the study, and it can constitute the basis for a more structured and extensive activity of collecting vocal samples that can be used as a reference population.

Starting from this dataset, signal filtering, cutting and noise contamination has been used to simulate telephone-quality speech recordings, at different duration and different Signal-to-Noise Ratio (SNR) with Additive White Gaussian Noise (AWGN) and babble noise. Noise contamination at a given SNR value has been achieved by the FaNT tool (Filtering and Noise adding Tool)[2]. The whole information about the experimental conditions is summarized in Table 1. Different channels conditions and different types of noise will be experimented in future works.

Once the questioned- and known-speaker conditions have been chosen among those envisaged for the various tests, the questioned-speaker condition specifications are used to adapt all first recordings of the speakers in the dataset, and the known-speaker condition specifications are used to adapt the second recordings of the same speakers. This allows to build couples of recordings that can simulate actual caseworks in the selected operating conditions, in which same-speaker comparisons (when both the recordings belong to a same speaker) as well as differ-speakers comparisons (when the two recordings belong to different speakers) occur.

According to a one-two-speakers-out cross-validation paradigm, the pairs of recordings are extracted one at a time, and for each pair a new reference population is considered consisting of all the recordings that were not produced by the extracted speakers. By this way, for each couple, it is possible to compute an unbiased LR value, following the procedure illustrated in the **§** 3. The set of LR values is finally used to compute the evaluation metrics.

This process is reiterated for each operating conditions, to achieve evaluation metrics for each one.

Table 1 - *Summary of the parameters considered in the experimental design*

Spectral feature computation	80 mel filter bank, in charge to pre-trained ECAPA-TDNN models
Embeddings extraction	based on ECAPA-TDNN (192 embeddings)
Similarity	cosine distance
Score normalization	Adapted Symmetric Norm (AS-Norm)
Reference population	127 Italian male adults (audiobook narrators)
Channel emulation	filtering in the telephone bandwidth (300-3400 Hz)
Noise contamination	Additive White Gaussian Noise (AWGN); Babble noise; SNR levels: 6, 12, 18, 24 dB
Questioned- recording duration	5, 10, 20, 30, 40, 50, 60 s
Suspect-recording duration	10, 20, 30, 40, 50, 60 s
Metric for evaluation:	Cllr (Log-likelihood ratio cost), EER (Equal Error Rate)

[2] http://dnt.kr.hs-niederrhein.de/indexbd2f.html.

5. *Experimental results.*

Figure 3 shows the EER and C_{llr} trends, where the audio were filtered in the telephone band (300-3400 Hz), the known-speaker recordings have 24 dB SNR and different durations, while questioned-speaker recordings have various SNR values and various duration. All the audio samples we contaminated by AWGN. Figure 4 shows similar results achieved when all the audio samples we contaminated by a babble noise.

Figure 3 - *EER (upper plots) and C_{llr} (lower plots) as a function of the duration of the questioned-speaker recordings (horizontal axes), at different SNR values of the questioned-speaker recordings (colored lines) and different durations of the known-speaker recordings (from 10 s to 60 s, at the top of each plot). The known-speaker recordings have 24 dB SNR. All recordings have 300-3400 Hz frequency bands. Noise contamination with Additive White Gaussian Noise (AWGN)*

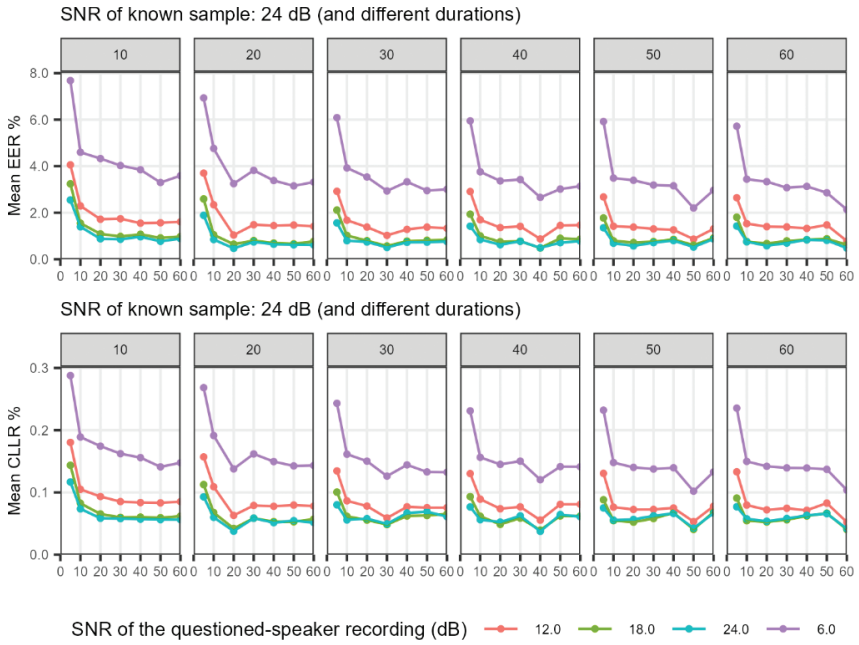
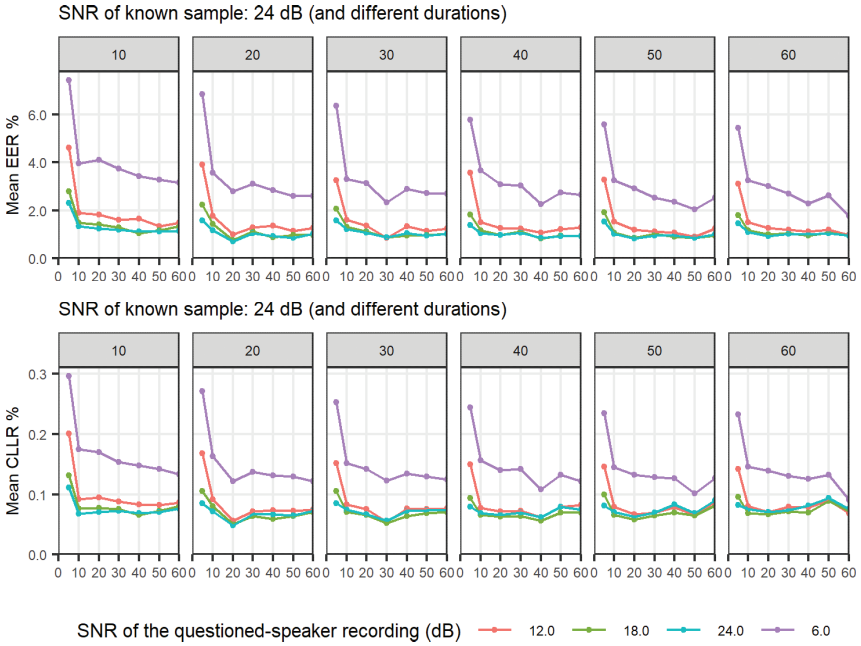


Figure 4 - *The same as Figure 3, but audio samples were contaminated by babble noise*



6. *Discussion*

The tests conducted and the results represented in Figures 3-4, although obtained with an unstructured dataset of speakers, not informed about the sociolinguistic conditions of the subjects, offer numerous insights for the forensic application of the tested system.

a) Duration of the suspect recording. Usually, suspect recording is carried out with a controlled procedure to obtain high quality audio, at most adapted, or adaptable in post-processing, to the recording channel of the questioned recording. In the tests carried out here, this situation can be traced back to the conditions of high SNR and long duration of the suspect recording.

b) Duration of questioned-speaker recording. There is a tendency to improve EER and C_{llr} as the duration of the questioned-speaker recording increases, going from 5 seconds to 60 seconds. This is an expected behaviour, because vocal samples that are too short may express only a small part of the characteristics of the speaker's voice, and therefore may lead to lower accuracy levels. As the net duration of the speech sample increases, so does the variety of these characteristics exhibited, and thus the system becomes able to build a more accurate model of the speaker's voice.

c) Effect of the loudness. The rate at which system accuracy improves as the voice sample length of the anonymous voice increases depends on the loudness of the signal (and, to some extent, on the duration of the known-speaker recording).

Indeed, when the anonymous sample has high SNRs, the accuracy improvement occurs suddenly, as early as 10 seconds, or at 20 seconds when the duration of the known sample is only 10 seconds. In general, increasing the duration of the known sample tends to improve performance when the conditions of the anonymous sample are not optimal (low duration, low SNR). As the SNR of the anonymous sample increases, it becomes less and less important to have high durations of the same sample. It should also be noted that the Cllr values remain below 0.1 already with values of 12 dB, for all the durations of the known sample. In particular, considering that the known sample is very often the result of a special recording, carried out in good recording conditions and without time limits, it can be stated that, except in the case of very noisy anonymous samples, the Cllr values settle comfortably below 0.05 in many operating conditions.

d) The system seems to perform quite similarly in both AWGN and babble noise conditions, at least in the scope of the tests conducted.

7. *Conclusions*

Our study provided preliminary results of the application of the ECAPA-TDNN model in the field of forensic voice comparison. Specifying "forensic" in the voice comparison task indicates that the speaker recognition procedure rises to a higher level of complexity. In addition to the similarity between the two voice recordings compared, their typicality, with respect to a reference population, must be taken into account, applying the Likelihodd Ratio Framework. The variety of conditions that can arise in a real case is such that it generally makes the task of testing an FVC system in all conditions very difficult. For this reason, FVC systems are most frequently tested under a few conditions, which are assumed to be more or less representative of one possible condition among many. One of the major difficulties in this sense is the lack, or in any case the difficulty, of building structured databases from a sociolinguistic point of view and of other factors susceptible to altering the speaking style. These difficulties also exist in particular for the Italian area. To overcome this problem, at least to a minimal extent, a dataset of speakers was specifically collected for the study presented here, without adopting particular selection criteria, and devoid of sociolinguistic information. Such a dataset, despite all the limitations already highlighted, still proved functional in highlighting the performance of the tested model, which is the main purpose of this work.

The experimentation carried out shows that a FVC based on the representation of voice utterances by means of embeddings calculated with ECAPA-TDNN can provide excellent levels of accuracy in the tested operating conditions. The results achieved encourage us to continue experimenting by introducing various other types of noise (e.g. babble, car, etc.) and with other reference populations, distinguished by gender of speakers and different linguistic varieties.

References

Brümmer, N. & Du Preez, J.A. (2006). Application-independent evaluation of speaker detection. In *Computer Speech & Language*, 20(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N.& de Villiers, E. (2013). The BOSARIS toolkit: theory, algorithms and code for surviving the new DCF. arXiv:1304.2865

Dehak, N., Kenny, P. Dehak, R., Glembek, O. Dumouchel, P., Burget, L., Hubeika, V., & Castaldo, F. (2009). Support vector machines and joint factor analysis for speaker verification. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP'09), 4237–4240.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P. & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proc. Interspeech,* Brigthon, U.K.

Dehak, N., Kenny, P., Dehak, R., DumouchelP. & Ouellet, P. (2011). Front-end factor analysis for speaker verification. In *IEEE Trans. Audio Speech Lang. Process*. 19(2011) 788–798.

Desplanques, B., Thienpondt, J. & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, 3830-3834, doi: 10.21437/Interspeech.2020-2650.

Gao S., Cheng, M.M., Zhao K., Zhang, X., Yang, M.-H & Torr, P.H.S. (2019). Res2Net: A new multi-scale backbone architecture. In *IEEE TPAMI*, 2019.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2016). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises. *https://enfsi.eu/*

Enzinger, E., Morrison, G.S. & Ochoa F. (2106). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic voice-comparison case. In *Science & Justice*, 56, 1(2016): 42–57. DOI:10.1016/j.scijus.2015.06.005.

Jessen, M.E., Meir, G., & Solewicz, Y.A. (2019). Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). In *Speech Communication*, 110, 101–107. https://doi.org/10.1016/j.specom.2019.04.006

Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. In*Tech. Rep. CRIM-06/08-13,* CRIM, Montreal, Quebec, Canada, 2005.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigen channels in speaker recognition. In *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 4, pp. 1435–1447, 2007.

Van Leeuwen, D.A., & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In *Lecture Notes in Computer Science*, 330–353, Springer Science+Business Media. https://doi.org/10.1007/978-3-540-74200-5_19

Lei, Y., Scheffer, N., Ferrer, L.& McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 1695–1699. doi: 10.1109/ICASSP.2014.6853887

Morrison, G.S. (2014). Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. In*Science & Justice*, 54(2014): 245–256.

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. In *Science & Justice*, 51(3), 91–98. https://doi.org/10.1016/j.scijus.2011.03.002

Morrison, G.S. (2010). Forensic voice comparison. In Freckelton, I., Selby, H. (a cura di), *Expert Evidence*. Thomson Reuters, Sydney, Australia ch. 99. http://expert-evidence.forensic-voice-comparison.net

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. In *Science & Justice,* 49(2009): 298–308.

Nagrani, A., Chung, J.S., Xie, W. & Zisserman A. (2020). Voxceleb: Large-scale speaker verification in the wild. In *Comput. Speech Lang*., 60, 2020.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. In *Annals of Mathematical Statistics*, 33(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Ravanelli M., Parcollet T., Plantinga P., Rouhe A., Cornell S., Lugosch L., Subakan C., Dawalatabad N., Heba A, Zhong J., Chou JC., Yeh SL., Fu SW, Liao CF, Rastorgueva E., Grondin F., Aris W., Na H., Gao Y., De Mori R. & Bengio Y. (2021). In SpeechBrain: A general-purpose speech toolkit. *arXiv:2106.04624*.

Shum, S., Dehak, N., Dehak, R. & Glass, J. (2010). Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification. In *Odyssey*, 16. http://people.csail.mit.edu/sshum/papers/Odyssey10paper.pdf

Sigona, F. (2018). Voice Biometrics Technologies and Applications for Healthcare: an overview". In *Journal of interDisciplinaryREsearch Applied to Medicine (JDREAM)*, Vol.2, Issue 1, 2018, ESE - Salento University Publishing e-ISSN: 2532-7518, 5-15. http://siba-ese.unisalento.it/index.php/jdream; DOI: 10.1285/i25327518v2i1p5

Slavíček, J., Swart, A., Klčo, M., & Brümmer, N. (2021). The PhonexiaVoxCeleb Speaker Recognition Challenge 2021 SystemDescription. https://arxiv.org/abs/2109.02052

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y. & Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, 165–170. doi: 10.1109/SLT.2016.7846260

van der Vloed, D., Kelly, F. & Alexander, A. (2020). Exploring the Effects of Device Variability on Forensic Speaker Comparison Using VOCALISE and NFI-FRIDA, A Forensically Realistic Database. In P*roc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 402-407, doi: 10.21437/Odyssey.2020-57.