Effects of emotional speech on forensic voice comparison using deep speaker embeddings

Mohammed Hamzah Abed[1,2], Dávid Sztahó[1]

[1]Department of Telecommunications and Media Informatics,Budapest University of Technology and Economics,Magyar tudósok körútja 2, 1117 Budapest, Hungary

[2]Department of Computer Science, Faculty of Computer Science and Information Technology,Al-Qadisiyah University,Al Diwaneyah, Iraq
m.abed@tmit.bme.hu ,sztaho.david@vik.bme.hu

Abstract. Emotional conditions play a significant role in forensic voice comparison and speaker verification systems. When emotion is present in speech, the verification's performance will deteriorate. In this paper, speaker verification has been investigated and analyzed in the case of emotional speech using metrics evaluating the performance of forensic voice comparison using pre-trained speaker embedding models: x-vector and ECAPA-TDNN for embedded feature extraction. This study investigates whether emotional content affects the forensic voice comparison and verification performance evaluated on a Hungarian speech dataset. The speaker verification performance has been assessed using the likelihood-ratio framework using Cllr and $Cllr_{min}$ and Equal Error Rate. The ECAPA-TDNN achieved higher performance than the x-vector. In the same emotion scenario, the best EERs were 2.6% and 7.7% for ECAPA-TDNN and x-vector. Both models are sensitive to the emotional content of the speech samples.

Keywords: Forensic voice comparison, speaker verification ; x-vector, ECAPA-TDNN, likelihood-ratio framework.

1 Introduction

Speaker recognition is usually classified into two major fields: speaker verification (automatic speaker verification, ASV) and identification (Sztahó et al., 2021). Speaker verification aims to verify if the voice belongs to the claimed identity by comparing the voice with another voice in the dataset and verifying whether the same person produces it. Unlike speaker identification concept, which aims to identify the speaker by selecting one model from a set of enrolled speaker models (Sztahó & Fejes, 2022). One of the applications of speaker verification dealing with forensic voice comparison is based on comparing an unknown criminal's voice with a well-known suspect's voice like when a DNA sequence is matched with another known DNA profile (Sztahó & Fejes, 2022). Furthermore, the decision of speaker verification in such a forensics case is very critical. It must present a high level of confidence because the error-prone behavior is critical and the error is unacceptable. Due to the behavioral and biological differences between people and the way they speak, each person's voice contains unique information. This allows people to recognize each person from his\her voice (Arya et al., 2021). The main

aim of this study is to analyze and investigate how emotional content affects forensic voice comparison and verification performance. The experiments were done on a Hungarian speech dataset.

To compare two samples of voices and judge whether they belong to the same identity, we need to consider how the voices are similarly based on the features extracted from each sample. In the last few years, deep learning has been a significant tool in speaker verification and emotional speaker recognition. Studies show that these deep-learning embeddings outperform previous i-vector based features. However, a lot of training data is needed to get a highly efficient model (Sztahó & Fejes, 2022). In this paper, we have used two speaker embedding methods commonly used in speaker verification: x-vector and ECAPA-TDNN. Pre-trained models were applied that are available in the Huggingface repository. The speaker identities were evaluated in a forensic voice comparison framework by calculating the likelihood ratio based on cosine distance between sample pairs and logistic regression models.

Recently, in forensic voice comparison, there have been studies dealing with speaker verification where emotional content is present (Scherer et al., 2000). An acoustic analysis regarding the effect of the emotional content of possible automatic speaker verification systems shows that an evaluation of training ASV material on emotional speech requires in-depth analyses of the individual differences in vocal reactivity and further exploration of the link between acoustic changes under stress or emotion and verification results. In (Rusko et al., 2018), researchers investigate the weakness of voice as a biometric model and try to improve the performance of the verification system. In addition, they used the emotional speech dataset to increase the diversity of all cases belonging to the speaker. The suggested model has been evaluated based on CRISIS dataset with six levels of emotion per speaker and used i-vector as embedded features with PLDA. In (Shahin et al., 2021), the authors proposed a hybrid deep neural network for speaker verification in an emotional case study. Four DNN models (DNN-HMM, DNN-GMM, GMM-DNN and HMM-DNN) have been used. These models were evaluated using three different speech datasets: private Arabic and two English public datasets. Their result shows the HMM-DNN outperformed all other models in an emotional and stressful environment and shows high performance in terms of equal error rate (EER), which was 7.19%. In another work (Prasetio et al., 2020), the authors proposed a speaker verification model under stress conditions by applying i-vector and investigated the effect of the emotional speaking style in speech. Emotional Variability Analysis (EVA) has been proposed which is based on i-vector technique, but it considers the emotional effect as the channel variability component. Based on the experimental result, the proposed model was outperformed the standard i-vector. Biswajit and his colleagues (Dev Sarma & Kumar Das, 2020) tried to map i-vector embeddings to an emotionally invariant space. They obtained a slight performance increase in speaker identification using the IEMOCAP dataset compared to models trained only on neutral samples. Parthasarathy and his colleagues (Parthasarathy et al., 2017) also used i-vector features to test speaker verification with expressive speech. The results show that speaker verification errors increase when the values of the emotional attributes increase, but the overall results are reliable. Pappagari and his colleagues (Pappagari et al., 2020) applied a more novel, deep learning-based embedding method, the x-vector for speaker

verification with an emotional speech. They observed that speaker verification performance is prone to changes in testing speaker emotions. They found that trials with angry utterances performed worst in all three datasets.

The present study aims to extend the speaker verification topic by applying two speaker embedding models based on deep learning in a forensic voice comparison scenario. We investigate how the ECAPA-TDNN embeddings perform with emotional speech and how they relate to the x-vector embeddings. The evaluation of the workflow model was investigated in the likelihood-ratio framework. And the performance of the models was evaluated by equal error rate (EER) of speaker verification, log-likelihood-ratio cost (Cllr) and $Cllr_{min}$ (Sztahó & Fejes, 2022). Samples were compared in multiple emotion combination cases.

The rest of this paper is organized as the following: section 2 illustrates the related work of emotional speech verification and deep learning embedding in speaker recognition. The proposed model and the methodology are described in section 3. Section 4 included the result and analysis. Finally, section 5 shows the conclusion and future scope.

2 Materials and Methods

The workflow of speaker verification based on speaker embedding vectors (x-vector and ECAPA) has been evaluated on a Hungarian dataset containing emotional and neutral speech. Figure 1 illustrates the layout of the applied process. Two methods have extracted speaker embeddings from the audio samples: x-vector and ECAPA-TDNN. Feature extraction was implemented in the SpeechBrain toolkit (Ravanelli et al., 2021), and the pre-trained models were downloaded from Huggingface[1,2]. After extracting features from speech samples, cosine similarity was calculated between feature vector pairs. The likelihood-ratio scores were obtained by feeding the cosine similarity scores to a logistic regression model trained on the ForVoice120+ dataset (Sztahó & Fejes, 2022).

[1] x-vector: https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[2] ECAPA-TDNN: https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

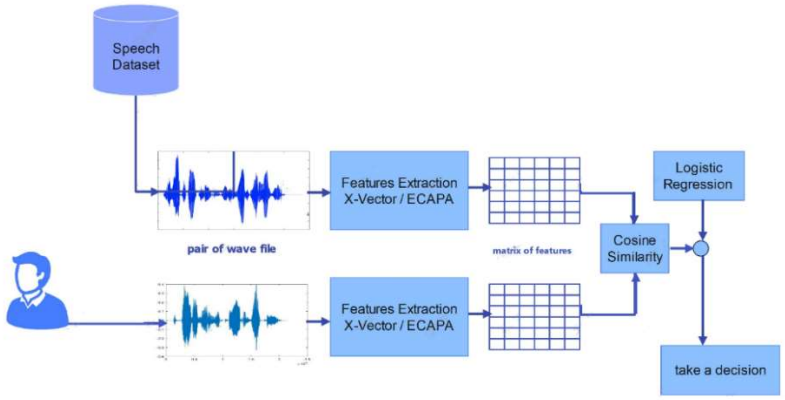

Fig. 1. Speaker verification process applied in the study

2.1 Dataset

The Hungarian Emotional Speech Database has been used in this work to evaluate the proposed models. This dataset consists of 38 volunteers (20 females and 18 males). The voices have been recorded in a quiet office room, and the recording equipment was a Sound Blaster NX 2 USB external sound card with a Monacor ECM-100 microphone. The recordings were PCM encoded with 16-bit quantization and 16 kHz sample rates. Three different sentences were recorded with each speaker in Hungarian with eight other emotions: sadness, anger, fear, excitement, disgust, surprise, joy and neutral. The linguistic content of the sentences are:

(1) "Kovács Katival szeretnék beszélni" (English: "I would like to speak with Kovács Kati.")
(2) "A falatozóban sört, bort, üdítőitalokat és finom malacsültet lehet kapni." (English: "In the snack bar you can get beer, wine, beverages and delicious pork steak.")
(3) "A jövő hétvégén megyek el." (English: "I will leave next weekend.")

2.2 Speaker embedding models

Two pre-trained models have been used in this work for embedding feature extraction: x-vector and ECAPA-TDNN.

2.2.1 The x-vector

The deep learning-based feature extraction method, x-vector was developed primarily for speaker verification (Egas-López et al., 2022; Snyder et al., 2018). It is based on a multiple-layered DNN architecture (with fully connected layers) with different temporal contexts at each layer (which they call ‘frames’). Due to the wider temporal context, the architecture is called time-delay NN (TDNN). The TDNN embedding architecture can be seen in Figure 2 and Table 1.

The first five layers operate on speech frames, with a slight temporal context centred at the current frame t. For example, the frame indexed as 3 sees a total of 15 frames, due to the temporal context of the earlier layers. After training with speaker ids as target vectors, the output of layer segment6 ('x-vector') is used as input to a classifier.

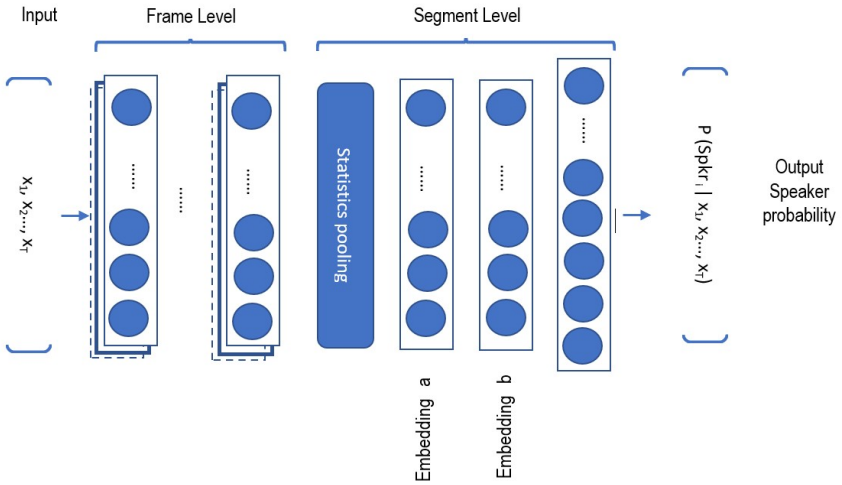

Fig. 2. The x-vector DNN embedding architecture in (Snyder et al., 2018). The two parts: frame level (with the 5 frame layers) and segment level (with segment6, segment7 and softmax).

Table 1. The x-vector embedding DNN architecture (Snyder et al., 2018)

Layer	Layer context	Total context	Input x output
Frame 1	{t-2,t+2}	5	120 x 512
Frame 2	{t-2,t,t+2}	9	1536 x 512
Frame 3	{t-3, t,t+3}	15	1536 x 512
Frame 4	{t}	15	512 x 512
Frame 5	{t}	15	512 x 1500
Stats pooling	[0,T)	T	1500 T x 3000
Segment 6	{0}	T	3000 x 512
Segment 7	{0}	T	512 x 512
softmax	{0}	T	512 x N

2.2.2 ECAPA-TDNN

The ECAPA-TDNN model is the extension of the x-vector model architecture in three ways (Desplanques et al., 2020): channel- and context-dependent statistics pooling, 1-Dimensional Squeeze-Excitation Res2Blocks (1D SE-Res2Block) and multi-layer feature aggregation and summation. The channel- and context-dependent statistics pooling enables the network to focus more on speaker characteristics that do not activate on identical or similar time instances, e.g. speaker-specific properties of vowels versus

speaker-specific properties of consonants. Using the SE-Res2Block (taken from the field of computer vision), the limited frame context of the x-vector (15) is extended to the global properties of the recording. The multi-layer feature aggregation means that not only the activation of the selected deep layer is used as a feature map (as in x-vector), but the shallower layers (here:SE-Res2Blocks) are also concatenated, because they also hold information about the speaker identity.The architecture is shown in Figure 3.

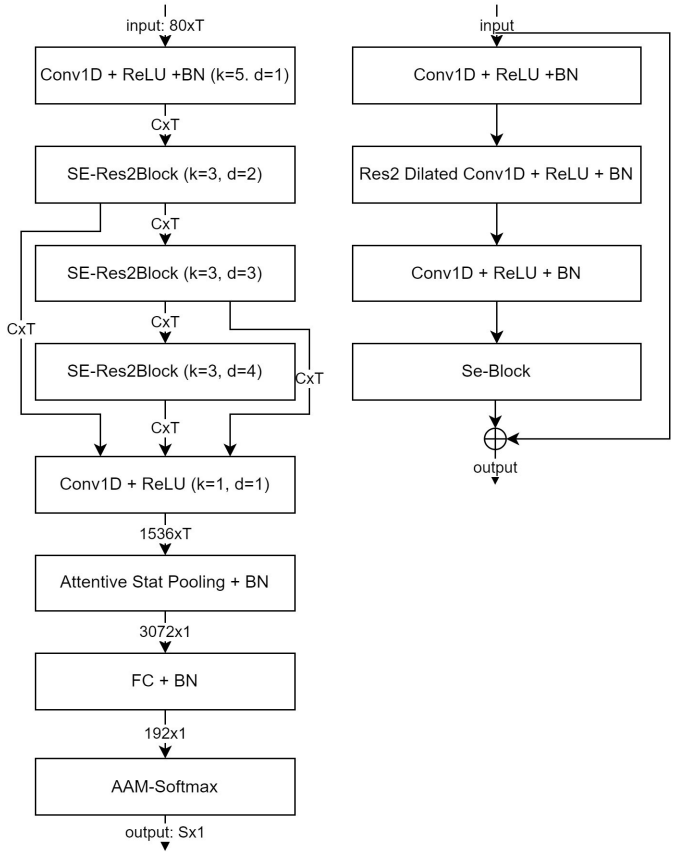

Fig. 3. The ECAPA-TDNN layer architecture and its SE-Res2Block (Desplanques et al., 2020)

2.3 Cosine similarity

Cosine similarity has been used to measure the similarity of pairs of embedded feature vectors. It is a measure by the cosine of the angle between two vectors calculated using Eq. 1 (Han et al., 2012). Figure 4 illustrates the cosine similarity mechanism between two vectors.

$$\text{Cosine similarty}(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

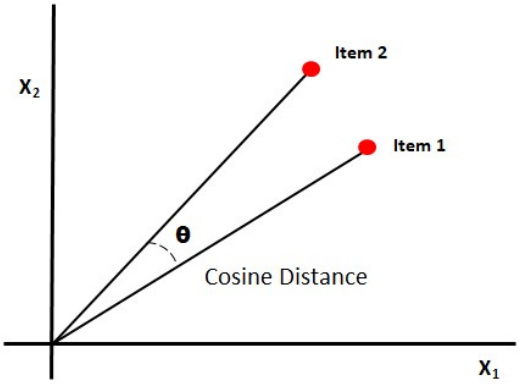

Fig. 4. Cosine similarity between pair of vectors (Han et al., 2012)

2.4 LR score calculation

Logistic regression has been used (implemented based on the python sklearn package) to calculate LR scores. All possible sample pairings (trials) were constructed with the target/non-target indicator (same speaker or different speaker). Each trial contains a speaker pair, the suspect and the offender and the indication if the speakers are the same or not. The probability of the same speaker decision was computed based on the logistic regression model using Eq. 2, where E is the evidence, H_{so} is the hypothesis of same-origin speakers and H_{do} is the hypothesis of different-origin speakers. The probability of different speaker origins can be calculated using Eq. 3. The applied logistic regression models (two separate models for x-vector and ECAPA-TDNN) were trained on the ForVoice120+ dataset using 2-10 second long speech samples (Sztahó & Fejes, 2022). Figure 5 illustrates an example of a trained logistic regression model. The distribution of the same and different origin vector pairs is shown in blue and yellow, respectively.

$$LR = \frac{P(E|H_{so})}{P(E|H_{do})} \quad \textbf{(2)}$$

$$P(E|H_{do}) = 1 - P(E|H_{so}) \quad \textbf{(3)}$$

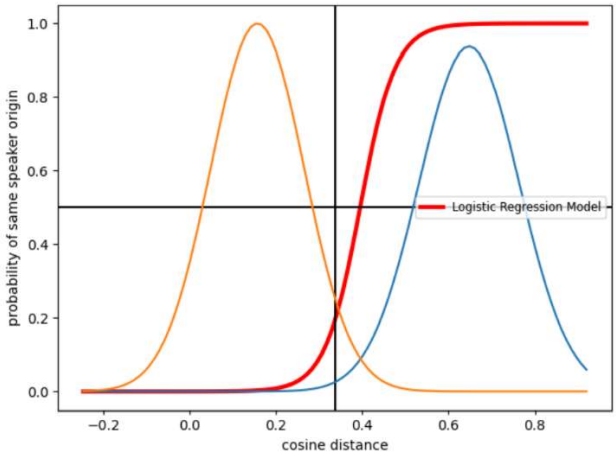

Fig. 5. Trained Logistic Regression model example. Blue and yellow lines show the distributions of cosine distances of embedding vector pairs of the same and different speaker origins, respectively. (Sztahó & Fejes, 2022)

2.5 Evaluation metric

The output of the two pre-trained models has been evaluated by using an Equal Error Rate (EER) and log-likelihood ratio cost (Eq. 4) between each sample pair. In speaker verification, the EER's level is where the false acceptance and rejection rates are equal.

$$Cllr = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\left(1+LR_{do_j}\right)\right) \quad \textbf{(4)}$$

where N_{so} and N_{do} are the number of same-origin and different-origin comparisons, LR_{so} and LR_{do} are the likelihood ratios derived from same-origin and different-origin comparisons.

Cllr is a function used to measure the balance of LR scores of same and different origin comparisons (Brümmer & du Preez, 2006). Ideal same-origin and different-origin comparisons have logLR>0 and logLR<0, respectively. Besides Cllr, the minimum Cllr value is also reported, which is the generalization of the original cost function and produces application-independent Cllr values.

Tippet plots, commonly used as a visualization in speaker verification, are used to display the proportion of correctly identified same and different speaker origin pairs.

3 Results

In this section, we evaluate the effectiveness of the forensic voice comparison and speaker verification system if emotional speech is present and how it affects performance. We investigated the models in multiple scenarios, considering how the emotions

are paired for the known and unknown speakers. Table 2 shows the results if the same or different emotions are used in the trials compared to the case when all samples are used together. The first row of the table can be considered as a baseline because all samples were used without any filtering. Higher performance was found using the same emotion (0.026 and 0.046 EER for ECAPA-TDNN with the same and different emotions, respectively). Both models are affected by emotional content. ECAPA-TDNN outperforms the x-vector, as was expected. Tippet plots for the two scenarios for ECAPA-TDNN are shown in Figure 6.

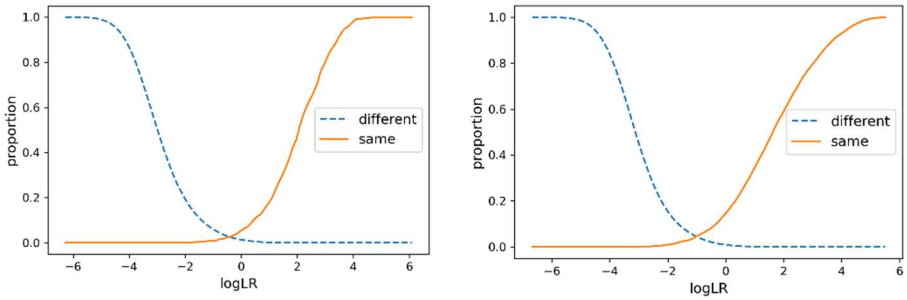

Fig. 6. Tippet plots for trials with same (left) and different (right) emotions for ECAPA-TDNN embeddings.

Table 2. Performance metrics of x-vector and ECAPA-TDNN when same and different emotions are used in the trials

Scenario	x-vector			ECAPA-TDNN		
	Cllr	$Cllr_{min}$	EER%	Cllr	$Cllr_{min}$	EER%
All samples	0.634	0.373	10.9	0.274	0.161	4.5
Same emotions	**0.387**	**0.270**	**7.7**	**0.122**	**0.092**	**2.6**
Different emotions	0.658	0.377	11.1	0.288	0.163	4.6

Table 3. Performance metrics of x-vector and ECAPA-TDNN when neutral samples were used for the suspect (first sample of the trials).

Emotion in offender samples	x-vector			ECAPA-TDNN		
	Cllr	$Cllr_{min}$	EER%	Cllr	$Cllr_{min}$	EER%
Neutral	**0.402**	**0.263**	**7.8**	**0.117**	**0.080**	**2.7**
All emotions	0.707	0.396	11.6	0.291	0.164	4.7
Anger	1.233	0.447	15.1	0.500	0.201	5.2
Sadness	0.536	0.324	10.8	0.149	0.083	3.1
Joy	0.632	0.355	11.5	0.314	0.159	4.6
Fear	0.703	0.362	10.8	0.369	0.174	5.4
Excitement	0.851	0.399	12.8	0.330	0.134	4.8
Disgust	0.713	0.333	12.9	0.311	0.120	4.3
Surprise	0.650	0.369	11.8	0.286	0.148	4.7

Table 3 shows the results when neutral samples were used as the first sample in the trials. These cases describe how a given emotion (and all emotions at once) affect the ASV performance if only neutral samples are recorded from the suspect. The results show that the ECAPA-TDNN outperformed the x-vector in this case also. The first row can be considered as a baseline because all samples were used without emotional content. Higher performance was achieved in the case of neutral vs. neutral as was expected. EERs were 2.7% and 7.8 % for ECAPA-TDNN and x-vector, respectively. In the case when emotional sentences were used as the second part of the trials, the EERs were slightly worse in each case.

Table 4. Performance metrics of x-vector and ECAPA-TDNN when all emotions samples were used for the suspect (first sample of the trials).

Emotion in offender samples	x-vector			ECAPA-TDNN		
	Cllr	$Cllr_{min}$	EER%	Cllr	$Cllr_{min}$	EER%
Anger	0.735	0.383	11.7	9.887	0.780	4.5
Sadness	**0.527**	0.345	**9.9**	**9.831**	**0.698**	**3.3**
Joy	0.645	0.352	10.7	9.939	0.834	4.9
Fear	0.633	0.379	11.3	9.883	0.739	5.4
Excitement	0.669	0.361	10.6	9.847	0.805	5.1
Disgust	0.596	0.346	10.4	9.943	0.796	3.6
Surprise	0.563	**0.340**	10.4	9.877	0.764	4.4

Table 4 shows the results when all emotional samples were used as the first sample in the trials. These cases describe how a given emotion and all emotions affect the ASV performance if all emotions samples are recorded from the suspect. The results show that the ECAPA-TDNN outperformed the x-vector in this case also. Higher performance was achieved in the case of sadness vs. all emotions for both models. EERs were 3.3% and 9.9 % for ECAPA-TDNN and x-vector, respectively.

4 Discussion and Conclusion

Based on the metrics obtained, we can conclude that the emotional content affects the pre-trained speaker embedding models' performance. Both ECAPA-TDNN and x-vector performed worse when the trials were composed of different emotional samples. A 2% decrease was found if different emotions were present compared to the case when the same emotions were applied. Considering the same emotional samples in the trials, the separate emotion did not have a large effect compared to the case when only neutral samples were used (2.6% and 2.7% EER for all emotions and neutral, respectively). Inspecting the separate emotions when the first sample in the trial was always neutral, the best results were obtained in the case of comparing them to neutral sentences (7.8% and 2.7% EER for x-vector and ECAPA-TDNN, respectively). As was expected, comparing neutral samples to emotional samples, worse performances were achieved. Table 3 shows that only sadness is close to the neutral case. There is no emotion that can be

said to have the highest effect on speaker verification. All emotions deteriorate the metrics. This implies that if the emotional content on the recording is not neutral, but is the same in the trial, the performance deterioration is not present.

In this work, the effects of emotions were investigated in a forensic voice comparison setting using deep speaker embeddings and a Hungarian dataset. The main goal of this study was to evaluate whether emotional characteristics affect ASV performance. Two pre-trained deep learning models have been used for feature extraction (x-vector and ECAPA-TDNN). The similarity of the embedding vectors was measured by cosine similarity. The performances were evaluated in the likelihood-ratio framework by calculating the LR and logLR based on the cosine similarities. It can be stated that for a public service expert, emotional content can be a significant factor during speaker verification. In future work, more emotion-robust models can be built, trained or fine-tuned to make ASV more reliable in such cases.

Acknowledgement

The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_18 funding scheme.

References

Arya, R., Pandey, D., Kalia, A., Zachariah, B. J., Sandhu, I., & Abrol, D. (2021). Speech based Emotion Recognition using Machine Learning. *2021 IEEE Mysore Sub Section International Conference, MysuruCon 2021*, 613–617. https://doi.org/10.1109/MysuruCon52639.2021.9641642

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

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

Dev Sarma, B., & Kumar Das, R. (2020). Emotion Invariant Speaker Embeddings for Speaker Identification with Emotional Speech; In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Egas-López, J. V., Kiss, G., Sztahó, D., & Gosztolya, G. (2022). AUTOMATIC ASSESSMENT OF THE DEGREE OF CLINICAL DEPRESSION FROM SPEECH USING X-VECTORS. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2022-May*, 8502–8506. https://doi.org/10.1109/ICASSP43922.2022.9746068

Han, J., Kamber, M., & Pei, J. (2012). 2 - Getting to Know Your Data., *Data Mining (Third Edition)* (Third Edition, pp. 39–82). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00002-2

Pappagari, R., Wang, T., Villalba, J., Chen, N., & Dehak, N. (2020). X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7169–7173. https://doi.org/10.1109/ICASSP40776.2020.9054317

Parthasarathy, S., Zhang, C., Hansen, J. H. L., & Busso, C. (2017). A study of speaker verification performance with expressive speech. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5540–5544. https://doi.org/10.1109/ICASSP.2017.7953216

Prasetio, B. H., Tamura, H., & Tanno, K. (2020). Emotional variability analysis based I-vector for speaker verification in under-stress conditions. *Electronics (Switzerland)*, *9*(9), 1–15. https://doi.org/10.3390/electronics9091420

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. http://arxiv.org/abs/2106.04624

Rusko, M., Trnka, M., Darjaa, S., Stelkens-Kobsch, T., & Finke, M. (2018). *WEAKNESSES OF VOICE BIOMETRICS-SENSITIVITY OF SPEAKER VERIFICATION TO EMOTIONAL AROUSAL*.

Scherer, K. R., Johnstone, T., Klasmeyer, G., & Bänziger, T. (2000). *CAN AUTOMATIC SPEAKER VERIFICATION BE IMPROVED BY TRAINING THE ALGORITHMS ON EMOTIONAL SPEECH?* http://www.iscaaspeech.org/archive

Shahin, I., Nassif, A. B., Nemmour, N., Elnagar, A., Alhudhaif, A., & Polat, K. (2021). Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments. *Neural Computing and Applications*, *33*(23), 16033–16055. https://doi.org/10.1007/s00521-021-06226-w

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Sztahó, D., & Fejes, A. (2022). *Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings*.

Sztahó, D., Szaszák, G., & Beke, A. (2021). Deep learning methods in speaker recognition: A review. *Periodica Polytechnica Electrical Engineering and Computer Science*, *65*(4), 310–328. https://doi.org/10.3311/PPee.17