Tandem: Spoof Aware Speaker Verification Challenge 2022

Ayush Agarwal[1,2], *Srikanth Nalluri*[2], *Dattatraya Kulkarni*[2]

[1]Indian Institute of Technology (IIT) Dharwad
[2]McAfee

160020008@iitdh.ac.in, ayush_agarwal@mcafee.com, srikanth_nalluri@mcafee.com, dattatraya_kulkarni@mcafee.com

1. Introduction

Automatic verification system (ASV) performs the task of verifying the already enrolled speaker based on the threshold. Spoof detection system has the task to detect the spoof and non-spoof speech based on the counter measures. These two systems have shown excellent performance when applied individually. The spoof aware speaker verification (SASV) challenge [1] is the first challenge the combines the both the systems into one system, such that combined system has the potential to reject both utterances spoken by different speakers as well as spoofed utterances [1].

In this work we have exploited the speech production features used to model both the speaker verification system and the spoof detection system. Previous systems for speaker verification and spoof detection have used both spectral and excitation source features for both the tasks separately. Hence, motivated from this we combined the source and spectral features for the combined system.

The chronological order of the paper is as follows. Section 2 very briefly describes the baseline systems provided by the organizers. Section 3 motivates and describes the proposed work. In section 4, we describe the experiments, results.

2. Existing SASV Baseline systems

The SASV challenge provides the two baseline systems [1]. The baseline 1 model does not involve any training and performs the score level fusion of the score from the speaker verification and countermeasure systems. Baseline 2 performs the embedding level fusion and trains the neural network on the fused embedding. The scores are further computed on these fused embeddings.

3. Proposed approach

3.1. Motivation of proposed approach

In literature it has been shown that the components of excitation source feature residual phase of LP residual have shown improved performance in the results of both automatic speaker verification and spoof attack. In this work we exploit these features for the tandem system. In [2], residual phase when combined with spectral features like MFCC at score level has significantly improved the EER by lowering it's value. Similarly in [3], the usefulness of cepstral coefficients of residual phase features to improve the detection of spoof attack have been shown. Therefore, excitation source features have been used to improve the performance of both speaker verification and spoof detection systems. This motivates us to use these features in the combined system where speaker verification and spoof detection will be done simultaneously.

3.2. Method

Figure 1 shows the proposed approach. In step 1 we extract the features from the speech and it's components. Further, model is finetuned in the to extract the speaker and countermeasure embeddings from the pre-trained networks. Later the scores obtained with respect to all the features are combined at the score level to give the various equal error rates (EER).

3.2.1. Feature extraction

For the input speech (target or non-target or spoof) spectral and source features are extracted. For the source features, LP residual is extracted using LP analysis [4, 5]. Later, LP residual is divided into amplitude and phase components. Amplitude component is captured using Hilbert envelope of LP residual and phase component is captured using residual phase [3]. To extract the speaker specific information of speech, Hilbert envelope and residual phase, we compute the cepstral coefficients. Cepstral coefficient of speech Mel frequency cepstral coefficient (MFCC) and for residual phase it is RPCC. These features are used modeling the speaker verification, spoof detection and the combined system.

3.2.2. Fine Tuning

Pre-trained networks of the speaker verification and spoof detection are used for learning the embeddings. Pre-trained speaker embedding is learnt from ECAPA-TDNN [6] and pre-trained counter measure embedding is learnt from ASSIST [7]. For this work the pre-trained embedding given in the github[1] repository of the SASV-2022 challenge were used. For learning the embedding of the new data-set the speaker embeddings and counter measure embeddings were adapted from the pre-trained model as done in github[2].

3.2.3. Score level combination

The likelihood score from all the three features are computed separately. These three features are combined at score level as done in [2]. The combined score S is obtained using the equation for as shown in equation 1.

$$S = \alpha S_1 + \beta S_2 \tag{1}$$

with the constraint that,

$$\alpha + \beta = 1 \tag{2}$$

3.2.4. Evaluation metric

The three evaluation metrics that were given in the guidelines of SASV challenge [1] were used to evaluate the performance of

[1] https://github.com/sasv-challenge/SASVC2022_Baseline

[2] https://github.com/sasv-challenge/SASVC2022_Baseline

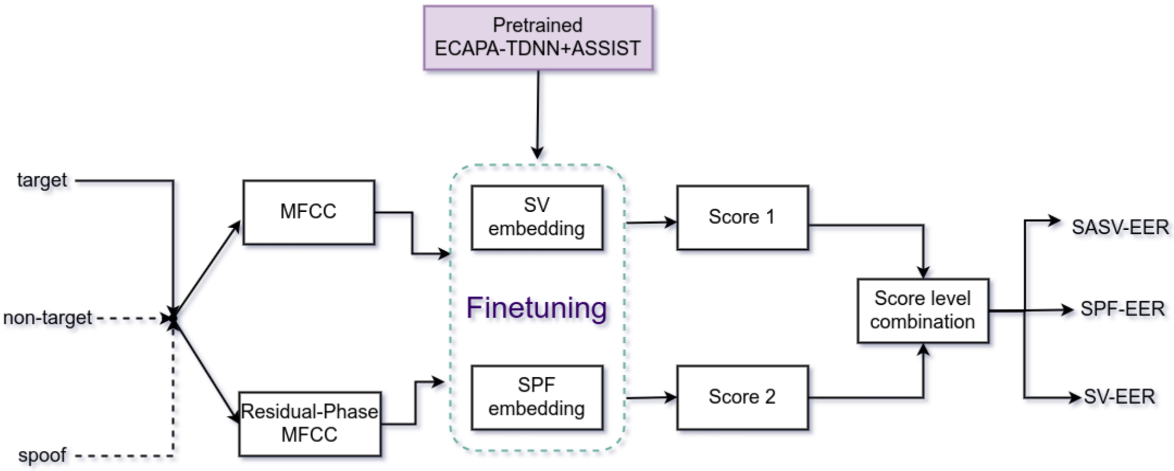

Figure 1: *Proposed method for spoof aware speaker verification. MFCC and residual phase MFCC (RPCC) are the features used to model speaker and countermeasure system. Score level combination of score 1 and 2 is done.*

the speaker verification, spoof detection and combined system. SV-EER is used to evaluate speaker verification, SPF-EER is used to evaluate spoof detection systems and SASV-EER is used to evaluate combined systems.

4. Experiments, results and discussion

4.1. Dataset

The dataset used to train the fine-tune network is the train set of ASVspoof 2019 LA challenge. Development and evaluation is done according to the development and evaluation protocols given in [1]. For development there are 29548 utterances and for evaluation there are 102579 utterances of target, non-target and spoof speakers.

4.2. Results

Table 1: *EER on the development (Dev) and evaluation (Eval) dataset according to the development and evaluation protocol of ASVspoof 2019 challenge is tabulated. SV-EER, SPF-EER and SASV-EER of Baseline 1, baseline 2, RPCC and RPCC+MFCC based systems are shown.*

Model	**SV-EER**		**SPF-EER**		**SASV-EER**	
	Dev	**Eval**	**Dev**	**Eval**	**Dev**	**Eval**
Baseline-1 (MFCC)	32.88	35.32	**0.06**	**0.67**	13.07	19.31
Baseline-2 (MFCC)	12.87	**11.48**	0.13	0.78	4.85	6.37
RPCC	17.04	26.62	33.22	39.50	20.44	31.17
MFCC+RPCC	**12.23**	11.91	0.20	0.76	**4.58**	**6.22**

Score level combination for the score obtained from MFCC and RPCC were done. On trying various values of α and β, we found that the system performs best by weighting MFCC with 0.9 and RPCC with 0.1.

From the table 1 it can be seen that the proposed method of using MFCC+RPCC as the feature has performed better for the combined system evaluation metric i.e. SASV-EER in both development and evaluation dataset. The MFCC+RPCC method has also performed better than baseline 2 on SPF-EER metric on the evaluation dataset and have given close scores in development dataset. From the table it can be also that MFCC+RPCC has shown the best SV-EER on the development data and nearly close to baseline 2 in the evaluation data.

5. References

[1] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "Sasv challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

[2] K. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[3] M. Singh and D. Pati, "Combining evidences from hilbert envelope and residual phase for detecting replay attacks," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 313–326, 2019.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[5] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[7] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *arXiv preprint arXiv:2110.01200*, 2021.