IEEE Copyright Notice

This work has been accepted for publication in 18th Asian Conference on Intelligent Information and Database Systems. The final published version will be available via IEEE Xplore.

Enhancing Speaker Verification with Whispered Speech via Post-Processing

Magdalena Gołębiowska[0000−0001−7507−6737] and Piotr Syga[0000−0002−0266−5802]

Department of Artificial Intelligence,
Wroclaw University of Science and Technology,
Wybrzeze Wyspianskiego 27, Wroclaw, 50-370, Poland
{magdalena.golebiowska,piotr.syga}@pwr.edu.pl

Abstract. Speaker verification is a task of confirming an individual's identity through the analysis of their voice. Whispered speech differs from phonated speech in acoustic characteristics, which degrades the performance of speaker verification systems in real-life scenarios, including avoiding fully phonated speech to protect privacy, disrupt others, or the lack of full vocalization may be dictated by a disease. In this paper we propose a model with a training recipe to obtain more robust representations against whispered speech hindrances. The proposed system employs an encoder–decoder structure built atop a fine-tuned speaker verification backbone, optimized jointly using cosine similarity–based classification and triplet loss. We gain relative improvement of 22.26% compared to the baseline (baseline 6.77% vs ours 5.27%) in normal vs whispered speech trials, achieving AUC of 98.16%. In tests comparing whispered to whispered, our model attains an EER of 1.88% with AUC equal to 99.73%, which represents a 15% relative enhancement over the prior leading ReDimNet-B2. We also offer a summary of the most popular and state-of-the-art speaker verification models in terms of their performance with whispered speech. Additionally, we evaluate how these models perform under noisy audios, obtaining that generally the same relative level of noise degrades the performance of speaker verification more significantly on whispered speech than on normal speech.

Keywords: Speaker verification · Whispered speech · Speaker embeddings.

1 Introduction

This study examines speaker verification (SV) with whispered speech. Speaker verification is a task of confirming an individual's identity through the analysis of their speech to determine whether to accept or deny the claimed identity of the speaker [5]. The testing sample from the speaker is compared to another previously obtained sample, usually recorded by a user in a neutral state, both emotionally and vocally, i.e., in fully phonated speech.

Acoustic characteristics of whispered speech are different from those of normal speech due to the absence of vocal cord vibrations [4], including an upward shift

of formant frequencies of vowels, lower energy on voiced consonants at low frequencies, and greater spectral flatness [3]. Naturally, these varying vocal efforts cause a mismatch between a previously obtained speaker sample in a neutral environment and a new testing sample. These are problematic for speaker verification systems that have not been designed and evaluated for whispered speech [13]. Considering whispered speech data in speaker verification is motivated by real-world use cases. A user might whisper contents to protect privacy, to avoid disrupting others, or the change is forced by disease or vocal cords are removed after an operation [4].

Previous studies on whispered speech speaker verification primarily relied on legacy architectures that predate recent advances in deep embedding systems.

First approaches were based on including whispered speech in the training data for a GMM-UBM model [12], which improved the performance of speaker verification with whispered speech. Additionally, the work proved that using speaking-style and gender dependent-models, and adding AM-FM signal representation-based features also improved verification reliability.

The study expanded into testing diverse approaches, including frequency warping, sub-band analysis, alternate feature representations, and feature combination. The findings revealed that these modifications were ineffective in improving whispered data, highlighting the necessity for representations tailored to whispered speech. The most successful model in the study achieved an EER of 22.77% on the CHAINS (Characterizing Individual Speakers) dataset [1] under normal versus whispered speech trials.

Such features were proposed by fusing information from spectral, modulation spectral and bottleneck features computed via deep neural networks at the feature- and score-levels [10]. The experimental findings indicated that relative enhancements could reach up to 79% for neutral speech and 60% for whispered speech on wTIMIT and TIMIT when compared to a baseline system trained using i-vectors derived from mel frequency cepstral coefficients. These findings were further improved by introducing three distinct features and implementing a score fusion method, which relied on systems trained on these three feature sets [11], resulting in 66% and 63% relative improvement on combined CHAINS, TIMIT, and wTIMIT databases for whispered and normal speech, respectively.

Another approach to choosing features robust to whispered speech was based on formant and formant gap (FoG) features thought to be more invariant to speech modes than MFCCs and AAMFs [7]. The system was trained on i-vectors and showed a 3.79% absolute value improvement in EER on CHAINS, TIMIT, and wTIMIT compared to the baseline AAMF in the mismatched trials; however, under normal conditions, the performance slightly degrades.

The important notion is that all mentioned research until now has been conducted on systems that required a development phase, i.e., they have seen the speaker data during training, which entails no generalizability to unseen speakers, which is not the correct approach nowadays.

The most recent studies explored the x-vector/PLDA/SPLICE system [8], which scored 17.8% EER in neutral-whispered conditions on CHAINS, resulting in a 0.5% relative improvement compared to the baseline.

Semi-supervised learning models were also explored in the context of SV with whispered speech. A system based on wav2vec 2.0 with augmentation based on whisper synthesis using Praat software and domain score normalization with whisper detection achieved 2.29% EER on CHAINS in whisper vs whisper trials and 46.31% relative improvement on the MSP-AVW in normal vs whisper conditions [6].

In this article, we propose a new model with a training framework for more robust embeddings against whispered speech. We provide a review of robustness against whispered speech of recent and state-of-the-art speaker verification encoders. Additionally, we check how noise affects the baseline models' performance. Our contributions are as follows:

1. We propose a new model and training for speaker verification which improves verification performance with whispered speech.
2. To our knowledge, this is the first study to evaluate speaker verification with whispered speech on state-of-the-art speaker verification systems, including ECAPA2, ECAPA-TDNN, and different versions of ReDimNet.
3. We evaluate how state-of-the-art speaker verification models act under noisy whispered speech.

2 System Architecture and Training

In this section, we present details of our proposed architecture and training recipe for speaker verification with whispered speech.

To improve speaker verification with whispered speech, we introduce a few fully-connected layers with ReLU that create an encoder-decoder architecture with bootleneck on top of the pre-trained speaker verification model chosen to be ReDimNet-B6 [18]. The encoder-decoder module is shallow by design with limited capacity, consisting of four fully-connected layers. The choice is dictated by the already employed strong speaker recognition model (ReDimNet-B6); thus, its objective is to only compensate for the systematic phonation mismatch between whispered and phonated speech. Preliminary experiments with deeper or wider architectures resulted in either no additional gains or degraded performance, indicating that higher model capacity tends to over-adapt to the training speakers. The bottleneck dimensionality was chosen to enforce a compact latent representation that preserves phonation-related variability while filtering out nuisance factors. This design encourages the network to act as a residual correction rather than a complete transformation of the embedding.

To preserve speaker identity, we add a speaker classification head that consists of a cosine-normalized fully-connected layer and outputs cosine similarity scores as seen in NormFace [17]. Fig. 1 presents the model training. The audio embeddings processed by the encoder-decoder are compared using triplet loss in triplets (phonated sentence, whispered sentence, random sentence chosen from

another speaker) L_{trip}. We also add a residual connection between encoder input and encoder output. The objective for the model is to learn to convert whispered representations into phonated ones while conserving speaker identity. To further keep speaker discerning capabilities, we add cosine softmax loss L_{ce} with similarities coming from the speaker classification head. The losses are combined with formula:

$$L = L_{trip} + \gamma L_{ce} \ , \tag{1}$$

where $\gamma = 10^{-4}$. The value was selected to scale L_{ce} accordingly to L_{trip} to influence model weights less than the triplet loss. The pre-trained speaker verification model is fine tuned with layers gradually unfrozen every 5 epochs.

After training, we remove the speaker classification head, as shown in Fig. 2, leaving only the encoder and decoder that produce embeddings used for speaker verification. At this point all weights are frozen.

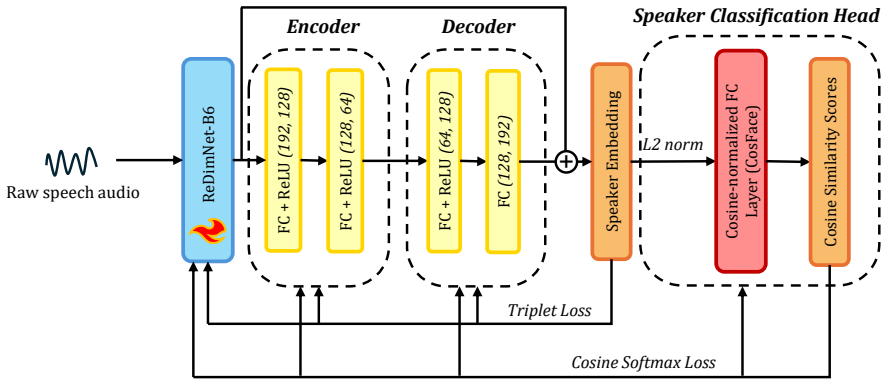


Fig. 1: Proposed architecture during training phase. The proposed architecture employs an encoder and decoder layers which map whispered representations into phonated ones and a speaker classification head with cosine-normalized linear projection scaled by a constant factor and optimized using cross-entropy loss and triplet loss.

3 Experimental Setup

3.1 Dataset

We evaluate the systems using the CHAINS (Characterizing Individual Speakers) database [1] released in 2008. This dataset features 36 English speakers (18

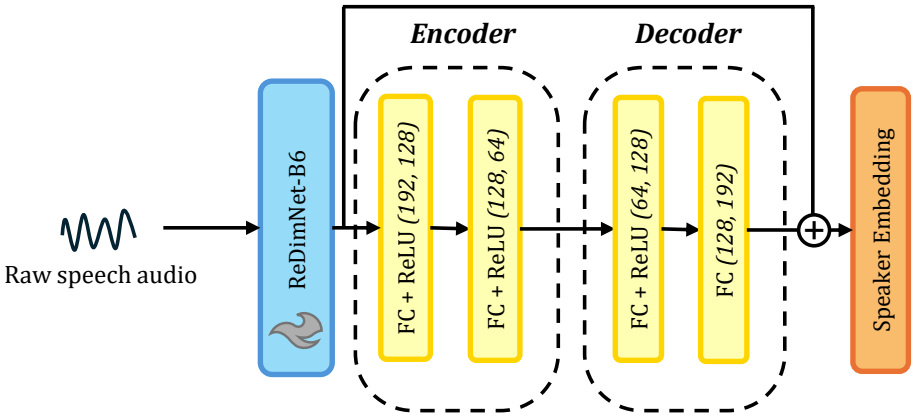


Fig. 2: Proposed architecture during evaluation phase. Speaker classification head is detached and the audio passes through fine-tuned ReDimNet whose representations are further processed by encoder and decoder to convert whispered representations into phonated ones.

male, 18 female) who read fables and selected sentences employing various speaking styles, including neutral and whispered. Data gathering occurred under laboratory conditions. We utilized subsets of the dataset described as solo and whisper readings of fables and sentences (text ids f01-f04, sentences s01-s33), resulting in a total of 5,860 samples (158-198 samples per speaker), balanced between neutral and whispered speech as we excluded audios without a pair from both speaking styles. Measured average root-mean-square energy (RMSE) value for normal speech $\mu_{norm} = 0.046$, $\sigma_{norm} = 0.014$ and for whispered speech $\mu_{whsp} = 0.0138$, $\sigma_{whsp} = 0.0032$. As expected, energy of whispered speech is lower indicating lower perceived loudness.

We split the dataset by speakers, using 70% for training and 30% for testing. We keep the split in the speaker verification tests.

For noise robustness experiments, we use MUSAN (Music, Speech, and Noise Corpus) [14]. It consists of approximately 109 hours of audio in the US Public Domain or under a Creative Commons license. The audios are categorized in music, speech and noise. We use the noise part of the corpus with the duration of 6 hours. Noises consist of technical sounds, including DTMF tones, dial tones, and fax machine sounds, as well as environmental sounds like idling cars, thunder, wind, footsteps, rustling paper, rain, various animal noises and crowd noises.

3.2 Environment and Training

We reran our experiments a total of 10 times to assess the model's performance and consistency, then computed the average of the outcomes. Each experiment was run on a single NVIDIA Hopper (H100) GPU with 955 GB RAM (not used to full capacity). Code base with details for this paper is available in a Github repository[1].

Audio samples underwent preprocessing to meet the front-end requirements, which included converting them to mono-channel and resampling to 16 kHz.

The system was trained with an Adam optimizer (learning rate of 10^{-4} for encoder-decoder and speaker classification head, and 10^{-5} for fine-tuning ReDimNet-B6, weight decay of 10^{-4}) for 100 epochs, batch size 128. We added dropout with value 0.3 to encoder-decoder layers.

In order to evaluate, we randomly select pairs for each dataset sample. For each instance, we randomly select one positive sample, originating from the same speaker, and one negative sample, which comes from a different speaker. The trial type being assessed determines the vocal style of these samples.

We evaluate our framework using the Equal Error Rate (EER), a standard metric in speaker verification tasks. EER corresponds to the point where the misclassification rates of positive and negative samples are equal (i.e., FPR=FNR), and lower values indicate better overall performance.

4 Results

Tab. 1 presents the EER scores for the testing subset of the dataset across various trials. These trials consider different speaking conditions between enrollment and testing, such as the common situation where a user enrolls with normal speech and later attempts to authenticate using whispered speech.

We include the most popular and state-of-the-art speaker verification systems, i.e. x-vector [15] as shared by SpeechBrain Toolkit [9], ECAPA-TDNN [2] also from SpeechBrain [9], ECAPA2 [16] shared by the original authors, ReDimNet-B0, ReDimNet-B2, and ReDimNet-B6 [18] pretrained by the original authors.

Among the models evaluated, x-vector showed the lowest robustness, with an EER of 29.93% in the All vs All trial. ReDimNet-B2 achieved an EER of 12.75%, which was similar to ECAPA-TDNN's 13.72%. Its successor, ECAPA2, performed marginally better, with an EER of 10.95%. The top performer was ReDimNet-B6, with an EER of 7.76%. Our model achieved second best result of EER 8.40%. The likely reason for this drop in performance is that the training dataset was not as large as the original datasets used during pre-training.

In Norm vs Norm conditions this trend also applies. X-vector as the oldest system, performs the worst with EER equal to 3.32%. ECAPA-TDNN and ReDimNet-B0 were comparable in this conditions achieving EER of 0.41% and 0.49% respectively. Similar results were achieved for ECAPA2, ReDimNet-B2, and by our model with EER of 0.21%, 0.23%, and 0.28% accordingly.

[1] Code base is available at https://github.com/mgraves236/sv-whispred-speech

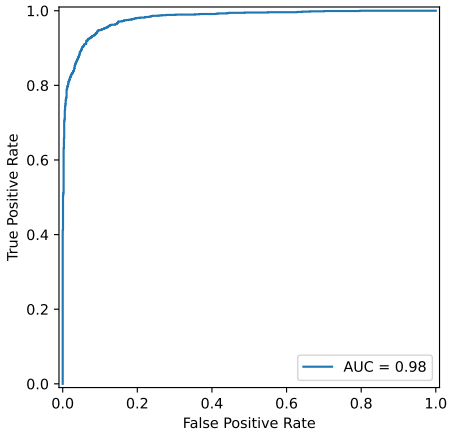
Table 1: The comparison of EER scores on the testing dataset of baselines and our model across trials of normal vs whispered speech (Norm vs Whsp), normal vs normal speech (Norm vs Norm), whispered vs whispered speech (Whsp vs Whsp), both whispered and normal vs whispered and normal speech (All vs All). The table lists mean and standard deviation from 10 independent reruns. The best scores are in bold.

Trial / **Model**	**Norm vs Whsp**	**Norm vs Norm**	**Whsp vs Whsp**	**All vs All**
x-vector	26.56% ± 1.44 %	3.32% ± 0.59 %	8.48% ± 0.94 %	29.93% ± 1.44 %
ECAPA-TDNN	10.49% ± 1.47 %	0.41 % ± 0.18 %	2.77% ± 0.51 %	13.72% ± 1.26 %
ECAPA2	8.28% ± 0.76 %	0.21 % ± 0.24 %	2.48% ± 0.53 %	10.95% ± 0.75 %
ReDimNet-B0	13.02% ± 1.56 %	0.49 % ± 0.31 %	2.38% ± 0.40 %	19.57% ± 0.96 %
ReDimNet-B2	8.80% ± 1.22 %	0.23 % ± 0.16 %	2.20% ± 0.38 %	12.75% ± 0.88 %
ReDimNet-B6	6.77% ± 1.51 %	**0.12%** ± 0.16 %	2.30% ± 0.43 %	**7.76%** ± 0.78 %
Ours	**5.27%** ± 0.82 %	0.28 % ± 0.20 %	**1.88%** ± 0.28 %	8.40% ± 0.90 %

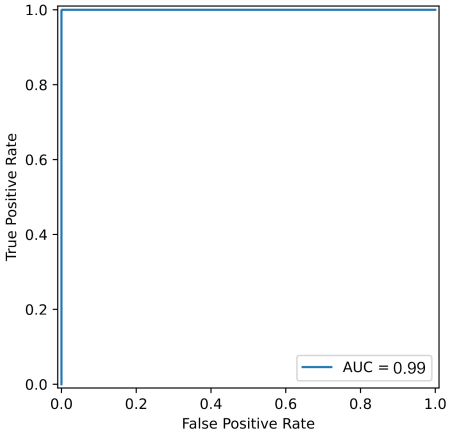
In Whsp vs Whsp trials, the EERs are lower compared to Norm vs Whsp tests due to the absence of mixed conditions, yet they remain higher than in Norm vs Norm trials. Again, the worst performance of 8.48% was noted in x-vector. Remaining models, ECAPA-TDNN, ECAPA2, ReDimNet-B0, ReDimNet-B2, and ReDimNet-B6 showed similar EER of 2.77%, 2.48%, 2.38%, 2.20%, and 2.30% respectively. Interestingly, ReDimNet-B6, which surpassed other ReDimNets in other trials, was 0.10 percentage points worse than its smaller version ReDimNet-B2. Our model achieved the best result in this category with EER equal to 1.88%, i.e, 15% relative improvement over the previous best ReDimNet-B2.

Among all evaluation conditions, the Normal vs Whispered trial is the most challenging and realistic verification scenario. Our fine-tuned ReDimNet-B6 reduces EER from 6.77% to 5.27%, a 22.16% relative improvement over the unadapted model, indicating that the proposed fine-tuning approach substantially improves robustness to cross-phonation variability.

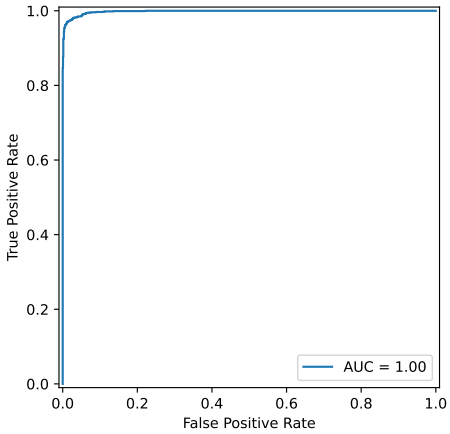
Although the proposed model slightly underperforms in the Norm vs Norm and All vs All conditions, this is consistent with its design objective: enhancing robustness to phonation mismatch. The trade-off is minimal and justified by the significant improvement in Normal vs Whispered performance.



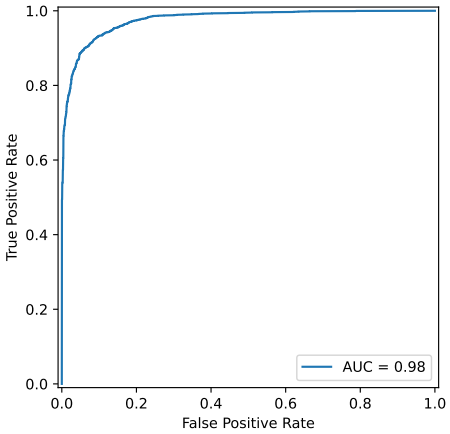
(a) ROC for Normal vs Whisper trial



(b) ROC for Normal vs Normal trial



(c) ROC for Whisper vs Whisper trial



(d) ROC for All vs All trial

Fig. 3: ROC curves illustrating performance of our model from one run (one seed) throughout all trials. Area under the curve is displayed in the bottom right of each figure. AUC close to 1 shows that our model acts almost perfectly.

To further assess the efficiency of our model, we inspected receiver operating characteristic curves across all trails (Fig. 3) and area under curve values (Tab. 2). For Normal vs Whisper tests, the curve is not ideal, however still very steep, reaching high True Positive Rate at low False Positive Rate, thus the model is able to identify most positives early while keeping false alarms at a very low level. Averaged AUC is equal to 98.16% which is very close to a perfect score. For Normal vs Normal, the ROC is almost ideal which is supported by the averaged AUC equal to 99.99%, meaning the model raises very few false alarms. In Whisper versus Whisper trials, the performance slightly declines, as evident in the form of

the ROC curve and the averaged AUC value of 99.73%. For the All vs All AUC, the lowest value is 97.72%, as reflected in the shape of the ROC curve. We have to take into the account that the mono-style trials i.e. Normal vs Normal and Whisper vs Whisper, due to low data availability cover 3,024 samples, which is a relatively mediocre amount.

Table 2: The comparison of AUC scores on the testing dataset across trials of normal vs whispered speech (Norm vs Whsp), normal vs normal speech (Norm vs Norm), whispered vs whispered speech (Whsp vs Whsp), both whispered and normal vs whispered and normal speech (All vs All). The table lists mean and standard deviation from 10 independent runs.

Model \ Trial	**Norm vs Whsp**	**Norm vs Norm**	**Whsp vs Whsp**	**All vs All**
Ours	**98.16%** ± 0.48 %	**100.0%** ± 0.00 %	**99.73%** ± 0.08 %	**97.72%** ± 0.36 %

4.1 Ablation Study

Additionally, we conducted a few experiments to investigate which part of the architecture provided the majority of gain. Results are listed in Tab. 3.

The first experiment involved changing the ReDimNet-B6 front end to ECAPA-TDNN. The experiments were run with learning rate equal to 10^{-4}, as it gave better performance than with previous learning rate of 10^{-5}. We also kept the idea of unfreezing the layers gradually every 5 epochs. Still, using ECAPA-TDNN as an embeddings extractor, gave worse results than ReDimNet-B6 (14.20% vs 5.27% for All vs All). Interestingly, the results are worse than the baseline (14.20% vs 13.72% for All vs All), suggesting that more parameter-tuning would be needed.

Next, we unfroze only the last two blocks of ReDimNet-B6, allowing only the last layers to adapt to the new domain. This way, the achieved results were overall worse than our proposed model (9.19% vs 8.40% in All vs All) but the model was slightly better in Norm vs Norm trial (0.27% vs 0.28%), showing that unfreezing more layers degrades performance for normal speech speaker verification. This is related to catastrophic forgetting of pretrained phonation representations, degrading neutral-speech verification performance and hindering the large dataset used for the original training.

The primary experiment sought to determine if incorporating the speaker classification head and encoder-decoder block was beneficial during training or redundant. The findings indicate that excluding this module during fine-tuning leads to the model losing valuable speaker verification structures, resulting in a substantial increase in EER compared to the baseline result of ReDimNet-B6, with figures rising from 7.76% to 17.85% in the All vs All trial.

Table 3: The comparison of EER scores on the testing dataset across trials of normal vs whispered speech (Norm vs Whsp), normal vs normal speech (Norm vs Norm), whispered vs whispered speech (Whsp vs Whsp), both whispered and normal vs whispered and normal speech (All vs All). The table lists mean and standard deviation from 10 independent runs. The best scores are in bold.

Model \ Trial	Norm vs Whsp	Norm vs Norm	Whsp vs Whsp	All vs All
ECAPA-TDNN	11.59%	0.64%	2.46%	14.20%
+ post-processing	± 1.04%	± 0.17%	± 0.24%	± 1.00%
ReDimNet-B6	5.98%	0.27%	2.55%	9.19%
unfrozen last 2 blocks	± 0.88%	± 0.18%	± 0.33%	± 0.96%
ReDimNet-B6	17.00%	5.32%	8.53%	17.85%
fine-tune	± 1.33%	± 1.06%	± 1.57%	± 1.55%
Ours	**5.27%**	**0.28%**	**1.88%**	**8.40%**
	± 0.82 %	± 0.20 %	± 0.28 %	± 0.90 %

4.2 Noise Robustness

Additionally, we evaluate how the best baseline models ECAPA-TDNN, ECAPA2, ReDimNet-B2, and ReDimNet-B6 are affected by noise. We combine MUSAN noises with normal and whispered speech so that the peak signal to noise ratio (PSNR) is comparable for both conditions, meaning the noise added to normal speech is louder than the noise added to whispered speech.

For normal speech we choose $SNR = \{5\,\text{dB}, 15\,\text{dB}\}$ encompassing both low and high noise. Then, SNR for whispered speech is scaled so that PSNR for normal and whispered speech are approximately the same. The measured PSNRs are $\text{PSNR}_{normal}(5\,\text{dB}) = 21.30$, $\text{PSNR}_{normal}(15\,\text{dB}) = 37.85$, $\text{PSNR}_{whisper}(5\,\text{dB}) = 22.87$, $\text{PSNR}_{whisper}(15\,\text{dB}) = 37.84$.

Tab. 4 lists results for $\text{PSNR}(15\,\text{dB}) \approx 38$. For all models adding noise acted detrimentally, with ReDimNet-B6 receiving the biggest drop in performance (All vs All trial) of 16.78 percentage points. The second biggest decline occured in ECAPA2 with performance being worse 13.97 percentage points than for clean samples. ECAPA-TDNN and ReDimNet-B2 dropped by 13.43 and 13.77 percentage points accordingly.

Tab. 5 presents values achieved for $\text{PSNR}(5\,\text{dB}) \approx 22$, so now the signal is more noisy. ECAPA2 and ReDimNet-B6 were not susceptible to the change in noise levels; the performance has not dropped. For ECAPA-TDNN ERR has declined by further 5.45 percentage points, totaling in 18.88 percentage points difference compared to EER with clean speech. Interestingly, ReDimNet-B2 is slightly better than for higher PSNR, improving its performance by 0.06 percentage points.

To gain more insights, we calculated relative change compared to EERs for clean speech in Norm vs Norm and Whsp vs Whsp trials for both PSNRs. The

relative change is defined as:

$$\Delta(EER) = \frac{|EER_{\{low\ \text{PSNR}, high\ \text{PSNR}\}} - EER_{clean}|}{EER_{clean}} \cdot 100\% \tag{2}$$

The results are listed in Tab. 6. The relative change compared to EERs on clean speech is mostly stronger for whispered speech (expect for ECAPA-TDNN with $PSNR \approx 22$). ECAPA2 showed almost the same level of degradation for normal and whispered speech but for other models the difference is noticable. We can conclude that generally adding noise of the same relative power to speech, degrades performance of speaker verification more strongly with whispered speech than with clean speech.

Table 4: The comparison of EER scores on the testing dataset with high SNR (PSNR(15 dB) $\approx$ 38) of baselines across trials of normal vs whispered speech (Norm vs Whsp), normal vs normal speech (Norm vs Norm), whispered vs whispered speech (Whsp vs Whsp), both whispered and normal vs whispered and normal speech (All vs All). The table lists mean and standard deviation from 10 independent runs.

Model \ Trial	Norm vs Whsp	Norm vs Norm	Whsp vs Whsp	All vs All
ECAPA-TDNN	27.53% ± 1.35 %	2.40% ± 0.53 %	24.18% ± 2.13 %	27.15% ± 1.02 %
ECAPA2	24.36% ± 0.91 %	1.70 % ± 0.56 %	21.81% ± 1.98 %	24.92% ± 1.23 %
ReDimNet-B2	26.51% ± 1.13 %	1.25% ± 0.59 %	24.08% ± 1.76 %	26.52% ± 0.68%
ReDimNet-B6	27.52% ± 0.94 %	0.85% ± 0.39 %	25.60% ± 1.66 %	24.54% ± 0.97 %

5 Conclusion

In this paper, we showed that adding an encoder-decoder-like architecture on top of a speaker verification model fine-tuned and trained with a speaker classification head with cosine similarity and triplet loss improves speaker verification with whispered speech. Specifically, using ReDimNet-B6 as the encoder enhances the EER by a relative 22.26% compared to the baseline (baseline 6.77% vs ours 5.27%) in Normal vs Whispered speech trials, achieving AUC of 98.16%. For Whispered vs Whispered tests our model achieves EER of 1.88% with AUC equal to 99.73%, showing 15% relative improvement over the previous best ReDimNet-B2. We provided a comparative evaluation of current state-of-the-art speaker

Table 5: The comparison of EER scores on the testing dataset with low SNR (PSNR(5 dB) $\approx$ 22) of baselines across trials of normal vs whispered speech (Norm vs Whsp), normal vs normal speech (Norm vs Norm), whispered vs whispered speech (Whsp vs Whsp), both whispered and normal vs whispered and normal speech (All vs All). The table lists mean and standard deviation from 10 independent runs.

Model \ Trial	Norm vs Whsp	Norm vs Norm	Whsp vs Whsp	All vs All
ECAPA-TDNN	35.63% ± 0.91 %	6.14% ± 0.62 %	34.85% ± 1.56 %	32.60% ± 0.85 %
ECAPA2	24.36% ± 0.91 %	1.70 % ± 0.56 %	21.81% ± 1.98 %	24.92% ± 1.23 %
ReDimNet-B2	26.30% ± 1.22 %	1.30 % ± 0.59 %	23.60% ± 1.87 %	26.46% ± 0.76 %
ReDimNet-B6	27.52% ± 0.94 %	0.85% ± 0.39 %	25.60% ± 1.66 %	24.54% ± 0.97 %

verification systems under whispered speech conditions. We discovered that even contemporary solutions must adapt to maintain their advanced performance, particularly with whispered speech. For instance, ReDimNet-B6's effectiveness in verifying normal vs normal speech compared to normal vs whispered speech decreased by relatively 55.42%. We also analyzed how noise affects the speaker verification models on normal and whispered speech, obtaining that generally the same relative level of noise has a more degrading influence on whispered speech than on normal speech.

Despite the promising improvements achieved by incorporating an encoder decoder architecture into the speaker verification pipeline, several limitations remain. First, our experiments were conducted on one dataset with whispered speech, which may not capture the full diversity of real-world whispering styles, recording environments, and speaker demographics. In fact, there is a limited number of datasets with English whispered speech which hinders obtaining a general robust model. As a next step, the model may be trained using both major whispered speech datasets, CHAINS and wTIMIT. Secondly, our framework requires fine-tuning a complex speaker verification model, which could be resource-intensive in terms of computation.

Future work should explore more data-efficient or lightweight architectures that preserve performance while reducing computational cost. Further experiments could explore multilingual and cross-lingual whispered datasets to help determine the model's adaptability across linguistic contexts. Moreover, due to scarce databases with whispered speech, collecting a larger dataset may be considered for future work to fundamentally improve the performance of speaker verification systems with whispered speech, including samples collected in real-

Table 6: Relative change of EER (defined as in Equation 2) in noisy conditions compared to results achieved with clean speech across normal vs normal (n vs n) and whispered vs whispered (w vs w) trials.

Trial / **Model**	**N vs N** PSNR ≈ 38	**W vs W** PSNR ≈ 38	**N vs N** PSNR ≈ 22	**W vs W** PSNR ≈ 22
ECAPA-TDNN	3.15	7.73	13.97	11.58
ECAPA2	7.01	7.79	7.01	7.79
ReDimNet-B2	4.43	9.94	4.65	9.72
ReDimNet-B6	6.08	10.13	6.08	10.13

world scenarios. Finally, adding synthetically generated data might help with generalization to new speakers and environmental conditions, which influence speaker verification with clean and whispered speech unfavorably as shown in experiments with added noise.

Acknowledgments. This work has been partially funded by Department of Artificial Intelligence, Wrocław University of Science. Created using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl), grant no. 1754073716.

References

1. Cummins, F., Grimaldi, M., Leonard, T., Simko, J.: The chains corpus: Characterizing individual speakers. Proc. SPECOM pp. 431–435 (01 2006)
2. Desplanques, B., Thienpondt, J., Demuynck, K.: Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification (10 2020). https://doi.org/10.21437/Interspeech.2020-2650
3. Ito, T., Takeda, K., Itakura, F.: Analysis and recognition of whispered speech. Speech Communication **45**(2), 139–152 (2005). https://doi.org/https://doi.org/10.1016/j.specom.2003.10.005, https://www.sciencedirect.com/science/article/pii/S0167639304000706
4. Jovičić, S.T., Šarić, Z.: Acoustic analysis of consonants in whispered speech. Journal of Voice **22**(3), 263–274 (2008). https://doi.org/https://doi.org/10.1016/j.jvoice.2006.08.012, https://www.sciencedirect.com/science/article/pii/S0892199706001159
5. Juang, B.H., Sondhi, M., Rabiner, L.R.: Digital speech processing. In: Meyers, R.A. (ed.) Encyclopedia of Physical Science and Technology (Third Edition), pp. 485–500. Academic Press, New York, third edition edn. (2003). https://doi.org/https://doi.org/10.1016/B0-12-227410-5/00178-2, https://www.sciencedirect.com/science/article/pii/B0122274105001782

6. Khmelev, N., Avdeeva, A., Novoselov, S., Chirkovskiy, A., Volkova, M.: Robust speaker recognition for whispered speech. In: 2025 27th International Conference on Digital Signal Processing and its Applications (DSPA). pp. 1–5 (2025). https://doi.org/10.1109/DSPA64310.2025.10977907
7. Naini, A.R., M. V., A.R., Ghosh, P.K.: Formant-gaps features for speaker verification using whispered speech. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6231–6235 (2019). https://doi.org/10.1109/ICASSP.2019.8682571
8. Prieto, S., Ortega, A., López-Espejo, I., Lleida, E.: Shouted and whispered speech compensation for speaker verification systems. Digital Signal Processing **127**, 103536 (2022). https://doi.org/https://doi.org/10.1016/j.dsp.2022.103536, https://www.sciencedirect.com/science/article/pii/S1051200422001531
9. 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., Mori, R.D., Bengio, Y.: SpeechBrain: A general-purpose speech toolkit (2021), arXiv:2106.04624
10. Sarria-Paja, M., Falk, T.H.: Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification. Computer Speech & Language **45**, 437–456 (2017). https://doi.org/https://doi.org/10.1016/j.csl.2017.04.004, https://www.sciencedirect.com/science/article/pii/S0885230816303382
11. Sarria-Paja, M., Falk, T.H.: Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech. Speech Communication **102**, 78–86 (2018). https://doi.org/https://doi.org/10.1016/j.specom.2018.07.005, https://www.sciencedirect.com/science/article/pii/S0167639317304703
12. Sarria-Paja, M., Falk, T.H., O'Shaughnessy, D.: Whispered speaker verification and gender detection using weighted instantaneous frequencies. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7209–7213 (2013). https://doi.org/10.1109/ICASSP.2013.6639062
13. Sarria-Paja, M.O., Falk, T.H.: Strategies to enhance whispered speech speaker verification: A comparative analysis. Canadian Acoustics **43**(4), 31–45 (Dec 2015), https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2670
14. Snyder, D., Chen, G., Povey, D.: MUSAN: A Music, Speech, and Noise Corpus (2015), arXiv:1510.08484v1
15. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333 (2018). https://doi.org/10.1109/ICASSP.2018.8461375
16. Thienpondt, J., Demuynck, K.: Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2023)
17. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. Proceedings of the 25th ACM international conference on Multimedia (2017), https://api.semanticscholar.org/CorpusID:7680631
18. Yakovlev, I., Makarov, R., Balykin, A., Malov, P., Okhotnikov, A., Torgashov, N.: Reshape dimensions network for speaker recognition. In: Interspeech 2024. pp. 3235–3239 (2024). https://doi.org/10.21437/Interspeech.2024-2116