PPX-Anon: Prosody, Pitch and X-vectors for De-Anonymization; the submisison of the SHADOW team for the Attacker Challenge 2024

1st Thomas Thebaud*
ECE dept. & CLSP
Johns Hopkins University
Baltimore, MD, USA
0000-0001-8953-7872

2nd Yaohan Guan*
ECE dept. & CLSP
Johns Hopkins University
Baltimore, MD, USA
0009-0007-6263-3232

3rd Nick Mehlman*
SAIL Lab
University of Southern California
Los Angeles, California
nmehlman@usc.edu

4th Jesus Villalba
ECE dept. & CLSP
Johns Hopkins University
Baltimore, MD, USA
0000-0001-9459-8426

5th Laureano Moro-Velazquez
ECE dept. & CLSP
Johns Hopkins University
Baltimore, MD, USA
0000-0002-3033-7005

6th Shrikanth Narayanan
SAIL Lab
University of Southern California
Los Angeles, California
shri@usc.edu

7th Najim Dehak
ECE dept. & CLSP
Johns Hopkins University
Baltimore, MD, USA
0000-0002-4489-5753

***Abstract*—We present a novel approach to de-anonymize speech that has been transformed by a voice privacy system. Inspired by the complex and multi-factorial nature of speaker identification, we extract three different identity-related features, namely x-vectors, pitch-based representations, and prosody embeddings. These features are then fused together and used to perform speaker verification on the anonymized data. By integrating multiple parallel streams of identity information, we increase the robustness of the system to different voice conversion methods, and also allow for easy fine-tuning to exploit the unique weaknesses of a specific anonymization method.**

***Index Terms*—de-anonymization, x-vectors, prosody, pitch estimation**

I. System Overview

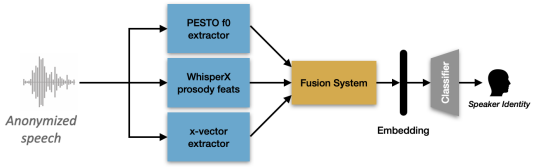

Fig. 1. Overview of the proposed de-anonymization system.

Speaker identification is known to be influenced by various features, including spectral elements [10], pitch information [8], [12], and prosody [3], [5]. While a voice conversion system may effectively obscure a subset of these attributes, it is likely that identifiable features will still leak through in other components. For example, a system that manipulates the timbre may fail to alter the pitch trajectory. This motivated our system's design philosophy, which emphasizes leveraging multiple knowledge-based sources of identity information in parallel. In particular, we extract an x-vector representation, a pitch-based feature, and a prosody embedding. This approach increases the likelihood of discovering salient identity features that the anonymization system has left exposed. It also offers improved robustness to different obfuscation methods since the fusion of the different information sources can be adapted to exploit the weakness of a specific system.

We train each of the three sub-systems separately on the anonymized training data and then use their trained models to generate feature vectors for each enrollment and test utterance. These feature vectors are then combined using a learnable fusion module. The fused output is used to perform speaker verification on the anonymized evaluation data.

A. X-vectors Features

Since they were first introduced by [9], x-vector systems have demonstrated great effectiveness for speaker recognition and verification tasks. X-vector models are generally trained to perform speaker identification on a labeled dataset, which causes the model's hidden representation space to adapt to capture speaker identity. As a result, the embeddings extracted from the trained model (the x-vectors) can be employed in down-steam identification and verification tasks. For this task, we fine-tune a **fw-SEResNet34** [4] on the anonymized speech of each model. The model used is first trained on VoxCeleb1-dev [6] and VoxCeleb2-dev [2], until it reaches an EER of 0.63% on Voxceleb1-O split.

Then, we fine-tune once this network for 15 epochs on all the available anonymized datasets, and repeat the process for each of the 7 anonymized datasets. For each given test dataset, we use the concatenated triplet of x-vectors, composed of :

1) The x-vector extracted from the base model
2) The x-vector extracted from the model fine-tuned on all the anonymized data available
3) The x-vector extracted from the model fine-tuned only on the related dataset at hand (e.g. fine-tune on the B3 model when the test data is extracted using the B3 model.).

TABLE I

EER(%) ON BOTH MALE (M) AND FEMALE (F) SUBSETS, FOR EACH OF OUR SUBSYSTEMS, AS WELL AS THE FUSION PROPOSED SYSTEM. "X-VECTOR FT. ALL" SHOW THE RESULT FOR THE X-VECTOR SYSTEM FINE-TUNE ON ALL DATASETS, WHILE "X-VECTOR FT. EACH" SHOW THE SYSTEM FINE-TUNED ON THE RELATED SYSTEM. SUBMITTED SYSTEMS ARE IN **BOLD**.

Set	System	B3		B4		B5		T8-5		BT10-2		T12-5		T25-1		orig	
		m	f	m	f	m	f	m	f	m	f	m	f	m	f	m	f
Dev	X-Vector pretrained	43.91	43.80	44.85	48.03	49.06	47.89	45.79	46.69	34.80	34.60	49.06	46.77	48.12	48.56	0.15	2.42
Test	X-Vector pretrained	44.88	44.85	47.11	45.12	48.58	47.57	44.80	44.96	34.79	31.23	49.26	48.77	49.24	48.58	0.40	0.17
Dev	X-Vector ft. all	49.01	49.50	49.23	49.53	49.62	49.32	49.55	49.66	48.18	46.74	49.43	46.89	48.60	48.42	49.93	47.75
Test	X-Vector ft. all	45.37	47.83	48.28	48.06	48.14	47.32	48.11	47.87	43.65	46.15	45.50	47.81	45.57	48.11	38.52	42.83
Dev	X-Vector ft. each	**33.19**	**36.14**	49.02	46.07	**37.65**	**44.71**	**36.87**	**34.93**	**33.47**	**37.36**	**39.91**	**45.69**	**46.62**	**48.19**	0.15	2.42
Test	X-Vector ft. each	**35.48**	**37.35**	46.23	47.06	**39.22**	**39.00**	**37.31**	**35.82**	**33.71**	**35.08**	**40.86**	**40.47**	**44.16**	**47.86**	0.40	0.17
Dev	Prosody Features	46.16	44.42	**45.42**	**44.85**	44.80	44.53	46.83	43.83	45.57	44.77	45.03	45.14	45.48	44.53	45.45	44.56
Test	Prosody Features	41.83	43.67	42.70	**44.46**	**43.49**	44.40	44.16	44.52	41.68	43.32	43.31	43.42	42.71	44.31	41.74	43.53
Dev	Pitch Features	49.92	46.56	49.94	46.37	49.70	45.75	45.70	45.80	44.55	50.00	45.84	49.88	46.12	47.03	49.99	50.00
Test	Pitch Features	49.05	47.06	48.56	48.64	49.54	49.30	48.81	47.86	47.38	49.60	48.90	49.27	48.41	48.13	49.37	49.99
Dev	Fusion	49.81	36.56	49.06	46.12	46.56	44.68	46.59	36.90	47.14	37.56	48.36	46.64	49.84	46.61	**0.28**	**9.53**
Test	Fusion	49.80	38.91	49.78	47.01	48.61	46.49	49.31	41.44	48.98	41.90	49.85	46.54	49.94	45.32	**0.53**	**23.45**

B. Prosodic Features

It has been widely shown (e.g. [3], [5], [11]) that prosodic information, related to the temporal dynamics of an individual's speech, plays an important role in speaker identification. Furthermore, prosody is more difficult to modify than other identify-related speech features such as spectral content. As a result, we have chosen to integrate temporal information into our attacker model.

Using WhisperX [1], we predict character-level aligned transcripts. This allows us to assign a unique character to each audio frame (including a null character for frames where no speech is present). These sequences capture the unique speaking dynamics of the individual, such as inter-word pauses and phoneme lengths, and so should be informative for modeling speaker identity. We thus train a transformer encoder equipped with a linear classification head on this task. Each character sequence is tokenized, and converted to 128 dimensional vectors via a learnable embedding layer. Positional encoding is then added and a linear projection layer is applied. Next, the sequences are passed to a 2 layer transformer encoder with model dimensions of 512. Mean pooling is applied to aggregate the encoder output over time to produce a single speaker-embeddings vector. Finally, a linear classification head is applied to predict speaker identity. This system is summarized in figure 2. We pretrained the model on the train split for all VC systems and then extracted speaker embeddings for the enrollment and trial utterances to be used in the fused de-identification system.

C. Pitch Features

Like prosody, pitch information also contributed significantly to speaker recognition [8], [12]. While static pitch straights (e.g. the mean fundamental frequency f_0) are fairly easy for a voice conversion system to modify, the dynamic pitch trajectories that occur over the course of an utterance are more challenging to obscure. We use the pre-trained PESTO model [7], a pitch-estimation framework trained using self-supervised learning, to extract dynamic pitch features from the spoken audio.

Once the pitch is extracted for each 10ms of speech, we compute the deltas and delta-deltas of the pitch, then compute a statistical pooling (mean and standard deviation) on the pitch, deltas, delta-deltas and self-estimated confidence of the model, resulting in a 8 dimensional vector for each utterance.

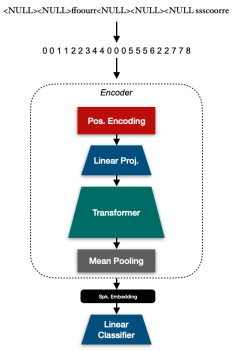

Fig. 2. Overview of prosody speaker-identification model. Per-frame character sequences are tokenized and passed to a transformer-based encoder that outputs a speaker embedding. During training, a linear head is applied to the embeddings to predict speaker labels

D. Fusion

Once the scores are obtained for each of the subsystems, we train a weighted average on the development set to maximize the variance between the positive and negative comparisons (practically training a Linear Discriminant Analysis on the scores of the development set). The weights obtained are used to compute the final scores for the test sets.

II. RESULTS

The de-anonymization results are presented using the Equal Error Rate (EER) in Table I. If we can observe significant improvement on each system from the base x-vector system used, most of the presented results are still over 40% EER, which either show a relatively weak attacker, or a very strong set of anonymization systems. Our best system seems to be the X-Vector system, trained on each model separately, followed closely by the prosody features.

III. CONCLUSION

In conclusion, we proposed a novel approach that demonstrates the effectiveness of combining x-vectors, prosody, and pitch features to de-anonymize speech after privacy transformations. This fusion of diverse identity cues enhances robustness across different anonymization methods, though results suggest further improvements are needed to address stronger anonymization techniques.

REFERENCES

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[3] Elina Helander and Jani Nurminen. On the importance of pure prosody in the perception of speaker identity. pages 2665–2668, 08 2007.

[4] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. Assert: Anti-spoofing with squeeze-excitation and residual networks. *arXiv preprint arXiv:1904.01120*, 2019.

[5] L. Mary. Significance of prosody for speaker, language, emotion, and speech recognition. In *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*, SpringerBriefs in Speech Technology. Springer, Cham, 2019.

[6] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

[7] Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. Pesto: Pitch estimation with self-supervised transposition-equivariant objective. In *International Society for Music Information Retrieval Conference (ISMIR 2023)*, 2023.

[8] J. Shen and J. Wu. Recognition of speech with dynamic pitch manipulation in noise: Effects of manipulation methods. *Journal of Speech, Language, and Hearing Research*, 67(1):269–281, Jan 2024. Epub 2023 Nov 20.

[9] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[10] William D. Voiers. Perceptual bases of speaker identity. *The Journal of the Acoustical Society of America*, 36(6):1065–1073, 06 1964.

[11] Frederick Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg. Using prosodic and lexical information for speaker identification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–141–I–144, 2002.

[12] Jian-wei Zhu, Shui-fa Sun, Xiao-li Liu, and Bang-jun Lei. Pitch in speaker recognition. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, volume 1, pages 33–36, 2009.