Supplementary Material for: SAMAF: Sequence-to-sequence Autoencoder Model for Audio Fingerprinting

ABRAHAM BÁEZ-SUÁREZ, ITESM, MX and Univ. of Houston, USA
NOLAN SHAH, Univ. of Houston, USA
JUAN ARTURO NOLAZCO-FLORES, ITESM, MX
SHOU-HSUAN S. HUANG, OMPRAKASH GNAWALI, and WEIDONG SHI, Univ. of Houston, USA

A SUPPLEMENTARY MATERIALS

A.1 Speech Transformations

Table A1 describes the transformations applied over the VoxCeleb1 dataset and their parameters. These transformations were chosen based on the literature of existing audio fingerprinting techniques to assess the robustness of the SA model in handling compression, distortion, and interference of audio.

A.2 Hash Space Structure vs. Audio Identification Analysis

The audio identification task performance against hash space structure was discussed in Section 4.4. Although we hypothesized that the more structured a hash space is, the higher its performance, it was found that it depends on the parameters of the model. The combined information of Tables A2, A3, and A4 shows that the model with a structured space and high identification performance is 2LST128M-Acc.

A.3 Audio Identification Performance

In this study, the creation of audio fingerprints was developed via an unsupervised deep learning architecture (Sequence-to-Sequence Autoencoder) with different model parameters such as the number of layers, hash size, strategy, loss function terms, and audio length. These model parameters were tested throughout a subset of the VoxCeleb1 dataset. An exhaustive testing per model and its different parameters are presented in the following tables.

Table A3 describes the audio identification performance in a subset of VoxCeleb1 dataset where the Mean Square Error (M) loss term and the Accuracy strategy achieved the best performance overall regardless of the model parameters. The improvement in performance between 128 and 256 bits is minimal, suggesting that, if required, a small hash size can be used without diminishing the overall performance. There is a significant difference when comparing the performance per layer; it is clear that 2-layer models achieve higher results.

Table A4 describes the transformation performance in a subset of VoxCeleb1 dataset. Once again, the Mean Square Error (M) loss term and the Accuracy strategy achieved the best performance overall regardless of the model parameters. There are significant differences between a small (128) and big (256) hash size. It is believed that certain transformations were either better or worse encoded, thereby affecting the classification performance. When comparing the number of layers, the difference is not clear, suggesting that the hash size is the one driving the transformation classification performance and that the number of layers has a minor effect.

1551-6857/2020/05-ART43 $15.00
https://doi.org/10.1145/3380828

Table A1. Speech Transformations

Transformation	Description
TruncateSilence	Removes silence (less than −60dB)
Noise0.05	Adds white noise of amplitude 5% to the audio
Echo	Repetition of audio with delay of 1 second and decay factor of 0.5
Reverb	Adds reverberation with room size 75%, reverberance of 50%, damping of 50%, wet gain of −1dB, stereo depth of 100%, and predelay of 10 ms.
HighPassFilter	Attenuates frequencies below 1 KHz
LowPassFilter	Attenuates frequencies above 3 KHz
Reverse	Audio was processed to be reversed
Pitch0.5	Pitch reduction to 50% of original
Pitch0.9	Pitch reduction to 90% of original
Pitch1.1	Pitch increase to 110% of original
Pitch1.5	Pitch increase to 150% of original
Speed0.5	Speed reduction to 50% of original
Speed0.9	Speed reduction to 90% of original
Speed0.95	Speed reduction to 95% of original
Speed1.05	Speed increase to 105% of original
Speed1.1	Speed increase to 110% of original
Speed1.5	Speed increase to 150% of original
Tempo0.5	Tempo reduction to 50% of original
Tempo0.9	Tempo reduction to 90% of original
Tempo1.1	Tempo increase to 110% of original
Tempo1.5	Tempo increase to 150% of original

Table A2. Average Hamming Distance in the VoxCeleb1 Dataset with the Different Model Parameters Such as the Number of Layers, Hash Size, Strategy, Loss Function Term, and Audio Length

Length	1 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	5	5	4	3	0	2	3	**0**	2	13	12	11	7	7	3	6	7	**2**
2 seconds	70	68	62	32	11	24	33	**9**	26	139	134	127	54	56	**29**	53	57	29
3 seconds	148	143	131	67	25	51	68	**21**	53	290	279	265	106	110	60	105	113	**59**
4 seconds	232	223	204	104	39	79	105	**33**	81	449	434	413	162	168	92	160	172	**91**
5 seconds	318	305	279	142	55	109	144	**47**	111	613	593	565	220	227	126	215	232	**124**
6 seconds	406	389	355	183	72	139	185	**61**	142	780	755	720	279	288	161	272	294	**157**
Length	2 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	6	**5**	7	8	9	10	9	10	13	**14**	16	15	25	35	36	28	38	36
2 seconds	72	65	78	**64**	71	76	72	81	94	**141**	154	148	161	228	230	178	244	232
3 seconds	151	136	161	**126**	140	150	142	158	184	**288**	314	301	309	437	441	341	466	444
4 seconds	235	211	248	**192**	213	227	214	240	275	**443**	484	463	464	653	658	509	696	662
5 seconds	322	290	338	**260**	286	305	289	322	368	**604**	660	631	622	871	876	680	928	883
6 seconds	411	371	431	**330**	360	383	364	405	460	**769**	842	803	784	1,091	1,097	855	1162	1105

Abbreviations: M - Mean Square Error Loss || MH - Mean Square Error Loss + Hash Loss ||
MHE - Mean Square Error Loss + Hash Loss + Bitwise Entropy Loss

Table A3. Audio Identification Performance in the VoxCeleb1 Dataset Showing All the Classification Results with the Different Model Parameters Such as the Number of Layers, Hash Size, Strategy, Loss Function Term, and Audio Length

Length	1 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	**14.4**	14.1	12.8	14.3	3.5	13.0	14.3	2.1	12.3	18.8	17.9	17.2	**24.0**	22.9	16.0	21.8	21.8	12.4
2 seconds	37.1	36.8	36.6	41.2	30.9	40.1	**41.7**	26.6	38.3	39.8	39.1	37.7	**45.4**	42.6	42.1	42.6	40.5	37.8
3 seconds	42.5	42.2	42.4	51.1	39.7	48.7	**52.0**	34.7	44.9	45.0	44.3	42.0	**51.8**	47.8	50.0	48.1	45.1	43.8
4 seconds	45.2	45.4	44.9	55.2	43.9	52.3	**56.3**	38.9	48.2	47.7	47.4	44.5	**54.6**	50.5	53.6	50.0	47.7	46.7
5 seconds	46.5	46.9	46.0	57.0	45.5	53.9	**57.8**	41.5	49.6	48.6	48.5	45.8	**55.8**	52.0	54.5	50.4	49.1	48.1
6 seconds	46.5	47.0	45.8	56.5	46.1	54.1	**57.7**	42.6	49.9	48.6	48.4	45.3	**55.8**	52.4	54.4	50.1	49.3	48.5
Length	2 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	14.9	13.9	16.5	**26.6**	23.8	23.8	26.3	24.3	24.5	21.2	21.7	21.9	**36.8**	34.0	33.9	36.7	34.1	33.8
2 seconds	39.6	39.5	43.5	**47.3**	43.1	42.1	45.2	42.5	40.5	44.4	45.0	45.6	**50.8**	44.9	44.5	50.5	44.3	44.8
3 seconds	46.8	46.3	51.0	**54.3**	47.0	45.9	51.8	46.2	43.1	51.2	51.2	52.3	**56.5**	48.0	47.8	55.6	47.4	48.1
4 seconds	50.5	49.5	54.1	**57.8**	48.6	47.1	54.7	47.6	44.0	53.9	54.0	55.1	**58.8**	49.1	49.0	57.9	48.8	49.3
5 seconds	51.9	50.2	54.7	**58.7**	48.6	47.4	55.4	47.7	43.9	54.5	54.3	55.9	**59.3**	49.0	48.9	58.4	49.0	49.2
6 seconds	51.6	49.5	53.6	**58.3**	47.9	46.8	55.4	47.3	43.3	53.9	53.7	54.8	**58.2**	48.1	48.0	57.2	48.4	48.4

Abbreviations: M - Mean Square Error Loss || MH - Mean Square Error Loss + Hash Loss ||
MHE - Mean Square Error Loss + Hash Loss + Bitwise Entropy Loss

Table A4. Transformation Performance in the VoxCeleb1 Dataset Showing All the Classification Results with the Different Model Parameters Such as the Number of Layers, Hash Size, Strategy, Loss Function Term, and Audio Length

Length	1 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	**15.3**	14.9	13.6	15.3	3.7	14.0	15.3	2.2	13.2	19.8	18.9	18.2	**25.7**	24.4	17.1	23.3	23.3	13.3
2 seconds	39.8	39.4	39.2	44.3	33.3	43.1	**44.8**	28.8	41.0	42.5	41.7	40.2	**48.5**	45.5	45.1	45.6	43.2	40.5
3 seconds	46.0	45.6	45.7	55.3	43.1	52.7	**56.2**	37.8	48.5	48.6	47.8	45.4	**55.9**	51.6	54.0	52.1	48.6	47.4
4 seconds	49.4	49.5	49.0	60.4	48.0	57.2	**61.4**	42.7	52.6	52.0	51.7	48.7	**59.7**	55.3	58.6	54.8	52.1	51.1
5 seconds	51.6	51.8	50.9	63.3	50.4	59.7	**64.0**	46.1	55.0	53.9	53.8	50.9	**62.1**	57.8	60.4	56.2	54.6	53.5
6 seconds	52.8	53.1	51.9	64.2	52.1	61.2	**65.1**	48.4	56.4	55.1	54.8	51.4	**63.4**	59.5	61.5	57.1	56.1	55.0
Length	2 Layer																	
	128 Bits									256 Bits								
	Rnd			Acc			Loss			Rnd			Acc			Loss		
	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE	M	MH	MHE
1 second	15.9	14.8	17.6	**28.4**	25.4	25.4	28.0	25.9	26.2	22.5	23.0	23.2	**39.1**	36.1	36.0	39.0	36.2	35.9
2 seconds	42.4	42.4	46.6	**50.7**	46.1	45.1	48.4	45.5	43.4	47.6	48.1	48.8	**54.4**	48.1	47.7	54.0	47.4	48.0
3 seconds	50.6	50.1	55.1	**58.7**	50.9	49.7	56.0	50.0	46.7	55.3	55.3	56.4	**61.2**	52.0	51.8	60.3	51.4	52.2
4 seconds	55.2	54.2	59.0	**63.3**	53.4	51.6	59.9	52.3	48.3	58.9	58.9	60.2	**64.4**	53.9	53.8	63.6	53.6	54.1
5 seconds	57.6	55.7	60.5	**65.1**	54.4	52.8	61.8	53.3	49.0	60.4	60.1	61.9	**66.2**	54.8	54.7	65.3	54.8	55.0
6 seconds	58.4	56.2	60.8	**66.1**	54.8	53.4	63.2	54.1	49.5	61.2	60.8	62.2	**66.6**	55.0	54.8	65.5	55.2	55.2

Abbreviations M - Mean Square Error Loss || MH - Mean Square Error Loss + Hash Loss ||
MHE - Mean Square Error Loss + Hash Loss + Bitwise Entropy Loss