EURECOM
Department of Digital Security
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-20-343

Speaker anonymisation using the McAdams coefficient

February 28th, 2020
Last update February 28th, 2020

Jose Patino, Massimiliano Todisco, Andreas Nautsch and Nicholas Evans

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : lastname@eurecom.fr

[1] EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Orange, Principauté de Monaco, SAP, Symantec.

Speaker anonymisation using the McAdams coefficient

Jose Patino, Massimiliano Todisco, Andreas Nautsch and Nicholas Evans

Abstract

This report presents an alternative to speaker anonymisation that does not require external training data and is based on a single processing block using basic signal processing techniques. The proposed method employs the McAdams coefficient to apply a slight contraction/expansion to the poles derived from linear predictive coding (LPC) coefficients of speech content on a frame-by-frame basis, consequently leading to a transformation of the related formants.

Index Terms

Speaker anonymisation, privacy preservation, signal processing, speaker recognition

Contents

List of Figures

1 Motivation

The official baseline for the VoicePrivacy 2020 Challenge, based on the work presented in [1], is currently available[1]. This x-vector based recipe to anonymisation increases the equal error rate (EER) of an automatic speaker verification (ASV) system at the cost of a relatively small increase in the word error rate (WER) of an automatic speech recognition system (ASR). However, it does so at the cost of expensive training and a complex pipeline. The first limitation, undesirable in certain use cases, can be managed by the sharing of pre-trained models. On the other hand, the second may constitute a limiting factor to the prospective participants to the challenge, as it requires familiarity with the Kaldi framework and numerous processing blocks at the anonymisation stage. In consequence, it could be of interest to provide participants with an easily approachable alternative that lowers the entry point to the anonymisation concept, as well as to the processing of the speech data involved in the assessment of the systems.

This report presents an alternative to speaker anonymisation that does not require external training data and is based on a single processing block using basic signal processing techniques. The proposed method employs the McAdams coefficient [2] to apply a slight contraction/expansion to the poles derived from linear predictive coding (LPC) coefficients of speech content on a frame-by-frame basis, consequently leading to a transformation of the related formants. The method is briefly introduced in Section 2, results are presented and compared to the existing baseline in Section 3, and final comments are made in Section 4.

2 The McAdams coefficient: a *poor man's* anonymisation

2.1 McAdams coefficient in music processing

In music signal processing, one of the most common synthetic sound generation techniques is that of additive synthesis [3]. The technique generates timbre through the addition of cosinusoidal oscillations. Such a process can be seen as the Fourier series consisting in multiple harmonic partials. The frequency of each partial may also be adjusted using the McAdams coefficient [2], that allows to change the distribution of the partials and consequently the resulting timbre. The McAdams coefficient, when applied upon the additive synthesis formula, consequently transforms harmonic to inharmonic partials or overtones, and is defined as follows:

$$y(t) = \sum_{k=1}^{K} r_k(t) \cos(2\pi(kf_0)^{\alpha} t + \phi_k) \tag{1}$$

where α is the McAdams coefficient, $r_k(t)$ is the amplitude of each harmonic, k is the partial number, f_0 is the fundamental frequency, ϕ_k is the phase and t is time.

[1] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

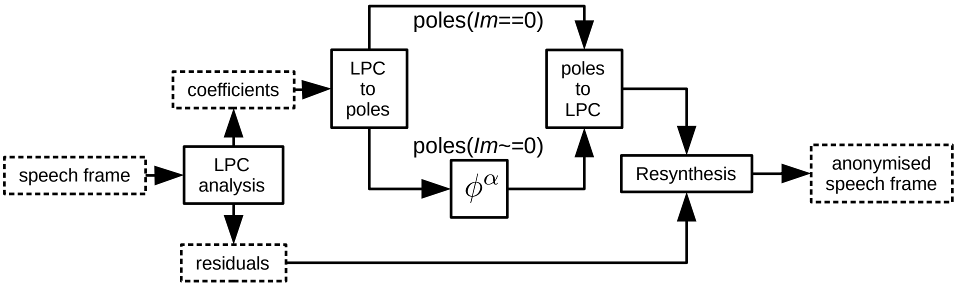

Figure 1: Pipeline of the application of the proposed McAdams coefficient-based approach to anonymisation on a speech frame basis. The angle ϕ of poles with a non-zero imaginary part are raised to the power of the McAdams coefficient α to provoke a expansion/contraction in frequency in its associated formant.

2.2 McAdams coefficient in anonymisation

Inspired by its use in music processing, this secondary baseline for speaker anonymisation is based on a formant transformation (as per its partials counterpart in additive synthesis) which employs the McAdams coefficient. The process, depicted in Figure 1, starts with the the application of LPC source-filter analysis to a speech signal, frame-by-frame. LPC coefficients and residuals are obtained. Residuals are set aside and retained for later resynthesis. The McAdams transformation is then applied to the angle of the poles that are derived from the LPC coefficients, each one of which corresponds - in first approximation - with a formant in the spectrum. While real-valued poles are left unmodified, the angles (ϕ) of the poles with a non-zero imaginary part (with values between 0 and π radians) are raised to the power of the McAdams coefficient α so that a transformed pole has an angle ϕ^α.

In consequence an angle in radians $\phi < 1$ contracts for an $\alpha > 1$, and expands for an $\alpha < 1$. And viceversa, an angle $\phi > 1$ contracts for an $\alpha < 1$, and expands for an $\alpha > 1$. Its effect upon the poles is visible in Figure 3 for values of $\alpha = \{0.9, 1.1\}$ and on the spectral envelope in Figure 2. For a sampling rate of 16kHz as the one processed in the challenge, this threshold dependent on the value of the angle is equivalent to around 2.5kHz, splitting the spectrum into two balanced parts with respect to the formant average values [4]. The final set of new poles, including the modified poles and the original poles with a zero imaginary part, are then converted back to LPC coefficients. Finally, LPC coefficients and residuals are used to resynthesise the speech frame. It is worth noting that the technique introduced here is similar in nature to the VoiceMask method [5] (recently studied within a privacy context in [6]).

3 Results

For the sake of comparison, results are presented for both the original x-vector baseline system and the proposed McAdams system. Table 1 shows the impact

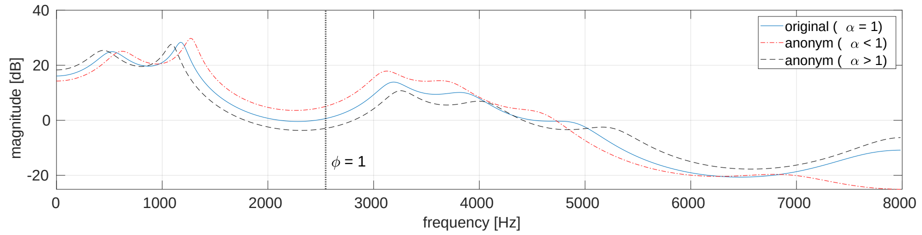

Figure 2: Example of the spectral envelope of a speech frame for both the original formants and the two anonymised versions. The effect of the McAdams coefficent α with regards to causing a expansion or contraction of the spectrum is relative to the value of ϕ.

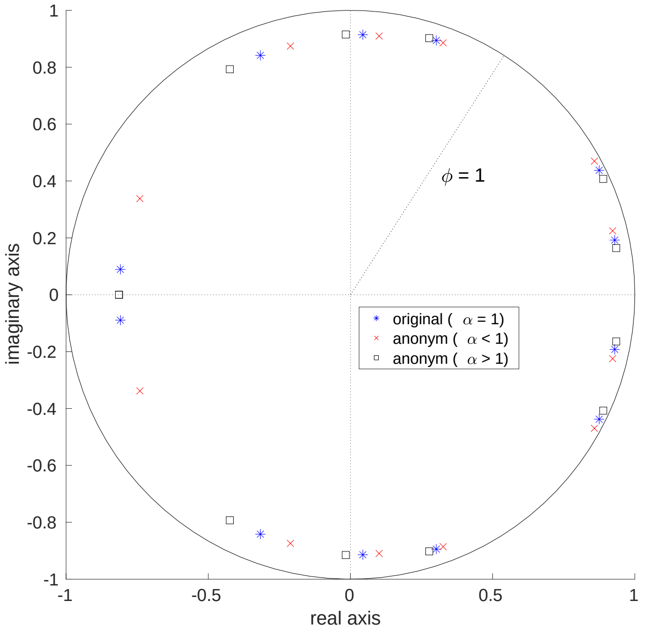

Figure 3: Example of pole-zero map as shown in Figure 2

of the x-vector system upon ASV whereas Table 2 shows the impact upon ASR, on both development and test sets. On the other hand, Tables 3 and 4 shows the corresponding results for the McAdams system.

As regards anonymisation (impact upon ASV for an $\alpha = 0.8$), results for the McAdams system are slightly behind those of the x-vector system for both Librispeech and VCTK data. The EER is lower than that of the x-vector baseline as possibly *less anonymous* speech is generated by using the McAdams coefficient method (comparison between Tables 1 and 3).

Results in terms of intelligibility (impact upon ASR for an $\alpha = 0.8$), show a similar trend (comparison between Tables 2 and 4). WER results for x-vector system are lower than those for the McAdams system. The difference between the two methods, is however, very small for the Librispeech data, while more substantial for the VCTK data, where the McAdams coefficient method seems to significantly degrade audio quality (albeit not excessively as judged from casual subjective listening tests).

The use of the McAdams coefficient for anonymisation is hence inferior to the x-vector system, yet reasonably competitive, given the comparatively simple approach. It is stressed also that the α coefficient and the number of LPC coefficients have not been optimised. They were set to 0.8 and 20, respectively, from experiments based upon a single speech file not belonging to any of the challenge partitions. It is also stressed that performance was not an objective in this work. Instead, the objective was to provide a comparatively simple approach to anonymisation that would lower the cost of entry for potential participants, and also show that participation does not necessarily require expertise in ASV, ASR and deep learning. Reasonable performance can be achieved with the application of basic signal processing techniques. It is hoped, therefore, that the introduction of this second baseline might serve as further inspiration for potential participants, while also broadening the appeal of the challenge to a wider audience.

4 Conclusions

The present document introduces the work done at EURECOM to produce a baseline less complex than the currently available x-vector based system for the VoicePrivacy 2020 Challenge. An alternative, single-block, signal-processing based approach to anonymisation based on the McAdams coefficient and LPC processing is proposed. The performance achieved by this new baseline system is, unsurprisingly, behind that of the x-vector system, but is worth sharing with potential participants in order to provide them with an alternative, simpler view of the task and provide additional inspiration.

Code is readily available to be shared in a fork of the current baseline[2].

[2] https://github.com/josepatino/Voice-Privacy-Challenge-2020/

#	Dev. set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	Enroll	Trial	Gen	Test set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$
1	libri_dev	8.665	0.304	42.857	o	o	f	libri_test	7.664	0.183	26.793
2	libri_dev	49.720	0.994	144.285	o	a	f	libri_test	50.000	0.996	149.432
3	libri_dev	35.510	0.887	15.014	a	a	f	libri_test	33.390	0.862	16.024
4	libri_dev	1.242	0.034	14.250	o	o	m	libri_test	1.114	0.041	15.303
5	libri_dev	57.140	0.999	164.327	o	a	m	libri_test	52.780	0.998	163.700
6	libri_dev	32.450	0.857	23.683	a	a	m	libri_test	33.630	0.878	31.056
7	vctk_dev_com	2.616	0.088	0.868	o	o	f	vctk_test_com	2.890	0.091	0.866
8	vctk_dev_com	48.550	0.984	161.083	o	a	f	vctk_test_com	50.000	0.995	156.089
9	vctk_dev_com	27.620	0.769	8.896	a	a	f	vctk_test_com	32.080	0.853	10.590
10	vctk_dev_com	1.425	0.050	1.559	o	o	m	vctk_test_com	1.130	0.036	1.041
11	vctk_dev_com	55.270	0.998	186.612	o	a	m	vctk_test_com	55.370	0.999	184.863
12	vctk_dev_com	30.200	0.799	21.623	a	a	m	vctk_test_com	26.840	0.744	17.896
13	vctk_dev_dif	2.864	0.100	1.134	o	o	f	vctk_test_dif	4.887	0.169	1.495
14	vctk_dev_dif	50.700	0.982	162.258	o	a	f	vctk_test_dif	49.180	0.999	141.784
15	vctk_dev_dif	27.790	0.793	8.711	a	a	f	vctk_test_dif	33.800	0.885	11.707
16	vctk_dev_dif	1.439	0.052	1.158	o	o	m	vctk_test_dif	2.067	0.072	1.817
17	vctk_dev_dif	54.190	1.000	163.566	o	a	m	vctk_test_dif	53.850	1.000	163.102
18	vctk_dev_dif	29.630	0.818	21.604	a	a	m	vctk_test_dif	28.070	0.801	20.364

Table 1: x-vector-based baseline ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		$\mathbf{LM}_s$	$\mathbf{LM}_l$			$\mathbf{LM}_s$	$\mathbf{LM}_l$
1	libri_dev	5.25	3.83	o	libri_test	5.55	4.14
2	libri_dev	9.49	6.96	a	libri_test	10.44	7.78
3	vctk_dev	14.00	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	19.68	15.96	a	vctk_test	19.52	15.74

Table 2: x-vector-based baseline ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	Enroll	Trial	Gen	Test set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$
1	libri_dev	8.807	0.305	42.903	o	o	f	libri_test	7.664	0.184	26.808
2	libri_dev	35.370	0.820	116.889	o	a	f	libri_test	26.090	0.686	115.572
3	libri_dev	23.580	0.620	11.765	a	a	f	libri_test	15.150	0.489	12.542
4	libri_dev	1.242	0.035	14.294	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	17.860	0.526	105.727	o	a	m	libri_test	17.820	0.498	106.444
6	libri_dev	10.560	0.359	11.959	a	a	m	libri_test	8.463	0.263	15.393
7	vctk_dev_com	2.616	0.088	0.869	o	o	f	vctk_test_com	2.890	0.092	0.861
8	vctk_dev_com	34.010	0.879	85.860	o	a	f	vctk_test_com	30.920	0.807	93.959
9	vctk_dev_com	11.630	0.368	43.488	a	a	f	vctk_test_com	14.450	0.464	42.745
10	vctk_dev_com	1.425	0.050	1.555	o	o	m	vctk_test_com	1.130	0.036	1.042
11	vctk_dev_com	23.930	0.669	90.705	o	a	m	vctk_test_com	24.290	0.713	99.336
12	vctk_dev_com	10.540	0.317	24.945	a	a	m	vctk_test_com	11.860	0.347	28.230
13	vctk_dev_dif	2.920	0.101	1.135	o	o	f	vctk_test_dif	4.938	0.169	1.492
14	vctk_dev_dif	35.430	0.907	90.524	o	a	f	vctk_test_dif	29.990	0.795	93.164
15	vctk_dev_dif	15.780	0.504	39.761	a	a	f	vctk_test_dif	16.980	0.546	41.337
16	vctk_dev_dif	1.439	0.052	1.155	o	o	m	vctk_test_dif	2.067	0.072	1.816
17	vctk_dev_dif	28.140	0.740	98.410	o	a	m	vctk_test_dif	28.300	0.720	101.697
18	vctk_dev_dif	11.120	0.384	23.024	a	a	m	vctk_test_dif	12.230	0.397	25.074

Table 3: McAdams coefficient-based baseline ASV results for both development and test partitions (o-original, a-anonymized speech) for an $\alpha = 0.8$.

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM$_s$	**LM**$_l$			**LM**$_s$	**LM**$_l$
1	libri_dev	5.24	3.84	o	libri_test	5.55	4.17
2	libri_dev	12.15	8.74	a	libri_test	11.75	8.90
3	vctk_dev	14.00	10.78	o	vctk_test	16.38	12.80
4	vctk_dev	30.05	25.56	a	vctk_test	33.30	28.15

Table 4: McAdams coefficient-based baseline ASR results for both development and test partitions (o-original, a-anonymized speech) for an $\alpha = 0.8$.

References

[1] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2018, pp. 155–160.

[2] S. McAdams, "Spectral fusion, spectral parsing and the formation of the auditory image," *Ph. D. Thesis, Stanford*, 1984.

[3] C. Dodge and T. A. Jerse, *Computer Music: Synthesis, Composition and Performance*, 2nd ed. Macmillan Library Reference, 1997.

[4] S. Ghorshi, S. Vaseghi, and Q. Yan, "Cross-entropic comparison of formants of british, australian and american english accents," vol. 50, pp. 564–579, 2008.

[5] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.

[6] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," *arXiv preprint arXiv:1911.03934*, 2019.