Design of Voice Privacy System using Linear Prediction

Priyanka Gupta, Gauri P. Prajapati, Shrishti Singh, Madhu R. Kamble, Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India.

{priyanka_gupta, gauri_prajapati, shrishti_singh, madhu_kamble, hemant_patil}@daiict.ac.in

Abstract

The most crucial information exploited by an Automatic Speaker Verification (ASV) system is the speaker's identity (although implicitly). If privacy preservation is exercised for a speaker's identity, numerous attacks can be obliterated simultaneously. The *baseline-2* of the Voice Privacy Challenge 2020 uses the Linear Prediction (LP) model of speech, and McAdam's coefficient for achieving speaker de-identification. It focuses on altering only the pole angles using McAdam's coefficient. However, from speech acoustics and digital resonator design, - $3dB$ bandwidths (and hence, z-domain pole radius) associated with formants capture information about various energy losses (that implicitly carry speaker-specific information) during speech production. To that effect, the authors have brought fine-tuned changes in both pole angle and pole radius, resulting in 18.98% higher value of EER for *Vctk-test-com* dataset, and 5% lower WER for *Libri-test* dataset compared to the baseline. This means privacy-preservation is indeed improved by our approach. Furthermore, gender-based analysis of the obtained results reveals that our approach leads to better speaker anonymization for females as compared to male speakers.

Index Terms: Voice Privacy, speaker de-identification, anonymization, linear prediction, design of digital resonator.

1. Introduction

An Automatic Speaker Verification (ASV) system is used for authentication of claimed identity of a speaker doing analysis on speech utterances with the help of machines [1]. Boldness in performance of an ASV system is desired mainly in terms of functionality of speaker verification and security (i.e., robustness from spoofing). With the evolution of various of spoofing attacks, such as voice conversion [2, 3], replay [4, 5], and mimicry attacks [6], development of several measures against spoofing attacks has also been prioritized in the recent years. Attacker exercises these attacks to impersonate and pretend to be a genuine speaker and hence, the attacker can successfully access sensitive information, where authentication via ASV is required to access it. If the speech data of users is published publicly without applying any privacy preservation measures [7], it is left susceptible to various spoofing attacks and then the attacker might gain illegal access to the information related to speakers' identities to attack the ASV system [8, 9]. Therefore, if the speech data is anonymized such that even if the attacker gains an illegal access to it, it would be impossible to extract any information about users' identities [8]. However, qualities of the speech signal such as *naturalness* and *intelligibility* along with the speakers' identities should remain intact. This can be achieved by effective Voice Privacy (VP) system, also called as speaker de-identification. *anonymization* [10]. For de-identification, two main approaches have been given as *baseline-1* and *baseline-2* in the Voice Privacy Challenge 2020 [11–13]. Baseline-1 is about anonymization using x-vectors and neural waveform models, while baseline-2 is about anonymization using McAdam's coefficient. It should be noted that, cryptography algorithms, can also be useful for achieving voice privacy, however, they are not used due to their difficulty to install, and their complexity increases the overall computational cost of implementation [11, 14–17].

Baseline-2 achieves speaker de-identification by shifting position of the formant frequencies. From speech acoustics, vocal tract walls are prone to bending and show movements under acoustic pressure induced by sound propagation. Thus due to presence of various energy losses (such as wall vibration, thermal and viscosity, lip radiation and glottal boundary), there is some movements in the vocal tract walls which leads to increase in $-3dB$ bandwidths of formants, and these losses contribute implicitly to speakers' identities. To that effect, we varied angles and radius of the complex z-domain poles, which contribute to shift in formant frequencies and widening of formant peaks, respectively.

The rest of this paper is organized as follows. Section 2 gives the details of the notion and logic used related to the proposed speaker anonymization method. Section 3 describe the experimental setup and present the results. Finally Section 5 summarizes and concludes the work done in the paper.

2. LP-Based Resonator Design

2.1. Speech Production Model

According to [18], production of the voiced speech can be modeled as $H(z) = G(z)V(z)R(z)$, where $G(z)$ is the transfer function of the glottal pulse system, $V(z)$ is the transfer function of the vocal tract system, and $R(z)$ is the lip radiation.

The vocal tract system, $V(z)$ is modeled as the cascading of 2^{nd} order resonators (equation (1)), and thus the overall $H(z)$ is given in equation (3) [19]:

$$V(z) = \frac{G}{\prod_{k=1}^{N/2}(1 - 2r_k cos\theta_k z^{-1} + r_k^2 z^{-2})}, \quad (1)$$

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \quad (2)$$

where G is the gain of H(z), r_k and θ_k are the pole radius and pole angle, respectively, of k^{th} pole-pair. One of the first study in speaker recognition by L. G. Kersta states that resonance is defined as *reinforcement* of spectral energy at or around a particular frequency [20]. And if we consider first four formant frequencies then corresponding vocal tract is a cascade of $4\ 2^{nd}$ order digital resonators. The shape of the vocal tract system can be specified with resonant frequencies. The spectrum of the vocal tract system, $H(z)$, consists of peaks located at the

formant frequencies (also called as *formants*) [21]. Mathematically, $H(z)$ is given by equation (3) and equation (4).

$$H(z) = \prod_{i=1}^{4} H_i(z), \tag{3}$$

where each $H_i(z)$ is a 2^{nd} order resonator. Transfer function for 2^{nd} order resonator is given by:

$$H_i(z) = \frac{1}{(1 - p_1 z^{-1})(1 - p_2 z^{-1})}, \tag{4}$$

p_1 and p_2 are the complex conjugate pole-pair of 2^{nd} order resonator transfer function. At resonance, $|H_i(e^{j\omega})|$ will be maximum therefore,

$$\frac{d|H_i(e^{j\omega})|}{d\omega} = 0, \tag{5}$$

We can solve equation (5) to get the resonant frequency, ω_r as,

$$\omega_r = \cos^{-1}[\frac{1 + r^2}{2r} \cos \omega_o]. \tag{6}$$

When $r \to 1$ the resonant frequency, ω_r is approximately equal to the pole angle. The impulse response of a 2^{nd} order digital resonator is given by,

$$h_i[n] = K r^n \sin \omega_o (n + 1) u[n], \tag{7}$$

where r is the pole radius, and K is the overall gain. The $-3dB$ bandwidth of the formant is inversely proportional to the pole radius. Hence, the quality (Q)-factor of the resonator is dependent on the pole radius. When radius of the pole is unity then the corresponding formant will have the edged resonance resulting in nearly zero $-3dB$ bandwidth. On the other hand practically all the resonators are considered to be stable (i.e., $r < 1$ in Z-plane). So we will not achieve a sharp resonance like an impulse, rather we will have some finite $-3dB$ bandwidth around the formants. This relation between pole radius and $-3dB$ bandwidth can be derived from mapping of stable Laplace domain pole to stable Z-domain pole as given below,

$$r = e^{-\pi B T}, \tag{8}$$

where B is the $-3dB$ bandwidth (in Hz), and T is the sampling interval (in seconds). For anonymization if radius is decreased, the bandwidth will increase compared to original bandwidth(without anonymization). The gain which was concentrated around their central (resonant) frequency before anonymization will now spread around the central frequency(i.e., will tend towards resonance breakdown) instead of the sharp peaks. Hence, the formants will not be easily discernible after anonymization using this method, thus speaker identification will be more effortful.

We can consider the speech model as an all-pole model. One such all pole model is LP model, which predicts the current sample of speech, $x[n]$ using the past p samples of the speech [22]. The LP model is given by

$$\tilde{x}[n] = a_1 x[n-1] + a_2 x[n-2] + ... + a_p x[n-p], \tag{9}$$

where $a_1, a_2, ..., a_p$ are called as LP coefficients. Thus a speech sample can be approximated as a linear combination of the past speech samples [23]. The system function for p^{th} order predictor is given as

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k}. \tag{10}$$

The prediction error or LP residual sequence is given by equation (11), and associated prediction error filter is defined in equation (12),

$$e[n] = x[n] - \tilde{x}[n] = x[n] - \sum_{k=1}^{p} \alpha_k x[n-k], \tag{11}$$

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = 1 - P(z). \tag{12}$$

We can recover the input sequence $s[n]$ by passing $Au_g[n]$ through $\frac{1}{A(z)}$, where $Au_g[n]$ is vocal tract input with gain A. This can be done when $\alpha_k \approx a_k$, so the prediction error filter, $A(z)$ is sometimes called the *inverse* filter. This inverse filtering at least suppress the formants of speech signal, and the remaining signal is called LP residual. It is used as excitation source signal that is used to excite a filter (representing formants) for speech generation (after anonymization in our case).

In this context using system theory, we can fine tune the residual or formants to change the resulting speech signal characteristics.

2.2. Formants and speaker de-identification

In an all-pole model of the vocal tract, a complex pole-pair at $r_0 e^{jw_0}$ and $r_0 e^{-jw_0}$ corresponds to a vocal tract formant. A male speaker tends to have lower formants than a female speaker [18]. Because an increase in the length of the vocal tract system corresponds to decrease in formant frequencies [24].

In a LP model, the LP coefficients $a_i's$ are responsible for pole locations. The formant frequency and bandwidth are governed by the pole locations [25]. Mathematically, formant frequency is given by $\frac{F_s \theta}{2\pi}$, where θ is the angle of the pole in radians, given F_s is the sampling frequency in Hz. The formant bandwidth is given by $\frac{F_s}{\pi}(-log(r))$, where r is the radius of the pole [18]. As proposed by M.R. Schroeder, the ability of human beings to emit and perceive sounds is more dependent on spectral peaks than spectral valleys [26]. The formants of speech signal are obtained from these spectral peaks. So by modifying the formant frequencies, we can achieve different modifications of formant spectrum and thus, leading to speaker de-identification along with naturalness and intelligibility.

To anonymize a speaker, a controlled shift in pole angle and radius can be done, such that intelligibility is not lost and the speaker identity is mapped to another voice. Since every complex pole conjugate pole-pair corresponds to one formant frequency [27], only one of the poles in the pair is considered for de-identification [28]. In the given baseline, pole angles are shifted by a McAdam's coefficient with a value of 0.8 initially [11, 12, 29].

3. Performance Evaluation

The baseline- 2 system along with the improved experimental results and analysis is included in this section. The objective performance is measured in terms of Equal Error Rate (EER), and Word Error Rate (WER) to evaluate anonymization and speech intelligibility, respectively [30]. An ASV system which relies on x-vector speaker embeddings and, Probabilistic Linear Discriminant Analysis (PLDA)is used to compute the EER scores [31].

#	\Dev. set	\EER, %	\C$_{llr}^{min}$	\C$_{llr}$	\Enroll	\Trial	\Gen	\Test set	\EER, %	\C$_{llr}^{min}$	\C$_{llr}$
1	libri_dev	8.665	0.304	42.891	o	o	f	libri_test	7.664	0.184	26.812
2	libri_dev	32.950	0.807	115.483	o	a	f	libri_test	25.730	0.691	119.399
3	libri_dev	24.290	0.652	15.379	a	a	f	libri_test	15.880	0.511	15.183
4	libri_dev	1.242	0.035	14.246	o	o	m	libri_test	1.114	0.041	15.340
5	libri_dev	19.570	0.579	112.062	o	a	m	libri_test	17.370	0.493	110.935
6	libri_dev	11.180	0.368	15.765	a	a	m	libri_test	8.909	0.275	21.850
7	vctk_dev_com	2.616	0.089	0.872	o	o	f	vctk_test_com	2.890	0.092	0.867
8	vctk_dev_com	33.140	0.864	100.451	o	a	f	vctk_test_com	29.770	0.797	107.716
9	vctk_dev_com	10.760	0.349	43.631	a	a	f	vctk_test_com	17.050	0.502	47.549
10	vctk_dev_com	1.425	0.049	1.560	o	o	m	vctk_test_com	1.130	0.036	1.029
11	vctk_dev_com	24.500	0.666	97.415	o	a	m	vctk_test_com	27.680	0.723	107.513
12	vctk_dev_com	12.540	0.393	34.154	a	a	m	vctk_test_com	12.990	0.389	36.018
13	vctk_dev_dif	2.864	0.101	1.150	o	o	f	vctk_test_dif	4.990	0.170	1.499
14	vctk_dev_dif	33.860	0.897	102.523	o	a	f	vctk_test_dif	29.420	0.798	103.744
15	vctk_dev_dif	13.870	0.450	44.237	a	a	f	vctk_test_dif	18.470	0.580	49.801
16	vctk_dev_dif	1.390	0.052	1.162	o	o	m	vctk_test_dif	2.067	0.072	1.826
17	vctk_dev_dif	26.450	0.732	101.214	o	a	m	vctk_test_dif	27.150	0.729	111.908
18	vctk_dev_dif	13.350	0.433	36.581	o	a	m	vctk_test_dif	12.630	0.425	35.185

Table 1: *EER results of the approach: Shifting radius to 0.975 to its value and McAdam's coefficient=0.8 for development and test data (o – original, a – anonymized speech data).*

#	Dev. set	WER, %		Data	Test set	WER, %	
		\LM$_s$	\LM$_l$			\LM$_s$	\LM$_l$
1	libri_dev	11.76	8.60	a	libri_test	11.37	8.43
2	vctk_dev	29.09	24.58	o	vctk_test	32.26	27.01

Table 2: *WER results of the approach: Shifting radius to* 0.975 *to its value and McAdam's coefficient*=0.8 *for development and test data (o-original, a-anonymized speech) for two trigram LMs:* LM_s *- small, and* LM_l *- large LM.*

3.1. The Baseline System and Proposed Improvement

In the baseline system LP analysis of speech is performed, which results in frame-by-frame (with 50% overlap) generation of LP coefficients and LP residual. Then using these LP coefficients poles are obtained. Anonymization is achieved by considering only one pole out of the complex pole-pair and shifting the poles' angle ϕ by a constant known as the McAdam's coefficient, α [29]. The new pole angle is ϕ^{α}. The residuals are unchanged to retain the naturalness and intelligibility, and are used in the reconstruction of the anonymized speech signal. Depending on the values of ϕ and α, the pole is shifted either in positive or the negative direction. The effect of pole shifting in z-plane is detailed in the evaluation plan for two values of α, greater and less than 1.

The radius of the poles is also reduced along with the modification in pole angles, so that intelligibility is not lost, while achieving speaker de-identification. However, The baseline involves shifting of only pole angles for anonymization. Two approaches of speaker anonymization is described in this section. First approach includes shifting of the pole locations by changing only the radius of the pole while keeping the pole angle untouched. In the second approach changing both the pole radius, and the pole angle is considered to shift the pole locations. In the first set of experiments, the radius of the each pole is decreased by arbitrarily chosen amount of 15%, 5%, and 2.5% of pole radius that is measured from the original utterances. In the second set of experiments, the radius is changed by the same amount as in the first set of experiments, and also the angle of the poles is shifted by McAdam's coefficient with values of 0.8 and 0.9. Tables 1 and 2 shows the results of the proposed approach. The detailed discussion on results are included in the next Section.

The amount of anonymization is evaluated using a x-vector speaker embedding-based ASV system, which gives two objective metrics : Equal-Error-Rate (EER) and Caliberation Cost (Cllr). For objective evaluation an assumption is made by the attacker's model that the attackers have access to a single anonymized trial utterance and several enrollment utterances. Also another assumption is made that the corresponding pseudo-speakers of trial and enrollment utterances are different [11, 12]. Therefore, a higher value of EER indicates better anonymization. A TDNN-F acoustic model and a tri-gram Language Model (LM) are used by an ASR system to measure the intelligibility. It gives the intelligibility score in terms of WER for small and large LMs. Lower value of WER indicates better intelligibility. Both of these systems are trained on the *LibriSpeech-train-clean-360* dataset using Kaldi speech recognition toolkit [32–34].

3.2. Experimental Results and Analysis

This Section presents the experimental results w.r.t. the baseline-2 system. In the experiments, the radius and/or phase of the poles of the speech signal derived using LP source-filter model are varied. It should be noted that a high value of EER, and a low value of WER is desired for speaker de-identification.

3.2.1. Pole Placement using only Pole Radius

Radius (r) was varied for three cases- $0.85r$, $0.95r$, and $0.975r$. No changes were introduced in the pole angles. It was observed that when the radius was changed to 0.85 times the original radius, slightly better values of EERs were obtained (increased by 3%) than the original baseline. However, WER values were considerably degraded. When the radius was changed to 0.95 times the actual radius, undesirable values of the EERs were obtained (decreased by 7 to 10) for most of the cases when compared to the original baseline system. However, better values of WER were obtained and were less by 15 for $vctk\_dev$ and $vctk\_test$ datasets.

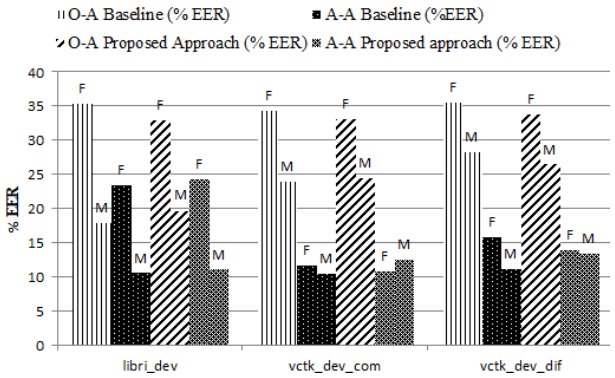

Figure 1: *%EER for development data (o–original, a–anonymized) for radius*$=0.975$ *to its value, and* $\alpha = 0.8$*, F-Female, M-male.*

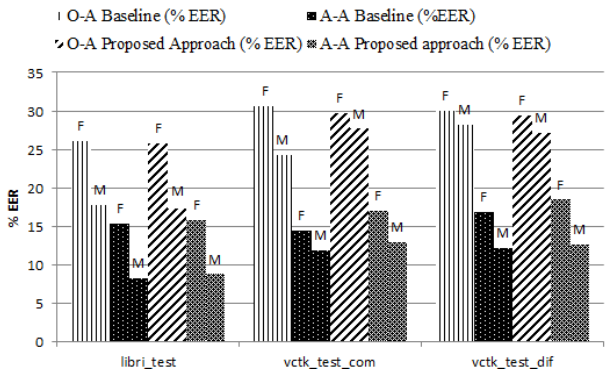

Figure 2: *%EER for test data (o–original, a–anonymized) for radius*$=0.975$ *to its value, and* $\alpha = 0.8$*, F-Female, M-male.*

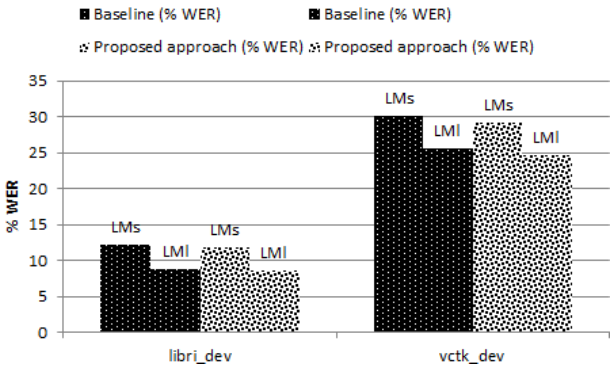

Figure 3: *% WER for development data (o–original, a–anonymized) for radius*$=0.975$ *to its value, and* $\alpha = 0.8$*, for two trigram LMs : LM_s-small, and LM_l-large LM.*

3.2.2. Pole Placement using Pole Radius and Angle

When only the radius was changed to shift the pole locations, it does not give appreciable results. Hence, the pole locations were shifted by decreasing the pole radius by 2.5% along with transformation of the pole angle from ϕ to ϕ^{α}, where $\alpha = 0.8$. In this case, improved results were obtained, both in terms of EER and WER, as shown in Figs. 1, 2, 3 and 4. Furthermore, if the pole radius was decreased by more than 2.5%, the performance of the WER degraded drastically. Hence, with a new

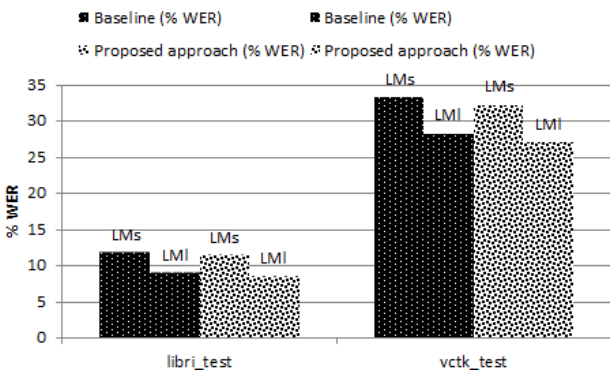

Figure 4: *% WER for test data (o–original, a–anonymized) for radius*$=0.975$ *to its value, and* $\alpha = 0.8$*, for two trigram LMs : LM_s-small, and LM_l-large LM.*

radius which is 0.975% of the original radius and McAdam's coefficient as 0.8, we obtained relatively best results in terms of EER and WER both.

The increase in EER we obtained by reducing the pole radius is justified by the relation of formant bandwidth with the pole radius as described in the Section 2.2. The formant bandwidth will increase when the radius is decreased since the pole radius and formant bandwidth shares a logarithmic relation with each other and also the value of r is less than 1. The quality factor (Q) of the speech signal will degrade due to the increase in formant bandwidth and thus, it will get difficult for the ASV system to identify the speaker. Hence, the ASV system will give high EER value indicating the efficient transformation of the speakers' identity in frequency-domain. Moreover, speaker-specific information is obtained from formant frequencies, therefore, pole angles were also shifted using McAdam's coefficient to improve the performance of the system. On shifting the pole angles along with the change in pole radius, improved results were obtained.

4. Summary and Conclusions

In the proposed work, authors have used LP model and McAdam's coefficient to achieve effective speaker anonymization. The baseline-2 have used method of shifting only the pole angles for achieving anonymization [11]. However, the pole radius is also significant in speaker de-identification. The pole radius is related to different energy losses during production of the natural speech and is related to -3dB bandwidth. Thus, the authors have varied the pole radius along with the shift in phase of the poles to get better anonymization. Along with the proposed methods, other signal processing techniques can be used for better anonymization, with the combination of neural network-based approaches.

5. References

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[3] Y. Stylianou, "Voice transformation: A survey," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April 2009, pp. 3585–3588.

[4] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 10-12 September 2014, pp. 1–6.

[5] A. Paul, R. K. Das, R. Sinha, and S. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *2016 International Conference on Signal Processing and Communications (SPCOM)*, IISc, Bengaluru, India, 12-15 June 2016, pp. 1–5.

[6] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *International Symposium on Intelligent Multimedia, Video, and Speech Processing*, Hong Kong, 20-22 October 2004, pp. 145–148.

[7] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," *arXiv preprint arXiv:1907.03458*, 2019, {Last Accessed: 2020-05-07}.

[8] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X.-Y. Li, "Towards privacy-preserving speech data publishing," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, Honolulu, HI, USA, 16-19 April 2018, pp. 1079–1087.

[9] 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, {Last Accessed: 2020-05-07}.

[10] 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*, Shenzhen , China, 4-7 November, 2018, pp. 82–94.

[11] "The voice privacy 2020 challenge evaluation plan," https://www.voiceprivacychallenge.org, {Last Accessed: 2020-05-02}.

[12] J. Patino, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdam's coefficient," Eurecom, Tech. Rep., February 2020. [Online]. Available: http://www.eurecom.fr/publication/6190{LastAccessed:2020-05-07}

[13] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. L. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the voiceprivacy initiative," in *INTERSPEECH*, Shanghai , China, 24-28 October, 2020, {Last Accessed: 2020-05-07}.

[14] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, Special issue, 2019.

[15] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 62–74, 2013.

[16] P. Smaragdis and M. V. S. Shashanka, "A framework for secure speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, New Orleans, USA, 5-9 March, 2007, pp. IV–969–IV–972.

[17] S.-X. Zhang, Y. Gong, and D. Yu, "Encrypted speech recognition using deep polynomial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 12-17 May, 2019, pp. 5691–5695.

[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. 2^{nd} Edition, Pearson Education India, 2004.

[19] H. A. Patil and S. Viswanath, "Energy separation algorithm based spectrum estimation for very short duration of speech," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruña , Spain, 2-6 September, 2019, pp. 1–5.

[20] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[21] G. Fant, *Acoustic Theory of Speech Production*. 2^{nd} Edition, Walter de Gruyter, 1970.

[22] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America (JASA)*, vol. 50, no. 2B, pp. 637–655, 1971.

[23] J. D. Markel and A. J. Gray, *Linear Prediction of Speech*. Springer Science & Business Media, 2013, vol. 12.

[24] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, vol. 1. Atlanta, Georgia, USA: IEEE, 7-10 May, 1996, pp. 346–348.

[25] H. Mizuno and M. Abe, "A formant frequency modification algorithm dealing with the pole interaction," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 79, no. 1, pp. 46–55, 1996.

[26] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proceedings of the IEEE*, vol. 54, no. 5, pp. 720–734, May 1966.

[27] J. Slifka and T. R. Anderson, "Speaker modification with lpc pole analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Detroit, Michigan, USA, 8-11 May, 1995, pp. 644–647.

[28] D. Rentzos, S. Vaseghi, Q. Yan, and C.-H. Ho, "Voice conversion through transformation of spectral and intonation features," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Montreal, Quebec, Canada: IEEE, 17-24 May, 2004, pp. I–21.

[29] S. McAdams, "Spectral fusion, spectral parsing and the formation of auditory image," *Ph.D. Thesis, Department of Hearing and Speech, Stanford University, California, USA*, May, 1984.

[30] W. M. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 4, pp. 205–212, 2002.

[31] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 15-20 April, 2018, pp. 5329–5333.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding(ASRU)*, no. CONF, Big Island, Hawaii, USA, 11-15 December, 2011.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Queensland, Australia, 19-24 April, 2015, pp. 5206–5210.

[34] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019, {Last Accessed: 2020-05-07}. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/3443