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.
E-mail: {priyanka_gupta, gauri_prajapati, shrishti_singh, madhu_kamble, hemant_patil}@daiict.ac.in

***Abstract*—Speaker's identity is the most crucial information exploited (implicitly) by an Automatic Speaker Verification (ASV) system. Numerous attacks can be obliterated simultaneously if privacy preservation is exercised for a speaker's identity. The baseline of the Voice Privacy Challenge 2020 by INTERSPEECH uses the Linear Prediction (LP) model of speech, and McAdam's coefficient for achieving speaker de-identification. The baseline approach focuses on altering only the pole angles using McAdam's coefficient. However, from speech acoustics and digital resonator design, the radius of the poles is associated with various energy losses. The energy losses 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, we have exploited the relatively poor spectral resolution of female speakers to our advantage for achieving effective anonymization. To that effect, gender-based analysis of the obtained results reveals that our approach leads to better speaker anonymization for females as compared to the male speakers.**

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

I. Introduction

An Automatic Speaker Verification (ASV) system is used to verify claimed identity of a speaker with the help of machines [1]. The robustness of an ASV system can be viewed with two perspectives- robustness in terms of functionality of speaker verification, and robustness in terms of security (i.e., robustness from spoofing). With the advent of various of spoofing attacks, such as speech synthesis, voice conversion [2], [3], replay [4], [5], and mimicry attacks [6], significant attention to develop countermeasures against spoofing attacks has been given in the recent years. By these attacks, an attacker can impersonate and pretend to be a genuine user (speaker). Hence, the attacker can successfully access sensitive information, wherever authentication via ASV is required to access it. However, it should be noted that we are far away from designing a versatile Spoofed Speech Detection (SSD) system, which would alleviate all the types of attacks. This leads to a serious vulnerability if speech data of users is published publicly without any privacy preservation [7]. The attacker (also called an *adversary*) can have illegal access to this data, and might further use information related to speakers' identities to attack the ASV system [8], [9], as shown in Fig. 1. Therefore, it would be impossible to infer any information about users' identities with anonymized speech data, even if the attacker gains illegal access to it, [8]. However, aspects of the speech signal such as *naturalness* and *intelligibility* should remain intact. This can be achieved by designing an effective Voice Privacy (VP) system [10].

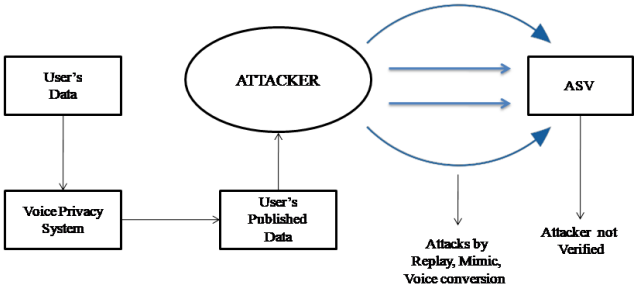

Fig. 1. Game Between Attacker and Voice Privacy System.

For de-identification, we have considered the signal processing based baseline system provided by the voice privacy challenge organized by INTERSPEECH 2020 [11], [12], [13]. In this paper, the authors have achieved slightly better results than the baseline system. Furthermore, gender-based analysis of the obtained results is shown in this paper. It is important to note that, to achieve voice privacy, cryptography algorithms, such as homomorphic encryption, and secure-multiparty computation can also be useful. However, they are not used due to their difficulty in deployment, and their complexity increases the overall computational cost of implementation [11], [14], [15], [16], [17].

The baseline system achieves speaker de-identification by shifting position of the formant frequencies. This is done by varying the pole angles with the help of McAdam's coefficient. However, from speech acoustics (in particular, vocal tract walls are pliant and can have movements under acoustic pressure), there are various energy losses (such as wall vibration, thermal and viscosity, lip radiation, and glottal boundary) that are mapped to increase in $-3dB$ bandwidths of formants. These losses contribute implicitly to speakers' identities. To that effect, we varied the radius of the complex z-domain poles also, instead of just varying the pole angle. This leads to the widening of the peaks in the Linear Prediction (LP) spectrum, and also shifting the position of those peaks. We have also included spectrogram analysis of male and female anonymized speech signals, which shows better anonymization of female as

compared to the male speakers. In particular spectral resolution problem associated with female speech is exploited to achieve better anonymization for female speakers.

The rest of the paper is as follows: Section-II describes the all-pole model of speech production and hence, LP based de-identification. Section-III contains the experimental setup followed by the obtained results of Equal Error Rate(EER) and Word Error Rate(WER). Furthermore, gender-based analysis is done on the results. The paper concludes with Section IV, describing the achievements of the experiment and some possible future directions.

II. LP-Based Resonator Design

A. Speech Production Model

For voiced speech, the overall transfer function of the speech production model is $H(z) = G(z)V(z)R(z)$, 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 [18].

$V(z)$ is the cascading of 2^{nd} order resonators given by (1), and the overall $H(z)$ is given by (3):

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

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

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 in z-domain [19]. Before moving into LP model analysis, let us discuss about modelling of vocal tract system using 2^{nd} order digital resonators. As per original investigations by L. G. Kersta, who reported one of the first studies in speaker recognition, resonance is defined as *reinforcement* of spectral energy at or around a particular frequency [20]. Vocal tract system is a cascade of four 2^{nd} order resonators with first four resonant frequencies. 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 (3) and (4).

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

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

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

p_1 and p_2 are the complex conjugate pole-pair of 2^{nd} order resonator transfer function.

For resonance, $|H_i(e^{j\omega})| \rightarrow max$, therefore,

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

solving the (5) will give resonant frequency, ω_r,

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

Resonant frequency, ω_r is approximately equal to the pole angle, ω_o as $r \rightarrow 1$. Impulse response of 2^{nd} order digital resonator is given by taking inverse z-transform of Eq. (4), i.e.,

$$h_i[n] = K r_i^n \sin\omega_{o_i}(n+1)u[n], \quad (7)$$

where ω_{oi} and r_i is angle and radius of i^{th} pole-pair, and K is the overall gain. The quality (Q)-factor is dependent on the pole radius. This means that $-3dB$ bandwidth of the formant is *inversely* proportional to the pole radius. We get the sharpest resonance with highest quality (i.e., $-3dB$ bandwidth = 0) when radius = 1. However, in practical cases, a stable resonator (i.e., $r < 1$ in Z-plane) is considered, because in real physical system, such as speech production, series RLC circuit, mass-spring-damper system, there is mechanism of energy dissipation via various energy losses. Hence, we cannot achieve sharpest impulse-like resonances. Thus, we get some effect of damping factor (r^n) in the form of $-3dB$ bandwidth. Relationship between $-3dB$ bandwidth and pole radius r is given by invoking impulse invariant transformation (IIT) to map stable Laplace-domain pole to stable Z-domain pole, i.e.,

$$r = e^{-\pi BT}, \quad (8)$$

where B is the $-3dB$ bandwidth (in Hz), and T is the sampling interval (in seconds). Therefore, considering our case, when the radius is more before anonymization, formants have sharpest peaks. Thus, the gain is concentrated around their central (resonant) frequency. For anonymization if pole radius is decreased, the bandwidth will increase and the sharpest peaks observed earlier will now become more flat, and therefore, the gain around the central frequency will spread (i.e., will tend towards resonance breakdown). Hence, the formant peaks will not be as distinct as in the original speech signal, which makes speaker identification more difficult.

Now, resuming to LP model analysis, we can say that speech modelling can be done by considering an all-pole model. Moreover, LP model which predicts the current sample of speech, $s[n]$ using the past p samples of the speech is based on an all-pole model [22]. The LP model is given by

$$\tilde{s}[n] = a_1 s[n-1] + a_2 s[n-2] + ... + a_p s[n-p], \quad (9)$$

where $a_1, a_2, ..., a_p$ are called as LP coefficients. This means that 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

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

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

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k], \quad (11)$$

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

When $\alpha_k \approx a_k$, the prediction error filter, $A(z)$ is sometimes called the *inverse* filter because 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 inverse filtering removes (or at least suppresses) the formants of the speech signal, and the remaining error signal is called LP residual. It is used as excitation source signal to excite a vocal tract filter (representing formants) for speech generation. In this context, using system theory we can fine tune the residual or formants (or corresponding source-filter *coupling*) to change the resulting speech signal characteristics.

B. Formants and speaker de-identification

In an all-pole model, a complex pole-pair at $r_0 e^{jw_0}$ and $r_0 e^{-jw_0}$ corresponds to one peak of the LP spectrum (i.e., one vocal tract formant). It is seen that the formant frequencies are lower if the length of the vocal tract system is longer [24]. Therefore, a male speaker tends to have lower formants than a female [18].

In a LP model, the LP coefficients $a_i's$ are responsible for pole locations. Pole locations govern the formant frequency and bandwidth [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 per M.R. Schroeder, human beings emit and perceive sounds by emitting spectral peaks more dominantly than the spectral valleys [26]. These spectral peaks are related to formant frequencies of speech signal.

In our experiments, we have considered all the frequencies corresponding to the poles generated by LP coefficients. This includes all the formant frequencies also. Shifting of all the poles locations by changing their respective *pole angles* is done using McAdam's coefficient. Shifting of all the poles locations by changing their respective *pole radius* is done by fine-tuning w.r.t. the EER and WER obtained so that intelligibility is not lost. Since every complex 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].

III. Experimental Setup and Performance Evaluation

This Section describes the baseline system along with the experimental results and analysis. The objective performance is measured in terms of EER, and WER to evaluate anonymization and speech intelligibility, respectively [30]. The EER is computed by ASV system, which relies on x-vector speaker embeddings, and Probabilistic Linear Discriminant Analysis (PLDA) [31].

A. Corpora Used

For development data, subsets from two corpora, namely, LibriSpeech-dev-clean and VCTK are provided [32], [33]. These subsets are further divided into trial and enrollment subsets. There are 40 speakers in LibriSpeech-dev-clean. There are 29 speakers in enrollment utterances and 40 speakers in trial utterances. From these 40 speakers of trial subset, 29 speakers are also included in enrollment subset.

In VCTK-dev dataset, there are total 30 speakers which are the same for both trial and enrollment utterances. Furthermore, for trial utterances, there are two parts, denoted as *common part* and *different part.* Both the parts are disjoint in terms of utterances, however, they have the same set of speakers. The *common part* of the trials has utterances from $\#1$ to $\#24$ in the VCTK corpus, which are the same for all the speakers. The *common part* of the trials is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. $\#25$ onward utterances are distinct and hence, are included in the *different part* of the VCTK-dev dataset.

For evaluation data, the structure is the same as that of development set, except for the number of utterances.

B. The Baseline System and Proposed Improvement

The baseline system uses LP analysis of speech, which results in frame-by-frame (with 50% overlap) generation of LP coefficients and LP residual. The LP coefficients are converted to poles. 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. To retain the naturalness and intelligibility, the residuals are unchanged. They are then used in the reconstruction of the anonymized speech signal. Depending on the values of ϕ and α, the pole is shifted either in clockwise or anti clockwise direction. The effect of pole shifting by varying only the pole angle in z-plane is shown in Fig.2.

Furthermore, to evaluate speaker variability, x-vector speaker embedding-based ASV system is used. To evaluate the intelligibility, an ASR system baed on a TDNN-F acoustic model and a tri-gram Language Model (LM) is used. 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 [34], [32], [33]. Fig.3 shows the schematic representation of the proposed approach for speaker anonymization. The authors have modified the pole radius along with pole angles to achieve better anonymization. Furthermore, the residual is kept intact for retaining naturalness and intelligibility in the anonymized speech signal.

Our experiments can be categorized in two parts. First, shifting of the pole locations is done by changing only the radius of the pole while keeping the pole angle intact. Second, shifting of the pole locations by changing both the pole radius, and the pole angle. In the first set of experiments, the radius of the each pole is decreased by arbitrarily chosen and fine tuned

TABLE I
ASV RESULTS OF THE APPROACH: SHIFTING RADIUS TO 0.975 TO ITS VALUE AND MCADAM'S COEFFICIENT=0.8 IN TERMS OF EER% FOR DEVELOPMENT AND TEST DATA (O – ORIGINAL, A – ANONYMIZED SPEECH DATA, F - FEMALE SPEAKER, M - MALE SPEAKER).

#	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	a	a	m	vctk_test_dif	12.630	0.425	35.185

TABLE II
ASR RESULTS OF THE APPROACH: SHIFTING RADIUS TO 0.975 TO ITS VALUE AND MCADAM'S COEFFICIENT=0.8 IN TERMS OF WER% FOR DEVELOPMENT AND TEST DATA (O-ORIGINAL, A-ANONYMIZED SPEECH, F - FEMALE SPEAKER, M - MALE SPEAKER) FOR TWO TRIGRAM LMS: LM_s - SMALL, AND LM_l - LARGE LM.

#	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	11.76	8.60	a	libri_test	11.37	8.43
3	vctk_dev	14.00	10.78	o	vctk_test	16.38	12.80
4	vctk_dev	29.09	24.58	a	vctk_test	32.26	27.01

amounts of 15%, 5%, and 2.5% of pole radius that is measured from the original utterances (i.e., before anonymization). In the second set of experiments, the radius is changed by the amount of 15%, 5%, and 2.5%, and the angle of the poles is shifted by McAdam's coefficient with values of 0.8 and 0.9. The results are given in the tables I and II. Their detailed analysis is given in the next sub-Section.

For objective evaluation, the attacker's model assumes that the attackers have access to a single anonymized trial utterance and several enrollment utterances. It is also assumed that the corresponding pseudo-speakers of trial and enrollment utterances are different [11], [12]. Therefore, a higher value of EER indicates better anonymization, and a lower value of WER is preferred to ensure intelligibility.

C. Experimental Results and Analysis

This Section presents the experimental results w.r.t. the baseline system given in the evaluation plan of voice Privacy challenge [11]. The experiments include varying the radius and/or phase of the poles of the speech signal derived using LP source-filter model.

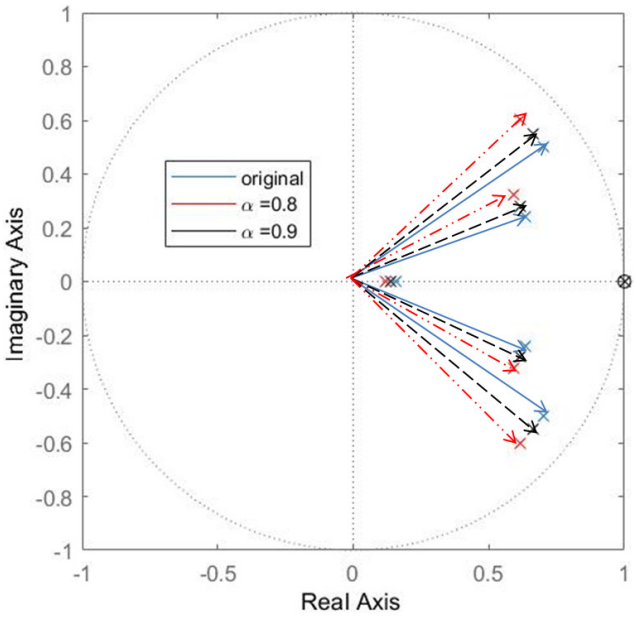

Fig. 2. Example of pole-zero plot for original, and two cases of pole placement.

1) Pole Placement using only Pole Radius: Keeping the pole angles unchanged, radius (r) was varied for three cases- $0.85r$, $0.95r$, and $0.975r$. It was observed that when the new radius was 0.85 times the original radius, the EERs obtained were slightly better (increased by 3%) than the baseline. However, the performance in terms of WER was degraded significantly. For the case when the new radius was 0.95 times the actual radius, the EERs obtained were undesirable (decreased by 7 to 10) for most of the cases when compared

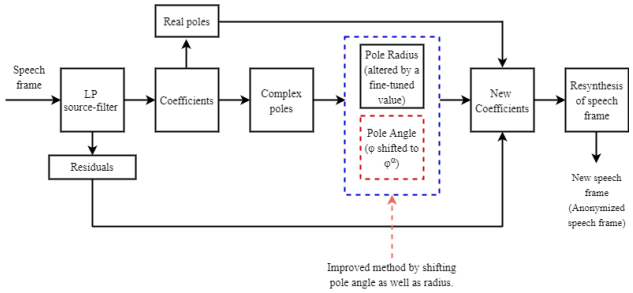

Fig. 3. Proposed approach for speaker anonymization : α is McAdam's coefficient, and ϕ is the pole angle in radian.

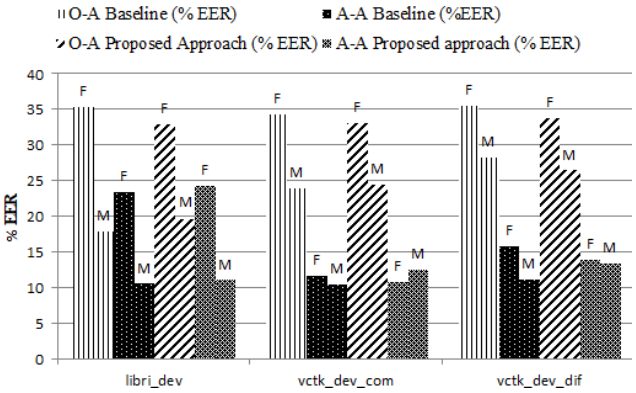

Fig. 4. %EER for development data (o–original, a–anonymized) for radius= 0.975 to its value, and $\alpha = 0.8$, F-Female, M-male.

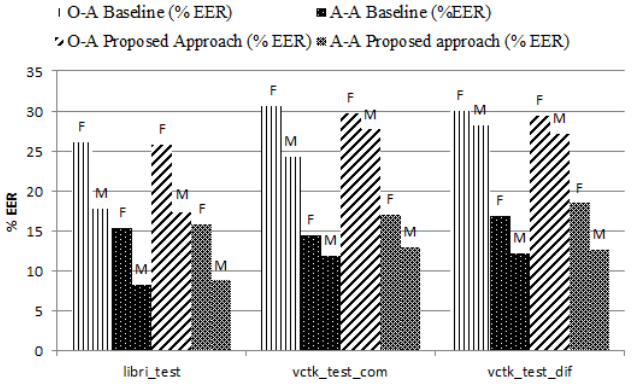

Fig. 5. %EER for test data (o–original, a–anonymized) for radius= 0.975 to its value, and $\alpha = 0.8$, F-Female, M-male.

to the baseline system. However, the WER values obtained were better and were less by 15 for $vctk\_dev$ and $vctk\_test$ datasets, when compared with the provided baseline.

2) Pole Placement using Pole Radius and Angle: We observed that shifting the pole locations by changing only the radius does not give appreciable results. Hence, the pole locations were changed by decreasing the pole radius by

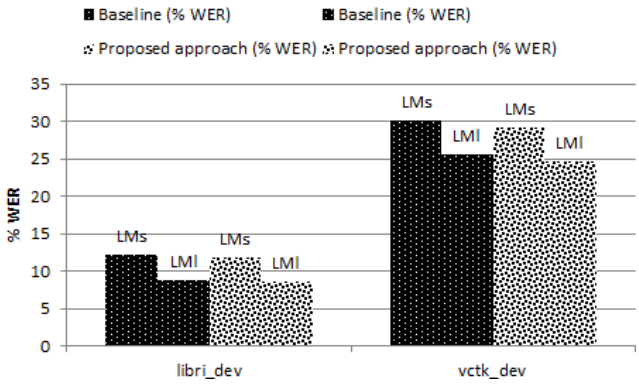

Fig. 6. % WER for development data (o–original, a–anonymized) for radius= 0.975 to its value, and $\alpha = 0.8$, for two trigram LMs : LM_s-small, and LM_l-large LM.

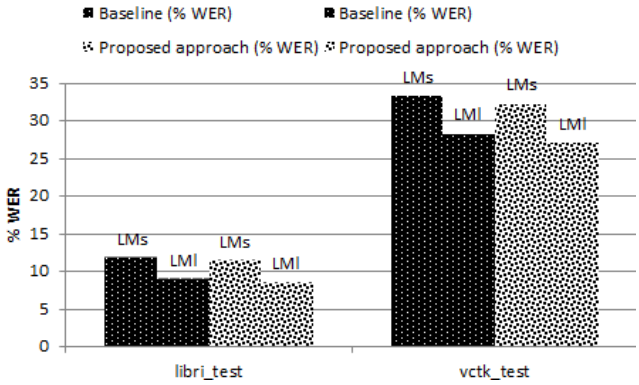

Fig. 7. % WER for test data (o–original, a–anonymized) for radius= 0.975 to its value, and $\alpha = 0.8$, for two trigram LMs : LM_s-small, and LM_l-large LM.

2.5% and transforming 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 (Tables I and II respectively). The graphical comparison of obtained results with baseline system is also provided in Figs. 4, 5, 6 and 7. It was also observed that if the pole radius was decreased by more than 2.5%, the WER performance degraded drastically. Hence, with a new radius of 0.975% of the original radius and McAdam's coefficient as 0.8, we get relatively best results in terms of both EER and WER. This increase in EER with the reduction in pole radius is justified by the relation of formant bandwidth with the radius of the pole as described in the Section II-B. Due to the logarithmic relation between pole radius and formant bandwidth, and the value of r is less than 1, the formant bandwidth will increase when the radius is decreased. This increase in formant bandwidth will degrade the quality factor (Q) of the resonance in speech spectrum. Hence, the ASV system will not be able to identify the speaker easily, thereby, giving a high value of EER, which further indicates the efficient transformation of the speaker's identity in the frequency-domain. For improving the anonymization further, pole angles are shifted using McAdam's coefficient. Since

formant frequencies contain speaker-specific information, shifting the pole angles gives improved results.

3) Gender-Based Analysis for Voice Privacy: Additional observation which can be inferred from the results is by comparing the EER scores of the female and male speakers. We observe that EER$\%$ for the female speaker is higher than the male speaker in almost every case of development and evaluation set, given that the anonymization technique on the utterances is the same for both the female and male speakers as shown in Fig. 4 and Fig. 5. This result suggests that the privacy preservation of female speakers is more efficient than the male speakers. This is also be supported by the fact that the spectral resolution of female speech is relatively lesser than the spectral resolution of male speech [1] (since the vocal tract speech spectrum gets uniformly sampled by high pitch source harmonics for females). Also, a slight variation in the glottal waveform can result in considerable amount of change in the voice characteristics. Therefore, due to the larger pitch duration in male speakers as shown in Fig. 8, they get sufficient time to perform activity near glottal closure which is not the scenario for female speakers due to the lower pitch duration (almost half the pitch duration of the male speakers). This large variation in the glottal waveform changes the speaker's characteristics drastically. The speaker recognition techniques uses information based on the 1 to 2 ms glottal closure period. Hence, tracking this large variation in 1 to 2 ms of glottal closure period becomes difficult for the ASV systems, which can lead to the higher $EER\%$ values. In particular, the poor spectral resolution of female speech is advantageous to achieve the voice anonymization by having lesser stricter intermittent with the other female speakers' speech [35] (similar to significance of over-smoothing of Gaussian Mixture Model (GMM) parameters for iterative combination of a Nearest Neighbour search step and a Conversion step Alignment (INCA)-based voice conversion [36]). In addition, we also observed and

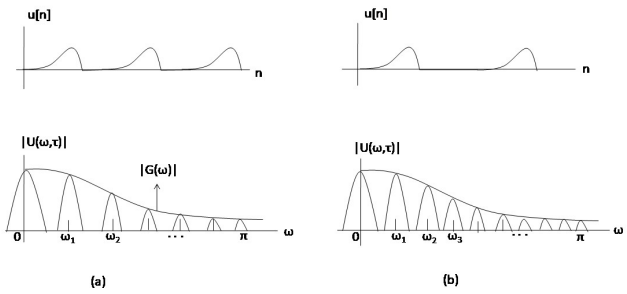

Fig. 8. Illustration of periodic glottal flow and its spectrum. (a)-higher pitch (Female Speaker), (b)-lower pitch (Male speaker).

compared the spectral energy densities using spectrogram for original and anonymized speech signals. Fig. 9 shows the comparison of the spectrogram for a female and a male speaker from the test dataset of *Libri-Dev*. These original speech signals are anonymized by the same method for both males and females which is by decreasing radius by 2.5% and changing

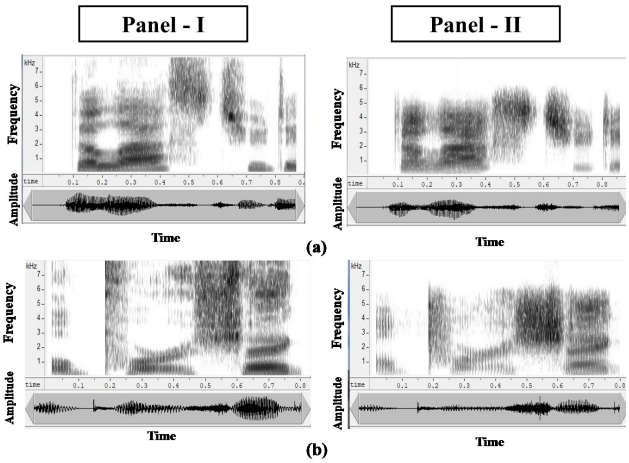

Fig. 9. Panel-I : Analysis for original speech signal. Panel-II: Analysis for anonymized speech signal. (a) spectrogram and speech signal for a female speaker, (b) spectrogram and speech signal for a male speaker.

angle of the poles ϕ to $\phi^{0.8}$. When $\phi < 1$, the formant will be shifted to a higher value, and vice-versa for $\phi > 1$. Since for a male speech, the formants are lower, they will be shifted to a higher value. Similarly, speech spectrum gets high pitch source harmonics for females. We observe that there is a denser frequency spectrum because the energy at lower frequency values increases, and the energy at higher frequency values decrease. This makes the spectrogram have more energy gathered below some frequency, here $< 6000Hz$. It is also observed that the energy distribution in the male speaker is more uniform as compared to the female speaker.

IV. Conclusions

In this paper, we have used LP model and McAdam's coefficient to achieve effective speaker anonymization. The baseline system in [11] attains anonymization by shifting the pole angles only. However, the pole radius also contributes to speaker identification (as pole radius contributes to various energy losses during natural speech production and hence, -3dB bandwidth). Thus, the authors have varied the pole radius along with the pole phases to get better anonymization. In addition, gender-based analysis is done by observing spectrograms, and it has been found that, better anonymization of female speakers is obtained as compared to the male speakers.

In future, better anonymization can be achieved by extracting speaker specific information even from the residual [37]. Furthermore, other signal processing techniques apart from LP based analysis, can be used for better anonymization, along with neural network-based approaches as given in baseline-1 of the Voice Privacy challenge 2020. It is based on x-vectors and neural waveform models which can provide better EER and WER results [12], but it does so at the cost of complex and expensive training based method.

ACKNOWLEDGMENT

The authors would like to thank the authorities of DA-IICT Gandhinagar, India for their support to carry out this research work.

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 Last Accessed: 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] 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.

[33] 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

[34] 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.

[35] H. A. Patil, P. K. Dutta, and T. K. Basu, "On the investigation of spectral resolution problem for identification of female speakers in Bengali," in *IEEE International Conference on Industrial Technology (ICIT)*, 15-17 December,2006, pp. 375–380.

[36] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, December, 2009.

[37] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2005.