A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender

Pierre Champion,[1] **Denis Jouvet,** [2] **Anthony Larcher** [1]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.
[2] Le Mans Université, LIUM, France
{pierre.champion, denis.jouvet}@inria.fr, anthony.larcher@univ-lemans.fr

Abstract

Speech pseudonymization aims at altering a speech signal to map the identifiable personal characteristics of a given speaker to another identity. In other words, it aims to hide the source speaker identity while preserving the intelligibility of the spoken content. This study takes place in the VoicePrivacy 2020 challenge framework, where the baseline system performs pseudonymization by modifying x-vector information to match a target speaker while keeping the fundamental frequency (F0) unchanged. We propose to alter other paralinguistic features, here F0, and analyze the impact of this modification across gender. We found that the proposed F0 modification always improves pseudonymization. We observed that both source and target speaker genders affect the performance gain when modifying the F0.

Introduction

In many applications, such as virtual assistants, speech signal is sent from the device to centralized servers in which data is collected, processed, and stored. Recent regulations, e.g., the General Data Protection Regulation (GDPR) (Parliament and Council 2016) in the EU, emphasize on privacy preservation and protection of personal data. As speech data can reflect both biological and behavioral characteristics of the speaker, it is qualified as personal data (Nautsch et al. 2019). The research reported in this paper has been done in the context of the VoicePrivacy challenge framework (Tomashenko et al. 2020), which is one of the first attempt of the speech community to encourage research on this topic, define the task, introduce metrics, datasets and protocols.

Anonymization is performed to suppress the personally identifiable paralinguistic information from a speech utterance while maintaining the linguistic content. The task of the VoicePrivacy challenge is to degrade automatic speaker verification performance, by removing speaker identity as much as possible, while keeping the linguistic content intelligible. This task is also referred to as *speaker anonymization*(Fang et al. 2019) or *de-identification* (Magariños et al. 2017).

Anonymization systems in the VoicePrivacy challenge should satisfy the following requirements:

- output a speech waveform;
- conceal the speaker's identity;
- keep the linguistic content intelligible;
- modify the speech signal of a given speaker to always sound like a unique target pseudo-speaker, while different speaker's speech must not be similar.

The fourth requirement constraints the system to have a one-to-one mapping between the real speaker identities and a pseudo-speaker. Such system can be considered as a voice conversion system where the output speaker identity resides in a pseudonymized space.

The GDPR defines pseudonymization as: *"processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"*(Art.4.5 of the GDPR (Parliament and Council 2016)). pseudonymization techniques differ from anonymization techniques. With anonymization, data is modified so that any information that may serve as an identifier to a subject is deleted. pseudonymization enhances privacy by replacing most identifying information within data by artificial identifiers. Per the requirements imposed by the VoicePrivacy challenge, and the above definition from GDPR, the challenge imposes contestants to build pseudonymization systems. The VoicePrivacy challenge focuses on modifying the speech characteristics; while keeping the linguistic content unchanged; hence removing personal information from the linguistic content is not part of that challenge.

Recently, Fang et al. (Fang et al. 2019) proposed a speech synthesis pipeline where only the continuous speaker representation (the x-vector (Snyder et al. 2018)) is modified. Linguistic related information necessary to generate anonymized speech is left untouched. The corresponding toolchain doesn't alter the fundamental frequency (F0) input values, and the articulation of speech sounds feature (the Phoneme Posterior-Grams (PPGs) (Sun et al. 2016)).

The F0 values of speech determine the perceived relative highness or lowness of the sound, it plays an indispens-

able role for the listener as it helps to perceive a variety of paralinguistic, and prosodic information (Gussenhoven 2004). Analysis of the F0, which is typically higher in female voices than in male voices, can be used to characterize speaker-related attributes.

In this paper, we use the pipeline proposed by Fang et al. (Fang et al. 2019) in the VoicePrivacy challenge 2020 (Tomashenko et al. 2020), and discuss what possible improvement may be obtained by modifying the F0 values.

The remainder of the paper is structured as follows. First, we review the baseline framework and explains the conversion process. Secondly we describes the experimental setup. Then we present and discuss the results. Finally, we concludes the paper.

Anonymization technique

The baseline system

The VoicePrivacy challenge provides two baseline systems: *Baseline-1* that anonymizes speech utterances using x-vectors and neural waveform models (Fang et al. 2019) and *Baseline-2* that performs anonymization using McAdams coefficient (McAdams 1984). Our contributions are based on *Baseline-1* which is referred to as the *baseline* system in this paper.

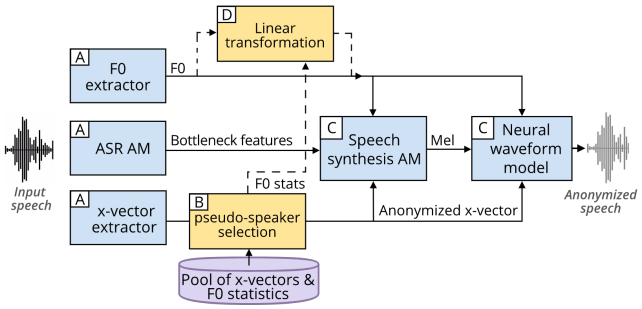

Figure 1: The speaker anonymization pipeline. Modules A, B and C are parts of the baseline model. We added module D to modify the F0 values, which are later used by modules C.

The central concept of the baseline system introduced in (Fang et al. 2019) is to separate speaker identity and linguistic content from an input speech utterance. Assuming that those information are disentangled, an anonymized speech waveform can be obtained by altering only the features that encode the speaker's identity. The anonymization system illustrated in Figure 1 breaks down the anonymization process into three groups of modules: *A - Feature extraction* comprises three modules that respectively extract fundamental frequency, PPGs like bottleneck features, and the speaker's x-vector from the input signal. Then, *B - Anonymization* derives a new pseudo-speaker identity using knowledge gleaned from a pool of external speakers. Finally, *C - Speech synthesis* synthesizes a speech waveform from the pseudo-speaker x-vector together with the original PPGs features, and the **original F0** using an acoustic model (Tomashenko et al. 2020) and a neural waveform model (Wang, Takaki, and Yamagishi 2020). For all utterances of a given speaker, a single target pseudo-speaker is used to modify the input speech. This strategy, described as *perm* in (Srivastava et al. 2020b), ensures that a one-to-one mapping exists between the source speaker identity and the target pseudo-speaker.

x-vector pseudonymization

Given the baseline system, where only the x-vector identity is changed, the selection algorithm used to derive a pseudo-identity plays an important role. Many criteria can be chosen to select the target pseudo-speaker identity. Recent research made by (Srivastava et al. 2020a) has outline multiple selection techniques for the VoicePrivacy Challenge. The baseline's pseudo-speaker selection is performed by averaging a set of x-vectors candidates from the speaker pool. The candidate x-vectors are selected by retrieving the 200 furthest speakers given the original x-vector. From this subset of 200 x-vectors, a set of 100 x-vectors is randomly chosen to create the pseudo-speaker x-vector. Speaker's distances are queried according to the probabilistic linear discriminant analysis (PLDA). The speaker pool is composed of speakers from the LibriTTS-train-other-500 (Zen et al. 2015) dataset. This dataset is not used elsewhere in our experiments.

Gender selection

Information conveyed by the x-vector embeddings can be used for other tasks than speaker recognition/verification. Work by (Raj, Snyder, and Povey 2019) has shown that session and gender information, along with other characteristics, are also encoded in x-vectors.

The aforementioned x-vector anonymization procedure is designed to select a pseudo-speaker identity from the same gender as the source speaker. Constraining the x-vector anonymization procedure to target x-vectors from same gender as the source is referred to as *Same*, While constraining the selection to target the opposite gender is referred to as *Opposite*. *Same*, and *Opposite* gender selection were experimentally studied by (Srivastava et al. 2020a). Work on *gender independent* selection still needs to be done.

In this paper, we focus our experience on *Same* and *Opposite* gender selections. We discuss the impact that F0 modification has on female and male speakers when using these two selection algorithms.

Speech synthesis

The speech synthesizer (cf. pipeline C in Figure 1) in the the VoicePrivacy baseline system is composed of a speech synthesis acoustic model, used to generate mel-fbanks features; and a vocoder, used to generate a speech signal. The vocoder used in the baseline is a Neural Source-Filter (NSF) Waveform model (Wang, Takaki, and Yamagishi 2020). NSF models uses the F0 information to produce a sine-based excitation signal that is later transformed by filters into a waveform. Manipulating the F0 values will impact both the speech synthesis acoustic model and vocoder models to transform the speech signal.

F0 modification

In the VoicePrivacy baseline, the F0 values extracted from the source speech are directly used (unchanged) by the speech synthesizer pipeline (acoustic model and neural vocoder), even though a different target pseudo-speaker was selected. Multiple works have investigated F0 conditioned voice conversion (Bahmaninezhad, Zhang, and Hansen 2018; Huang et al. 2020; Qian et al. 2020; Ueda et al. 2015). In some papers modifying the F0 improves the quality of the converted voice. Motivated by those results, we propose to modify the F0 values of a source utterance from a given speaker (cf. module D in Figure 1) by using the following linear transformation:

$$\hat{x}_t = \mu_y + \frac{\sigma_y}{\sigma_x}(x_t - \mu_x)$$

where x_t represents the log-scaled F0 of the source speaker at the frame t, μ_x and σ_x represent the mean and standard deviation for the source speaker. μ_y and σ_y represents the mean and standard deviation of the log-scaled F0 for the pseudo-speaker. The linear transformation and statistical calculation are only performed on voiced frames. The mean and standard deviation for the target pseudo speaker are calculated by averaging information from the same 100 speakers selected to derive the pseudo-speaker x-vector.

Experimental setup

Data

All experiments where based on the challenge publicly available baseline[1]. The development and evaluation sets are built from LibriSpeech *test-clean*. The pool of external speakers on which x-vectors and F0 statistics are computed is LibriSpeech *train-other-500*. Additional information on the number of speakers, and the gender distributions can be found in the evaluation plan (Tomashenko et al. 2020).

Attack models

One of the requirements of the VoicePrivacy challenge is to *conceal the speaker's identity*. To assess the robustness of anonymization systems, two attack models were designed (cf. evaluation plan). The first scenario consists of a user who publishes anonymized speech and an attacker who uses one enrollment utterance of non-anonymized (original) speech to compute a linkability score. In this scenario (referred as **o-a** in Figure 2), the goal is to ensure the **o**riginal speaker identity is not the same as the one in the generated **a**nonymized speech. Performant systems are expected to show low linkability. The second scenario consists of a user who also publishes anonymized speech, but this time, the attacker has itself anonymized an enrollment utterance using the same exact anonymization pipeline except for the random seed. This scenario (referred as **a-a** in Figure 2) is defined as a *Semi-Informed* attacker in work done by (Srivastava et al. 2020b). **Hence, the pseudo-speaker corresponding to a given speaker in the enrollment set is different from the pseudo-speaker corresponding to that same speaker in the trial set, as mentioned in Section 3.3 of the evaluation plan**. Consequently, we also expect to have low linkability in this **a-a** scenario even through the attacker has gained some knowledge about the anonymization system.

Utility and linkability metrics

To evaluate the performance of the system in both linkability (*speaker's concealing* capability) and utility (*content intelligibility*) two systems are used. To assess the linkability, a pre-trained x-vector-PLDA based Automatic Speaker Verification (ASV) system provided by the challenge organizers is used. The privacy protection is measured in terms of C_{llr}^{min} as this measure provides an application-independent (Brummer and Preez 2006) evaluation score. As the Equal Error Rate (EER) measure is more often used in speaker verification, we present our result in terms of both EER and C_{llr}^{min}. Those metrics are computed using the cllr toolkit[2] of the challenge. For the utility, a pre-trained Automatic Speech Recognition (ASR) system provided by the challenge organizers is used to decode the anonymized speech and compute the Word Error Rate ($WER_{\%}$). In this challenge, the $WER_{\%}$ measure is used to evaluate how the content is kept intelligible. Both ASR and ASV systems are trained on LibriSpeech *train-clean-360* using Kaldi (Povey et al. 2011). The higher the EER/C_{llr}^{min}, the better the systems are capable of *"concealing a speaker identity"*. The lower the $WER_{\%}$ is, the more intelligible the anonymized speech is.

Experimental results

All results are compared to the VoicePrivacy baseline system. The pseudonymization pipeline with F0 modification contribution is publicly available[3]. Figure 2 details the speaker linkability scores for **o**riginal to **a**nonymized (o-a) ASV tests, and for **a**nonymized to **a**nonymized (a-a) ASV tests in different gender selection and F0 modification setups. The **o**riginal to **a**nonymized test case helps to assess how capable systems are at modifying the original speech to make it sound like another speaker's speech. As the system used to evaluate the linkability between **o**riginal and **a**nonymized speech is domain-dependent (Srivastava et al. 2020b), and only trained on the original speech, it is thus of no surprise that the baseline provided in the challenge already shows great results. As for the **a**nonymized to **a**nonymized test, enrolling the ASV system with anonymized data brings some speaker information in the process, although the pseudo-speaker x-vector is not exactly the same between random and trial utterances, because of the random part of the x-vector selection process (see Section on x-vector pseudonymization above). Given this evaluation framework, our goal is to further degrade the linkability in both attacks models. For each anonymization pipeline setups, the corresponding $WER_{\%}$ values are reported in Table 1.

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

[2] https://gitlab.eurecom.fr/nautsch/cllr/

[3] https://github.com/deep-privacy/Voice-Privacy-Challenge-2020

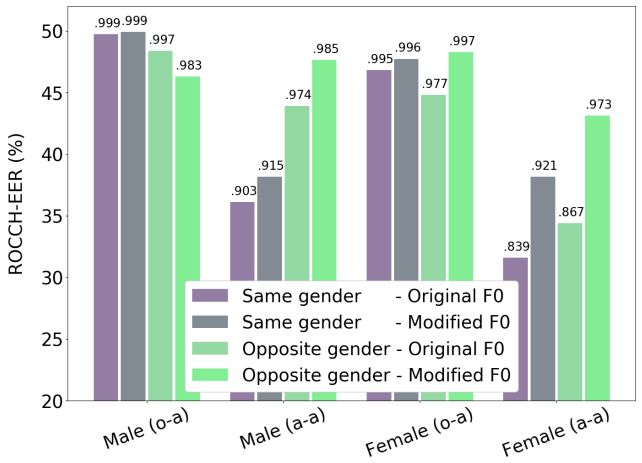

Figure 2: EER (%) score obtained by the ASV evaluation system on Librispeech tests sets. The C_{llr}^{min} score is displayed on the top of each bar. Multiple pipelines setups are reported for the gender selection and F0 modification. o – original, a – anonymized speech data for enrollment and trial parts. Entry "Same gender - Original F0" corresponds to the challenge baseline system.

Male linkability

In the **o**riginal to **a**nonymized attack scenario (o-a in Figure 2), we can observe that the proposed F0 modification doesn't affect the already good male un-linkability performance when compared to the challenge's baseline ("Same gender - Modified F0" compared to "Same gender - Original F0"). It appears that selecting an x-vector from the opposite gender without applying the F0 modification always degrades the pseudonymization un-linkability ("Opposite gender - Original F0" compared to "Same gender - Original F0"). Applying the F0 modification together with the opposite gender x-vector selection doesn't improves performance. This limitation might come from the x-vector selection algorithm, where the furthest speakers are selected to derive the pseudo-identity.

Regarding the **a**nonymized to **a**nonymized attack scenario (a-a in Figure 2). Using the baseline anonymization setup, the attacker is able to re-identify the user at a much higher degree. On their own, the F0 modification always improves compared to the baseline performance. Jointly selecting the opposite gender and applying the F0 modification appears to be an excellent design choice against this attacker.

Female linkability

Contrary to the male results, the proposed F0 modification always improves the pseudonymization for female speaker in the **o**riginal to **a**nonymized attack scenario. This effect is observed regardless of the gender's x-vector selection ("Same gender - Modified F0" compared to "Same gender - Original F0" and "Opposite gender - Modified F0" compared to "Opposite gender - Original F0"). Applying both the F0 modification and the opposite x-vector selection beats the baseline system.

The **a**nonymized to **a**nonymized attack scenario draws similar conclusions as for the male speaker. Jointly modifying gender for the x-vector selection and applying the F0 modification always improves pseudonymization. It is worth noting that female speakers are more sensitive to F0 modification than males. Meaning, the source's gender information plays a role in choosing the best anonymization procedure.

Speech intelligibility

Gender-selection	**F0**	**Test** $\text{WER}_{\%}$
Same	Original	6.73
	Modified	6.92
Opposite	Original	7.24
	Modified	6.74

Table 1: Speech recognition results in terms of $\text{WER}_{\%}$ for the LibriSpeech test set.

Across all experiments, the utility (Table 1) is not tremendously affected by the gender x-vector selection, F0 modification, or the two modifications applied together. The high $\text{WER}_{\%}$ score (7.24) reported with the opposite x-vector gender selection, and no F0 modification might come from the fact that the ASR model used for the evaluation was trained on audiobooks data; and the fact that selecting opposite gender without modifying F0 might leads to some inconsistencies in the speech signal.

Conclusions

In this work, we proposed to alter the F0 paralinguistic information in an x-vector based speech pseudonymization system. We evaluated this modification against the *Opposite* and *Same* gender x-vector target selection to obtain various anonymization setups. We objectively evaluated the F0 modification using the VoicePrivacy 2020 challenge tools. The performance was assessed in terms of EER/C_{llr}^{min} to measure privacy protection and $\text{WER}_{\%}$ to measure utility. We observed that keeping the original F0 values retains some information about the original speaker. The experiments show that applying the F0 modification and selecting an x-vector from the *Opposite* gender allows for better privacy protection against attackers who has access to the anonymization pipeline. Our results also show that the performance of anonymization depends on the gender of the source. This raises the question of the importance of personalized modification in a privacy context. In future work, we plan to subjectively evaluate the naturalness of the generated speech. We think the F0 modification helps to produce a more natural speech when an *Opposite* gender's x-vector is selected. Because the F0 features will be coherent with the selected gender.

Acknowledgments

This work was supported in part by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018) and Région Lorraine.

References

Bahmaninezhad, F.; Zhang, C.; and Hansen, J. H. L. 2018. Convolutional Neural Network Based Speaker De-Identification. In *Odyssey*.

Brummer, N.; and Preez, J. 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language* .

Fang, F.; Wang, X.; Yamagishi, J.; Echizen, I.; Todisco, M.; Evans, N.; and Bonastre, J.-F. 2019. Speaker Anonymization Using X-vector and Neural Waveform Models. In *Proc. 10th ISCA Speech Synthesis Workshop*.

Gussenhoven, C. 2004. *Pitch in Language I: Stress and Intonation*. Research Surveys in Linguistics. Cambridge University Press.

Huang, W.-C.; Luo, H.; Hwang, H.-T.; Lo, C.-C.; Peng, Y.-H.; Tsao, Y.; and Wang, H.-M. 2020. Unsupervised Representation Disentanglement Using Cross Domain Features and Adversarial Learning in Variational Autoencoder Based Voice Conversion. *IEEE Transactions on Emerging Topics in Computational Intelligence* .

Magariños, C.; Lopez-Otero, P.; Docio-Fernandez, L.; Rodriguez-Banga, E.; Erro, D.; and Garcia-Mateo, C. 2017. Reversible speaker de-identification using pre-trained transformation functions. *Computer Speech & Language* .

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

Nautsch, A.; Jasserand, C.; Kindt, E.; Todisco, M.; Trancoso, I.; and Evans, N. 2019. The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. In *Proc. Interspeech*.

Parliament, E.; and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. *General Data Protection Regulation* .

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlíček, P.; Qian, Y.; Schwarz, P.; Silovský, J.; Stemmer, G.; and Vesel, K. 2011. The Kaldi speech recognition toolkit. *IEEE Workshop on Automatic Speech Recognition and Understanding* .

Qian, K.; Jin, Z.; Hasegawa-Johnson, M.; and Mysore, G. J. 2020. F0-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder. *IEEE ICASSP* .

Raj, D.; Snyder, D.; and Povey, D. 2019. Probing the Information Encoded in X-Vectors. In *IEEE ASRU*.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust DNN Embeddings for Speaker Recognition. In *IEEE ICASSP*.

Srivastava, B. M. L.; Tomashenko, N.; Wang, X.; Vincent, E.; Yamagishi, J.; Maouche, M.; Bellet, A.; and Tommasi, M. 2020a. Design Choices for X-vector Based Speaker Anonymization. *Proc. Interspeech* .

Srivastava, B. M. L.; Vauquier, N.; Sahidullah, M.; Bellet, A.; Tommasi, M.; and Vincent, E. 2020b. Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. In *IEEE ICASSP*.

Sun, L.; Li, K.; Wang, H.; Kang, S.; and Meng, H. 2016. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *IEEE ICME*.

Tomashenko, N.; Srivastava, B. M. L.; Wang, X.; Vincent, E.; Nautsch, A.; Yamagishi, J.; Evans, N.; Patino, J.; Bonastre, J.-F.; Noé, P.-G.; and Todisco, M. 2020. Introducing the VoicePrivacy Initiative. *Proc. Interspeech* .

Ueda, R.; Aihara, R.; Takiguchi, T.; and Ariki, Y. 2015. Individuality-Preserving Spectrum Modification for Articulation Disorders Using Phone Selective Synthesis. In *Proc. Interspeech*.

Wang, X.; Takaki, S.; and Yamagishi, J. 2020. Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. *IEEE TASLP* .

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2015. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*.