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

Pierre Champion, Denis Jouvet, Anthony Larcher

A STUDY OF F0 MODIFICATION FOR X-VECTOR BASED SPEECH PSEUDO-ANONYMIZATION ACROSS GENDER

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

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.
[2]Le Mans Université, LIUM, France

ABSTRACT

Speech pseudo-anonymization 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 pseudo-anonymization 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 pseudo-anonymization. We observed that both source and target speaker genders affect the performance gain when modifying the F0.

Index Terms— VoicePrivacy 2020 Challenge, Speaker anonymization, F0 modification

1. 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) [1] 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 [2]. This research has been done in the VoicePrivacy challenge framework [3], 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*[4] or *de-identification* [5].

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 pseudo-anonymization 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 [1]). Pseudo-anonymization 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. Pseudo-anonymization 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 pseudo-anonymization 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. [4] proposed a speech synthesis pipeline where only the continuous speaker representation (the x-vector [6]) 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) [7]).

The F0 values of speech determine the perceived relative highness or lowness of the sound, it plays an indispensable role for the listener as it helps to perceive a variety of paralinguistic, and prosodic information [8]. Analysis of the F0, which is typically higher in female voices than in male voices, can be used to characterize speaker-related attributes. For example, a simple gender classifier [9] can be driven by only using F0 features.

In this paper, we use the pipeline proposed by Fang et al. [4] in the VoicePrivacy challenge 2020 [3], and discuss what possible improvement may be obtained by modifying the F0 values.

The remainder of the paper is structured as follows. Section 2 reviews the baseline framework and explains the conversion process. Section 3 describes the experimental setup. Section 4 summarizes the results. Section 5 discusses the need to develop such effort. Finally, Section 6 concludes the paper.

2. ANONYMIZATION TECHNIQUE

2.1. The baseline system

The VoicePrivacy challenge provides two baseline systems: *Baseline-1* that anonymizes speech utterances using x-vectors and neural waveform models [4] and *Baseline-2* that performs anonymization using McAdams coefficient [10]. Our contributions are based on *Baseline-1* which is referred to as the *baseline* system in this paper.

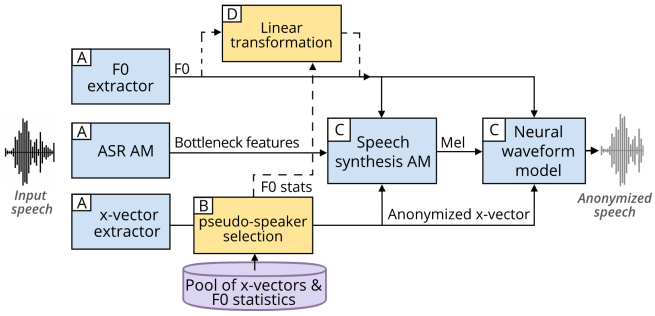


Fig. 1. The speaker anonymization pipeline. Modules A, B and C are parts of the baseline model. We added module D to modify the F0 values, which are later used by modules C.

The central concept of the baseline system introduced in [4] is to separate speaker identity and linguistic content from an input speech utterance. Assuming that those information can be 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 together with the original PPGs features, and the **original F0** using an acoustic model [3] and a neural waveform model [11]. 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 [12], ensures that a one-to-one mapping exists between the source speaker identity and the target pseudo-speaker.

2.2. x-vector pseudo-anonymization

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 [13] 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 [14] dataset. This dataset is not used elsewhere in our experiments.

2.3. Gender selection

Information conveyed by the x-vector embeddings can be used for other tasks than speaker recognition/verification. Work by [15] 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 [13]. 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 speaker when using these two selection algorithms.

2.4. Speech synthesis

The speech synthesizer **Figure 1 - C** pipeline in 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 [11]. While this architecture wasn't initially created to be conditioned on a speaker embedding, the F0 contour is a critical element. 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.

2.5. F0 modification

The VoicePrivacy baseline system uses the same F0 inputs values as the source speech, even through a different target pseudo-speaker was selected. Multiples works have investigated F0 conditioned voice conversion [16, 17, 18, 19] and shown that they improve the conversion system to separate content, F0, and speaker identity. Motivated by those results, we propose to modify the F0 values of a source utterance from a given speaker (**Figure 1 - D**) 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, respectively. μ_y and σ_y represents the mean and standard deviation of the log-scaled F0 for the pseudo-speaker, respectively. 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 taking the same 100 speakers selected to derive the pseudo-speaker x-vector (please refer to Section 2.2).

3. EXPERIMENTAL SETUP

3.1. Data

All experiments where based on the challenge publicly available baseline[1]. The development and evaluation sets are built from LibriSpeech *test-clean* [20] and *VCTK* [21]. 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 [3].

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

3.2. Attack models

One of the requirements of the VoicePrivacy challenge is to *conceal the speaker's identity* (please refer to Section 1). 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 also has access to an anonymized enrollment utterance. This scenario (referred as **a-a** in **Figure 2**) is defined as a *Semi-Informed* attacker in work done by Brij Mohan Lal Srivastava et al. [12]. In the *Semi-Informed* attack, the user is more vulnerable as the attacker has gained some knowledge about the anonymization system. If the attacker has access to the real identity hidden behind pseudo-speaker, the privacy of the user is heavily compromised. With this specific motivation, performant systems are expected to show low linkability.

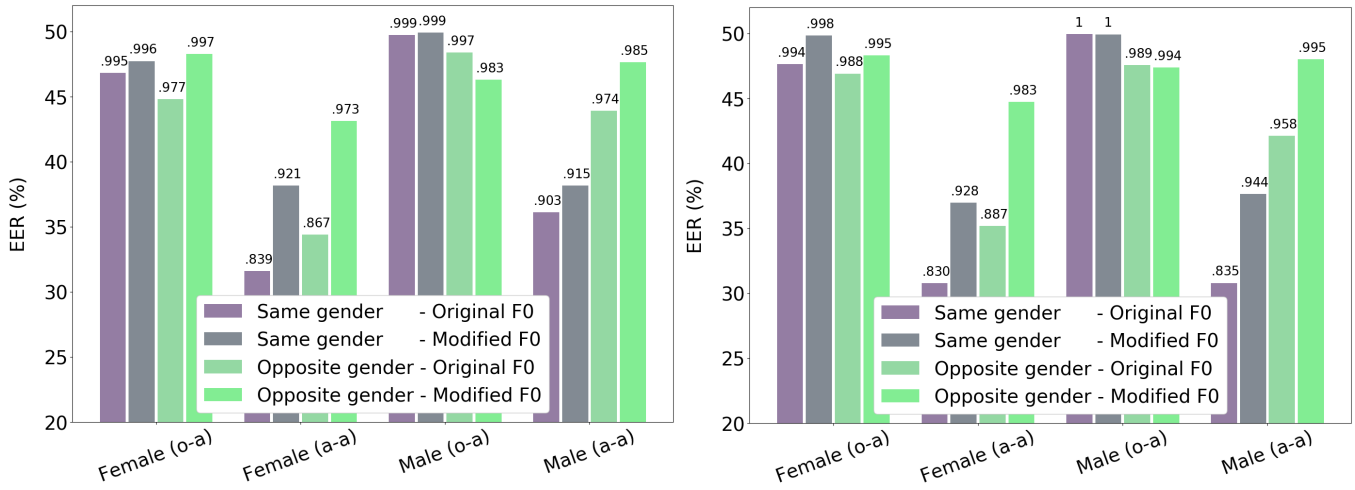
3.3. Utility and linkability metrics

To evaluate the performance of the system in both linkability (*speaker's concealing* capability) and utility (*content intelligibility*) two systems are used. To access 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 [22] evaluation. As the 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 $WER_{\%}$. 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 [23]. 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.

4. EXPERIMENTAL RESULTS

All results are compared to the VoicePrivacy baseline system. The pseudo-anonymization pipeline with F0 modifi-

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



(a) Librispeech test-set

(b) VCTK-common test-set

Fig. 2. EER (%) score obtained by the ASV evaluation system on Librispeech and VCTK tests sets. The C_{llr}^{min} score is displayed on the top of each bar, for additional information. Multiple pipelines setup 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.

cation contribution is publicly available[3]. **Figure 2** details the speaker linkability scores for **o**riginal to **a**nonymized ASV tests, and for **a**nonymized to **a**nonymized ASV tests in different gender selection and F0 modification setup. 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 [12], 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, the model used is still not trained on anonymized speech, but having the same anonymized enrollments and trials utterances helps the attacker to re-identify users at a more significant degree. Given this evaluation framework, our goal is to further degrade the linkability in both attacks models. For each anonymization pipeline setups, the corresponding $\text{WER}_{\%}$ values are reported in **Table 1**.

4.1. Male linkability

In the **o**riginal to **a**nonymized attack scenario (o-a in Figure 2), we can observe that on both Librispeech and VCTK dataset 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 selectionning an x-vector from the opposite gender without applying the F0 modification always degrades the pseudo-anonymization un-linkability ("Opposite gender - Original F0" compared to "Same gender - Original F0"). Applying the F0 modification together with the opposite x-vector selection doesn't allow to recover the performance lost from the opposite gender selection ("Opposite gender - Modified F0" compared to "Same gender - Original F0"). This limitation might come from the x-vector selection algorithm, where the furthest speakers are selected to derive the pseudo-identity. Experiments done in [13] shows that choosing a speaker in a dense x-vector region could overcome this limitation.

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, both opposite gender selection and F0 modification show improvements over the baseline system. Jointly selecting the opposite gender and applying the F0 modification appears to be an excellent design choice against this attacker.

4.2. Female linkability

Contrary to the male results, the proposed F0 modification always improves the pseudo-anonymization 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"). On the Librispeech test-set, applying both the F0 modification and the opposite x-vector se-

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

lection beats the baseline system, but this conclusion doesn't apply to the VCTK dataset. Again, this limitation might come from the x-vector selection algorithm and not the F0 modification.

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 pseudo-anonymization. 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.

4.3. Speech intelligibility

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 on LibriSpeech with the opposite x-vector gender selection, and no F0 modification might come from the fact that the ASR model used to evaluate is trained on similar data as it was tested, i.e., audiobooks (please refer to section 3.3). Meaning the evaluation model is more sensitive to slight speech distortion within the same dataset type.

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

Dataset	**Gender-selection**	**F0**	**Test** $\text{WER}_{\%}$
LibriSpeech	Same	Original	6.73
		Modified	6.92
	Opposite	Original	7.24
		Modified	6.74
VCTK	Same	Original	15.23
		Modified	15.29
	Opposite	Original	15.46
		Modified	15.28

5. DISCUSSION

This study takes place in the framework of the VoicePrivacy 2020 challenge. One of the requirements imposed by the challenge is to "*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*". This requirement is motivated by the fact that "*in a multi-party human conversation, each speaker cannot change his/her anonymized voice over time, and the anonymized voices of all speakers must be distinguishable from each other.*" (VoicePrivacy evaluation plan [3]). Speech data is a personal data [2], and the GDPR article definition exposed in section 1 applies to it.

In the context of the challenge, pseudo-anonymization should be evaluated according to an additional metric. Indeed in the scenario where original and anonymized speech are compared, the EER must always be maximized, as it reflects that the pseudo-anonymization process fools the reference speaker verification system. However, in the scenario where enrollment and trial anonymized speech are compared, the question of whether the EER should be minimized or maximized is not clear.

In case an attacker has access to one anonymized session from a given speaker, maximizing the EER protects other sessions as no link can be established between them. Though, this goal doesn't allow any automatic processing in the arrival space; for instance: no automatic speaker diarization can be performed anymore. Despite this, the challenge rule requires that subjective speaker linkability should be preserved. Being capable of measuring how the EER score reflects the actual subjective speaker linkability would benefit the development of speaker verification systems by highlighting the weaknesses of current automated approaches.

6. CONCLUSIONS

In this work, we proposed to alter the F0 paralinguistic information in an x-vector based speech pseudo-anonymization system. We tested this modification against the *Opposite* and *Same* x-vector target selection to obtain various anonymization setup. We objectively evaluated the F0 modification using the VoicePrivacy 2020 challenge. The performance was assessed in terms of EER/C_{llr}^{min} to measure privacy protection and $\text{WER}_{\%}$ to measure utility. We observed that the F0 retains some information about the original speaker. It showed 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 anonymized speech. Our results also show that the performance of anonymization depends on the gender of the source. This raises the question of the importance of customized modification in a privacy context. In future work, we plan to subjectively evaluate the naturalness of our modification. We think the F0 modification helps to produce a more natural speech when an *Opposite* gender's x-vector is selected.

7. ACKNOWLEDGMENTS

This work was supported in part by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018) and Région Lorraine. Experiments were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.

8. REFERENCES

[1] European Parliament and Council, "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*, 2016.

[2] Andreas Nautsch, Catherine Jasserand, Els Kindt, Massimiliano Todisco, Isabel Trancoso, and Nicholas Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding," in *Proc. Interspeech*, 2019.

[3] Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco, "Introducing the VoicePrivacy Initiative," *ArXiv*, 2020.

[4] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019.

[5] Carmen Magariños, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo Rodriguez-Banga, Daniel Erro, and Carmen Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," *Computer Speech & Language*, 2017.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018.

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

[8] Carlos Gussenhoven, *Pitch in Language I: Stress and Intonation*, Research Surveys in Linguistics. Cambridge University Press, 2004.

[9] Jerzy SAS and Aleksander SAS, "Gender recognition using neural networks and asr techniques," *Journal of MIT*, vol. 22, 2013.

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

[11] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE TASLP*, vol. 28, 2020.

[12] Brij Mohan Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *IEEE ICASSP*, 2020.

[13] Brij Mohan Lal Srivastava, Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, and Marc Tommasi, "Design choices for x-vector based speaker anonymization," *ArXiv*, 2020.

[14] H. Zen, V. Dang, R. Clark, Yu Zhang, Ron J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[15] Desh Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," *IEEE ASRU*, 2019.

[16] Fahimeh Bahmaninezhad, Chunlei Zhang, and John H. L. Hansen, "Convolutional neural network based speaker de-identification," in *Odyssey*, 2018.

[17] Wen-Chin Huang, Haiyan Luo, Hsin-Te Hwang, Chen-Chou Lo, Yu-Huai Peng, Yu Tsao, and Hsin-Min Wang, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, 2020.

[18] Kaizhi Qian, Zeyu Jin, Mark Hasegawa-Johnson, and Gautham J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," *IEEE ICASSP*, 2020.

[19] Reina Ueda, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, "Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis," in *Proc. Interspeech*, 2015.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015.

[21] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[22] N. Brummer and J.A. Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, 2006.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel, "The kaldi speech recognition toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.