Evaluation of Speaker Anonymization on Emotional Speech

Hubert Nourtel, Pierre Champion, Denis Jouvet, Anthony Larcher, Marie Tahon

Evaluation of Speaker Anonymization on Emotional Speech

Hubert Nourtel[1⋆] *Pierre Champion*[1,2⋆] *Denis Jouvet*[1] *Anthony Larcher*[2] *Marie Tahon*[2]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] LIUM - EA4023, Le Mans Université, Avenue Olivier Messiaen,72085 LE MANS CEDEX 9, France
⋆ equal contribution from authors
hubert.nourtel@inria.fr

Abstract

Speech data carries a range of personal information, such as the speaker's identity and emotional state. These attributes can be used for malicious purposes. With the development of virtual assistants, a new generation of privacy threats has emerged. Current studies have addressed the topic of preserving speech privacy. One of them, the VoicePrivacy initiative aims to promote the development of privacy preservation tools for speech technology. The task selected for the VoicePrivacy 2020 Challenge (VPC) is about speaker anonymization. The goal is to hide the source speaker's identity while preserving the linguistic information. The baseline of the VPC makes use of a voice conversion. This paper studies the impact of the speaker anonymization baseline system of the VPC on emotional information present in speech utterances. Evaluation is performed following the VPC rules regarding the attackers' knowledge about the anonymization system. Our results show that the VPC baseline system does not suppress speakers' emotions against informed attackers. When comparing anonymized speech to original speech, the emotion recognition performance is degraded by 15% relative to IEMOCAP data, similar to the degradation observed for automatic speech recognition used to evaluate the preservation of the linguistic information.

Index Terms: Speaker Anonymization, Voice Privacy, Emotion Recognition

1. Introduction

Voice-controlled applications, such as smart speakers, have become widely popular. Large amount of data is required to train such applications. This motivates service providers to collect, process, and store personal data in centralized servers. Voice is one of the most sensitive modalities as it encapsulates many discernible attributes of a speaker such as age, gender, health, personality traits, socioeconomic status, geographical origin, biometric identity, moods, and emotions [1, 2]. Given that speech data falls under the category of personal data [3], speech privacy-preserving solutions are becoming increasingly important. Additionally, recent regulations, e.g., the General Data Protection Regulation (GDPR) [4] in the European Union, emphasize on privacy preservation and protection of personal data. The research reported in this article has been done using the Voice privacy Challenge (VPC) framework [5] which is one of the first attempts of the speech community to evaluate research on this topic by producing dedicated protocols, metrics, datasets, and baselines.

The goal of the VPC system is to anonymize the speaker. This task is performed to suppress the personally identifiable paralinguistic information from a speaker's speech utterance while maintaining the linguistic content. The VPC baseline system uses a speaker anonymization approach [6] based on x-vectors and voice conversion. The quality of anonymization in the VPC is assessed using a speaker verification system, which evaluates the speaker concealing capability (privacy metric) and using an automatic speech recognition system to evaluate the preservation and intelligibility of the linguistic content (utility metric) [5]. In this work, we investigate the extent to which an utterance's emotional content can be retrieved after anonymization.

Speaker recognition and voice privacy usually focus on so-called "neutral" speech. However, in spontaneous expressive speech, the audio signal carries speaker information, linguistic content and emotional cues. The anonymization process can be altered by emotional speech, for which the speech signal strongly differs from the "neutral" speech. Human emotion is usually described in psychological theories using diverse and complementary theories [7, 8]. For a long time, the collection of emotional data mainly focused on acted and semantically controlled data from a few speakers [9]. However, the actual trend is to capture the diversity of humans expressively in real-life conditions in order to model social aspects or induced interactions such as laughter [10] or disfluencies [11]. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [12] is in-between acted and spontaneous speech and has the advantage of being used as a benchmark in the community.

Linguistic cues mainly rely on the words pronounced by the speaker, while paralinguistic cues are directly related to the acoustic content of the speech signal. More precisely, prosodic features such as the fundamental frequency (F0), intensity and rhythm, are often considered as the most important cues in the field of speech emotion recognition (SER). Although most SER systems intend to capture prosody in input, another option is to extract Mel frequency cepstral coefficients (MFCCs) [13] or even spectrograms [14]. The HUMAINE association also took an inventory of acoustic features in the CEICES initiative [15] which conducts to a set of a hundred descriptors selected over several corpora with various techniques [16]. These features have the advantage of being easily interpretable. However, their extraction in degraded signals is error-prone. The use of input pre-trained features, i.e., embeddings extracted with neural models trained for speech processing tasks different from SER, are currently extensively used. The advantage of such an approach is to benefit from a large amount of data from a different task such as automatic speech recognition (ASR) [17] or speaker recognition [18].

Remarkably few works have handled the problem of privacy preservation in the context of emotional speech. In [19], the authors proposed distance-preserving hashing techniques and homomorphic encryption to protect sensitive data such as emotions. Generative adversarial networks have been used as an intermediate layer between users and cloud services to sanitize the input speech [20]. We aim to investigate how applying

an anonymization process on emotional data, which is supposed to hide the speaker identity, impacts SER performance.

The paper is organized as follows. Section 2 presents the anonymization framework. It recalls the VPC baseline system, details F0 transformation enhancements, and presents the attack scenarios. Section 3 details the experiments conducted and the evaluation protocol with respect to the emotions and discusses the results. A conclusion ends the paper.

2. Anonymization framework

This section presents the speaker anonymization baseline system of the Voice Privacy Challenge and the attack scenarios.

2.1. The VPC speaker anonymization system

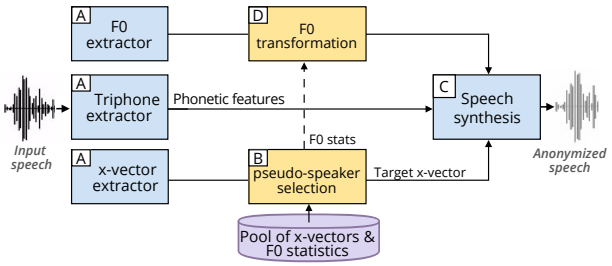

Figure 1: *The Voice Privacy speaker anonymization pipeline. Modules A, B, and C are parts of the baseline model. Module D is an enhancement used to transform the F0 values.*

The baseline system introduced in [6] aims at separating speaker identity and linguistic content from an input speech utterance. Assuming that those features are disentangled, an anonymized speech waveform is generated by altering only the features that encode the speaker's identity. The anonymization system depicted in Figure 1 can be decomposed into three groups of modules. Modules from the *group A* extract different features from the source signal: the fundamental frequency (F0), the phonetic features encoding articulation of speech sounds and the speaker's x-vector. *The module B* derives a new pseudo-speaker target x-vector identity. The x-vector from each source input speaker is compared to a pool of external x-vectors to select the 200 furthest x-vectors; 100 of them are randomly selected and averaged to create an anonymized pseudo-speaker x-vector identity. Finally, *the module C* synthesizes a speech waveform from the target x-vector together with the original phonetic features and F0. Speaker anonymization is achieved through the selection of the pseudo-speaker target x-vector. The triphone extractor has been trained on the *train-clean-100* and *train-other-500* subsets of LibriSpeech. The x-vector extractor has been trained on VoxCeleb-1,2. The speech synthesis system has been trained on the *train-clean-100* subset of LibriTTS. Finally, the *train-other-500* subset of LibriTTS has been used to create a pool of x-vector and F0 statistics.

2.2. The F0 transformations

In the original VPC anonymization system, the F0 values extracted from the source speech are directly used (unchanged) by the speech synthesizer, even though a different pseudo-speaker target x-vector is selected. Multiple studies have investigated F0 conditioned voice conversion [21, 22, 23, 24, 25]. They conclude that modifying the F0 improves the quality of the converted voice. Motivated by those results, and also by the fact that emotions undoubtedly affect intonation, we propose to modify the F0 values of a source utterance from a given speaker (cf. module D in Figure 1) by using a linear transformation and a random warping.

2.2.1. F0 linear transformation

In this method [21, 22], to feed the synthesizer with F0 values close to the pseudo-speaker selected, the F0 features of the source speaker are transformed using a linear transformation:

$$\log \hat{x}_t = \mu_y + \frac{\sigma_y}{\sigma_x} (\log x_t - \mu_x) \quad (1)$$

where x_t represents the F0 of the source speaker at frame t, μ_x and σ_x represent the mean and standard deviation of the log-scaled F0 for the source speaker computed on all his/her utterances. μ_y and σ_y represent 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 is calculated by averaging the F0 of voiced frames from the 100 speakers selected to derive the pseudo-speaker x-vector.

2.2.2. F0 random warping

In this method [25, 26], the contour of the F0 values are randomly modified to increase or decrease the range of the F0 variation using a warping factor:

$$\hat{x}_t = \mu_x + (x_t - \mu_x) \times \alpha \quad (2)$$

where α is sampled from a uniform distribution between 0.8 and 1.2, x_t represents the F0 value at frame t and μ_x represents the mean F0 of the utterance to transform. The α warping factor is randomly sampled for each utterance. This F0 random warping is applied after the F0 linear transformation.

2.3. The VPC attack scenarios

In the Voice Privacy Challenge, multiple sets of tests were performed depending on the attacker's knowledge of the anonymization algorithm. In this work, we focus on the *Ignorant*, and the *Informed* attacker scenarios [27, 28]. In the *Ignorant* scenario, the attacker is unaware that speech is transformed. Thus, privacy measurement is assessed using models trained on original, non-anonymized data, while the evaluation is performed using anonymized data. This mismatch leads to the measurement of a rather good anonymization performance. On the opposite, the *Informed* attacker scenario is entirely aware of the anonymization algorithm. Such attackers are able to anonymize a training dataset in the same manner as the service provider. This anonymized dataset is later used to train the evaluation model.

3. Experiments

The global aim here is to assess the emotion recognition performance once a speaker anonymization system has transformed the voices. Thus, the following sections present the emotional dataset, the evaluation protocol (based on the VPC attack scenarios), and the evaluation results.

3.1. Dataset

The IEMOCAP dataset [12] is an emotional dataset used for experiment purposes such as emotion recognition and speech

Table 1: *WER and UAR results on IEMOCAP. The LibriSpeech results from VPC are presented for comparison purposes. The first line shows the WER and UAR results when no anonymization is performed. The second line shows the corresponding results when speech is anonymized and evaluated using an ASR system retrained on anonymized speech and an Informed attacker scenario.*

	$WER_{\%}$		$UAR_{\%}$
	LibriSpeech	IEMOCAP	IEMOCAP
Original speech data Model trained on original speech	4.15	34.62	44.48
Anonymized speech data Model trained on anonymized speech	4.77	38.97	37.92
Difference Anonymized / Original	15% degradation	13% degradation	15% degradation

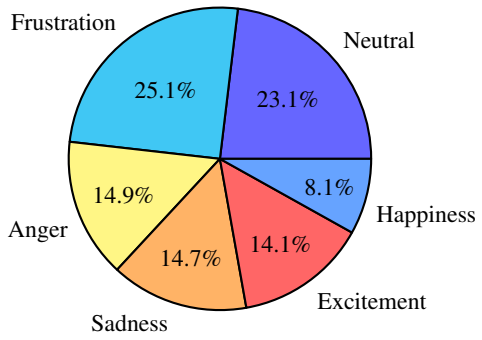

Figure 2: *Original distribution of emotion categories in the IEMOCAP dataset. Percentages are displayed with respect to number of utterances.*

recognition. It is composed of 12h of audio-visual data. Improvised and scripted dialogues between 10 female and male actors in the English language were recorded. Directional microphones have been used to capture each speaker's speech. It implies that in audio files, the two speakers can appear simultaneously (overlap). In the case of overlapping speech, the closest speaker to the microphone is considered dominant, and its speech will only be transcribed in the reference transcriptions. Because of directional microphones used, the level of the overlapping voice is much lower than the level of the "dominant" speaker voice.

The data is segmented by (dominant) speaker turns. Each turn has been annotated with emotion categories by six human annotators. Only the recordings that had the majority of annotators agreed on were used. Figure 2 shows the distribution of emotional labels in the dataset. Following previous works [18, 29], we consider only five emotions: neutral, frustration, sadness, anger, and happiness. Happiness combines the original annotations of happiness and excitement to balance the number of utterances in each emotion class.

3.2. Evaluation protocol

In this paper, we focus on evaluating emotional information present in the speech signal, both in the original utterances and in the corresponding anonymized utterances. These evaluations are carried out using a speech emotion recognition system.

The emotion recognition system used is based on a Support Vector Machine (SVM) model. In the SER literature, SVM with the radial basis function non linear kernel is widely used [30, 31] as baseline. As input features, the eGeMAPS features are used as they provide a minimalist yet efficient representation of emotion [16]. We also experimented MFCCs as input, and the results are similar. Following the VPC attack scenarios, the emotion recognition is evaluated using the *Ignorant* and the *Informed* attacker scenarios. In the *Ignorant* scenario, the SVM model used to evaluate the emotion information is trained on non-anonymized original speech data. For the *Informed* scenario, the SVM is trained on anonymized speech data. The standard Unweighted Average Recall (UAR) metric score (defined in Equation 3) is used to measure the emotion recognition performance. High UAR values means good emotion recognition. Regardless of the attack scenario, and to accommodate for the small dataset, the training and evaluation are performed using leave-one-session-out cross-validation protocol. The global performance is obtained by computing the UAR globally on the five test folds.

$$UAR = \frac{\sum \text{Recall per class}}{\#\ class} \tag{3}$$

Utility evaluation, which assesses the preservation and intelligibility of the linguistic content, is performed using two Automatic Speech Recognition (ASR) systems provided by the VPC organizers. Results are reported with the Word Error Rate (WER). The lower the WER is, the more intelligible the speech is. Evaluation results are obtained using an ASR system trained on original non-anonymized LibriSpeech *train-clean-360* data and a second one trained on the corresponding anonymized data. Retraining the ASR system on anonymized speech significantly decreases the WER when decoding anonymized speech data in comparison to the case when the ASR model is trained on the original data.

3.3. Results

This section presents the experimental results for the evaluation protocol detailed in Section 3.2.

Table 1 shows WER results on the IEMOCAP dataset, both for the original speech data, and for the anonymized speech data. Original speech is evaluated using the ASR model trained on original speech, while anonymized speech is evaluated with the model retrained on anonymized speech. The corresponding LibriSpeech scores from the VPC post-evaluation analysis [32] are presented for comparison purposes. In the IEMOCAP dataset, from original to anonymized, the utility score drops from 34.62 $WER_{\%}$ to 38.97 $WER_{\%}$, which represents a relative degradation of 13%. Similar behavior is observed on the LibriSpeech dataset, meaning the VPC baseline anonymization system performs the transformation properly on emotional speech. The high WER on IEMOCAP can be explained by the presence of overlaps, and non-neutral speech, which is not present in the LibriSpeech data.

Table 2: *Speech emotion recognition ($UAR_{\%}$) performance for the considered attack scenarios and the 95% confidence interval.*

		F0 linear transformation	F0 random warping	SVM model	$UAR_{\%}$
1	Baseline (original speech)			original	44.48 ± 1.14
2	*Ignorant* attacker			original	21.97 ± 0.95
3		✓		original	21.80 ± 0.94
4		✓	✓	original	22.14 ± 0.95
5	*Informed* attacker			retrained	37.92 ± 1.11
6		✓		retrained	37.88 ± 1.11
7		✓	✓	retrained	38.84 ± 1.11

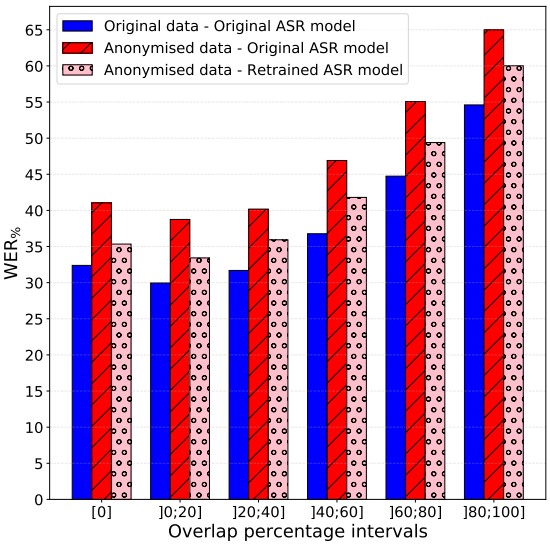

Figure 3: *WER performance on IEMOCAP with respect to the amount of overlap speech.*

Following the attack scenarios defined by the VPC (see Section 2.3), the emotion recognition scores under the *Ignorant* and *Informed* attackers are summarized in Table 2. For an *Ignorant* attacker, the UAR is 21.97%, which is nearly equal to random guessing. For an *Informed* attacker, the UAR is 37.92%, which is a 15% degradation compared to the UAR measured on original speech. The emotion recognition performance in terms of UAR seems to be impacted in the same manner as the utility performance (measured by WER).

3.4. Prosodic parameters

If one wants to hide emotional information in the anonymized speech, modification of the prosodic parameters (i.e., fundamental frequency, intensity, and rhythm) are needed. Note that F0 mean, variability, and range are included in the eGeMAPS [16] input features used in our experimental setup. As F0 was available in the anonymization baseline system, we carried out some experiments involving random modifications of the F0 values (as described in Section 2.2.2).

Results are shown in Table 2 together with the 95% confidence interval. We can see that, regardless of the attacker scenario, applying or not the F0 linear transformation alone, or the F0 linear transformation followed by a F0 random warping, leads to very similar results in terms of emotion recognition scores (UAR). In the *Informed* attacker scenario, where the SVM classifier is trained on anonymized speech, applying the F0 linear transformation does not modify the emotion recognition performance. In the future, we will investigate other modifications of the F0 curves in the anonymization process.

4. Conclusion

In this paper, we have evaluated the application of the Voice Privacy baseline system on emotional speech. Concerning the utility metric, based on automatic speech recognition performance, we observed a 13% degradation of the Word Error Rate (WER) on the IEMOCAP anonymized data compared to the WER measured on original speech. This degradation is similar to the one reported on the LibriSpeech data. However, the WER is much higher on the emotional data (IEMOCAP), also impacted by the presence of overlapping speech, than on the clean neutral data (LibriSpeech).

For what concerns the emotion information carried by the speech signal, we have measured it through the standard Unweighted Average Recall (UAR) metric. We have observed a 15% degradation of the UAR when measured on the anonymized data compared to its measure on original data. The degradation observed for emotion recognition is similar to the degradation observed on the Word Error Rate (that measures the utility), which is fine if one considers the emotion a valuable information to be kept in the anonymized speech signal.

However, one can also consider emotion as personal information that the anonymization system should remove. The preliminary experiments reported in this paper regarding simple random modifications of the F0 values show that such simple modifications are not enough to hide emotional information. Hence, further research will investigate other modifications of the F0 values and modifications of the duration and energy, which are other prosodic parameters that carry emotion information.

5. Acknowledgements

This work was made with the support of the French National Research Agency, in the framework of the project ANR DEEP-PRIVACY (18-CE23-0018) and Région Grand Est. Experiments presented in this paper 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 (see https://www.grid5000.fr).

6. References

[1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Comput. Speech Lang.*, 2019.

[2] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy implications of voice and speech analysis - information disclosure by inference," in *Privacy and Identity Management*, 2019.

[3] 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," in *Proc. Interspeech*, 2019.

[4] E. 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.

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

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

[7] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, 1980.

[8] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, 2005.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[10] L. Devillers, S. Rosset, G. D. Duplessis, M. A. Sehili, L. Béchade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Turker, M. Sezgin, K. E. Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell, "Multimodal data collection of human-robot humorous interactions in the joker project," in *Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xian, China, 2015.

[11] E. Gilmartin and N. Campbell, "Capturing chat: Annotation and tools for multiparty casual conversation." in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, 2016.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, 2008.

[13] M. Macary, M. Lebourdais, M. Tahon, Y. Estève, and A. Rousseau, "Multi-corpus Experiment on Continuous Speech Emotion Recognition: Convolution or Recurrence?" in *22nd International Conference on Speech and Computer SPECOM*, 2020.

[14] Z. Li, L. He, J. Li, L. Wang, and W.-Q. Zhang, "Towards Discriminative Representations and Unbiased Predictions: Class-Specific Angular Softmax for Speech Emotion Recognition," in *Proc. Interspeech*, 2019.

[15] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, and et al., "CEICES : Combining efforts for improving automatic classification of emotional user states: a forced cooperation initiative," in *Language and Technologies Conference*, 2006.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, 2015.

[17] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "Speech Representation Learning for Emotion Recognition Using End-to-End ASR with Factorized Adaptation," in *Proc. Interspeech*, 2020.

[18] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[19] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[20] A. Ranya, H. Hamed, and B. David, "Emotionless: Privacy-preserving speech analysis for voice assistants," in *Privacy Preserving in Machine Learning (CCS19) Workshop*, London, UK, 2019.

[21] F. Bahmaninezhad, C. Zhang, and J. H. L. Hansen, "Convolutional neural network based speaker de-identification," in *Odyssey*, 2018.

[22] W.-C. Huang, H. Luo, H.-T. Hwang, C.-C. Lo, Y.-H. Peng, Y. Tsao, and H.-M. 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*, 2020.

[23] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[24] R. Ueda, R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis," in *Proc. Interspeech*, 2015.

[25] D. Chappell and J. Hansen, "Speaker-specific pitch contour modeling and modification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[26] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," 1993.

[27] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[28] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, and A. e. a. Bellet, "Privacy and utility of x-vector based speaker anonymization," *Transactions on Audio, Speech and Language Processing*, 2021.

[29] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. Interspeech*, 2018.

[30] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, 2011.

[31] M. Tahon and L. Devillers, "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[32] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Post-evaluation analysis for the voice privacy 2020 challenge: Using anonymized speech data to train attack models and asr," 2020.