Privacy Risks in Speech Emotion Recognition: A Systematic Study on Gender Inference Attack

Basmah Alsenani[1,2], *Tanaya Guha*[1], *Alessandro Vinciarelli*[1]

[1]University of Glasgow, United Kingdom
[2]Umm Al-Qura University, Mecca, Saudi Arabia

Abstract

Increasingly more applications now use deep networks to analyse speaker's affective states. An undesirable side effect is that models trained to perform one task (e.g, emotion from speech) can be *attacked* to infer other, possibly privacy-sensitive attributes (e.g., gender) of the speaker. The amount of information an attacker can infer through such attacks is called *leakage*, and this article presents the first systematic study of the interplay between gender leakage and the main characteristics of the attacker model (family, architecture and training condition). To this end, we define various attack scenarios, and perform extensive experiments to analyse privacy risks in Speech Emotion Recognition (SER). Results show that SER models can leak a speaker's gender with an accuracy of 51% to 95% (upper bound) depending on the attack condition. Furthermore, our results provide fresh insights on how to limit the effectiveness of possible attacks and, thereby, to ensure privacy preservation.

Index Terms: speech privacy, speech emotion recognition, inference attack

1. Introduction

Applications that infer users' affective states from speech are increasingly more common [1, 2]. This include personal assistants such as *Siri* and *Alexa*, mobile applications monitoring mental health, and virtual conversational agents or systems for interactive entertainment. This poses a risk to user's privacy because speech signals convey sensitive information about the speakers (e.g., identity and gender). Therefore, approaches designed to detect only affect from speech can be *attacked* to obtain information that should remain private. The term 'attack' here refers to the attempt of inferring privacy-sensitive information from intermediate representations learned by deep models that are originally trained to perform an unrelated task [3, 4, 5].

The majority of approaches dealing with mitigating speech privacy are driven by the need of automatic speech recognition [5]. Our work, however, focuses on Speech Emotion Recognition (SER), because affective aspects of speech are shown to interplay with privacy-sensitive demographic information. Common approaches to preserving privacy in SER include differential privacy [6], noise injection [7, 8] and cryptographic techniques [9]. Differential privacy corresponds to the practice of information sharing where an individual's information can not be distinguished from the group's. Noise injection involves learning to perturb speech without degrading SER accuracy [7]. Traditional cryptographic approaches [9] encode data such that they can be accessed only with an appropriate decoding key. The cryptographic approaches, though provide theoretical guarantees, often have high computational overhead. Another approach uses federated learning [10] - a methodology that distributes data over multiple devices making it difficult to access complete information about the data in use. Replacement autoencoders [8] and adversarial learning [11] have also been used to enhance privacy in SER.

All above efforts focus on the development of new deep models or on using deep networks in a way that preserves privacy. However, no attempt has been made to understand the relationship between the leakage of privacy-sensitive attributes and the major model aspects (family, type, complexity, etc.), training approaches or data conditions. Our current work addresses this gap by *systematically investigating how sensitive demographic information (such as gender) is compromised under various risk scenarios in SER* corresponding to the changes in the model architecture, training and data conditions.

In this work, we present a systematic framework to evaluate and quantify privacy risks in SER. Our experiments are based on an SER model using a Convolutional Neural Network (CNN) that runs on a device transmitting a representation of the data (the output of one of the hidden layers) to the cloud. An attacker *intercepts* this representation to infer the gender of the device's user in *five* different conditions or risk scenarios we propose (see Table 1). These conditions depend on whether the attacker's model family, model architecture and/or training data match with those of the SER model. Consequently, the outcome of these experiments provide insights on the conditions that are more likely to keep gender information private i.e., the conditions at which the gender recognition performance during attacks is lower. Overall, the main **contribution** of this article is twofold:

1. To the best of our knowledge, this is the first work proposing a systematic experimental framework for evaluating and quantifying SER privacy risks in various attack conditions;
2. The results from the experiments provide insights to better inform the development and evaluation of privacy-aware SER models, at least when it comes to gender leakage.

The rest of this article is organized as follows: Section 2 describes the models and various risk scenarios considered in this work, Section 3 presents extensive experiments and results, and Section 4 draws insights from the study.

Table 1: *Conditions for investigating privacy risks in SER*

Scenario	Closeness of the attacker model to the SER model		
	Model family	**Architecture**	**Training data**
Worst case	✓	✓	✓
Scenario 1	✓	✓	×
Scenario 2	✓	×	✓
Scenario 3	✓	×	×
Scenario 4	×	×	×

2. Models and Attack Scenarios

2.1. Primary Task model and Attacker models

The goal of the experiments is to show how the gender leakage varies depending on what an attacker might 'know' about the target SER model, referred to as the *Primary Task model*, $\mathcal{T}$. The model $\mathcal{T}$ here is a Convolutional Neural Network (CNN) (see Fig. 1a) that takes speech as input and outputs the emotion (e.g., *happy*, *sad*) expressed by the speaker. The assumption is that $\mathcal{T}$ is part of a wider system and must transmit the output of one of its hidden layers (an intermediate representation of speech), $\mathbf{h}_{\mathcal{T}}$, over a transmission channel to work. It is such an intermediate representation, $\mathbf{h}_{\mathcal{T}}$, that the attacker takes as input to recognize the gender of the speaker, an information supposed to remain private (see Fig. 1c).

An *Attacker model*, $\mathcal{A}$, can be thought of as a speech-based gender detector. Our work considers two types of attacker models (see Fig. 1b):
(1) **Generic attacker** ($\mathcal{A}_g$): A speech-based gender detector that does not know the goal of the primary task model $\mathcal{T}$. This is trained as a binary classifier to detect speaker's gender.
(2) **Application-aware attacker** ($\mathcal{A}_a$): A speech-based gender detector that knows the primary task of $\mathcal{T}$. This is trained following a *multi-task learning* paradigm that jointly recognizes the gender of the speaker and performs the same task as the model $\mathcal{T}$ i.e., recognition of speech emotion.

2.2. Attack scenarios

Table 1 presents various attack scenarios we address. Each scenario corresponds to a different risk setting based on what information is available to an attacker for the gender inference attack. This essentially means how closely the $\mathcal{A}$ matches various aspects of $\mathcal{T}$: (1) family of the primary task model $\mathcal{T}$ (convolutional network or recurrent), (2) architecture of $\mathcal{T}$ (e.g., VGG16, ResNet50) and (3) the corpus used to train $\mathcal{T}$. The case in which an attacker model matches all aspects of the primary task model is called the **worst case scenario**, because it is expected to provide an *upper bound* on the gender leakage, i.e., the highest possible leakage that can be obtained.

3. Experiments and Results

This section presents details of all experiments performed in this work and the results we obtained.

3.1. Datasets

Our experiments use two corpora: **IEMOCAP** [12] and **RAVDESS** [13]. The first is used to train the primary task model $\mathcal{T}$ and to evaluate the gender leakage under inference attack. RAVDESS is used to train the attacker models $\mathcal{A}_g$ and $\mathcal{A}_a$ when the training data is considered unavailable (for scenarios 1, 3 and 4 in Table 1).

IEMOCAP contains five dyadic sessions (one male and one female speaker) with 10039 utterances in total, and an average speech length of 4.5 sec. In order to keep the experiments comparable with past work, only four emotion categories were used: sad (1084), angry (1103), neutral (1708), and happy (595). The numbers of male and female utterances are 2284 and 2206.

RAVDESS contains 1440 speech utterances with average length of 3.5 sec, for eight different emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The dataset is gender-balanced, with 24 subjects in total. The participants utter identical statements in different emotional expression.

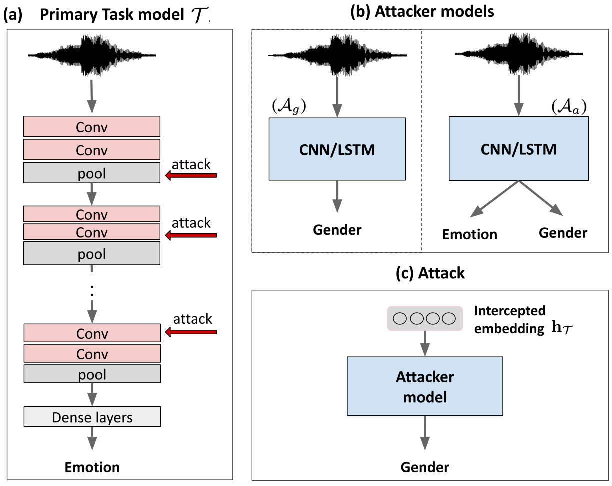

Figure 1: *Overview of the attack scenario: a) The primary task model is a Speech Emotion Recognition (SER) model trained to predict categorical emotion; b) Two types of attacker models we investigate: Generic and Application-aware; c) attacker intercepts an intermediate layer of the SER model to infer gender from the intercepted embedding.*

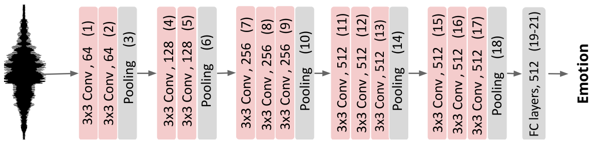

Figure 2: *VGG16 model used as the primary task (SER) Model.*

3.2. Features and Evaluation Metric

Our experiments use two sets of **features**: All convolutional models were trained with *mel-spectrograms* of speech samples using the following parameters: Window length for computing STFT is 30 ms, and the overlap length is 10 ms. The mel-spectrograms are directly fed to the CNNs. For the Long Short Term Memory Network (LSTM) [14] used in scenario 4, we used *Low Level Descriptors* (LLDs). They were extracted using OpenSMILE [15]. They include Energy (1 feature), Mel Frequency Cepstral Coefficients (12 MFCCs), Zero Crossing Rate (1 feature), voicing probability (1 feature), and Fundamental Frequency (1 feature). The features were extracted from 25 ms long windows starting at regular time steps of 10 ms. The delta coefficients of the features were concatenated to the original features, thus resulting into a feature dimension of 32. This feature set was originally designed for the Interspeech 2009 Emotion Recognition Challenge [16], and was shown to be effective not only for the inference of emotions from speech, but also for the recognition of a wide spectrum of other social and psychological phenomena.

The gender leakage, denoted as L, is measured in terms of the effectiveness of the attacker model in recognizing the gender of the speaker taking $\mathbf{h}_{\mathcal{T}}$ as input (see Figure 1). Therefore, the **performance metric** here is the gender recognition accuracy of the attacker models $\mathbf{A}_g(\mathbf{h}_{\mathcal{T}})$ or $\mathbf{A}_a(\mathbf{h}_{\mathcal{T}})$. The further L is from the accuracy of a random classifier (0.49 in Table 2), the more vulnerable model $\mathcal{T}$ is to the gender inference attack.

3.3. Experiments Design

The evaluation of gender leakage is done on IEMOCAP dataset using a 5-fold cross validation. The VGG16 architecture [17] (see Fig. 2) is used as the primary task model $\mathcal{T}$. Its weights were initialized using ImageNet [18] pretraining. The VGG16 model is then trained on IEMOCAP that achieved an accuracy of 0.63 in recognizing four emotion categories (see Section 3.1). The learning rate was set to $\lambda = 10^{-4}$ and the training was performed using Adam optimizer with a cross entropy loss.

The attacker models $\mathcal{A}_g$ are trained to predict only gender (Male or Female). They achieve accuracy between 0.94 and 0.98 on RAVDESS and ranging between 0.93 and 0.98 on IEMOCAP. The models $\mathcal{A}_a$ are trained to predict both gender and emotion. They achieve an accuracy between 0.94 to 0.97 in classifying gender and between 0.56 to 0.60 in classifying 8 emotion classes on RAVDESS. When trained on IEMOCAP, the accuracy of $\mathcal{A}_a$ models ranges from 0.91 to 0.95 for gender and from 0.55 to 0.60 for emotion (4 classes). All CNN-based attacker models (both $\mathcal{A}_g$ and $\mathcal{A}_a$) are initialized with ImageNet pretraining, and trained with a learning rate $\lambda = 10^{-4}$. The LSTM-based attacker models are trained directly on RAVDESS (no pretraining) with a learning rate of $\lambda = 10^{-3}$. All experiments were conducted using Keras 2.11.0 and executed using Google Colab that uses NVIDIA-SMI 510.47.03.

In the *Worst case scenario*, attacker models ($\mathcal{A}_g$ or $\mathcal{A}_a$) have the same architecture (VGG16) and training data as $\mathcal{T}$. For Scenario 1, the attacker models are still VGG16, but trained on RAVDESS, which has different emotion labels than IEMOCAP (relevant to $\mathcal{A}_a$). Scenario 2 varies the architecture of the attackers to VGG19 [17], ResNet50 [19], and DenseNet121 [20], all pretrained on ImageNet. Only the dimensions of the fully connected (FC) layers of ResNet50 were modified to have sizes of 1024 and 4096. Scenario 3 uses the same attacker models as in Scenario 2 but were trained on RAVDESS. In Scenario 4, the attacker model family is changed to a recurrent model (LSTM), which was trained on RAVDESS. The LSTM uses 4 layers with 256, 256, 128 and 64 units (1.06 million parameters). Unlike the convolutional models, the LSTM is fed with LLDs as described in Section 3.2.

Table 2: *Gender Leakage (L) (mean± standard deviation) on IEMOCAP measured in terms of the attacker model's accuracy to infer gender from the intercepted embedding $\mathbf{h}_{\mathcal{T}}$, where $\mathcal{T}$ is the SER model (VGG16). Higher L indicates higher leakage.*

Generic attacker $\mathcal{A}_g$			
Scenario	**Model**	**Training**	L
Worst case	VGG16	IEMOCAP	0.95 ± 0.04
1	VGG16	RAVDESS	0.82 ± 0.06
2	VGG19	IEMOCAP	0.78 ± 0.12
	ResNet50	IEMOCAP	0.59 ± 0.04
	DenseNet121	IEMOCAP	0.53 ± 0.05
3	VGG19	RAVDESS	0.73 ± 0.07
	ResNet50	RAVDESS	0.63 ± 0.03
	DenseNet121	RAVDESS	0.51 ± 0.04
4	LSTM	RAVDESS	0.52 ± 0.04
Application-aware attacker $\mathcal{A}_a$			
Scenario	**Model**	**Training**	L
Worst case	VGG16	IEMOCAP	0.95 ± 0.06
1	VGG16	RAVDESS	0.83 ± 0.07
2	VGG19	IEMOCAP	0.84 ± 0.12
	ResNet50	IEMOCAP	0.55 ± 0.03
	DenseNet121	IEMOCAP	0.52 ± 0.04
3	VGG19	RAVDESS	0.63 ± 0.04
	ResNet50	RAVDESS	0.58 ± 0.04
	DenseNet121	RAVDESS	0.53 ± 0.06
4	LSTM	RAVDESS	0.52 ± 0.04
	Random gender classifier		0.49

3.4. Results and Analysis

With the primary task model $\mathcal{T}$ set as VGG16, we obtain different attacker models by varying their model family, model architecture and training data, in line with the attack scenarios listed in Table 1. For all scenarios in Table 2, input to the attacker models $\mathbf{h}_{\mathcal{T}}$ = output of the first pooling layer of $\mathcal{T}$ (i.e., layer #3 in Fig. 2). Table 2 shows gender leakage L for all attack scenarios for both generic and application-aware cases evaluated on IEMOCAP. Note that all scenarios have gender leakage higher than the random classifier (0.49). The trends are similar for generic and application-aware attacks, with similar upper bound on gender leakage i.e., 95%. Changes in training data reduces the leakage by 12-13%, but leakage is still considerably high. Attack by DenseNet121 results in lower gender leakage than others, most likely due its architecture being fundamentally different from VGG16 used as the primary task model. Note that when VGG19, an architecture similar to $\mathcal{T}$, is used, the leakage is much higher compared to ResNet50 or DenseNet121; in fact, for the case of application-aware attack, this leakage is even higher than that in Scenario 1, despite differences in training condition. This suggests that *attacks tend to be more effective and robust to differences in training condition than that in model architecture*.

3.4.1. Effect of emotion category

Further analysis are considered to investigate if any specific emotion category is prone to more privacy risks than others. Table 3 shows that all emotions are vulnerable (all above random baseline), but *sadness* and *neutral* suffer more gender leakage than others. This suggests that expressions with lower arousal may be more prone to privacy risks, at least when it comes to gender leakage.

3.4.2. Effect of network depth

Another aspect that we analyzed is the effect of network depth on leakage, i.e., the relationship between the layers of the primary task model ($\mathcal{T}$ = VGG16) that is intercepted and the gender leakage. Fig. 2 shows how leakage varies when attacking different layers of $\mathcal{T}$ with different generic attackers $\mathcal{A}_g$ following Scenario 2 in Table 1. Attack by DenseNet121 results in lower leakage than others in general, possibly due the fundamental difference in its architecture with VGG16. In general, the leakage tends to decrease with the depth, and remains above chance. This suggests that while most layers can give away privacy sensitive information such as leakage, the *early layers of a deep model retain more information about speakers' gender, and possibly other attributes* that are not directly related to the

Table 3: *Mean gender leakage (L) per emotion category on IEMOCAP. The Attacker models considered are VGG16 (scenario 1) and ResNet50 (scenarios 2, 3).*

Generic attacker $\mathcal{A}_g$				
Scenario	**Angry**	**Happy**	**Neutral**	**Sad**
1	0.66	0.86	0.87	**0.88**
2	0.50	0.56	0.63	**0.64**
3	0.63	0.59	**0.65**	0.62
Application-aware attacker $\mathcal{A}_a$				
1	0.69	0.86	0.87	**0.89**
2	0.45	0.54	**0.61**	0.57
3	0.59	0.53	**0.62**	0.55
Random classifier	0.50	0.50	0.51	0.50

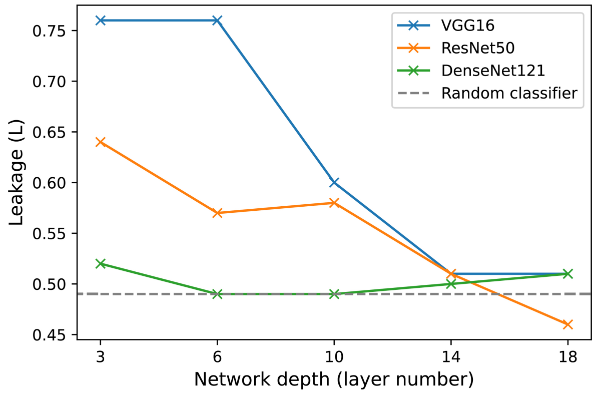

Figure 3: *Effect of network depth on gender leakage (L) measured in terms of gender recognition accuracy of the Attackers ($\mathcal{A}_g$) when different layers of the primary task model (VGG16, see Fig. 2) is intercepted. Early layers are prone to more privacy risks than the later ones.*

main task. Similar trend was observed for the other scenarios.

3.4.3. Effect of feature type

To investigate if feature type affects gender leakage, we conducted experiments using both mel-spectrogram and handcrafted features (see Section 3.3). The primary task model $\mathcal{T}$ was trained separately with two types of features and the same layer $\mathbf{h}_{\mathcal{T}}$ was intercepted by different application-aware attackers. Results in Table 4 show that when $\mathcal{T}$ is trained with LLDs, $\mathbf{h}_{\mathcal{T}}$ incur higher leakage than compared to mel-spectrogram. This is expected because even at input level LLDs are more informative than spectrograms.

3.4.4. Effect of spontaneity

We also examined if spontaneity of speech affects gender leakage in SER, as spontaneity has been shown to significantly effect SER performance [21]. We studied the gender leakage within the improvised and scripted utterances available in the IEMOCAP corpus. Both types of utterances were prone to similar level of privacy risks across all scenarios. We observed *no statistically significant difference* between the spontaneous and scripted utterances.

Table 4: *Gender leakage (L) (mean± standard deviation) comparison for different feature types used to train the primary task model $\mathcal{T}$ on IEMOCAP. Bold fonts are used where the difference is statistically significant.*

Scenario	**Attacker** ($\mathcal{A}_a$)	**Training**	L	
			LLDs	Spectrogram
2	ResNet50	IEMOCAP	**0.62**	0.55
	DenseNet121	IEMOCAP	**0.58**	0.52
3	ResNet50	RAVDESS	**0.64**	0.58
	DenseNet121	RAVDESS	0.51	0.52
4	LSTM	RAVDESS	0.51	0.52

4. Conclusions

To the best of our knowledge, this is the first work proposing a systematic framework to analyse the relationship between the main characteristics of an attacker model (family, architecture and material used for training) and the vulnerability of an SER model (primary task) measured in terms of gender leakage. The key observations and findings of this study are as follows:

1. Gender leakage is fairly similar for both the generic and the application-aware attacks. Therefore, the knowledge of primary task has little importance to gender inference attack.
2. The highest gender leakage is observed when the attacker and the SER models are of the same architecture (i.e., both VGG16). The training data not being available to the attacker, can help limit the leakage, but not substantially.
3. In general, the closer an attacker model is to the primary task model, the higher is the leakage. When the attacker and the primary task model are of very different architectures (e.g., VGG16 and DenseNet121 or LSTM), the difference of training corpora makes little difference. This means that attacks tend to be more effective and robust to differences in training condition than than that in model architecture.
4. The initial layers of the primary task model (at least in the case of the convolutional model used in our experiments) tend to contain more information that is not task-relevant and, therefore, lead to higher leakages when intercepted. This suggests that more on-device processing can help reducing privacy risks.
5. We note that the emotion category also affects gender leakage. Speech expressing *happiness* and *anger show lower gender leakage* compared to speech expressing *sadness* or no emotion (i.e., *neutral*).

The insights above, though obtained in a specific application domain (categorical emotion recognition in speech), are likely to also generalize to other domains, because the families and architectures of the models used in our experiments are among the most common ones. In this respect, future work will try to confirm the observations above by targeting different application areas (such as depression detection) and by further extending the range of models and the attack scenarios used in the current work.

5. Acknowledgement

This research was supported by UKRI grant EP/S02266X/1.

6. References

[1] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology 2020*, vol. 5, no. 1, pp. 96–116.

[2] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion 2022*, vol. 83–84, pp. 19–52.

[3] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Proc. EMNLP 2018*, pp. 11–21.

[4] C. Biswas, D. Ganguly, P. Mukherjee, U. Bhattacharya, and Y. Hou, "Privacy-aware supervised classification: An informative subspace based multi-objective approach," *Pattern Recognition 2022*, vol. 122, p. 108301.

[5] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in asr: Reality or illusion?" in *Proc. Interspeech 2019*, pp. 3700–3704.

[6] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning," in *Proc. Interspeech 2022*, pp. 5055–5059.

[7] T. Feng, H. Hashemi, M. Annavaram, and S. Narayanan, "Enhancing privacy through domain adaptive noise injection for speech emotion recognition," in *Proc. ICASSP 2022*, pp. 7702–7706.

[8] T. Feng and S. Narayanan, "Privacy and utility preserving data transformation for speech emotion recognition," in *Proc. ACII 2021*, pp. 1–7.

[9] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *Proc. ICASSP 2018*, pp. 2057–2061.

[10] S. Bn and S. Abdullah, "Privacy sensitive speech analysis using federated learning to assess depression," in *Proc. ICASSP 2022*, pp. 6272–6276.

[11] M. Jaiswal and E. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proc. AAAI 2020*, vol. 34, no. 05, pp. 7985–7993.

[12] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation 2008*, vol. 42, pp. 335–359.

[13] S. Livingstone and F. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS One 2018*, vol. 13, no. 5.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation 1997*, vol. 9, no. 8, pp. 1735–1780.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. ACM Multimedia 2013*, pp. 835–838.

[16] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech 2009*, pp. 312–315.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR 2015*, pp. 1–14.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE conference on computer vision and pattern recognition 2009*, pp. 248–255.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016*, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[21] K. Mangalam and T. Guha, "Learning spontaneity to improve emotion recognition in speech," *Proc. Interspeech 2018*, pp. 946–950.