MASKING SPEECH CONTENTS BY RANDOM SPLICING: IS EMOTIONAL EXPRESSION PRESERVED?

Felix Burkhardt[1], *Anna Derington*[1], *Matthias Kahlau*[1], *Klaus Scherer*[2], *Florian Eyben*[1], *Björn Schuller,*[1,3,4]

[1]audEERING GmbH, Germany, [2]University of Geneva, Switzerland,
[3]Chair EIHW, University of Augsburg, Germany, [4]GLAM, Imperial College London, UK

ABSTRACT

We discuss the influence of random splicing on the perception of emotional expression in speech signals. Random splicing is the randomized reconstruction of short audio snippets with the aim to obfuscate the speech contents. A part of the German parliament recordings has been random spliced and both versions – the original and the scrambled ones – manually labeled with respect to the arousal, valence and dominance dimensions. Additionally, we run a state-of-the-art transformer-based pretrained emotional model on the data. We find sufficiently high correlation for the annotations and predictions of emotional dimensions between both sample versions to be confident that machine learners can be trained with random spliced data.

Index Terms— speech, emotional, random splicing, anonymization, masking

1. INTRODUCTION

Machine learning models need data to be trained. This data might be artificial laboratory data or natural, real world data. Especially within the domain of emotion processing, real world data has many advantages: firstly, acted emotions rarely occur in the real world [1] and secondly, real world data sets are usually much larger in size.

Nonetheless, one disadvantage of real world data, which is especially relevant in the emotion domain, are potential privacy risks. Emotional expression is often something we do not want to reveal to the public as it might be private and even embarrassing.

This is especially important in the scope of projects like the ECoWeB project [2], which deals with a mobile smartphone app that potentially depressed young adults can use to learn how to deal with negative emotions. One part of the app lets the users record audio files that are analyzed with respect to an emotional estimate and can be sent for monitoring and tuning to a data server. It is important that the users understand that no-one will be able to listen to the data, firstly to keep the data private, but also to avoid users believing that a potential listener might react to inherent messages.

We developed a software to obscure speech content to preserve privacy by "scrambling" or "random splicing" the audio file, by splitting it into segments which then get concatenated in random order, so that the words cannot be understood any more.

We aim at two use cases:

1. anonymizing the speech for privacy protection
2. removing the linguistics to force human annotators of emotional expression to focus on the extra linguistic features and not on the linguistic content, thereby enabling training annotation that is valid across languages

We discuss the literature in the next section and then two experiments conducted to validate the approach: on the one hand, manually labeling the original and the corresponding scrambled data to find differences and on the other hand, by predicting emotional dimensions for the data with a trained machine learning model. For the experiments, we use a database collected from the German parliament recordings. Note that cultural biases may occur in the data because the annotator group is biased towards some culture.

Contributions of this paper are as follows:

- We present a software that automatically random splices audio data to obfuscate the lexical content. To our best knowledge, such a software has not been presented yet.
- We evaluate the approach with respect to the preservation of emotional expression with both manual and machine learning methods.

2. LITERATURE REVIEW

Generally, speaker characteristics that can be detected from the voice fall into two categories: on the one hand, *speaker traits* which do not change fastly, like age, gender, physical size, or state of health and on the other hand *speaker states* which do like emotional expression, fatigue or stress [3]. All of these, and especially the textual content, fall into the user private domain and might need to be protected from misuse.

According to [4], methods to preserve user privacy fall into four categories: deletion, encryption, distributed learning, and anonymization. Deletion can be used when speech is just part of a signal but not crucial for the training, for example in soundscape detection like traffic estimation from sound. Encryption can shelter the signal on its way from user to processor. Distributed, or federated learning is often used to train speech recognition, for example within the COMPRISE project[1] and means that acoustic features get extracted or model parameters updated on the user device and then transferred to a data server. Of course these features or model updates must then make it impossible to reconstruct the original audio signal or model output from them.

Speech anonymization has two aspects: obscuring the speech and obscuring the speaker, meaning the lexical content versus the identity of the speaker. Most of the approaches in the literature are focused on masking the speaker identity while preserving the phonetics, because the main goal is to train automatic speech recognition (ASR) systems, for example via voice conversion [5, 6]. Some approaches preserve speaker identity by substituting sensitive words [7].

There is a voice privacy challenge organized by the university of Avignon in the scope of Interspeech[2] which focused on obfuscating the speaker identity while at the same time keep as many speech characteristics as possible. [8] anonymize the speaker identity by a complete re-synthesis pipeline. Nourtel et al. [9] measure a 15 % degradation of emotion recognition for the standard Voice privacy conversion on IEMOCAP.

In our case, we focus on keeping as much acoustic properties of the speech like prosody, articulation or voice quality as possible, while removing the phonetic structure. Likewise, our primary focus is not to obfuscate the speaker identity, but the spoken content. Approaches that do this in the literature generally rely on the generation of artificial speech with similar features, for example by Generative Adversarial Nets (GANs) [10]. If the speech is not needed at all but a byproduct of public audio recordings, it can of course simply be detected and blurred [11].

3. RANDOM SPLICING

The random splicing algorithm we implemented was inspired by the method Scherer [12] had already described in 1971. Random-Splicing was compared with four different content-filtering techniques in [13].

For processing the digital audio recordings of the German parliament, our random splicing algorithm was first prototyped as a Python notebook and then implemented in C++ as an internal openSMILE [14] component called “AudioScrambler”, which was also used for the ECoWeB project [2]. openSMILE is an open source framework to extract acoustic features from audio.

Basically, our random splicing algorithm is divided into the following two steps: **Segmentation** and **Rearrangement**. Unlike the algorithm described by Scherer [12], no silent pauses of all kinds are removed from the data before segmentation. In addition, the segment length in our algorithm is not fixed, but is determined by configuring a minimum and maximum fragment length within which to cut, so that the intersections of the segments lie within this so-called region of interest (ROI). The intersection point within each ROI is determined using the root-mean-square (RMS) values, which are calculated for each frame of the audio files in the time domain for this purpose. Thus, no resource-intensive STFT calculation is needed. The cut is performed starting from the lowest RMS value within the ROI at the nearest zero-crossing to reduce the potential of pops in the resulting audio signal. This cut point is also the starting point for the next segment.

In the second step, the resulting segments are rearranged in a pseudo-random order by shuffling their list indices, in such a way that no segment is connected to any segment that was already connected there before, unless it is the last remaining segment. This ordering is intended to enhance the masking effect. Optionally, each segment is also reversed with a configurable probability, i. e., played backwards during playback, to further enhance the masking effect in a different way.

In context of this study, segment inversion was not applied in order to keep the unnatural modification of the audio data within certain limits. The configured minimum and maximum fragment lengths were 300 ms and 1000 ms. In the case of the scrambled German parliament recordings, this means that individual words can often still be recognized or guessed, while original partial sentences occur only in small numbers. An informal listening test showed that the rendered audio was incomprehensible.

4. THE DATASET

We tested the approach on a database of German parliament speeches. The database contains data from 9 German politicians. After a manual segmentation the data consists of 1198 segments spoken by the nine politicians. The age span was from 40 to 77 years, with 6 men and 3 women. All segments were then random spliced as described in Section 3 and manually annotated for emotional expression. For reproducibility, the data can be accessed via Zenodo[3]

5. MANUAL EVALUATION

12 annotators employed by audEERING GmbH rated the whole set of original segments with respect to the three dimensions *arousal*, *valence* and *dominance*. 10 annotators rated the random-spliced set, 6 of them the whole set and the remaining 4 a substantial part of the data (33 to 66 %). The annotators were instructed to rate the samples on a scale from -10 to 10 for each dimension. We used the evaluator weighted estimator (EWE) [15] of the labels as ground truth. No offset correction

[1] https://www.compriseh2020.eu/

[2] https://www.voiceprivacychallenge.org/

[3] https://zenodo.org/record/7224678

	PCC	CCC	pairwise t-test
Arousal	.785	.548	$> .001$
Valence	.524	.519	$> .001$
Dominance	.603	.545	$> .001$

Table 1. Results of the manual annotations

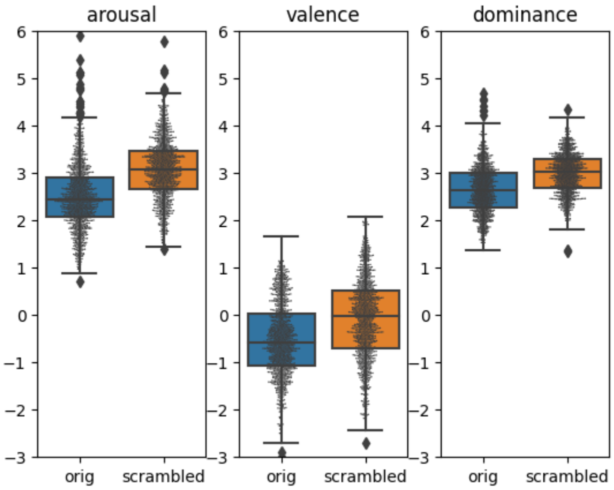

Fig. 1. Box- and swarmplots for arousal, valence and dominance in the orginal and scrambled versions

was performed. In Figure 1, we depict the distributions of the labels for arousal, valence and dominance for random-spliced and original samples, respectively. It can be easily seen that the labels differentiate mainly for the arousal dimension. The majority of the valence labels are negative which is probably due to the domain: politicians speaking in parliament. Arousal and dominance are both clearly on the positive side, which also makes sense with respect to the domain. It seems that arousal also gets overestimated for the random spliced versions.

Table 1 summarizes our statistics to estimate the differences: we computed Pearson's correlation coefficient as well as Concordance Correlation Coefficient and run pairwise T-tests on a 95 % confidence interval. Although all T-tests clearly show that the influence of random splicing on the samples is a highly significant one, the correlation between the corresponding samples is quite high, especially for arousal and least for valence, which tends to be over-estimated for the random spliced samples.

	CCC s_{10}	CCC s_6	PCC
Arousal (original)	.134	.409	.757
Arousal (scrambled)	.101	.405	.748
Valence (original)	.055	.080	.132
Valence (scrambled)	.090	.135	.258
Dominance (original)	.091	.351	.712
Dominance (scrambled)	.075	.319	.656

Table 2. Results of the model predictions

6. MODEL EVALUATION

We evaluate a dimensional SER model on both the original and the scrambled data. The model training follows the procedure described in [16]. We use a model with the wav2vec 2.0 architecture [17] that has been pre-trained on 4 different speech corpora with a total duration of 63 k hours [18]. The number of transformer layers is reduced from 24 to 11, a number which greatly reduces the required computing resources without sacrificing a significant loss in performance. The pruned model is then fine-tuned on the emotional dimensions arousal, dominance and valence by freezing the convolutional layers and re-training only the 11 transformer layers. The embeddings of the last hidden layer are aggregated by an average pooling layer, and then forwarded to a linear output layer and a sigmoid function. The training is run for 3 epochs, with a batch size of 32, a fixed learning rate of $1e-4$, and with a concordance correlation coefficient (CCC) loss. We use the official *train* and *dev* splits of MSP-Podcast Corpus version 1.7 [19] for fine-tuning.

As previously discussed, the annotators of the German politician data set were instructed to rate the segments out of the range $(-10, 10)$. Scaling the data via min-max scaling from the annotator range to the range of $(0, 1)$, (abbreviated as s_{10}), results in a narrow spread of ratings. Another option for the scaling is to look at the actually occurring minimum and maximum annotations. Since no labels below -6 and above 6 occur, we also evaluate the model predictions using min-max scaling from the range $(-6, 6)$ to $(0, 1)$ (abbreviated as s_6).

The model achieves a CCC of 0.745, 0.646, and 0.635 for arousal, dominance and valence on the *test-1* split of the MSP-Podcast version 1.7 dataset. The PCC results (in the same order) on the same data are given by 0.750, 0.668, and 0.637.

7. RESULTS AND DISCUSSION

The correlation of the model prediction with the true labels is shown in Figures 2, 3 and 4, using scaling s_6 for the ground truth. The ranges of the model predictions are generally wider than the scaled politician dataset, especially for valence. In Table 2, we show the model performance in terms of CCC when using scales s_{10} and s_6, as well as PCC, which is scale-invariant. In terms of CCC, for either scaling version, performance is considerably lower than the in-domain model results. Looking at the PCC of arousal and dominance however, comparable results are achieved. On the original sets, the PCC values for arousal, 0.757, and dominance, 0.712 are even better than on the in-domain results on MSP-Podcast. On valence, results are significantly worse in terms of both CCC

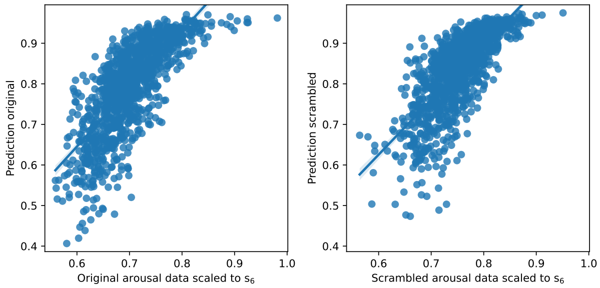

Fig. 2. Correlation of the original and scrambled arousal data scaled to s_6 with the respective model predictions

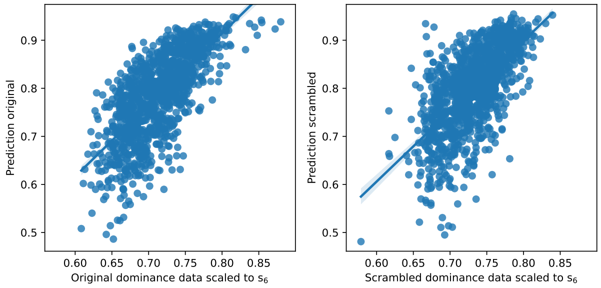

Fig. 3. Correlation of the original and scrambled dominance data scaled to s_6 with the respective model predictions

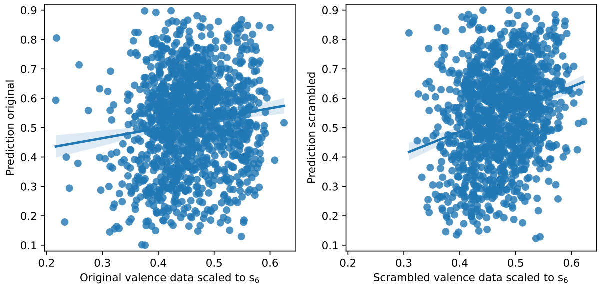

Fig. 4. Correlation of the original and scrambled valence data scaled to s_6 with the respective model predictions

and PCC on the politician datasets, reaching at best a PCC of 0.258 on the scrambled data. For both arousal and dominance, the model predictions are better on the original data than the scrambled data. In contrast, on valence, the model predictions have a higher CCC and PCC than on the original dataset.

As shown in Table 2, the model predictions have a higher CCC and PCC on scrambled data compared to the original data. This can be interpreted in that any linguistic information is no longer perceived by the annotators, and valence is rated solely based on paralinguistic characteristics. Although wav2vec2.0 models that have been pre-trained on large amounts of data have been shown to exploit linguistic data [20], the evaluation model has been trained on English data only. Therefore, it is understandable that the model performs better on the valence task, where the linguistic component has been removed. Still, a comparable value compared to the in-domain results is not achieved for either of the two versions.

8. CONCLUSION

We investigated the influence of random splicing on the emotional expression of German parliament speech, on the hand by an analysis of manually labeled samples and on the other by predicting the emotional dimensions with a pre-trained machine learning model. It has been shown that there are differences between the original and random spliced samples, but not to a degree that would hinder the assessment of emotional expression.

The model prediction did not really work with respect to valence, but this was true irrespective of random splicing and probably due to the language difference between the test data and the pre-trained model. Future investigations could deal with more elaborate splicing algorithms which may be informed by linguistic embeddings in order to less disrupt valence aspects.

9. ACKNOWLEDGEMENTS

This research has been partly funded by the European ECoWeB (Assessing and Enhancing Emotional Competence for Well-Being in the Young) project (Grant Agreement number: 754657) as well as the European MARVEL (Multimodal Extreme Scale Data Analytics for Smart Cities Environments) project (Grant Agreement ID: 975337).

10. REFERENCES

[1] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, no. 1, pp. 33–60, 2003.

[2] A. Newbold, F. C. Warren, R. S. Taylor, C. Hulme, S. Burnett, B. Aas, C. Botella, F. Burkhardt, T. Ehring, J. R. J.

Fontaine, M. Frost, A. Garcia-Palacios, E. Greimel, C. Hoessle, A. Hovasapian, V. E. I. Huyghe, J. Lochner, G. Molinari, R. Pekrun, B. Platt, T. Rosenkranz, K. R. Scherer, K. Schlegel, G. Schulte-Korne, C. Suso, V. Voigt, and E. R. Watkins, "Promotion of mental health in young adults via mobile phone app: study protocol of the ecoweb (emotional competence for well-being in young adults) cohort multiple randomised trials," *BMC Psychiatry*, vol. 20, no. 1, pp. 458, Sep 2020.

[3] Björn Schuller and Anton Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 10 2013.

[4] N. 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," in *Proc. Interspeech 2020*, 2020, pp. 1693–1697.

[5] Hiroto Kai, Shinnosuke Takamichi, Sayaka Shiota, and Hitoshi Kiya, "Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 560–566.

[6] 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," 2019.

[7] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li, "Speech sanitizer: Speech content desensitization and voice anonymization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 2631–2642, 2021.

[8] Sarina Meyer, Florian Lux, Pavel Denisov, Julia Koch, Pascal Tilli, and Ngoc Thang Vu, "Speaker anonymization with phonetic intermediate representations," 2022.

[9] Hubert Nourtel, Pierre Champion, Denis Jouvet, Anthony Larcher, and Marie Tahon, "Evaluation of Speaker Anonymization on Emotional Speech," in *SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication*, Virtual, Germany, Nov. 2021.

[10] Korosh Vatanparvar, Viswam Nathan, Ebrahim Nemati, Md Mahbubur Rahman, and Jilong Kuang, "A generative model for speech segmentation and obfuscation for remote health monitoring," in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2019, pp. 1–4.

[11] Alice Cohen-Hadria, Mark Cartwright, Brian McFee, and Juan Pablo Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.

[12] K. R. Scherer, "Randomized splicing: A note on a simple technique for masking speech content," *Journal of Experimental Research in Personality*, vol. 5, pp. 155–159, 1971.

[13] K. R. Scherer, S. Feldstein, R.N. Bond, and R. Rosenthal, "Vocal cues to deception: A comparative channel approach," *Journal of Psycholinguistic Research*, vol. 14, pp. 409–425, 1985.

[14] Florian Eyben, Martin Wöllmer, and Björn W. Schuller, "opensmile — the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[15] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, 2005, pp. 381–385.

[16] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn W Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[17] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[18] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[19] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 08 2017.

[20] Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," *arXiv preprint arXiv:2204.00400*, 2022.