INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Ensemble-within-ensemble classification for escalation prediction from speech

Oxana Verkholyak[1], *Denis Dresvyanskiy*[2,3], *Anastasia Dvoynikova*[1], *Denis Kotov*[2,3], *Elena Ryumina*[1], *Alena Velichko*[1], *Danila Mamontov*[2,3], *Wolfgang Minker*[2], *Alexey Karpov*[1]

[1]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia
[2]Ulm University, Ulm, Germany
[3]ITMO University, St. Petersburg, Russia

overkholyak@gmail.com, denis.dresvyanskiy@uni-ulm.de, dvoynikova.a@iias.spb.su, preductor@gmail.com, ryumina_ev@mail.ru, alena.n.velichko@gmail.com, danila.mamontov@uni-ulm.de, wolfgang.minker@uni-ulm.de, karpov@iias.spb.su

Abstract

Conflict situations arise frequently in our daily life and often require timely response to resolve the issues. In order to automatically classify conflict (also referred to as escalation) speech utterances we propose ensemble learning as it improves prediction performance by combining several heterogeneous models that compensate for each other's weaknesses. However, the effectiveness of the classification ensemble greatly depends on its constituents and their fusion strategy. This paper provides experimental evidence for effectiveness of different prediction-level fusion strategies and demonstrates the performance of each proposed ensemble on the Escalation Sub-Challenge (ESS) in the framework of the Computational Paralinguistics Challenge (ComParE-2021). The ensembles comprise various machine learning approaches based on acoustic and linguistic characteristics of speech. The training strategy is specifically designed to increase the generalization performance on the unseen data, while the diverse nature of ensemble candidates ensures high prediction power and accurate classification.
Index Terms: paralinguistic analysis, escalation prediction, ensemble classification

1. Introduction

Solving a classification task requires building and evaluating a great number of models before choosing the best one. But how to choose the best model among many competing candidates? Model selection has always been a difficult task with no trivial answer. The most commonly adopted strategy – to choose the algorithm that gives the smallest error on the training (validation) dataset, even when using cross-validation – is often misleading. This poses a risk of choosing a model with particularly poor generalization ability that shows a high accuracy on the validation set and low accuracy on the test set.

One powerful concept that allows us to avoid choosing a particularly poor model is ensemble learning. It combines predictions from several classification algorithms, reducing the impact of each individual learner's mistakes. The choice of ensemble candidates and the fusion strategy greatly influences the outcome of the whole ensemble, and there is no guarantee that the combination will always outperform each individual classifier. Nevertheless, careful integration of multiple learning algorithms has been shown effective in various areas of computational paralinguistics [1, 2, 3].

This article provides experimental evidence for the effectiveness of different prediction-level fusion strategies and demonstrates the performance of each proposed ensemble on the Escalation Sub-Challenge (ESS) in the framework of the Computational Paralinguistic Challenge (ComParE-2021) [4]. We propose an ensemble training strategy that is specifically designed to increase the generalization performance on the task of 3-way recognition of escalation in speech: low, medium and high.

2. Background

The series of Computational Paralinguistics Challenges (ComParE) [5] are organized annually to advance the state-of-the-art performance for speech paralinguistic systems and provide standards for evaluation of such systems. The organizers of the challenge present a variety of baseline features including OpenSmile [6], Bag-of-Audio-words (BoAW) [7], Deep Spectrum [8], auDeep [9], DiFE [4] and introduce other acoustic and linguistic feature representations on a recurrent basis. Each year the baseline systems become stronger and require greater efforts to overcome the baseline performance. The Speech Escalation Sub-Challenge offers a particularly difficult task due to the fact that training and testing samples come from different corpora and therefore are expected to have quite different feature distributions.

Following recent trends in machine learning, speech paralinguistics community has seen the rise of transfer learning approaches to mitigate the lack of training data, which is a pressing problem in many research areas, and the training sample presented for the ESS [10, 11] is no exception. Transfer learning is the transferring of knowledge from a model trained on one problem to a new task. The approach requires less training data and allows neural networks to hit the global minimum much faster. Having emerged from the computer vision field, transfer learning based on image processing of spectrograms has been proven effective in various areas of computational paralinguistics [2, 12]. Such neural network architectures as VGG16 [13], ResNet-50 [14], EfficientNetB3 [15], MobileNetV2 [16] pre-trained on the ImageNet dataset [17] have been in use for some time [18]. Recently, the first large-scale pre-trained audio neural networks (PANNs) [19] were introduced and successfully used for ComParE challenge 2020 [20].

Different from image-based approaches, audio models often suffer from necessity of preprocessing audio sequences via zero-padding to align audio sequence length to model input shape, which includes undesirable bias in data. In addition, when preprocessing long audios, researchers always should neatly choose the length and the step of windows on which

http://dx.doi.org/10.21437/Interspeech.2021-1821

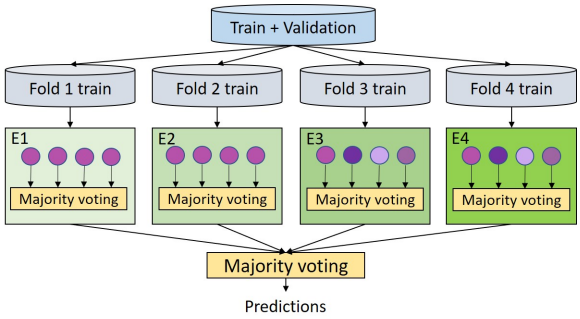

Figure 1: *Ensemble-within-ensemble training approach. Left-most systems (E1, E2) represent homogeneous ensembles, right-most (E3, E4) - heterogeneous.*

the audio sequence will be cut. Moreover, cut chunks of audio may lose a piece of information contained in other chunks (therefore it is better to consider all the utterance, even if it is cut on chunks). Lin and Busso solved such problem [21] by applying chunk-based segmentation procedure, which dynamically varies the shift of the cut window, thus splitting audio sequences into prior defined number of chunks (windows) with fixed duration (length of a window). All what users of such system need is specifying the maximal sequence duration and the length of chunks.

3. Proposed approach

We design a system for ComParE 2021 Escalation Speech Sub-Challenge [4] that mitigates the negative effects of model selection based on train/development data split. This is achieved by first, training the system on various data samples, and second, allowing for several estimators to vote for the final prediction and thus avoid choosing the model that performs particularly poor on the test set. The details of the proposed approach are described below.

3.1. Ensemble-within-ensemble approach

To assure high generalization performance on the new data we specifically build the system to be trained using ensemble-within-ensemble approach on a combined train + development dataset. All the available data is split into 4 cross-validation folds such that the distribution of classes and number of words in speech transcriptions in each fold are similar to the original distributions. On each fold, we train an ensemble of various classifiers (acoustic, linguistic and mixed) with a majority voting strategy. The best candidates for each ensemble are chosen experimentally by brute-forcing all possible combinations and ensemble sizes. Additionally, we look for the best model to resolve the ties. Therefore, the number of trials are $\prod_{i=0}^{N-1} N * C(N-1, i)$, where N is the total number of models considered, and $C(n, r)$ is the number of possible combinations of r models out of n models. At the end, we obtain 4 ensembles that in turn vote for the final prediction. The ties during the last voting procedure are regulated by hand-crafted rules. These ensembles can be homogeneous (containing same estimators in each fold) or heterogeneous (containing different estimators in each fold). The pipeline of the training process is depicted in Figure 1.

3.2. Ensemble candidates

The efficiency of an ensemble is guaranteed by the multiformity of its constituents. Therefore, we prepared diversified classifiers that take advantage of various acoustic and linguistic speech representations to make sure we cover multiple facets of the phenomena of interest. The list of ensemble candidates is obtained from the pool of independently trained estimators that are drawn from both traditional ML approaches and Neural Network architectures. The proposed features and classifiers are described below.

3.2.1. Language-based systems

Dutch is a low-resource language, so the extracted linguistic features from Dutch transcriptions may be unrepresentative. To solve this problem, all transcriptions were translated into English and German using the Google-cloud-translate library[1]. Before the text vectorization, all data was preprocessed by removing the punctuation and stop-words, tokenization and lemmatization. For all languages (Dutch, English and German), the following features were extracted: Bag-of-Words, Word2Vec [22], FastText [23], BERT [24], and ELMO [25]. For the last 4 methods we used pretrained models. Additionally, we used tonal-dictionary-based features as proposed in [26]. The length of the obtained feature vectors varied as following: tonal dictionaries - 19, Word2Vec[2] - 300 and 100, FastText[3] - 300, BERT[4] - 768, ELMO[5] - 1024. With the use of the Bag-of-Words method, the length of the feature vector changed depending on whether the stop-words were deleted or not: Dutch without deleting stop-words - 1061, with deleting stop-words - 985, English - 895 and 800, German - 1122 and 1020, respectively. Totally 37 language-based models were tested (12 Dutch, 12 English, 12 German, 1 mixed).

We used Logistic Regression (LR) to model the proposed linguistic features. For each feature type, the optimal regression parameters were selected using 4-fold cross-validation. The tuning parameters included the regularization parameter C, the weights balancing parameter, the parameter responsible for adding a constant to the decision function, the type of loss: binary or multinomial, the norm used in the penalization, and the solver type that defines the algorithm of the optimizer.

3.2.2. Audio-based systems

The audio-based ensemble candidates were obtained by using two types of modelling: traditional ML approach and Neural-Network-based approach using transfer learning.

SVM and LR classifiers were used with the baseline features (OpenSmile, BoAW, DeepSpectrum), except for auDeep, which wasn't included due to its relative poor performance on the test set, as reported by the challenge organizers. Additionally, Principle Component Analysis (PCA) was used together with OpenSmile and DeepSpectrum feature representations to reduce the dimensionality and decorrelate the features. Z-normalization was applied to all the features before fitting the SVM. Additionally, Random Forest Classifier was trained on openSMILE acoustic features provided by the organizers. The most effective parameter found during the fine-tuning of the model was the number of trees in the Random Forest equal

[1]https://pypi.org/project/google-cloud-translate/
[2]https://wikipedia2vec.github.io/wikipedia2vec/pretrained/
[3]https://fasttext.cc/docs/en/pretrained-vectors.html
[4]https://github.com/maknotavailable/pytorch-pretrained-BERT
[5]https://github.com/HIT-SCIR/ELMoForManyLangs

10000. To measure the quality of a split we used parameter 'entropy' for the information gain.

A Convolutional Neural Network (CNN) with PANN architecture [27] pre-trained on the large-scale AudioSet dataset [28] was used together with Log-scaled Mel-Spectrograms (LMS) extracted at 64 Mel-scaled filter banks and a maximum frequency of 14 kHz. These features were generated on the fly from the raw audio data during training process. For the purpose of fine-tuning on the challenge data we added one more fully-connected layer to the initial architecture. The data augmentation technique known as mixup [29] with parameter alpha equal to 1 was tested and achieved higher results on the development set. The training was performed using the Adam optimizer with a learning rate of 0.002.

Another CNN model with VGG16 [13] architecture pre-trained on the ImageNet dataset [17] was used with Mel-Frequency Cepstral Coefficients (MFCC) and LMS features. The choice of the model architecture comes from preliminary experiments on the ESS data that show on average 7.39% advantage as compared to other model architectures, such as Resnet-50, EfficientNetB3 and MobileNetV2. The MFCC features were extracted at 80 Mel-coefficients, a maximum frequency of 8 kHz and a type III Discrete cosine transform. The LMS features were extracted at 128 Mel-scaled filter banks and a maximum frequency of 8 kHz. Before starting the training process, we removed the last classification layer into 1000 classes from the pre-trained VGG16 and added three fully-connected layers for 512, 256 and 3 neurons, with random dropout of neurons by 50% after 1 and 2 fully-connected layers. The training was performed on 50 epochs using the Adam optimizer with a learning rate of 0.000005. We also applied inversely proportional to class frequencies class weighting and random affine transformations (shift in width and horizontal flip). Linearly normalized pixels of images with a resolution of 224 × 224 were fed to the input of the network.

We also constructed a Long Short-Term Memory (LSTM-LSTM) network following the idea of aggregation the temporal information on utterance level [21]: the first two LSTM layers independently process every chunk using time-distributed modification, outputting the last hidden states. Combining last hidden states of every chunk, we obtain matrix of features with shape $N_c \times N_n$, where N_c is the number of chunks and N_n is the number of neurons in the last layer of the first-level LSTM. Next, obtained feature matrix is fed into 2 further LSTM layers (each has 256 neurons), and the final fully-connected layer with 3 neurons outputs the probabilities of class categories. According to the lengths of audios in ESS challenge, we chose the length of chunks to be equal to 0.5 seconds and the maximal length of audio to be 12 seconds, hereby obtaining from every audio (no matter, how long it is) 24 chunks. Next, for every chunk we extracted 128 Mel-cepstral coefficients with 512 ms window length and 256 ms window shifting. Thus, every chunk has become a feature matrix with the shape 32 x 128.

3.3. Ensemble combination rules

3.3.1. Majority voting

The majority voting is a procedure of choosing the most frequent element in a list, which has been successfully applied as a fusion strategy in classification ensembles [3]. However, since the test set is collected in-the-wild and the ties happen frequently, we defined the following hand-crafted rules for breaking the ties: if the tie happens between the extreme opposite classes, choose the middle; otherwise, choose the less represented class (which is the class with higher level of escalation).

3.3.2. Weighted majority voting

The majority voting strategy has a big disadvantage - it does not take into account the uncertainty degree of ensemble members. If the classifier has class probabilities such as [0.4, 0.3, 0.3], the first class will be chosen; however, the classifier uncertainty is very high - it cannot strongly agree on any class.

To eliminate this drawback, we apply the Model and Class-based Weighted Fusion (MCWF), which assigns an "importance" weight to every ensemble member in each class. The weights are generated according to Dirichlet distribution as a $C \times L$ matrix, where C is the number of classes and L is the number of members in an ensemble; the sum of all weights distributed among ensemble members within each class is equal to 1. The final predicted class is the one with the maximum value of the sum of weighted probabilities.

4. Experiments and Discussion

The experiments were conducted in several stages. First, we evaluated the cross-validation (CV) performance in terms of Unweighted Average Recall (UAR) of each individual system outlined in the previous section independently. Next, we experimented with unimodal ensembles and assessed their effectiveness on the test set. After that, we built a mixed ensemble combining best acoustic and linguistic models and also evaluated its performance on the test set. Finally, we investigated different fusion strategies, such as simple majority voting and weighted majority voting, and examined performance of homogeneous ensembles vs. heterogeneous ensembles. The detailed results and analysis follow below.

4.1. Single system CV experiments

The individual performance of the best ensemble candidates in a set of CV experiments is shown in Table 1. The results of poorly performing linguistic methods are omitted for brevity.

Table 1: *Average 4-fold CV performance of the individual proposed systems, UAR (%)*

Acoustic	UAR	Linguistic	UAR
PANN	78.70	German BERT	52.04
VGG-16	76.38	German W2V	45.52
LSTM-LSTM	70.36	Dutch FT	44.23
Random Forest	70.70	Tonal Dictionary	41.77

The table reveals that audio-based systems have a decisive superiority compared to language-based systems. This is easily explained by the nature of speech utterances, which are often too short to contain any meaningful words. Moreover, the analysis of unique words in data shows that more than 60% of the words in test set are not seen during the training. This makes language-based ensemble candidates highly unreliable. The best single acoustic model turned out to be the fine-tuned PANN architecture with 78.70% average CV performance.

4.2. First ensembles and test trials

The challenge allows for 5 blind-test trials. The first 3 test trials were dedicated to homogeneous ensembles with simple majority voting strategy. The ensemble candidates were drawn from

the pool of proposed approaches and baseline systems. The ties inside the inner ensembles (E1, E2, E3, E4 in Figure 1) were resolved by the strongest ensemble candidate, which was experimentally determined during the CV experiments. The mean UAR performance on cross-validation against ensemble size (number of candidates within ensemble) is depicted in Figure 2. As can be seen from the Figure 2, the optimal number of candidates in an ensemble varies for audio-based and language-based systems, however it remains in the range of 4-6 estimators with a notable decrease in performance with further increase of ensemble size.

The best possible mean CV UAR 93.18% was achieved with a mixed ensemble of 4 systems: 2 acoustic (proposed VGG-16 architecture and baseline BoAW) and 2 linguistic (baseline DiFE plain-BlAt and DiFE sent-BlAt). However, this system was expected to perform poorly on the test set due to the fact that systems trained on DiFE features showed mediocre performance (45.2% and 47.2%) as reported by challenge organizers. Therefore, the ensembles that were tested on the first 3 test trials were the following: 1) Acoustic - 3 proposed systems (PANN, VGG-16, Random Forest) and baseline DeepSpectrum-based SVM; 2) Linguistic - 2 proposed systems (German BERT, Dutch FT) and 2 baseline models (DiFE plain-BlAt, DiFE sent-BlAt); 3) Mixed - 3 proposed audio-based methods (PANN, VGG-16, LSTM-LSTM) and 2 language-based models (proposed German BERT and baseline DiFE sent-BlAt). The results are shown in the first 3 rows of the Table 2.

Table 2: *Classification performance (UAR, %) of the proposed ensembles. H - Heterogeneous, W - weighted fusion (MCWF)*

Ensemble	CV	Test
Acoustic	80.77	58.5
Linguistic	72.70	43.6
Mixed	85.89	49.9
H(Acoustic)	87.35	56.7
W(Acoustic)	85.05	59.2
Baseline [4]		59.8

Different from the expected, the test set performance of the mixed ensemble was lower than the acoustic ensemble alone. This proves that linguistic modality is less informative in the given circumstances and all further experiments were conducted using audio modality only.

4.3. Choosing fusion strategy

The 4th and 5th trials were dedicated for different ensemble combination strategies. First, we allowed the inner ensembles (E1, E2, E3, and E4 in Figure 1) to have different estimators at each training iteration with the hope to better fit the data and diversify the ensemble. The CV performance of such ensembles is higher than the relative performance of the homogeneous ensembles, however it turned out to generalize poorly on the test set, reaching only 56.70%, presumably due to overfitting.

For the last submission, we applied MCWF to class probabilities of the acoustic-only ensemble. It consisted of our proposed LSTM-LSTM, PANN, and VGG-16 models and baseline BoAW, DeepSpectrum, and OpenSmile systems. As earlier, we formed 4 ensembles based on CV and generated weights for each of them separately (each time weights were generated 10000 times and the best ones were chosen). As an example, the generated weights for the ensemble of the third CV Fold are

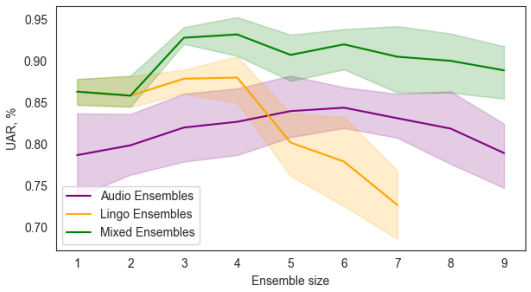

Figure 2: *Cross-validation performance vs. number of estimators in homogeneous ensemble networks*

presented in Figure 3. Such an approach outperformed all our former submissions, reaching UAR of 59.20 % on the test set.

Analyzing generated weights, it is interesting to see, that in every fold certain models are responsible for a particular class. For instance, Figure 3 shows that in case of class 2 LSTM-LSTM and PANN models have the highest contribution, while for class 1, this role performs DeepSpectrum model. Due to space limitations, we do not show the weights of all four ensembles. However, it is important to note that the weights distribution differs significantly across the folds.

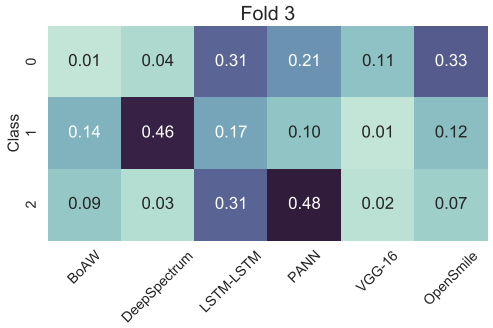

Figure 3: *Model and class weights used to generate final predictions of an ensemble trained on the 3rd fold during training*

5. Conclusions

Building a classification ensemble is an art that requires careful investigation of ensemble candidates and their combination strategies. We proposed acoustic, linguistic, and mixed ensembles with a number of prediction fusion approaches and evaluated their efficiency on the Escalation Speech ComParE Sub-Challenge 2021. Even though the test set performance on an external corpus has not exceeded the baseline, the cross-validation results on the training and validation data indicate high performance of 93.18% in terms of UAR for the 3-way escalation prediction from speech. We hope the presented experimental findings will serve as guidelines for future practices in ensemble learning research field.

6. Acknowledgements

This study was partially supported by the Russian Foundation for Basic Research (projects No. 20-37-90144 and 19-29-09081) and the Russian state research (No. 0073-2019-0005).

7. References

[1] G. Gosztolya and R. Busa-Fekete, "Ensemble bag-of-audio-words representation improves paralinguistic classification accuracy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 477–488, 2020.

[2] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," in *Proceedings INTERSPEECH*, 2020, pp. 2072–2076. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2666

[3] J. Szep and S. Hariri, "Paralinguistic classification of mask wearing by image classifiers and fusion," in *Proceedings INTERSPEECH*, 2020, pp. 2087–2091. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2857

[4] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, M. R. Leon J. J. Zwerts, J. Treep, and C. Kaandorp, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," in *Proceedings INTERSPEECH*, 2021, to appear.

[5] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings INTERSPEECH*, 2020, pp. 2042–2046. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-0032

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH*, 2013, pp. 148–152.

[7] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proceedings INTERSPEECH*, 2016, pp. 495–499. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1124

[8] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings INTERSPEECH*, 2017, pp. 3512–3516. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-434

[9] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[10] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "An audio-visual dataset of human–human interactions in stressful situations," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 29–41, 2014.

[11] I. Lefter, L. J. Rothkrantz, and G. J. Burghouts, "A comparative study on automatic audio–visual fusion for aggression detection using meta-information," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1953–1963, 2013.

[12] Z. Ren, J. Han, N. Cummins, and B. W. Schuller, "Enhancing Transferability of Black-Box Adversarial Attacks via Lifelong Learning for Speech Emotion Recognition Models," in *Proceedings INTERSPEECH*, 2020, pp. 496–500. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1869

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[17] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[18] S. Amiriparian, M. Gerczuk, S. Ottl, L. Stappen, A. Baird, L. Koebe, and B. Schuller, "Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–11, 2020.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[20] T. Koike, K. Qian, B. W. Schuller, and Y. Yamamoto, "Learning higher representations from pre-trained deep models with data augmentation for the compare 2020 challenge mask task," in *Proceedings INTERSPEECH*, 2020, pp. 2047–2051. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1552

[21] W.-C. Lin and C. Busso, "An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks," in *Proceedings INTERSPEECH*, 2020, pp. 2322–2326. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2636

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, vol. 2, 2013, p. 3111–3119.

[23] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, 2018, pp. 4171–4186.

[25] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[26] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *Proceedings INTERSPEECH*, 2020, pp. 2097–2101. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-3160

[27] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[28] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings ICASSP*, 2017, pp. 776–780.

[29] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.