The role of task and acoustic similarity in audio transfer learning: insights from the speech emotion recognition case

Andreas Triantafyllopoulos, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, and Björn W. Schuller. 2021. "The role of task and acoustic similarity in audio transfer learning: insights from the speech emotion recognition case." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6-11 June 2021*, edited by Dimitri Androutsos, Kostas Plataniotis, Xiao-Ping (Steven) Zhang, and Boulgouris Nikolaos, 7268–72. New York, NY: IEEE. https://doi.org/10.1109/icassp39728.2021.9414896.

THE ROLE OF TASK AND ACOUSTIC SIMILARITY IN AUDIO TRANSFER LEARNING: INSIGHTS FROM THE SPEECH EMOTION RECOGNITION CASE

Andreas Triantafyllopoulos[1,2], *Björn W. Schuller*[1,2,3]

[1]audEERING GmbH, Gilching, Germany
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

ABSTRACT

With the rise of deep learning, deep knowledge transfer has emerged as one of the most effective techniques for getting state-of-the-art performance using deep neural networks. A lot of recent research has focused on understanding the mechanisms of transfer learning in the image and language domains. We perform a similar investigation for the case of speech emotion recognition (SER), and conclude that transfer learning for SER is influenced both by the choice of pre-training task and by the differences in acoustic conditions between the upstream and downstream data sets, with the former having a bigger impact. The effect of each factor is isolated by first transferring knowledge between different tasks on the same data, and then from the original data to corrupted versions of it but for the same task. We also demonstrate that layers closer to the input see more adaptation than ones closer to the output in both cases, a finding which explains why previous works often found it necessary to fine-tune all layers during transfer learning.

Index Terms— Speech emotion recognition, transfer learning, representation learning

1. INTRODUCTION

Deep learning (DL) approaches have gained significant prominence in latter years. One of the most commonly accepted explanations for their effectiveness is that deep neural network (DNN) architectures learn generic representations that are transferable across different tasks [1, 2]. This has led to a wide amount of literature on leveraging past information to improve performance and increase convergence on new tasks and/or data sets, most recently with the advent of self-supervised learning [3, 4].

However, this technique does not always yield better performance, resulting in the well-documented effect of "negative transfer" [5, 6]. This raises the question of when transfer learning is successful, and, specifically, to which extent it is influenced by the juxtaposition between pre-training and downstream tasks, input features, architecture type, and the different data sets at play. To this end, Neyshabur *et al.* [7] investigated transfer learning in the visual domain and posited that downstream tasks benefit from pre-training both because the models are learning transferable high-level features, and because they learn low-level statistics. These observations can be attributed to the notion of compositionality that is extensively studied in the vision domain [8].

A lot of prior work has focused on learning generalisable representations in the audio domain as well, either using supervised [9, 10, 11], unsupervised [12, 13, 14], or self-supervised approaches [15, 16, 17, 18]. The topic is becoming increasingly relevant with the rise of numerous applications where data is scarce, e. g., in the medical domain [19].

In this work, we focus on speech emotion recognition (SER) as the downstream application, an area that is receiving considerable attention in the community, but for which big data is not yet widely available [20]. Although DNNs are already outperforming traditional approaches [21], that is not true for all tasks and data sets [22]. This has led the community to adopt transfer learning approaches, starting from feature-based [23] and recently moving to DL approaches [24, 25, 26]. Hence, understanding how transfer learning works could lead to the design of more powerful algorithms that unlock the full potential of DL for SER, and other low-resource audio tasks.

Our main contribution lies in disentangling the effects of the pre-training task from those of acoustic mismatches between the respective data sets; two factors we expect to play a big role in a successful transfer. To this end, we first utilize three different data sets and tasks for pre-training to illustrate the relative importance of both factors in an actual application. We then exploit the fact that our SER data set has been annotated for multiple, distinct emotional schemes with different degrees of similarity between them. This allows us to isolate the effect of task similarity by training on one scheme and transferring knowledge to another on the same data set. Finally, we isolate the effect of acoustic similarity by transferring knowledge from clean to corrupted versions of the same data. Our experiments show that while both factors are important, it is primarily the lack of task similarity that leads to negative transfer, whereas even extreme acoustic divergence can be overcome.

Moreover, we make the surprising observation that layers closer to the input are more susceptible to adaptation, a finding that could explain why authors in prior works have found it necessary to fine-tune all layers rather than the last ones [10, 17]. This is an important finding as fine-tuning more layers adds an overhead to optimisation since more parameters need to be adapted.

2. ARCHITECTURE

Our experiments are carried out using the *Cnn14* architecture recently introduced by Kong *et al.* [10]. It has been trained for the task of audio tagging on AudioSet [27]. The authors have open-sourced their code and trained weights[1]. We use the 16 kHz variant, because the data sets we use also come in 16 kHz. As features, we used log-Mel spectrograms computed with 64 Mel bins, a window size of 32 ms, and a hop size of 10 ms.

Cnn14 follows the VGG architecture design [28]. After the last convolution layer, the features are pooled across the feature dimensions using both max and mean pooling, and subsequently fed into two linear layers. Dropout with a probability of 0.2 is applied after every second convolution layer. The architecture is shown in Figure 1.

3. DATA

As a *downstream* data set for transfer learning, we use MSP-Podcast (v1.7) [29], a recently-introduced data set for SER. It is split in speaker independent partitions:

- a training set consisting of 38 179 segments
- a development set made of 7 538 segments, collected from 44 speakers (22 male – 22 female)
- a 12 902 segment test set, consisting of 60 speakers (30 male – 30 female)

MSP-Podcast has been annotated for the emotional dimensions of *arousal*, *valence*, *dominance*, as well as for 8 emotional categories, plus an extra *other* category. In the present work, we focus on the emotional dimensions. These have been annotated on a 7-point Likert scale on the utterance level, and scores by individual annotators have been averaged to obtain a consensus vote. Similar to other approaches in the literature [30, 31], we bin the continuous values to a 3-point scale. We use the following mapping:

- *low*: [1-3]
- *mid*: (3-5]
- *high*: (5-7]

This results in a heavily unbalanced distribution, with 9% of the data in the low range, 67% in the mid range, and 24% in the high range.

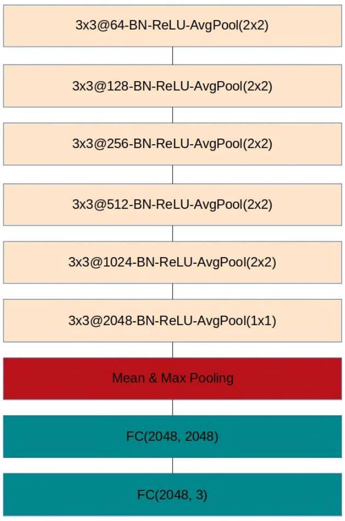

3x3@64-BN-ReLU-AvgPool(2x2)
3x3@128-BN-ReLU-AvgPool(2x2)
3x3@256-BN-ReLU-AvgPool(2x2)
3x3@512-BN-ReLU-AvgPool(2x2)
3x3@1024-BN-ReLU-AvgPool(2x2)
3x3@2048-BN-ReLU-AvgPool(1x1)
Mean & Max Pooling
FC(2048, 2048)
FC(2048, 3)

Fig. 1: A schematic of the *Cnn14* architecture. “@” designates the number of feature maps in each convolution block, and is preceded by the kernel size. Each block consists of two identical convolution layers, followed by average pooling, rectified linear unit (ReLU), and batch normalisation (BN). All filters are applied with a stride of 1.

These dimensions are well-validated constructs used to define affect [32], and have been shown to manifest through different acoustic cues [33]. Although they quantify different aspects of emotional expression, they are not completely unrelated to one another. Their similarity can be measured by the Pearson correlation between the three emotional dimensions. On the MSP-Podcast training set, arousal shows a 0.2412 correlation to valence, and a 0.7953 correlation to dominance. This is indicative of the degree of task similarity across those three schemes, a fact we exploit in our transfer learning experiments.

We also make use of three additional data sets for pre-training, which we will refer to as *upstream* data sets: *AudioSet* [27] which the original authors show helps for SER, *VoxCeleb1* [34], where we pre-train for speaker identification as it has been shown to translate well to SER [26], and *IEMOCAP* [35], which is annotated for the same emotional dimensions. Due to space limitations, we refer to the original publications for a detailed description of each data set.

4. EXPERIMENTS

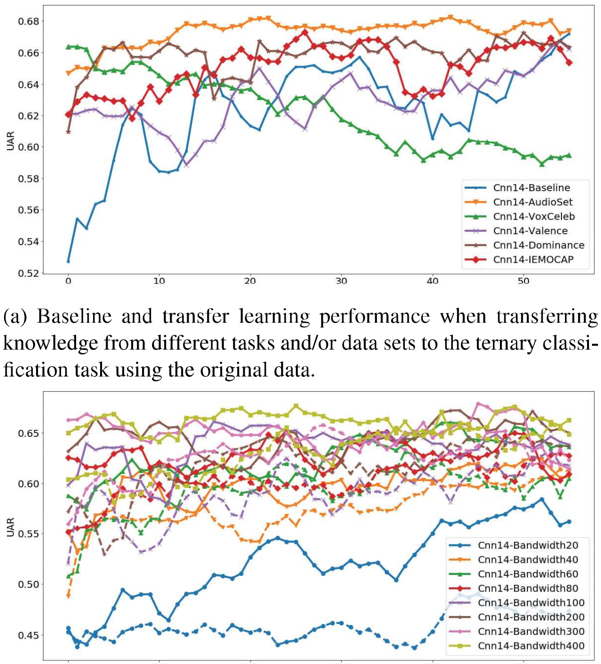
We perform three groups of experiments:

- training on different data sets; fine-tuning on MSP-Podcast for the arousal task
- training on different tasks on MSP-Podcast; fine-tuning on arousal
- training on clean arousal data; fine-tuning on bandlimited arousal data

[1] https://github.com/qiuqiangkong/audioset_tagging_cnn

Table 1: Unweighted average recall (UAR)% results on MSP-Podcast for ternary arousal classification on the utterance level. In parentheses, we show the epoch on which it was achieved, defined as the one where peak performance was reached on the validation set. Column name refers to the data set and task on which the network was pre-trained prior to fine-tuning on the arousal task. Random refers to training the model from random initialization. We report performance after re-training either all layers, or just the linear ones (last two).

Initialisation	**Random**	**IEMOCAP (arousal)**	**AudioSet (audio tagging)**	**VoxCeleb1 (speaker identification)**	**MSP-Podcast (valence)**	**MSP-Podcast (dominance)**
All layers	68.76 (58)	67.95 (16)	65.47 (28)	67.36 (6)	62.24 (56)	67.97 (13)
Linear only	-	61.15 (23)	58.62 (2)	62.45 (9)	44.61 (53)	65.29 (20)



(a) Baseline and transfer learning performance when transferring knowledge from different tasks and/or data sets to the ternary classification task using the original data.

(b) Baseline (dashed) and transfer learning (continuous) performance when transferring knowledge from clean to corrupted data for the arousal classification task.

Fig. 2: Test set UAR results computed on each epoch.

Unless otherwise mentioned, all experiments were run using a standard stochastic gradient descent (SGD) optimiser with a constant learning rate of 0.001, Nesterov momentum of 0.9 [36], and a batch size of 8. The networks were trained for a total of 60 epochs. We only show results for the epoch that yielded the best performance on the validation set. In order to deal with the imbalance in target labels, we use a balanced non-negative likelihood loss (NLLoss), obtained by multiplying each term with the inverse of the frequency of the corresponding class in the training set.

We first train a standard baseline by training the model from a randomised initialization. This gives us 68.76% UAR after 58 epochs of training. We refer to this model as *Cnn14-Baseline*.

For all transfer learning experiments, we try two variants:

- Fine-tuning all layers: with this experiment, we are interested in seeing how the network adapts to new data.
- Fine-tuning only the linear layers: with this experiment, we are interested in seeing how the network is able to leverage the features learnt during pre-training.

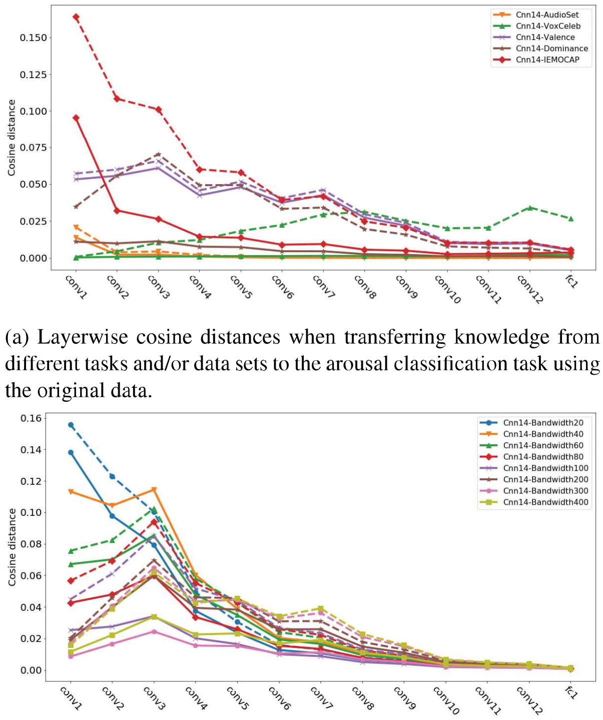
In order to measure how the model is adapting to the new task, a good proxy is the distance between the weights of the layers before and after training, as shown by Neyshabur *et al.* [7]. To this end, we use the *cosine distance*:

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \, \|y\|}. \tag{1}$$

Our first experiments are performed using models pre-trained on the upstream data sets mentioned in Section 3. For AudioSet, we use the weights released by Kong *et al.* [10] and refer to this pre-trained model as *Cnn14-AudioSet*. For VoxCeleb1, we train the network for 100 epochs using a standard NLLoss, and select the checkpoint that gives the best accuracy on the validation set (64% on epoch 92), a model we will refer to as *Cnn14-VoxCeleb*. Finally, for IEMOCAP, we train the network for 60 epochs, and select the checkpoint that gives the best UAR on the validation set (64% on epoch 33). We use the same binning and train/dev/test split as Zhang *et al.* [31]. We refer to this model as *Cnn14-IEMOCAP*.

In order to disentangle the effects of task and acoustic differences between the upstream and downstream tasks, we make use of the fact that MSP-Podcast has been annotated for three different emotional attributes, as described in Section 3. We thus pre-train one model on valence, and another on dominance, which we refer to as *Cnn14-Valence* and *Cnn14-Dominance*, respectively, and fine-tune them for arousal. During pre-training on the original tasks, the valence model reaches a peak performance of 63.56% UAR on epoch 52, whereas the dominance model reaches 54.87% on epoch 33.

Finally, we are interested in systematically studying the effects of acoustic similarity isolated from the effects task similarity. We simulate different degrees of acoustic similarity by passing the data through narrow bandpass filters that



(a) Layerwise cosine distances when transferring knowledge from different tasks and/or data sets to the arousal classification task using the original data.

(b) Layerwise cosine distances when transferring knowledge from clean to corrupted data for the arousal classification task.

Fig. 3: Layerwise cosine distance for the best performing (continuous) and last checkpoints (dashed) compared to their respective initial states.

remove a large part of the spectrum, whereas task similarity is kept constant by transferring knowledge to and from the arousal classification task. We use 4th order Butterworth filters with a central frequency of 500 Hz, and test out the following frequency bandwidths: [20, 40, 60, 80, 100, 200, 300, 400].

5. RESULTS AND DISCUSSION

Results on the effectiveness of transfer learning are mixed: although we were not able to surpass baseline performance when using the original data, Table 1 and Figure 2a both show that pre-training accelerates convergence. In addition, fine-tuning all layers was always more beneficial to fine-tuning only the linear ones, a finding consistent with previous work [10, 17]. This is also in accord with the trend exhibited by layerwise cosine distance in Figure 3a. Earlier layers see more adaptation that latter ones with respect to their initialization, with the trend exacerbated by further training for all models except *CNN14-VoxCeleb*.

Our main focus is on distinguishing between the effects of task and acoustic similarity. Initial experiments using different upstream data sets shown in Table 1 illustrate that both effects are at play. The fact that *Cnn14-AudioSet* performs worse than *Cnn14-IEMOCAP* and *Cnn14-VoxCeleb* for arousal classification indicates that task similarity is more important than acoustic similarity; although AudioSet is bigger and presents more acoustic diversity than the other two data sets, the audio tagging task is also substantially different from arousal classification.

Experiments on knowledge transfer from different labelling schemes to arousal on MSP-Podcast further illustrate the importance of task similarity. Pre-training on valence, which shows substantially less correlation to arousal than dominance, also results in substantially worse performance and takes longer to converge. The relative importance of task similarity vs acoustic similarity is also highlighted by a direct comparison between *Cnn14-Valence* and *Cnn14-IEMOCAP*. The latter is pre-trained on overall less data, coming from a different data set recorded in a very narrow set of conditions, but on the same task. Nevertheless, it performs substantially better than the former which was pre-trained on the same data, but for a different task.

Finally, our experiments on transferring knowledge to corrupted versions of the same data but for the same task (arousal), show that successful knowledge transfer, while dependant on the degree of acoustic mismatch, is possible. UAR results in Figure 2b demonstrate that starting from pre-trained networks results in better performance, faster convergence, and better initialisation, even though the corrupted acoustic signals are fundamentally different from the original. As expected, a wider bandwidth leads to overall better performance, both with, and without pre-training. Weight-space distances shown in Figure 3b show that fine-tuning primarily affects the earlier layers, an indication that these layers are also responsible for adapting to changes in the acoustic conditions between the upstream and downstream data sets.

6. CONCLUSION

We have experimentally shown that transfer learning for SER depends on both acoustic and task similarity, with the latter being the deciding factor. Results show that the wrong choice of task can be detrimental to transfer learning performance. In addition, we have ascertained that layers closer to the input are subject to more adaptation than those closer to the output; a fact explaining why numerous previous works show better performance when fine-tuning all layers rather than simply the last ones.

These findings should guide the design of architectures and pre-training strategies for SER. As shown, different emotional dimensions can lead to fundamentally different representations, a fact that makes the quest for a universal representation very challenging. In order to obtain a holistic emotional characterisation of a speech segment, we need representations that can generalise across several distinct tasks, and that should be reflected on the pre-training regiment.

7. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[5] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[6] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proc. CVPR 2019*, 2019, pp. 11 293–11 302.

[7] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" *arXiv preprint arXiv:2008.11687*, 2020.

[8] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *International Conference on Algorithmic Learning Theory*, Springer, 2011, pp. 18–36.

[9] A. Diment and T. Virtanen, "Transfer learning of weakly labelled audio," in *2017 ieee workshop on applications of signal processing to audio and acoustics (waspaa)*, IEEE, 2017, pp. 6–10.

[10] 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," *arXiv preprint arXiv:1912.10211*, 2019.

[11] T. Koike, K. Qian, Q. Kong, M. D. Plumbley, B. W. Schuller, and Y. Yamamoto, "Audio for audio is better? an investigation on transfer learning models for heart sound classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 74–77.

[12] 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.

[13] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[14] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019, pp. 146–150. DOI: 10.21437/Interspeech.2019-1473. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1473.

[15] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *Proc. Interspeech 2019*, pp. 161–165, 2019.

[16] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *Proc. ICASSP 2020*, IEEE, 2020, pp. 7694–7698.

[17] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. ICASSP 2020*, IEEE, 2020, pp. 6419–6423.

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[19] S. Amiriparian, M. Schmitt, S. Ottl, M. Gerczuk, and B. Schuller, "Deep unsupervised representation learning for audio-based medical applications," in *Deep Learners and Deep Learner Descriptors for Medical Applications*, Springer, 2020, pp. 137–164.

[20] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[21] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP 2016*, IEEE, 2016, pp. 5200–5204.

[22] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" *Proc. Interspeech 2018*, pp. 147–151, 2018.

[23] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. ICASSP 2016*, IEEE, 2016, pp. 5800–5804.

[24] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, 2013, pp. 511–516.

[25] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. ICASSP 2019*, IEEE, 2019, pp. 7390–7394.

[26] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. ICASSP 2020*, IEEE, 2020, pp. 7169–7173.

[27] 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 *Proc. ICASSP 2017*, IEEE, 2017, pp. 776–780.

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

[29] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/TAFFC.2017.2736999.

[30] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang, "Improving emotion classification through variational inference of latent variables," in *Proc. ICASSP 2019*, IEEE, 2019, pp. 7410–7414.

[31] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. ICASSP 2019*, IEEE, 2019, pp. 6705–6709.

[32] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[33] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[35] 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," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[36] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML 2013*, 2013, pp. 1139–1147.