



Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?

Johannes Wagner, Dominik Schiller, Andreas Seiderer, Elisabeth André

Human-Centered Multimedia, Augsburg University, Germany

name@hcm-lab.de

Abstract

In the past, the performance of machine learning algorithms depended heavily on the representation of the data. Well-designed features therefore played a key role in speech and paralinguistic recognition tasks. Consequently, engineers have put a great deal of work into manually designing large and complex acoustic feature sets. With the emergence of Deep Neural Networks (DNNs), however, it is now possible to automatically infer higher abstractions from simple spectral representations or even learn directly from raw waveforms. This raises the question if (complex) hand-crafted features will still be needed in the future. We take this year's INTERSPEECH Computational Paralinguistic Challenge as an opportunity to approach this issue by means of two corpora – Atypical Affect and Crying. At first, we train a Recurrent Neural Network (RNN) to evaluate the performance of several hand-crafted feature sets of varying complexity. Afterwards, we make the network do the feature engineering all on its own by prefixing a stack of convolutional layers. Our results show that there is no clear winner (yet). This creates room to discuss chances and limits of either approach.

Index Terms: Computational Paralinguistics, Deep Neural Networks, Hand-crafted Features, End-to-end Learning

1. Introduction

Machine learning deals with the problem of designing clever algorithms that allow a computer to learn from and make predictions on data. Since the representation of the data defines the "playground" on which an algorithm operates, it plays a key role for success or failure [1]. In paralinguistic computation finding an optimal set of the most important acoustic features was therefore declared as the "holy grail" [2]. Yet, the huge number of possible features that can be extracted from a speech signal make it a challenging task. And in fact, during the last decade we have seen increasingly large feature sets. The baseline set of this challenge, for instance, has grown from initially 384 attributes [3] to more than 6k [4]. Sets of this kind achieve considerable results across various recognition tasks like emotion, affect, and personality [5]. As a downside, usually only a subset of the parameters is actually relevant to a specific task.

In the recent years, the availability of massive labeled data sets along with increased computational power and improved algorithms, have helped Deep Learning (DL) to a breakthrough in various fields like object detection or speech recognition [1]. In paralinguistic computation, Deep Neural Networks (DNNs) now provide an alternative to Support Vector Machines (SVMs), which had dominated the field for many years. Their use, however, is not limited to the categorization of sound samples based on pre-extracted features. DNNs are even able to discover a suited representation of the data itself – a problem known as Representation or Feature Learning [6]. The advantage is obvious: instead of putting effort into the development

of hand-crafted features (which may or may not prove itself in practice), we make the machine do the job of finding the "holy grail".

Against this background and in view of the rapid progress we currently see in DL applications, we can naturally ask what role hand-crafted feature sets will play in paralinguistic computation in the future. In the following we want to approach this question by comparing the performance of manually designed features versus automatically learned representations within the INTERSPEECH Computational Paralinguistic Challenge (ComParE), namely the Atypical Affect and Crying sub-challenge [7].

2. Related Work

In 2004 several research groups (including our lab) started a cooperation under the named CEICES (Combining Efforts for Improving automatic Classification of Emotional user States) [8]. Amongst other things, the cooperation had set itself the task of finding an optimal set of the most important independent features for emotional speech recognition. Extensive testing on base of more than 4k features showed a comparable performance of the examined feature types [2]. Therefore, it became common practice to extract rather more than too little features. Nowadays, open-source tools like Emovoice [9] and OpenSMILE [10] extract thousands of acoustic parameters. The official feature set of ComParE, for instance, bundles more than 6k features parameterizing a speech chunk in terms of voice quality, loudness, harmonicity, spectral sharpness, pitch and many others [4]. An attempt for a more compact, yet generic feature set is the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [11]. It comprises 56 parameters (88 in the extended version), which have been selected based on their proven value in former studies, as well as, their theoretical significance. Tests show that GeMAPS achieves results comparable (sometimes superior) to the performance of larger sets.

At the Eating Condition sub-challenge in 2015, Milde and Biemann [12] applied Deep Learning (DL) to predict the type of food a speaker is eating and achieved a 15 % improvement over the baseline. Particularly interesting here is that – whereas the baseline was computed on basis of 6k features – the authors computed only a 40-dimensional spectrogram, which they fed into a Convolutional Neural Network (CNN). Although, a spectrogram is still a hand-crafted feature set, it has been used in sound analysis since the 1970s¹ and does not take any task-specific knowledge into account. Systems that go straight from a simple spectral representation (or the raw signal as we will soon see) to the target class are called *end-to-end*². The key

¹The breakthrough of spectral analysis began in 1965 with the discovery of the fast Fourier transform by Cooley and Tukey.

²Note that the term *end-to-end* is relatively loosely defined and depending on context can mean both: learning from the raw audio or from

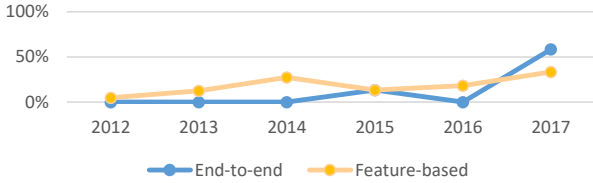


Figure 1: Percentage of DL systems presented at ComParE between 2012 and 2017. End-2-end systems use raw or a simple spectral representation as input. Feature-based systems a complex feature set like ComParE.

innovation here is that the classifier discovers a suited representation of the data by itself.

The finding that DNNs have the potential to learn robust feature representations [13, 14] has led to breakthroughs in various audio classification tasks such as speech recognition [15], sound event classification [16] and automatic music transcriptions [17]. Actually, the step of transforming the audio signal into frequency space can be skipped, too. SoundNet [18], for instance, applies a student-teacher training procedure to transfer visual knowledge from a large number of videos into the sound modality. The learned features can outperform state-of-the-art features in classifying acoustic scenes. Like in visual processing, the ability to learn from raw input is achieved through a stack of convolutional layers. The networks proposed by Wei and colleagues [19] consist of up to 34 of such layers and match the performance of models using spectral features in an environmental sound recognition task [20].

In light of the above, it is not surprising that the popularity of end-to-end systems has increased tremendously. A trend well reflected by the ComParE challenge: last year, for the first time, more DL systems opted for an end-to-end approach rather than using a complex feature set (see Figure 1). Yet, we need to realize that – just like there is not ‘the’ set of hand-crafted features – there is not ‘the’ network (yet). The DL system proposed by Gosztolya and colleagues [21], for instance, significantly improved the fusion baseline at last year’s Cold sub-challenge. Applied to the Addressee task, however, it did not even surpass the standard baseline (ComParE set + SVM). Instead of learning a new representation from the audio itself, Amiriparian et al. [22] extract spectrograms, which they pass into networks trained for image classification and use the activations of the fully-connected layers as features. This yielded an improvement of almost 10 % at last year’s Snore sub-challenge. Looking at this year’s baselines [7] we can see that the performance of the proposed end-to-end approach (End2You [23]) stays behind the other systems in all sub-challenges. AuDeep [24], another DL approach, performs better, but only beats the other systems on the Heartbeat dataset. Such diverse results make it desirable to estimate under which conditions end-to-end learning seems promising and when it is better to draw on a conventional feature set.

A systematic comparison of learned versus hand-engineered features in the context of emotional speech recognition is presented in [25]. In their work, Trigeorgis and colleagues opt for a time-continuous prediction of arousal and valence with Long-Short-Term Memory (LSTM) networks and let features learned by a CNN compete against two standard sets, namely eGeMAPS [11] and a simplified version of ComParE [4]. On both dimensions, the end-to-end approach

a spectral decomposition, usually a spectrogram.

Table 1: Overview of the feature sets considered in this study ordered by complexity (LLD: low-level descriptors, SSF: supra-segmental features). Features are input to the classifiers mentioned in the last row (RNN: Recurrent Neural Network, SVM: Support Vector Machines). Frame step and window size in ms.

		LLD				SSF		
		lin	mel	voc	map	cmp	map	cmp
dim	64	64	17	24	65	88	6373	
frame	20	20	10	10	10	-	-	
win	40	40	25	60	60	-	-	
					RNN			
					SVM			

showed a significantly better performance in comparison to the manually engineered features.

In the work at hand, we follow a similar approach, but within a discrete recognition task. Also, to gain a deeper understanding under which conditions either approach performs better, we compare results from two different datasets and test a wider range of frame- and chunk-based features.

3. Methodology

3.1. Datasets

We run our study on the Atypical Affect and Crying sub-challenge [7]. The two databases are well qualified for our study as they differ greatly in terms of size and content. Atypical contains 9:10 h (6 h for training) speech from disabled individuals split into more than 10k files labeled as one of four basic emotional classes (anger, fear, happiness, sadness). Crying, on the other hand, contains 2:50 h (1:30 h for training) of vocalisations from infants and roughly half as many chunks classified into three categories (neutral/positive, fussing, and crying). Segment length varies greatly in both datasets and ranges from less than a second up to a minute. For details please refer to [7].

3.2. Inputs

The study at hand approaches the question if hand-crafted features are still relevant for paralinguistic tasks. To answer this, we use different data representations in our experiments. On the one hand, we learn directly from raw audio waveforms (*raw*). On the other hand, we investigate hand-crafted feature sets of varying complexity on frame- and turn-level. In the simplest form we apply a basic spectral decomposition³ (spectrogram) by mapping the short-term power spectrum on a linear (*lin*) or a Mel scale (*mel*). More complex are those sets that take *Low-level Descriptors* (LLDs) like loudness or pitch into account. Here we consider the (still relatively simple) Vocalization set (*voc*), which was used in the Social Signals sub-challenge in 2013 [4], as well as, the advanced, yet minimalistic eGeMAPS set (*map*) [11] and the large ComParE set (*cmp*) [7]⁴. Finally, we also apply the latter two on turn-level to extract *Supra-segmental Features* (SSFs). Table 1 gives a summary of the feature sets. Note the range of input dimensions from less than 100 to more than 6k.

³We use the ‘`scipy.signal.spectrogram`’ routine [v1.0.0] and clip amplitudes below -75 dB.

⁴We compute the three sets with OpenSMILE [10] using the provided standard scripts, but in case of LLDs exclude deltas, as we count on the ability of recurrent neural networks to model the temporal dependencies.

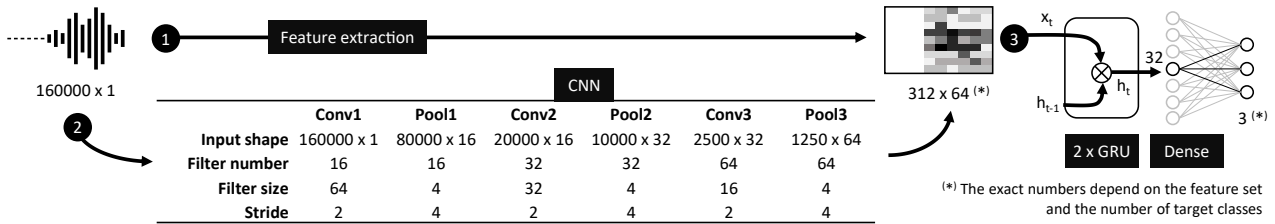


Figure 2: DL system: Audio files are cut or zero-padded to have an equal length of 10 seconds. The network takes as input pre-extracted features (1) or raw waveforms. The latter are processed by a three-layer convolution network (2). The result of either operation is then passed through two Gated Recurrent Units (GRU) layers followed by a full-connected layer (3). The output layer has a size equal to the number of classes.

Table 2: Results on the development set for supra-segmental features. C: Complexity parameter of SVM.

C	Crying		Atypical	
	map	cmp	map	cmp
10^{-6}	0.679	0.707	0.434	0.336
10^{-5}	0.679	0.752	0.434	0.377
10^{-4}	0.729	0.750	0.419	0.281
10^{-3}	0.762	0.710	0.421	0.285
10^{-2}	0.744	0.732	0.444	0.308

3.3. Architecture

Due to the different granularities, we use two classifiers: one that accepts raw audio input or a sequence of LLDs, and one to work with SSFs. For the latter, we stick to the procedure suggested by the challenge organizers. That is, we train a Support Vector Machine (SVM) classifier with WEKA [26] (see [7]).

For all other tests and to have a fair comparison, we apply the same two-layer recurrent neural network (RNN) architecture to learn the temporal structure of the input sequences. For a unified input length, we cut files to an equal length of 10 seconds⁵. The setup is similar to the end-to-end baseline system [7], however, we reduce the number of GRUs to 32 while at the same time we increase the batch-size to 32 (we found that this had a positive effect in terms of stability and performance in our experiments). In case of raw audio, we extend the RNN with three convolutional layers (CRNN). Here, we rely on the configuration proposed by SoundNet [18], i.e. each convolutional layer is followed by batch normalization, rectified linear activation units, and max-pooling.

To train the network we use Adam optimization [27] with a learning rate of 0.001 and a momentum term of 0.9. As loss function we rely on weighted cross entropy. Finally, we scale all inputs to the interval $[-1..1]$ (limits determined on the training data). The described architecture is implemented in Tensorflow [v1.4]. See Figure 2 for an illustration.

4. Results

In the following, we present results measured in terms of Un-weighted Average Recall (UAR). On the Crying dataset, instead of a LOSO validation, we hold out three subjects as a validation partition. In case of Atypical, we stick to the splitting suggested by the organizers. To compensate for underrepresented classes, we apply upsampling in all cases.

Table 2 lists results for the supra-segmental feature sets. We see that *map* outperforms *cmp* on both corpora, but particularly

⁵Shorter files are zero-padded from the front.

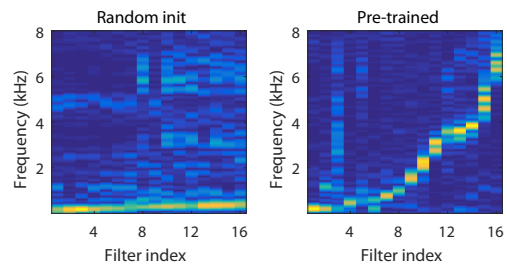


Figure 3: Magnitude responses of the filters of the first convolutional layer after training on the Crying corpus: when randomly initialized (left), when pre-training was applied (right). Filters are sorted by peak response.

on Atypical (by almost 7 %).

Table 3 lists results of the DL approach with respect to the number of training epochs. Training is stopped after 20 epochs. Since it takes a couple of epochs until a stable prediction is reached, we omit values for the first six epochs. Taking the average across a whole epoch we observe a maximum at 15 epochs (last column). To compare results of individual inputs at a glance, averaged values across all epochs are given, too (last row). Here, we can see that on the Crying corpus *lin* achieves the highest mean score (0.793), followed by *cmp* (0.787). Other hand-crafted features, as well as, *raw* audio perform significantly worse. The best individual accuracy is achieved with *cmp* (0.818). Interestingly, we have a quite different picture on Atypical. Here, raw audio clearly surpasses the other inputs in average (0.461) and also yields the best individual accuracy (0.478). Among the LLDs, again *lin* gives the best mean performance (0.415).

5. Discussion

To begin with, there is no clear winner among the examined inputs. A hand-crafted feature set (*cmp*) wins on Crying, while learned features from *raw* audio wins on Atypical. Interestingly, both winners perform rather poorly on the other corpus. Comparing the performance of hand-crafted features, we can say that more complexity does not necessarily lead to better results. On both corpora the simple spectral representation (*lin* and *mel*) show an equally good or better performance compared to the advanced sets. There is one exception, though: on Atypical the supra-segmental *map* set outranks other feature sets by at least 3 %. We explain this with the fact that GeMAPS was especially designed for affective speech tasks [11].

Of course, we would like to know, why the learned fea-

Table 3: Results on the development set for end-to-end and low-level descriptors with respect to the number of training epochs.

Epoch	Crying							Atypical					Mean	
	raw	raw*	lin	mel	voc	map	cmp	raw	lin	mel	voc	map		cmp
7	0.714	0.749	0.803	0.765	0.691	0.699	0.755	0.455	0.371	0.432	0.315	0.418	0.373	0.580
8	0.718	0.755	0.798	0.773	0.718	0.690	0.804	0.456	0.390	0.390	0.314	0.393	0.365	0.582
9	0.713	0.749	0.806	0.765	0.715	0.705	0.769	0.459	0.405	0.433	0.335	0.386	0.367	0.585
10	0.737	0.762	0.798	0.778	0.694	0.729	0.780	0.451	0.419	0.406	0.337	0.392	0.366	0.588
11	0.727	0.762	0.803	0.783	0.708	0.738	0.788	0.442	0.416	0.400	0.331	0.400	0.352	0.588
12	0.721	0.780	0.799	0.782	0.718	0.739	0.758	0.464	0.426	0.409	0.348	0.388	0.354	0.591
13	0.728	0.767	0.790	0.785	0.728	0.733	0.806	0.473	0.419	0.400	0.353	0.394	0.376	0.596
14	0.719	0.796	0.794	0.781	0.692	0.732	0.805	0.472	0.430	0.395	0.363	0.395	0.360	0.595
15	0.745	0.764	0.798	0.781	0.741	0.741	0.791	0.459	0.418	0.399	0.354	0.400	0.380	0.598
16	0.737	0.780	0.780	0.785	0.725	0.736	0.791	0.456	0.415	0.396	0.356	0.417	0.368	0.596
17	0.700	0.790	0.786	0.784	0.719	0.753	0.791	0.468	0.428	0.398	0.351	0.404	0.362	0.595
18	0.752	0.783	0.785	0.789	0.718	0.723	0.818	0.467	0.421	0.377	0.347	0.406	0.369	0.597
19	0.714	0.805	0.779	0.790	0.707	0.746	0.773	0.453	0.428	0.369	0.352	0.392	0.378	0.591
20	0.715	0.710	0.779	0.771	0.725	0.721	0.791	0.478	0.421	0.378	0.331	0.417	0.358	0.584
Mean	0.724	0.768	0.793	0.779	0.714	0.728	0.787	0.461	0.415	0.399	0.342	0.400	0.366	

Table 4: Summary and results on test set (baseline results for Crying with LOSO).

		Crying		Atypical	
		Devel	Test	Devel	Test
<i>Own Approaches</i>					
raw* raw	RAW	0.752	0.678	0.478	0.420
	LLD	0.806	0.675	0.430	0.365
cmp map	LLD	0.818	0.708	0.418	0.388
	SSF	0.762	0.729	0.444	0.411
<i>Baseline Approaches</i>					
End2You	RAW	-	0.635	0.418	0.280
AuDeep	LLD	0.744	0.711	0.404	0.356
OpenXBOW	SSF	0.769	0.732	0.405	0.413
OpenSMILE	SSF	0.756	0.719	0.378	0.431
Fusion	-	-	0.746	-	0.434

tures work well on Atypical, but fail on Crying. Since the latter corpus is about four times smaller, a plausible explanation is that there is not enough data to properly train the convolutional layers. Especially, if the network fails to response to high frequencies this could be problematic since crying sounds are often high-pitched. If this is the case, we should be able to improve the model by finding better filter weights. Hoshen et al. [28] showed that convolution layers can be manually configured to compute a Mel scale and after training their network with 400 hours of speech, it had even learned a similar representation when the filters were randomly initialized. To quicken the learning process, we decided to apply an approach suggested by Tax and colleagues [29]: we force the CNN layers to learn the output of a spectrogram transformation. To avoid over-fitting, we downloaded 6 hours of baby and infant cry from AudioSet⁶. After pre-training the CNN layers with the data, we keep the weights and continue with the normal training procedure on our target corpus. The plots in Figure 3 show that this improves the sensitivity of the network to higher frequencies. As suspected, this also leads to better recognition rates. The new results are listed in Table 3 under *raw**.

Finally, we would like to discuss performance on the challenge test set⁷, which is always a good indicator for the generalizability of the individual approaches. It also allows us to incorporate the results of the baseline systems into our conclusions. Looking at Table 4 we can see that OpenXBOW and

⁶<https://research.google.com/audioset/>

⁷To predict labels on the test set we train our network for 15 epochs on the full set (training and development).

OpenSMILE are the only systems, whose performances do not drop on the test set. This may suggest that supra-segmental features can better generalize to unseen data. Also notable is that our end-to-end system significantly outperforms End2You (especially on Atypical) even though the networks share a similar architecture. This may be for two reasons: we have reduced the number of units in the RNN layers and exchanged the frontal part of the network with the architecture proposed by SoundNet. The actual winner of the challenge, however, is the fusion system, which seems to profit from the synergy of hand-crafted features and learned representations.

6. Conclusion

In this paper, we have approached the question if – now that end-to-end learning is becoming more and more popular – there is still a place for hand-crafted features in paralinguistic computation. We believe a discussion on this topic is of great importance as it may directly influence the direction we want to shift our preferences in the future. To this end, we have investigated feature sets of varying complexity on frame- and chunk level and compared the results with an end-to-end system that learns the data representation directly from the raw waveforms. Experiments have been conducted on two corpora containing emotional speech and infant crying.

Our results show that there is no clear winner amongst the tested inputs. Perhaps, learning from spectrograms provides a reasonable middle way. In fact, with respect to the hand-crafted feature sets we found that more complexity did not necessarily yield better results. On the other hand, supra-segmental features generalized better to the test sets. While learning from raw audio outperformed hand-crafted features on the emotional corpus, it performed worse on the smaller crying database. However, the difference disappeared when the layers were pre-trained on spectrograms we had extracted from another dataset. This shows that feature engineering can still help improve the robustness of end-to-end systems. In the future, we plan to increase the amount of training data in an effective and unsupervised fashion by using generative adversarial networks [30, 31].

The diverse results suggest that we are not yet at a stage where we could afford to cast away hand-crafted features. Especially, in tasks with sparse data it may still be safer to rely on conventional feature sets. Given the rapid progress of deep learning, however, we can expect to see advances in automatic feature learning. Hence, on the long run the classic feature engineer will probably become more of a network architect.

7. References

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [2] A. Batliner, S. Steidl, B. W. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, “Whodunnit - searching for the most important feature types signalling emotion-related user states in speech,” *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.
- [3] B. W. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in *Proceedings of Interspeech*, Brighton, United Kingdom, 2009, pp. 312–315.
- [4] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, “The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism,” in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 148–152.
- [5] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013.
- [6] 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, Aug. 2013.
- [7] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, B. Harald, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. B. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, “The INTERSPEECH 2018 Computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats,” in *Proceedings of Interspeech*, Hyderabad, India, 2018.
- [8] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, “Combining efforts for improving automatic classification of emotional user states,” in *Proceedings of IS-LTC*, Ljubljana, 2006, pp. 240–245.
- [9] T. Vogt, E. André, and N. Bee, “Emovoice - A framework for online recognition of emotions from voice,” in *Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT*, Kloster Irsee, Germany, 2008, pp. 188–199.
- [10] F. Eyben, F. Weninger, F. Groß, and B. W. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the ACM Multimedia Conference, MM*, Barcelona, Spain, 2013, pp. 835–838.
- [11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transaction on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [12] B. Milde and C. Biemann, “Using representation learning and out-of-domain data for a paralinguistic speech task,” in *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 904–908.
- [13] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2009, pp. 1096–1104.
- [14] N. Jaitly and G. E. Hinton, “Learning a better representation of speech soundwaves using restricted boltzmann machines,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Prague, Czech Republic, 2011, pp. 5884–5887.
- [15] A. Mohamed, G. E. Dahl, and G. E. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.
- [16] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, “Robust sound event classification using deep neural networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, Mar. 2015.
- [17] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 5, pp. 927–939, May 2016.
- [18] Y. Aytaç, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.
- [19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [20] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2015.
- [21] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, “DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification,” in *Proceedings of Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3522–3526.
- [22] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, “Snore sound classification using image-based deep spectrum features,” in *Proceedings of Interspeech*, Stockholm, Sweden, 2017, pp. 3512–3516.
- [23] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, “End2you—the imperial toolkit for multimodal profiling by end-to-end learning,” *arXiv preprint arXiv:1802.01115*, 2018.
- [24] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, “audeep: Unsupervised learning of representations from audio with deep recurrent neural networks,” *arXiv preprint arXiv:1712.04382*, 2017.
- [25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Shanghai, China, 2016, pp. 5200–5204.
- [26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009.
- [27] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [28] Y. Hoshen, R. J. Weiss, and K. W. Wilson, “Speech acoustic modeling from raw multichannel waveforms,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4624–4628.
- [29] T. M. S. Tax, J. L. D. Antich, H. Purwins, and L. Maaløe, “Utilizing domain knowledge in end-to-end audio processing,” *arXiv preprint arXiv:1712.00254*, 2017.
- [30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [31] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, “Speech-based diagnosis of autism spectrum condition by generative adversarial network representations,” in *Proceedings of the International Conference on Digital Health*. ACM, 2017, pp. 53–57.