Interspeech 2018
2-6 September 2018, Hyderabad

What Do Classifiers Actually Learn? A Case Study on Emotion Recognition Datasets

Patrick Meyer, Eric Buschermöhle, Tim Fingscheidt

Institute for Communications Technology
Technische Universität Braunschweig
38106 Braunschweig, Germany
{patrick.meyer, t.fingscheidt}@tu-bs.de

Abstract

In supervised learning, a typical method to ensure that a classifier has desirable generalization properties, is to split the available data into training, validation and test subsets. Given a proper data split, we typically then trust our results on the test data. But what do classifiers actually learn? In this case study we show how important it is to analyze precisely the available data, its inherent dependencies w.r.t. class labels, and present an example of a popular database for speech emotion recognition, where a minor change of the data split results in an accuracy decrease of about $55\,\%$ absolute, leading to the conclusion that linguistic content has been learned instead of the desired speech emotions.

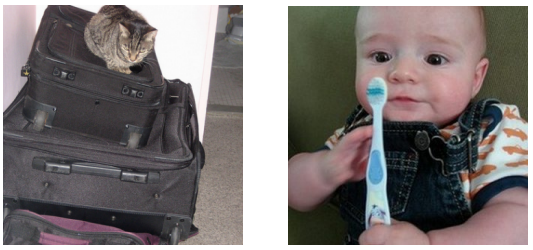
Index Terms: speech emotion recognition, databases

1. Introduction

Deep learning approaches have taken the field of pattern recognition to the next level. This became already obvious, when Hinton et al. raised the benchmarks of automatic speech recognition [1] and image classification [2] in 2012. Since this breakthrough, a large number of impressive applications have been published, such as automatic colorization of black and white images [3], automatic machine translation [4], automatic description of images [5], or networks which are able to play games like Go [6], just to name a few.

Despite this great success of deep learning, pattern recognition methods in general still make mistakes that human subjects would never do. An example of the automatic image description tool NeuralTalk [5] in Fig. 1 mirrors this very clearly: While the automatically generated caption of the left picture is almost perfect, classifying a toothbrush as a baseball bat in the right picture is funny, but most of all a big problem. Furthermore, besides the case that a trained model is not able to recognize or classify an object correctly, it also may happen that it learns undesired characteristics of the training data. This was shown by Bolukbasi et al. [7], who found out that the tool Word2vec [8], which was trained on Google News texts, has learned sexism in addition to the desired word embeddings.

Even more critical are the investigations of Das et al. [9] on actions of visual question answering systems. In this study, a trained model had to answer questions about an input image. During a question about the kind of sight protection of a window in a bedroom, it turned out that the system surprisingly focused not on the window, but instead on the bed in order to answer this question. The problem is that the unexpected focus means not necessarily that the answer is wrong, since, depending on the training data, maybe each image of a bedroom has curtains on the window. The critical point in this case is that we expect



(a) "black cat is sitting on top of suitcase."

(b) "a young boy is holding a baseball bat."

Figure 1: *Examples of the image description tool NeuralTalk [5]. Captions of both images*[1] *correspond to the output of NeuralTalk after analyzing the respective image.*

a focus on the right object, but even if the answer is correct, the focus might be somewhere else. This illustrates the hidden inner life of a trained model and the strong dependency on the training data.

A research field which is very familiar with a challenging basis of data is speech emotion recognition (SER). Since emotions can have a significant influence on the meaning of an utterance [10], SER has become in the meanwhile an important topic to improve the human-to-computer interaction with a long research history [11–14]. While in the past the best classical approaches [12, 13] typically consist of a feature extraction [15, 16] plus an extra classifier, like a support vector machine [17], a hidden Markov model [18], or a neural network [10], modern approaches achieve promising results by applying end-to-end deep learning models [14, 19–22].

However, SER research is in general biased by two issues: First: "Listeners' recognition and interpretation of emotions from recorded speech varies substantially" [23], which results in an ambiguous ground truth. And second: "Most speech emotional databases do not well enough simulate emotions in a natural and clear way" [12], which is caused by the major challenge of recording a large amount of natural and spontaneous emotions. Douglas-Cowie et al. [24] did already comprehensive investigations in order to record *genuine emotions* and defined some guidelines about scope, naturalness, context and description for developing emotional databases.

As a consequence of these issues, emotional databases vary widely regarding emotions, descriptions and structures. Thereby, the kind of emotions are typically divided into *simulated*, *induced* and *natural* emotions [25]. Further characteristics are the number of speakers, the number and length of utterances per speaker, the language, the kind of emotion description, the number of emotions, scripted and unscripted data, syn-

[1] Images from https://cs.stanford.edu/people/karpathy/deepimagesent/

10.21437/Interspeech.2018-1851

thetic and recorded data, or the linguistic nature of the material. A good overview of SER databases is given in [13, 24, 25].

Building upon the examples in [7, 9], we will show in this paper, how important it is, to precisely study the content and structure of (emotional) speech databases, since classifiers in general do not always learn what we believe, even if the achieved results seem plausible to us. To be specific, we first consider two well-known SER databases and reproduce state-of-the-art SER results by means of three SER classifiers. Subsequently, we apply a little change to the training, validation, and test data split for one of the two databases. The resulting dramatical decrease of the recognition accuracy of all classifiers will be shown to prove that none of the classifiers actually learned the emotion labels. Finally, we will disclose the construction mistake of this database and will verify, whether the second database is also concerned of this mistake.

The rest of the paper is organized as follows: The applied databases and approaches are presented in Sections 2 and 3, respectively. Afterwards, we provide our experimental analyses of the databases in Section 4, before we finally conclude this paper in Section 5 with some remarks.

2. Example Databases

We consider the freely available databases eNTERFACE [26] and IEMOCAP [27] for our investigations, which will be briefly introduced in the following.

2.1. The eNTERFACE Database

The audio-visual eNTERFACE database [26] deals with the six *archetypal* emotions *fear, anger, disgust, sadness, surprise* and *happiness* [11], and includes a high number of in total 42 subjects (34 men, 8 women) from 14 different nationalities all speaking English. Recordings were done with 48 kHz sampling rate in a 16-bit stereo format, with the microphone being positioned 30 cm below the mouth.

In order to elicit the desired emotions, all participants have first listened to a short story for each emotion before they expressed a predefined scripted reaction that fits the story. For each emotion, five different reactions have been defined and recorded from each person. An example of the predefined reactions for two emotions is given in Table 1. Different to other emotion databases (e.g., EmoDB [28]), the predefined reactions $\mathrm{R1}-\mathrm{R5}$ are different between all emotion classes. Overall, the database contains 1257 speech samples with an average length of 2.78 seconds.

The typical data split in the literature to ensure speaker independence is the leave-one-speaker-group-out (LOSGO) method [29], whereby a certain number of speakers are clustered in a group. This simplifies an implementation of a K-fold cross-validation strategy. Training, validation, and test (TVT) partitions can then be defined by sets of speaker groups.

2.2. The IEMOCAP Database

The creation of the interactive emotional dyadic motion capture database (IEMOCAP) [27] was motivated by generating a large multi-modal emotional corpus with genuine emotions including audio, video, as well as motion captures of both face and hands. It is with a total length of approximately 12 hours one of the larger datasets for SER and consists of 10 actors (5 men, 5 women) and nine emotion classes (*happiness, anger, sadness, neutral, frustration, disgust, fear, excitement* and *surprise*). All utterances are recorded in English at 48 kHz sampling rate, with a tube shotgun microphone, positioned a few meters in front of the actors.

Table 1: *The five predefined reactions for two emotion classes of the eNTERFACE database.*

Reaction	Emotion happiness	Emotion surprise
R1	That's great, I'm rich now!	You have never told me that!
R2	I won: This is great! I'm so happy!!	I didn't expect that!
R3	Wahoo...This is so great.	Wahoo, I would never have believed this!
R4	I'm so lucky!	I never saw that coming!
R5	I'm excited!	Oh my God, that's so weird!

Following Douglas-Cowie et al. [24], the data was recorded with plays (scripted sessions) and improvisations (spontaneous sessions) instead of using reading material to obtain genuine emotions. Both approaches were carried out with the aid of five dyadic sessions, whereby in each session a dialog between a man and a woman with a length of about 5 minutes took place. Afterwards, all dialogs were segmented into sentences and labeled by three persons. This means that the emotions were not predefined even in the scripted sessions, but the dialogs are chosen to elicit the desired emotions in a natural way.

Since the expressed emotions are not predefined, some emotion classes are underrepresented and hence, the dataset is very unbalanced. For that reason, only four classes (*anger, happiness, sadness* and *neutral*) are typically considered in publications, whereby *happiness* is additionally combined with the emotion *excited* [21, 30, 31]. This results in a total of 5531 speech samples (1103 anger, 1636 happiness, 1084 sadness, 1708 neutral) with an average length of 4.5 seconds.
Most publications split the data by means of the leave-one-speaker-out (LOSO) strategy [29] in order to ensure speaker independence. In [21], TVT partitions use four sessions for training, one of the two speakers of the remaining session for validation, and the other one for test.

3. Example Approaches

In order to exclude dependencies of a classifier by specific characteristics, we consider both classical and modern methods for our investigations. First, we choose a support vector machine (SVM) in combination with the openSmile feature extraction toolkit [29], which has already been applied as baseline in a large number of publications, as well as on several INTERSPEECH challenges [32–34]. Second, we select a simple convolutional network, which has already been successfully applied for classifying written sentences of variable length [35, 36], and was used in [21] for the IEMOCAP database. Finally, we employ a recurrent model, which achieved excellent results on the eNTERFACE in [22]. Below, we describe all applied approaches in detail. Note, for the purpose of simulations, both databases were downsampled from 48 kHz to 16 kHz.

3.1. OS/SVM

In line with Schuller et al. [29], we extract for each speech sample a 6552-dimensional supra-segmental feature vector with the predefined *emotion features large* set of the openSmile (OS) toolkit [15] as input for a support vector machine (SVM). We applied an SVM with RBF kernel and a pairwise discrimination for classifying multi-class emotions. The SVM is implemented by applying the `Scikit-learn` Python library [37]. This approach will henceforth be called **OS/SVM**.

3.2. CNN/DNN

This simple model called **CNN/DNN** is published in [21] and obtains as input a 2-dimensional log-mel spectrogram (cf. 3.4). It contains three different components: First, a convolutional layer extracts emotionally salient features of the speech sample. The filters have a size of $J_{\text{mel}} \times w$, whereby J_{mel} is the number of applied mel filters, and $w \in \mathbb{N}$ is the filter width that corresponds to the number of considered frames along the time axis. Pursuant to the best proposed model in [21], we use four different sizes of the filter width $w \in \{8, 16, 32, 64\}$ considering different contextual dependencies in parallel. Furthermore, each filter size is applied 384 times, resulting in a total number of $F = 1536$ filters. A rectified linear unit (ReLU) is used as activation function for the output of the convolutional layer.

The second component is a max-pooling-over-time layer. Since the applied filter widths w, as well as the length T of different log mel spectrograms vary, we obtain feature maps of size $F \times (T - w + 1)$, which thus also vary in length. In the process of the max-pooling over time we pick the output with the highest activity of each kernel and collect all values in an output vector with fixed dimension F. Thereby we obtain time invariance and can feed the third component of our model: A DNN, which comprises two layers with 1024 neurons each plus a final layer with N neurons. Thereby, N is the desired number of emotion classes regarding the considered database (IEMOCAP: $N = 4$, eNTERFACE: $N = 6$). As for the convolutional layer, a ReLU activation function is used for the first and second DNN layer, while a softmax activation function is applied to the final layer. The model was implemented and trained with the aid of `TensorFlow` [38].

3.3. CLDNN

The third model is implemented in accordance with [22] and expects also a 2-dimensional log-mel spectrogram. The processing takes place frame-wise by considering a left and right context of $l_{\text{in}} = 10$ and $r_{\text{in}} = 5$ frames, respectively. This results for each frame in an input feature matrix (image) of size $J_{\text{mel}} \times (l_{\text{in}} + r_{\text{in}} + 1)$. The model is composed of two convolutional layers (spectral modeling), a bi-directional long short-term memory (BLSTM) layer (temporal modeling) and four fully connected layers including the softmax output layer (classification), and will therefore be called **CLDNN**.

Each of the two convolutional layers have 32 filters with a time-frequency kernel of size 4×5 and 2×3 for the first and second layer, respectively. To each layer a stride of 1 and a ReLU activation function is applied, followed by a max-pooling of size 1×3 and a stride of 2 only on the frequency axis. The output of the convolutional layers is processed by a BLSTM with 128 cells for each direction. As a result we obtain a 256 dimensional feature vector for each considered frame. In order to recover time invariance, similar to the **CNN/DNN** model, an average pooling over time is carried out and thus, a 1×256 dimensional feature vector can be fed into the DNN component. The four fully connected layers have 128, 32, 32 and N neurons (IEMOCAP: $N = 4$, eNTERFACE: $N = 6$). All fully connected layers use a ReLU activation function except the last layer, which uses a softmax function. Again, implementation and training of the model was carried out with `TensorFlow` [38].

3.4. Feature Extraction and Data Augmentation

Different as published in [21], but in line with [22], we extract the $(J_{\text{mel}} \times T)$-dimensional log-mel filterbank (MFB) spectrograms for the **CNN/DNN** and **CLDNN** approach with the `Kaldi` toolkit [39]. For the generation, we applied $J_{\text{mel}} = 40$ mel filters, a frame length of 25 ms, a frame shift of 10 ms and a Hamming window. The feature extraction and statistical functionals behind **OS/SVM** are taken from [15, 29]. The sampling rate of the speech signals was 16 kHz.

Table 2: *Three different TVT split strategies on the eNTERFACE database w.r.t. expressed reactions (*R*) and ensured speaker independence* (SI). $\text{SI}_{\text{R1–R3}}$ *and* $\text{SIRI}_{\text{R4,R5}}$ *denote a reduced* SI *set and a reaction-independent* SI *set, respectively.*

Split	Training	Validation	Test
SI	R1-R5	R1-R5	R1-R5
$\text{SI}_{\text{R1–R3}}$	R1-R3	R1-R3	R1-R3
$\text{SIRI}_{\text{R4,R5}}$	R1-R3	R1-R3	R4,R5

Inspired by [21, 22] we augmented the training and validation subsets of the eNTERFACE database for both the **CNN/DNN** and **CLDNN** approach in two ways: First, we generated two copies of each speech sample and applied the `Sox`[2] sound processing toolkit in order to manipulate the speed of the copies. The input variable "factor" of the speed effect was set to 0.9 and 1.1. This results in tripling the size of our training data. Second, by means of the MUSAN corpus [40], which contains 929 noise samples like rain, paper rustling, footsteps, animal noises, etc., we augmented our data by a further factor of 18. For this, we created 18 copies of our augmented data and mixed respectively two copies with the same noise sample, but two different SNR values in the interval of $[-5, 15]$ dBov according to the ITU-T Recommendation P.56 [41]. Noise samples and SNR ranges are chosen randomly. This results in $1257 \times (3 \times 18 + 3) = 71649$ speech samples of the eNTERFACE database, which is a similar amount as in [22]. A data augmentation of the IEMOCAP database was not necessary, since our deep learning classifiers obtained already good results.

4. Experiments and Discussions

All experiments are carried out with a TVT partition of the datasets and a K-fold cross validation, whereby K is the number of speaker groups[3]. Different as published in [29], but according to [21, 22], we implemented a speaker-based z-normalization regarding mean and standard deviation. Explicit splits are described in detail in Sections 4.1 and 4.2. The results are reported by means of the unweighted average recall (UAR) as demanded in [32].

[2] http://sox.sourceforge.net

[3] For the **OS/SVM** approach, SVM hyper-parameters C and γ for the RBF kernel of the `SciKit-learn` library were determined for each fold on the validation set with the aid of a grid search in the ranges $C \in \{2^0, 2^2, \ldots 2^{12}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{-3}\}$ in line with [21]. Further, the cost parameter C was weighted with a factor for each class in order to counteract unbalanced data in the training. We choose a batch size of $B = 50$ and a learning rate of $\eta = 0.0001$ for the training of the **CNN/DNN** approach. In order to balance the data, oversampling was deployed. Additionally, we applied a dropout of $\theta = 0.8$ and an L2 regularization of $\lambda = 0.0001$ for the fully connected layers, used an Adam optimizer, and stopped the training, if the UAR on the validation set has not improved for more than 12 epochs. According to [22], the training procedure of the **CLDNN** model was implemented with $B = 10$, $\eta = 0.0005$, a dropout of $\theta = 0.8$ was applied to the output of the recurrent layer and an L2 regularization with $\lambda = 0.00001$ was used for the fully connected layers. Again, oversampling for balancing the data, an Adam optimizer and early stopping with 12 epochs were used.

Table 3: *UAR [%] of the three classifiers for the eNTERFACE database.* SI, $\mathrm{SI}_{\mathrm{R1–R3}}$ *and* $\mathrm{SIRI}_{\mathrm{R4,R5}}$ *denote different TVT split strategies.* **Taken from [29],* ***taken from [22].*

Approach	SI	SI	$\mathrm{SI}_{\mathrm{R1–R3}}$	$\mathrm{SIRI}_{\mathrm{R4,R5}}$
OS/SVM	72.5^{*}	77.2	83.3	36.3
CNN/DNN	–	86.6	89.0	24.0
CLDNN	91.7^{**}	87.7	89.3	25.0
Average	–	83.8	87.2	28.4

4.1. eNTERFACE Database

Huang et al. [22] published results on the eNTERFACE, which outperformed Schuller et al. [29] by almost 20 % absolute regarding the UAR. In order to understand this dramatic improvement, we investigated the eNTERFACE database and noticed that unlike other databases, the expressed utterances (here reactions) per emotion are not the same between different emotion classes (cf. Tab. 1). On top of that, according to [22, 29], common splits of the database do not consider reaction independence but only speaker independence. Therefore, we investigated three different split strategies on the eNTERFACE.

Before any TVT split of the data, we defined six groups each containing 7 speakers, respectively, in order to ensure speaker independence for all experiments in accordance to LOSGO. For each fold and for each experiment, the following allocation of the speaker groups (SGs) was applied: Training $= 5$ SGs, validation $= 1$ SG, test $= 1$ SG. The first TVT split is in line with [22, 29] and considers only speaker independence (SI). In the second split we cover a reduced dataset ($\mathrm{SI}_{\mathrm{R1–R3}}$). It is the same as the SI split, but we reduced the amount of data by removing reaction R4 and R5 of each emotion class for TVT. The last split considers a speaker-independent and *reaction-independent* dataset ($\mathrm{SIRI}_{\mathrm{R4,R5}}$). Here, we only changed the test set of the $\mathrm{SI}_{\mathrm{R1–R3}}$ set by exchanging the reactions R1-R3 by reactions R4 and R5. Thereby, we obtain a reaction-independent data split. An overview of all three splits is given in Tab. 2. Note, in line with [29] and to prove that data augmentation is not the cause of any surprising results, we consider the **OS/SVM** method without data augmentation.

The UAR results of the three investigated methods in all different TVT split strategies for the eNTERFACE database are summarized in Table 3. Results marked with one and two asterisks are taken from [29] and [22], respectively. It is clearly evident that the best results of all applied methods are reached for the split regarding $\mathrm{SI}_{\mathrm{R1–R3}}$. This is interesting, since we reduced the number of training data for $\mathrm{SI}_{\mathrm{R1–R3}}$ by $3/5$ regarding SI, while the number of emotion classes remains constant. The answer is given in the enormous performance decrease in the results for the $\mathrm{SIRI}_{\mathrm{R4,R5}}$ split. Since we replaced in this case only the exact utterances, but not the expressed (and, as supposed, from the models learned) emotions, it can be concluded that *the network did not learn the emotions, but instead the exact linguistic content of the utterances (reactions)*. In consequence, best results are obtained for $\mathrm{SI}_{\mathrm{R1–R3}}$, because it is easier for the models to distinguish 18 reactions ($\mathrm{SI}_{\mathrm{R1–R3}}$) than 30 reactions (SI). This result is also surprising in regards to the **OS/SVM** method, since supra-segmental features are said to be relatively independent on phonetic content and have a natural focus on the emotional content [32]. However, since the **OS/SVM** approach does not rely on augmented data, we can exclude that data augmentation has any influence on these conclusions.

As a result of our investigations, two key findings can be presented: First, the design of the eNTERFACE database does not allow a usage as training data for an SER model. Please note that this fact does not exclude the employment of the database for test and evaluation purposes. Second, published results of SER methods (e.g., [22, 29]), which were both trained and tested on the eNTERFACE database, are not conclusive for speech *emotion* recognition, since it has been shown that *linguistic* content has been learned.

Table 4: *UAR [%] of the three classifiers for the IEMOCAP database.* SI, $\mathrm{SI}_{\mathrm{scripted}}$ *and* $\mathrm{SI}_{\mathrm{unscripted}}$ *denote different parts of the data material.* **Taken from [21].*

Approach	SI	SI	$\mathrm{SI}_{\mathrm{scripted}}$	$\mathrm{SI}_{\mathrm{unscripted}}$
OS/SVM	–	61.3	55.7	63.2
CNN/DNN	59.5^{*}	59.7	56.0	63.6
CLDNN	–	59.4	54.6	64.9

4.2. IEMOCAP Database

The following experiments deal with two aspects: On the one hand, we investigate whether the IEMOCAP has a similar design problem as the eNTERFACE, since besides unscripted emotional dialogs the IEMOCAP deals also with scripted dialogs of five sessions. On the other hand, we verify the performance of the applied methods on an emotional database without data dependencies. Assuming that dependencies within the utterances of the scripted sessions would have an impact of the UAR, we carried out three different data splits for these purposes: "SI" includes the whole dataset, "$\mathrm{SI}_{\mathrm{scripted}}$" takes only the scripted sessions into account, and "$\mathrm{SI}_{\mathrm{unscripted}}$" considers only the unscripted sessions. Different to [21], we applied a LOSGO strategy, whereby each session forms a speaker group (SG) in oder to exclude dependencies within a session. Four SGs divided into 90 % and 10 % were used for training and validation, respectively, as well as the remaining SG for test.

Tab. 4 contains the UAR results for the IEMOCAP experiments. Again, asterisks denote published results, here taken from [21]. For all three approaches: The best results are clearly achieved for $\mathrm{SI}_{\mathrm{unscripted}}$, while the worst results are obtained for $\mathrm{SI}_{\mathrm{scripted}}$. In consequence, we can conclude that there are no dependencies regarding the utterances in the scripted sessions. In total, all applied methods achieve a UAR of about 60 % for SI, i.e., a quite good result on the IEMOCAP.

5. Conclusions

In this case study we analyzed two speech emotion recognition databases regarding different data split strategies for training, validation, and test. The obtained results of three published classifiers on the eNTERFACE database demonstrate that each classifier did not learn the desired emotion classes, but instead the linguistic content of the sentences. In consequence, the eNTERFACE database turns out to be not useful for training an emotion recognition classifier. However, since the database includes still emotional content, it can be used for test or validation on cross-database scenarios. The second database (IEMOCAP) did not show similar construction problems and was used to verify the performance of the three published classifiers.

6. References

[1] Mohamed, A.-R.; Dahl, G. E.; Hinton, G. E., "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[2] Krizhevsky, A.; Sutskever, I.; Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. of NIPS*, Lake Thao, NV, USA, Dec. 2012, pp. 1097–1105.

[3] Zhang, R.; Isola, P.; Efros, A. A., "Colorful Image Colorization," in *Proc. of ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 649–666.

[4] Sutskever, I.; Vinyals, O.; Le, Q. V., "Sequence to Sequence Learning with Neural Networks," in *Proc. of NIPS*, Montréal, Canada, Dec. 2014, pp. 3104–3112.

[5] Karpathy, A.; Fei-Fei, L., "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *Proc. of CVPR*, Boston, MA, USA, June 2015, pp. 3128–3137.

[6] Silver, D. et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[7] Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; Kalai, A., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Proc. of NIPS*, Barcelona, Spain, Dec. 2016, pp. 4349–4357.

[8] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J., "Distributed Representations of Words and Phrases and Their Compositionality," in *Proc. of NIPS*, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.

[9] Das, A.; Agrawal, H.; Zitnick, C. L.; Parikh, D.; Batra, D., "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, Oct. 2017.

[10] Nicholson, J.; Takahashi, K.; Nakatsu, R., "Emotion Recognition in Speech Using Neural Networks," *Neural Computing & Applications*, vol. 9, no. 4, pp. 290–296, Dec. 2000.

[11] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[12] El Ayadi, M.; Kamel, M. S.; Karray, F., "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[13] Ververidis, D.; Kotropoulos, C., "Emotional Speech Recognition: Resources, Features, and Methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sept. 2006.

[14] Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A., "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," *Information Fusion*, vol. 37, pp. 98–125, Sept. 2017.

[15] Eyben, F.; Wöllmer, M.; Schuller, B., "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of the 18th ACM Int. Conf. on Multimedia*, Firenze, Italy, Oct. 2010, pp. 1459–1462.

[16] Rong, J.; Li, G.; Chen, Y.-P. P., "Acoustic Feature Selection for Automatic Emotion Recognition from Speech," *Information Processing and Management*, vol. 45, no. 3, pp. 315–328, May 2009.

[17] Schuller, B; Rigoll, G.; Lang, M., "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture," in *Proc. of ICASSP*, Montreal, QC, Canada, May 2004, pp. 577–580.

[18] Nwe, T.L.; Foo, S.W.; De Silva, L.C., "Speech Emotion Recognition Using Hidden Markov Models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[19] Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y., "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," *IEEE Trans. on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[20] Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M. A.; Schuller, B.; Zafeiriou, S., "Adieu Features? End-to-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network," in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 5200–5204.

[21] Aldeneh, Z.; Provost, E. M., "Using Regional Saliency for Speech Emotion Recognition," in *Proc. of ICASSP*, New Orleans, LA, USA, Mar. 2017, pp. 2741–2745.

[22] Huang, C.-W.; Narayanan, S. S., "Deep Convolutional Recurrent Neural Network with Attention Mechanism for Robust Speech Emotion Recognition," in *Proc. of ICME*, Hong Kong, China, July 2017, pp. 583–588.

[23] Murray, I. R.; Arnott, J. L., "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature in Human Vocal Emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, Feb. 1993.

[24] Douglas-Cowie, E.; Campbell, N.; Cowie, R.; Roach, P., "Emotional Speech: Towards a New Generation of Databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, Apr. 2003.

[25] Koolagudi, S. G.; Rao, K. S., "Emotion Recognition from Speech: A Review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, June 2012.

[26] Martin, O.; Kotsia, I.; Macq, B.; Pitas, I., "The eNTERFACE'05 Audio-Visual Emotion Database," in *Proc. of Int. Conf. on Data Engineering Workshops*, Atlanta, GA, USA, Apr. 2006, pp. 1–8.

[27] Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; Narayanan, S. S., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[28] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B., "A Database of German Emotional Speech," in *Proc. of Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 1517–1520.

[29] Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G.; Wendemuth, A., "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. of ASRU*, Merano, Italy, Nov. 2009, pp. 552–557.

[30] Xia, R.; Yang, L., "A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space," *IEEE Trans. on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan.-Mar. 2017.

[31] Jin, Q.; Li, C.; Chen, S.; Wu, H., "Speech Emotion Recognition with Acoustic and Lexical Features," in *Proc. of ICASSP*, Brisbane, QLD, Australia, Apr. 2015, pp. 4749–4753.

[32] Schuller, B; Steidl, S.; Batlinger, A., "The INTERSPEECH 2009 Emotion Challenge," in *Proc. of Interspeech*, Brighton, United Kingdom, Sept. 2009, pp. 312–315.

[33] Schuller, B; Steidl, S.; Batlinger, A.; Schiel, F.; Krajewski, J., "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. of Interspeech*, Florence, Italy, Aug. 2011, pp. 3201–3204.

[34] Schuller, B; Steidl, S.; Batlinger, et al., "The INTERSPEECH 2013 Computational Paralinguistic Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of Interspeech*, Lyon, France, Aug. 2013, pp. 148–152.

[35] Kim, Y., "Convolutional Neural Networks for Sentence Classification," in *Proc. of Empirical Methods in Natural Language Processing*, Doha, Qatar, Oct. 2014, pp. 1746–1751.

[36] Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P., "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2493–2537, Aug. 2011.

[37] Pedregosa, F. et al., "SciKit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2825–2830, Oct. 2011.

[38] Abadi, M. et al., "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. of OSDI*, Savannah, GA, USA, Nov. 2016, pp. 265–283.

[39] Povey, D. et al., "The Kaldi Speech Recognition Toolkit," in *Proc. of ASRU*, Waikoloa, HI, USA, Dec. 2011.

[40] Snyder, D.; Chen, G.; Povey, D., "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.0848v1*, pp. 1–4, Oct. 2015.

[41] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.