In: *Proceedings of the International Conference on Intelligent Robots*, IROS 2018, pp. 854-860, Madrid, Spain, October 1-5, 2018.

On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks

Egor Lakomkin[1], Mohammad Ali Zamani[1], Cornelius Weber[1], Sven Magg[1] and Stefan Wermter[1]

***Abstract*— Speech emotion recognition (SER) is an important aspect of effective human-robot collaboration and received a lot of attention from the research community. For example, many neural network-based architectures were proposed recently and pushed the performance to a new level. However, the applicability of such neural SER models trained only on in-domain data to noisy conditions is currently under-researched. In this work, we evaluate the robustness of state-of-the-art neural acoustic emotion recognition models in human-robot interaction scenarios. We hypothesize that a robot's ego noise, room conditions, and various acoustic events that can occur in a home environment can significantly affect the performance of a model. We conduct several experiments on the iCub robot platform and propose several novel ways to reduce the gap between the model's performance during training and testing in real-world conditions. Furthermore, we observe large improvements in the model performance on the robot and demonstrate the necessity of introducing several data augmentation techniques like overlaying background noise and loudness variations to improve the robustness of the neural approaches.**

I. INTRODUCTION

Emotions and, in general, affective state recognition play an important role in communication between humans and they allow us to evaluate intentions and urgency of a message quickly. Emotions can be expressed very differently given the speaker's cultural background, context and environmental conditions and, as a result, models that learn from unstructured data are very effective in this case. Complex pattern recognition models like deep neural models led to many recent advances in solving difficult problems like speech recognition [1], image classification [2] or robot motion planning [3].

An essential ingredient to train neural networks is the training data. One common problem that affects the performance of the machine learning model is overfitting, when a model with high capacity like neural networks can fit the training data very well while performing poorly on the unseen data. Several regularization techniques were proposed to mitigate overfitting like dropout [4], batch-normalization [5], [6] and layer normalization [7]. Another possible solution is data augmentation which introduces deformations to the input while not changing its label, for example, by varying the tempo or pitch of a spoken utterance. For example, Tarvainen and Valpola demonstrated that enforcing consistency between the prediction of the original and the corrupted sample greatly improves the performance and robustness of the model [8].

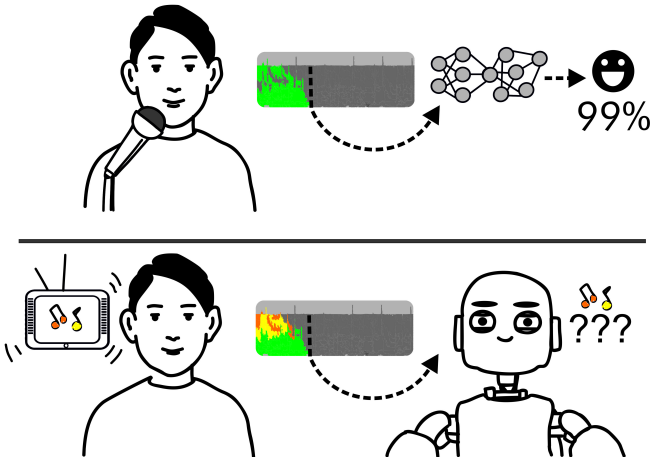

Fig. 1. Illustration of the core problem that we address in the paper. Neural speech emotion recognition models are commonly trained and evaluated on in-domain data collected in a constrained environment. When deployed on a robot, these models can expect a performance degradation due to the environmental conditions and presence of the robot's ego-noise.

The problem of discrepancy between training and testing conditions is especially relevant for robotic applications. For example, different types of microphones can be mounted on a robot or the robot can be present in very different environmental conditions, like small rooms or larger halls. The source of the sound can also have an arbitrary distance to the robot, which affects the signal-to-noise ratio. In addition, there are several sources of ego noise coming from fans, hardware, and joint movements, which complicate speech and emotion recognition. Also, robots will be naturally present in homes or offices where the important speech signal can be masked by various external noise and audio events. We hypothesize that noise can greatly affect the model's performance [9], especially, since many emotion-labelled datasets are collected in the controlled condition of a lab environment. In our work, we evaluate the performance of state-of-the-art neural network models on acoustic emotion recognition tasks in a realistic noise environment and perform the experiments on the iCub robot.

Our contribution is two-fold: a) We evaluate the performance of state-of-the-art neural acoustic emotion recognition models in a set-up that simulate real-world scenarios, like recognizing emotion of a human when the sound is overlaid with noise (e.g. robot's ego noise). b) We propose and evaluate several speech data augmentation techniques and

[1]University of Hamburg, Department of Informatics, Knowledge Technology Institute. Vogt-Koelln-Strasse 30, 22527 Hamburg, Germany {lakomkin, zamani, weber, magg, wermter}@informatik.uni-hamburg.de

analyze their effects on the performance of the neural models in acoustically clean conditions and when recorded by the iCub.

The paper is organized as follows: section II introduces related work and section III describes the methodology including feature extraction from the acoustic signal and data augmentation steps to improve the robustness of the model. We describe several neural models used in our experiments such as recurrent and convolutional neural networks. Section IV outlines the conducted experiments on the iCub robot head.

We found that the models that are trained on clean data, can fail when they are deployed on the robot due to reasons like ego noise or room conditions. However, we could reduce this performance drop using different data augmentations such as changing tempo and loudness, adding Gaussian and background noise, and convolving signals with different room impulse response functions. Moreover, we observed a significant increase in the performance on the iCub-recorded data when we included data and noise augmentations in our training pipeline. We achieved these improvements on the iCub neither using any training data samples from the robot (e.g. recording robot's ego-noise) nor the real lab impulse responses during training and validation. This result shows the importance of adding data augmentations during training for the SER task to make a model robot independent and more robust in general.

II. RELATED WORK

Deep neural networks significantly boosted the performance of acoustic emotion recognition models. The majority of recent work focuses on learning to extract useful input representations and searching for neural architectures for emotion recognition, as neural approaches outperform traditional ones like support vector machines and decision trees [10].

Recurrent neural networks have an ability to model long-term context information and were successfully applied to emotion recognition [11], [12]. Convolutional neural networks can capture only a local context, but have an ability to model longer dependencies when their architecture was designed with a deep hierarchy [10]. Commonly, these methods train neural networks on pre-extracted features: MFCC coefficients, spectrograms and high-level information like formants, pitch, and voice probability. Alternatively, Trigeorgis et al. demonstrate a model that learns how to recognize the affective state of a person directly from the raw waveform [13]. Another explored direction is transfer learning: adapting audio representations trained initially for other auxiliary tasks, like gender and speaker identification [14] or speech recognition [15], [16].

Robustness to noise was a subject of several previous work. Attention mechanisms [11], [17] aim to identify useful regions for emotion classification automatically by assigning a low importance to irrelevant inputs, for example, non-speech or silence frames. Adding background noise during training improved the robustness of neural models in several acoustic classification tasks [18]. Different types of data augmentation methods were explored by Zhou et al. [19] to improve the performance of speech recognition. Supervised domain adaptation was proposed by Abdelwahab et al. [9] to mitigate the problem of training and testing mismatch conditions by tuning the model on the small set of test samples.

Our work is close to Lane et al. [18] and our main difference is that our testing conditions are not synthetically constructed by overlaying clean samples with additive noise, but recorded on the iCub robot which adds a significant amount of ego-noise. We argue that distortions introduced by playing a sample through speakers, changing room conditions and distance from the speech source to the robot, reverberations, added external acoustic events and the robot's internal noise introduce non-linear deformations which are challenging for the neural network to deal with.

III. METHODOLOGY

In this section, we describe a feature extraction procedure from the acoustic signal and outline several proposed data augmentation methods. Then, we introduce the neural architectures that we used in our experiments and the training details.

A. Feature Extraction

We extracted 32 low-level features from the eGEMAPS low-level descriptors [20] using the OpenSMILE[1][21] toolkit. It contains frequency-related features (pitch, jitter, formant information), energy-related features (shimmer, loudness), and spectrum-related features (13 mel-frequency cepstrum coefficients, spectral flux) extracted from a 25ms window with 10ms stride. The original eGEMAPS feature set includes only the first four MFCC coefficients, but in our experiments, we found that using 13 coefficients improves the performance. All coefficients were smoothed with a window size of 3. We calculated mean and standard deviation for each feature value over the whole training set and used them to normalize the samples during training and testing.

B. Data Augmentation

Previous research in end-to-end speech recognition demonstrated the importance of introducing random perturbations into the speech signal like a change of pitch, tempo, loudness, and adding noise [1], [22], [19]. As such perturbations do not alter the target label (spoken text in the case of speech recognition or an emotion category), they can be conveniently applied with some occurrence probability during training. Data augmentation can be considered also as a way to increase the training data size. We will show that this procedure already reduces overfitting. In the case of acoustic emotion recognition, neural models overfit, since the available labelled data is sparse in terms of the number of samples, the variety of speakers and recording conditions.

In our experiments, we 1) changed the tempo of the recording by sampling the speed factor uniformly in a range

[1]https://audeering.com/technology/opensmile/

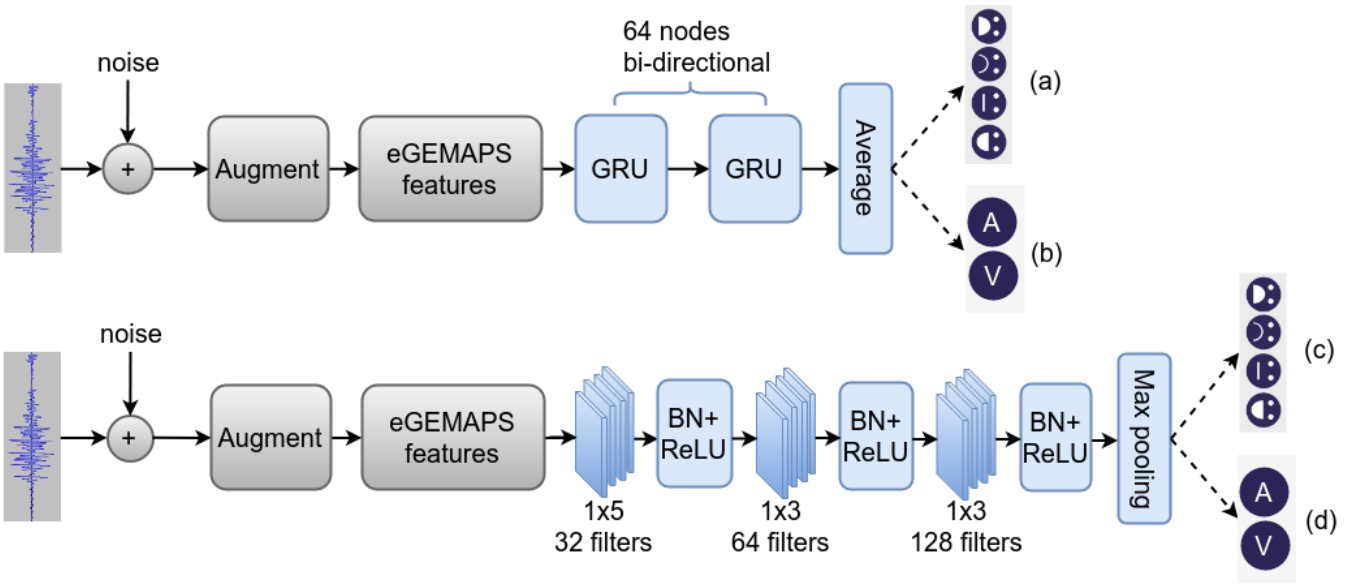

Fig. 2. A recurrent neural network (RNN) model, containing two bi-directional GRU layers, followed by a temporal averaging layer (e.g. uniform attention) for categorical (a) and dimensional cases (b). A convolutional neural network (CNN) model, containing three 1-dimensional convolution layers, followed by a batch-normalization layer and ReLU activation function for categorical (c) and dimensional (d) cases. Both architectures are trained with the same feature set: extended GEMAPS, optionally performing random data augmentation and noise injection before the feature extraction step.

of [85, 120] percent, 2) changed the loudness of the recording by sampling gain uniformly in a range of [-6, 3] dB, 3) added random background noise (more in section IV.B) by sampling the noise-to-signal ratio uniformly in the range [0.5, 0.9], and 4) applied room filter impulse responses selected randomly from the Aachen Impulse Response database [23]. We used the SoX[2] utility to perform all data augmentation steps. It is important to note that we did not use any sample of data corresponding to the iCub's ego-noise during training or validation. Our goal was to identify the model's performance in conditions it never observed during training.

C. Emotion Recognition Model

Neural network-based models recently demonstrated state-of-the-art performance in different affective modelling tasks: multi-modal sentiment analysis [24], facial expression [25] and emotion recognition [11]. In our experiments, we compared convolutional neural network (CNN) and recurrent neural networks (RNNs), as the two most popular approaches to process variable-length speech sequences.

D. Recurrent Neural Network Model

Recurrent neural networks demonstrated state-of-the-art performance in the affective modelling tasks of sentiment analysis [24] and emotion recognition [11], [16]. In our experiments, we evaluated a two-layer bi-directional recurrent neural network with Gated Recurrent Unit (GRU) [26] followed by a softmax layer with four output nodes, in the categorical case, modelling a distribution over four emotion classes, which is overall similar to the model used in Huang et al. [11]. In the dimensional case, the GRU is followed by two linear nodes for arousal and valence. We use the same hyperparameters and training set-up as presented in Huang et al. [11] in all our experiments (*RNN* in the results table I), and the scheme of the model is presented in Fig. 2.

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

E. Convolutional Neural Network Model

Convolutional neural networks were successfully applied to acoustic emotion recognition problems [17] and as demonstrated in Fayek et al. [10] can show competitive results compared with RNNs. Though CNNs, as opposed to RNNs, do no have an explicit memory vector to retain useful contextual information while processing the whole sequence, they are capable of modelling long sequences by using progressively large receptive fields by stacking several CNN layers on top of each other [27] and using various types of pooling to encode information in the sequence [17]. We used a 3-layer convolutional neural network architecture with 1-dimensional filters (see Fig. 2), where each convolution layer is followed by the batch normalization layer and ReLU as the non-linear activation function. Max temporal pooling was applied to encode the sequence into a fixed-length vector.

F. Training details

We used the Adam optimizer [28] with a learning rate of 3e-4, clipped the gradient values to keep them in the interval [-1, 1] and used a batch size of 32. If the results on the validation set did not improve over the course of training, we reduced the learning rate by a factor of 2. In addition, we followed the SortaGrad [1] training routine by presenting samples to the network in a sorted way during the first epoch.

IV. EXPERIMENTAL RESULTS

We conducted several experiments to evaluate the impact of data augmentation on the performance of the proposed

methods on the iCub robot. In this section, we describe the dataset that we used for training, the evaluation procedure, and achieved results.

A. Data

We used the IEMOCAP [29] dataset for our experiments, which contains five sessions with two actors in each, performing either scripted dialogues or improvising on several pre-defined topics (e.g. unsatisfied customer at the bank or sharing a happy moment with a friend) resulting in 10,030 utterances and 12 hours of speech overall. Each utterance is labelled by three to five annotators with categorical labels (*Angry, Happy, Neutral, Sad, Excited, Fear*, and *Disgust*) and dimensional values: valence and arousal. Valence represents a value from 1 (very negative) to 5 (very positive). The arousal value reflects the degree of excitement of the person, where 1 is very calm and 5 is very active speech. In our experiments, we only used samples labelled as *Angry, Happy, Neutral*, and *Sad*, as they are the most often occurring ones in the dataset and also to be consistent with previous work in HRI. As in multiple previous work [17], [10], we merged the *Happy* and *Excited* classes together. The data distribution was 1,103 *Angry*, 1,708 *Neutral*, 1,636 *Happy* and 1,084 *Sad* samples resulting in 5,531 utterances overall.

B. iCub Data

To evaluate the model deployed on the robot, we recorded the IEMOCAP dataset in the Knowledge Technology human-robot interaction lab. We played IEMOCAP utterances through the speakers and recorded the signal captured by two microphones on the iCub head (mounted in the left and right ears with realistic pinnae). We do not do any signal processing (like changing loudness) before playing the recording.

1) Lab Setup: Our goal is to simulate a scenario close to a real-life situation. The experimental setup which is shown in Fig. 3, consists of a humanoid robot head (iCub) immersed in a display to create a virtual-reality environment for the robot [30]. Loudspeakers are located behind the display between 0° and 180° along the azimuth plane with the same elevation. The iCub head is 1.6 meters away from the speakers. The setup introduces background noise generated by the projectors, computers, power sources as well as ego noise from the iCub head.

2) Background Noise and Acoustic Events: As we expect that robots will eventually share home environments with humans, we are interested in evaluating acoustic emotion recognition models in conditions similar to real-life situations. To this end, we play different natural noise samples: air conditioner, babbling noise (TV), and salient loud events like door knock or cell phone ring-tone. When such noise overlays speech, it can introduce distortions that might influence the neural network performance. We utilize two resources to select noise samples: Freesound[3] and UrbanSound 8k [31]. Freesound is a collection of audio samples uploaded by users with a short textual description and a list of tags. We fetched samples labelled with tags like *clap, click, crash*, or *kitchen* to collect a set of audio events that can occur at a home environment. In addition, we manually constructed an "unwanted" list of tags (like, *sfx*) and we filtered samples annotated with any tag from it. The full list of "desired" tags is: *mash, break, crash, accident, shatter, crack, cracking, kitchen, knock, knocking, domestic-sounds, collapse, alarm, warning, horn, fire-alarm, alert, gunfire, siren, tap, beep, falling, snapping, household*, and *falling*. The unwanted list is composed of human-produced vocal sound that can interfere with the task (e.g. *speech, voice, cry, scream, shout, pain, crying*, and *cough*), irrelevant (e.g. *nature*, and *field-recording*) and synthesized sounds (e.g. *special-effects, synthesizer*, and *sound-effect*).

[3]https://freesound.org/

C. Experiments

We followed the leave-one-speaker-out cross validation to report the performance of the model in all our experiments. As the IEMOCAP dataset contains five sessions with two speakers in each, we used four sessions for training and the remaining one for testing and validation (samples from one speaker were used for validation and from the other one for testing), resulting in ten folds overall. Original IEMOCAP samples were used when testing the models in clean conditions and re-recorded on the iCub in the *IEMOCAP-iCub* experiments. Thus we can evaluate the robustness of the models by testing them in two different testing conditions.

As the IEMOCAP dataset provides both categorical and dimensional labels, we also used them both in our experiments, and the models were trained independently for categorical and dimensional labels. We present the results in Table I. We report unweighted accuracy, unweighted average recall and macro F-score for categorical labels and mean-absolute error and Pearson's correlation coefficient of arousal and valence for dimensional labels.

We tested the trained model on two setups: samples from the original IEMOCAP dataset which we refer to as clean data (IEMOCAP in table I), and samples from the IEMOCAP dataset re-recorded on the iCub (*IEMOCAP-iCub* in table

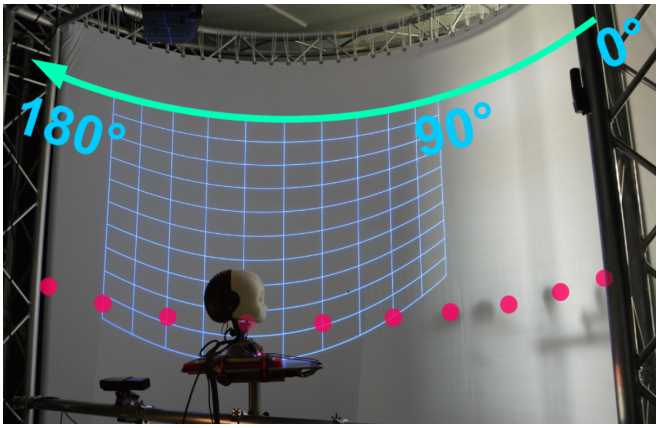

Fig. 3. Lab setup of the iCub in front of loudspeakers behind a screen. See also [30].

TABLE I

Evaluation results. We evaluated two neural network architectures (the RNN and CNN), trained on the IEMOCAP dataset and with data augmentation (IEMOCAP + augmentation and noise) and tested on the IEMOCAP original samples and re-recorded on the iCub (*IEMOCAP-iCub*). Metrics reported: unweighted accuracy, unweighted average recall, macro F-score, arousal and valence mean absolute error, and Pearson's correlation coefficient. Also, we reported a gap in performance for each model evaluated in clean and noisy conditions (higher is better for UW Acc, UAR, F-score, Arousal and valence corr and lower is better for Arousal and Valence MAE) and relative performance improvement on the *IEMOCAP-iCub* by adding augmentations during training.

			Categorical			Dimensional			
Model	Train conditions	Test conditions	UW Acc	UAR	F-score	Arousal MAE	Valence MAE	Arousal corr	Valence corr
RNN	IEMOCAP	IEMOCAP	0.533	0.559	0.531	0.430	0.655	0.715	0.525
		IEMOCAP-iCub	0.203	0.303	0.144	0.493	1.004	0.572	0.076
		Gap %	-61.9%	-45.8%	-72.9%	+14.6%	+53.2%	-20.1%	-85.5%
	IEMOCAP + Augmentation and noise	IEMOCAP	0.545	0.563	0.54	0.422	0.727	0.675	0.426
		IEMOCAP-iCub	0.475	0.418	0.411	0.431	0.762	0.658	0.33
		Gap %	-12.8%	-25.71%	-23.9%	+2.1%	+4.8%	-2.5%	-22.5%
		Improvement %	**+134.0%**	**+37.9%**	**+185.4%**	**+12.5%**	**+24.1%**	**+15.0%**	**+334.2%**
CNN	IEMOCAP	IEMOCAP	0.511	0.532	0.505	1.351	1.150	0.687	0.412
		IEMOCAP-iCub	0.360	0.342	0.247	1.419	1.116	0.647	0.155
		Gap	-29.5%	-35.7%	-51.1%	+5%	+2.9%	-5.4%	-62.3%
	IEMOCAP + Augmentation and noise	IEMOCAP	0.495	0.521	0.48	1.320	1.184	0.638	0.214
		IEMOCAP-iCub	0.400	0.401	0.312	1.399	1.164	0.605	0.145
		Gap %	-19.2%	-23%	-35%	+5.9%	-1.6%	-5.1%	-32.2%
		Improvement %	**+11.1%**	**+17.2%**	**+26.3%**	**+1.4%**	-4.3%	-6.49%	-6.45%

TABLE II

Ablation study on different augmentation techniques. A relative difference in the F-score when one augmentation method is excluded from the training pipeline of the RNN model (base F-scores are 0.54 and 0.411 for IEMOCAP and *IEMOCAP-iCub*, respectively).

Augmentation type	IEMOCAP	*IEMOCAP-iCub*
-tempo	+2.3%	-6%
-loudness	+5.1%	-2.7%
-background noise	+9.7%	**-36.8%**
-filter impulse response	+5.3%	+1.9%

I). The sample IDs in all these two sets were identical and taken from the IEMOCAP dataset, but the recording conditions were different as described above. We hypothesize that emotion detection in the *IEMOCAP-iCub* dataset should be a more challenging task, as it contains the robot's ego noise that can significantly corrupt the original sample.

D. Results

We present our results in Table I. The performance of our RNN model on the IEMOCAP test dataset match previously reported results by Huang et al. [11]. We observe a significant difference in the performance between IEMOCAP and *IEMOCAP-iCub* test conditions and consistent improvements on the *IEMOCAP-iCub* when we add augmentations. This result proves that neural acoustic emotion recognition is very sensitive to the recording conditions. The F-score of the RNN model without any augmentations drops from 0.531 on IEMOCAP down to 0.144 on *IEMOCAP-iCub*. The F-score for the CNN also drops from 0.505 to 0.247. When we add augmentations and additive noise during training, the F-scores on the *iCub* test set rise to 0.411 for the RNN and 0.312 for the CNN, which are significant improvements.

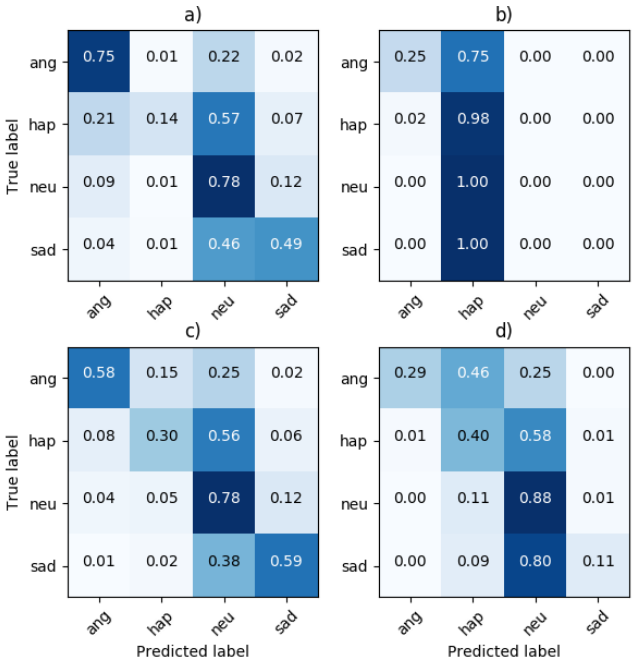

Fig. 4. A confusion matrix for our RNN model trained on the original IEMOCAP dataset and evaluated on the IEMOCAP test set (a), *IEMOCAP-iCub* samples (b), trained with data augmentations and noise injection and evaluated on the IEMOCAP dataset (c), and *IEMOCAP-iCub* samples (d).

We noted that *Sadness* and *Neutral* samples are particularly difficult to recognize (see Figure 4) and we hypothesize that *Sadness* samples have a lower signal-to-noise ratio when recorded on the iCub, as in the majority of samples belonging to the *Sadness* class an actor is speaking quietly in the IEMOCAP dataset. Adding augmentations mitigates the problems for the *Neutral* class, but it remains still very

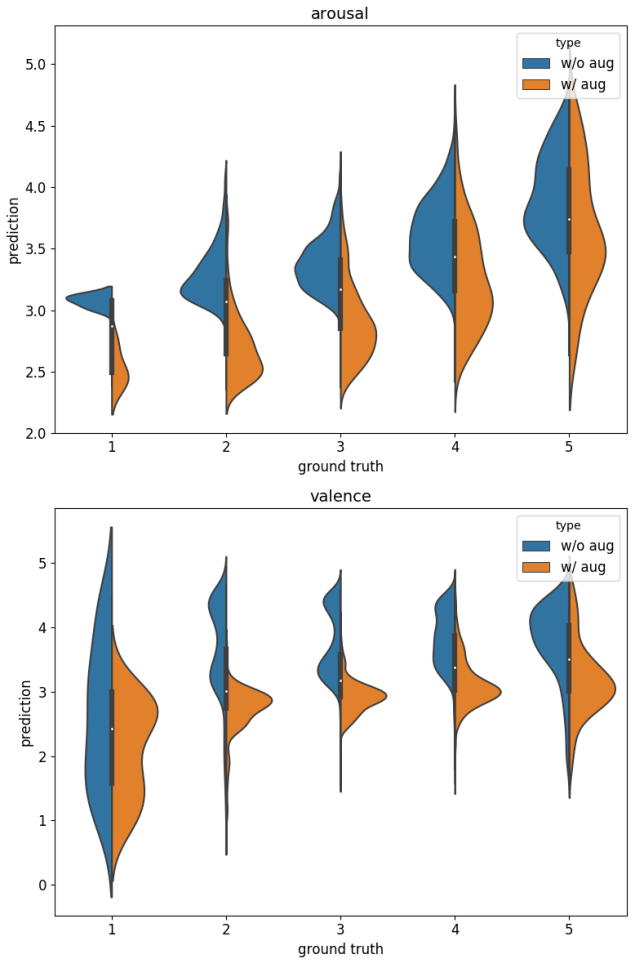

Fig. 5. Violin plot of ground-truth arousal (top) and valence (bottom) (x-axis) vs predicted arousal and valence (y-axis) tested on the iCub recordings of our RNN model with no augmentation (blue) and augmentations turned on (orange). The plot shows full distribution of model's predictions given the ground truth value. Adding augmentations leads to more neutral arousal compared to w/o augmentation which predicts too high arousal, which could be explained by that the network has not learnt noise and thus does not ignore acoustic frames containing only noise and can interpret them as high arousal speech. We observe similar behavior for the valence parameter as well.

difficult to classify *Sadness* samples correctly.

Among the two-dimensional parameters, valence is the most affected one (see Fig. 5) when evaluated on the iCub. The model without augmentations tend to overshoot the value of valence and have higher variance in its predictions compared to the model trained with augmentations. Valence (or degree of negativity/positivity) is naturally very difficult to evaluate given only acoustic signals as it depends also on the content of spoken text.

In addition, we note that there is no significant difference in the performance on the IEMOCAP test set, when we add noise augmentations during training, which leads us to the conclusion that it is safe to add noise augmentation to the training pipeline to improve the robustness of the neural model without jeopardizing the expected performance on the original data. Compared with the models that are trained on the IEMOCAP dataset, we gained 0.009 and lost 0.025 of F-score for the RNN and CNN, respectively, when we trained with data augmentations and tested on IEMOCAP (clean samples).

We conducted an ablation study (see Table II) to evaluate effects of different data augmentation techniques on the performance of the clean samples (IEMOCAP data) and the noisy samples (*IEMOCAP-iCub* data). We measured the performance difference when one of the data augmentation methods was excluded compared with the training regime when all of them were included. In our ablation study, we found that adding background noise was the most effective augmentation type and excluding it led to a 36% F-score drop on the *IEMOCAP-iCub* dataset. Interestingly, varying speech tempo led to the least increase in the performance of the IEMOCAP dataset while it was ranked as the second most important augmentation type according to our ablation study.

V. CONCLUSIONS

In this paper, we evaluated two neural speech emotion recognition models and showed that they perform significantly worse when trained only on in-domain clean data and tested on the iCub robot. We demonstrated that data augmentation reduced this performance loss and overall significantly improved the robustness of the model even without using any real robot ego noise or room conditions. We observed significant performance gains on the *IEMOCAP-iCub* data by injecting noise during training. We thus can conclude that it is crucial to include training data augmentations to prepare models for deployment on a robot.

In future work, we plan to investigate further ways to enhance the data augmentation pipeline. For example, data-driven generative models, like generative adversarial networks, can produce realistic speech samples, which potentially can be useful during training. Also, we plan to evaluate an option to enrich input representation with the information on the spoken text under noisy conditions as it appears to be difficult to analyze valence without it.

ACKNOWLEDGMENT

The authors thank Erik Strahl for his continuous support with the experimental setup and the iCub and Julia Lakomkina for her help with illustrations. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (SECURE) and Cross-modal Learning (TRR 169).

REFERENCES

[1] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, and et al, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 173–182, 2016. [Online]. Available: https://arxiv.org/pdf/1512.02595.pdf

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

[3] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2786–2793.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[6] T. Cooijmans, N. Ballas, C. Laurent, g. Gülçehre, and A. Courville, "Recurrent Batch Normalization," *International Conference on Learning Representations*, 2017.

[7] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5285–5294, 2017.

[8] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1195–1204, 2017.

[9] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4 2015, pp. 5058–5062. [Online]. Available: http://ieeexplore.ieee.org/document/7178934/

[10] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 8 2017.

[11] C.-W. Huang and S. Narayanan, "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition," *Proceedings of Interspeech*, pp. 1387–1391, 2016.

[12] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," *Interspeech*, 2015.

[13] 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3 2016, pp. 5200–5204.

[14] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, E. M. Provost, and A. Arbor, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," *Interspeech*, pp. 1098–1102, 2017.

[15] H. M. Fayek, M. Lech, and L. Cavedon, "On the Correlation and Transferability of Features between Automatic Speech Recognition and Speech Emotion Recognition," *Interspeech*, pp. 3618–362, 2016.

[16] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing Neural Speech Representations for Auditory Emotion Recognition," *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, vol. 1, pp. 423–430, 2017. [Online]. Available: http://www.aclweb.org/anthology/I17-1043

[17] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," *Interspeech*, pp. 1263–1267, 2017.

[18] N. D. Lane, P. Georgiev, L. Qendro, and B. Labs, "DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments using Deep Learning," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 283–294, 2015.

[19] Y. Zhou, C. Xiong, and R. Socher, "Improved Regularization Techniques for End-to-End Speech Recognition," *CoRR*, vol. abs/1712.07108, 2017. [Online]. Available: http://arxiv.org/abs/1712.07108

[20] F. Eyben, K. Scherer, J. Sundberg, E. And, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 190–202, 2016. [Online]. Available: http://sail.usc.edu/publications/files/eyben-preprinttaffc-2015.pdf

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. New York, New York, USA: ACM Press, 2013, pp. 835–838. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2502081.2502224

[22] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep Speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[23] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 7 2009, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/5201259/

[24] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Context-Dependent Sentiment Analysis in User-Generated Videos," *Association for Computational Linguistics*, pp. 873–883, 2017.

[25] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 279–283, 2016.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to align and translate," *International Conference on Learning Representations*, 2015.

[27] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP*, pp. 1746–1751, 2014.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2014.

[29] 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, pp. 335–359, 12 2008. [Online]. Available: http://link.springer.com/10.1007/s10579-008-9076-6

[30] J. Bauer, J. Davila-Chacon, E. Strahl, and S. Wermter, "Smoke and mirrors Virtual realities for sensor fusion experiments in biomimetic robotics," in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 9 2012, pp. 114–119. [Online]. Available: http://ieeexplore.ieee.org/document/6343022/

[31] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014.