Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2000–2006
Marseille, 20-25 June 2022

Towards Speech-only Opinion-level Sentiment Analysis

Annalena Aicher[1], Alisa Vinogradova[2], Aleksey Gusev[2], Yury Matveev[2], Wolfgang Minker[1]
[1] Ulm University, Albert-Einstein-Allee 43, 89075 Ulm
[2] Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia
annlena.aicher@uni-ulm.de

Abstract

The growing popularity of various forms of Spoken Dialogue Systems (SDS) raises the demand for their capability of implicitly assessing the speaker's sentiment from speech only. Mapping the latter on user preferences enables to adapt to the user and individualize the requested information to increase user satisfaction.
In this paper, we explore the integration of rank consistent ordinal regression into a speech-only sentiment prediction task performed by ResNet-like systems. Furthermore, we use speaker verification extractors trained on larger datasets as low-level feature extractors. An improvement of performance is shown by fusing sentiment and pre-extracted speaker embeddings reducing the speaker bias of sentiment predictions. Numerous experiments on CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) databases show that we beat the baselines of state-of-the-art unimodal approaches. Using speech as the only modality combined with optimizing an order-sensitive objective function gets significantly closer to the sentiment analysis results of state-of-the-art multimodal systems.

Keywords: speech sentiment analysis, sentiment intensity prediction, (speech) emotion recognition, acoustic emotion recognition, CMU-MOSEI, COnsistent RAnk Logits (CORAL), spoken dialogue systems, HCI, computational paralinguistics

1. Introduction

In the past decade, the development of spoken dialogue systems and the variety of techniques to meet the users' needs have grown rapidly. For instance, natural language displays a straightforward way to communicate for humans and thus, also interactions with computers via spoken language are perceived more comfortable and natural for users. Therefore, it is required that SDS are capable of managing complex, human-like interactions and learn how to adapt to users and their preferences. There are various approaches to detect user preferences in SDS. Many literature references estimate user preferences only in a static manner e.g. through historical user behaviors and profiles or by exploiting the interaction history (Liu and Mazumder, 2021). In Gao et al. (2021) conversational recommender systems and their possibilities to explicitly obtain the exact preference of users are investigated. Another explicit approach is shown by Zeng et al. (2018) where the user's preferences on food are elicited by interviewing the user directly. Furthermore, we (Aicher et al., 2021) introduced the argumentative dialogue system BEA which estimates the users' preferences (preferring or rejecting a presented argument) via explicit user feedback by using weighted Bipolar Argumentation Graphs (wBAGs) (Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018). As explicit feedback entails the risk of a repetitive course of the dialogue, which might bore or annoy the user, instead we aim for an implicit preference detection. Rach et al. (2019) previously suggested to use the information on multimodal emotion recognition techniques to implicitly obtain user preferences on certain aspects of a topic. As the role of the user within this scenario is passive (non-speaking), they analyze facial expressions, gestures and postures as a source for affective cues that represent the current emotional state (Rach et al., 2019).
Since often only one modality is accessible in spoken dialogue systems, we herein investigate techniques that are able to retrieve sentiment information from audio input only. Especially, as we envision the use of our approach in an argumentative spoken dialogue system, visual features from video recordings are rather hard to come by.
Furthermore, textual cues might be misleading as argumentative scenarios might require the user to reference or quote system utterances (and argumentative content) which are likely to be misinterpreted as the user's own stance (preference of an argument (positive stance), rejection of an argument (negative stance)) and thus, sentiment. Thereby one may overcome the dependence on explicit feedback/preference user statements, by implicit analysis of the spoken user utterance within argumentative discourses[1]. Henceforth, in this work we introduce and investigate an approach to detect user sentiment analyzing the spoken speech-signal of a user which can in future work be used to be linked to user preferences.
The remainder of the paper is as follows: Section 2 gives a short overview over related work. Section 3 examines the architecture of our approach, followed by the objectives in Section 4. Section 5 introduces the experimental setup and Section 6 discusses our results. We conclude and give a short outlook in Section 7.

[1]For instance, positive sentiment indicates a rather positive user stance, indifferent sentiment indifference and negative sentiment a rather negative user stance towards a presented argument.

2. Related Work

The vast majority of recent literature introduces fusion approaches to multimodal sentiment analysis (Zadeh et al., 2017; Cambria et al., 2017). For example fused textual, visual and acoustic modalities on the CMU-MOSEI-Dataset[2] are investigated by Li et al. (2021). A unimodal approach with extensive studies on facial expressions is introduced by Ekman and Keltner (1997). There also exist approaches in visual sentiment analysis and emotion recognition (retrained for sentiment analysis) (Byeon and Kwak, 2014; Ebrahimi Kahou et al., 2015). Bertero et al. (2016) explore speech emotion and sentiment recognition from raw audio samples by using a Convolutional Neural Network (CNN) with a single filter. In contrast, Tian et al. (2015) study the performance of knowledge-inspired disfluency and non-verbal vocalization features. Whereas their emotional encoding was done by either SVM or LSTM-RNN, we focus on heavier neural network classifiers using Mel-scaled filter banks (Heo et al., 2020; Gusev et al., 2020b).

Siriwardhana et al. (2020) explores the techniques to fine-tune Speech-BERT and RoBERTa Transformers pre-trained on a large-scale text dataset to perform multimodal discrete sentiment intensity label prediction for a given pair of spoken utterance and its transcription. Similarly (Kumar and Vepa, 2020) treats sentiment labels as cardinal labels and explores cross-attention-based text and speech modalities fusion. Lu et al. (2020) use an ASR-based encoder to extract features rich in both acoustic and linguistic characteristics. Whereas we consider sentiment intensity levels, they treat the sentiment prediction as a cross-entropy-aided multi-class classification task considering neutral, positive or negative classes. Likewise, we reuse the model trained on the data of the same modality, however, the distillation and acoustic features post-processing is different. Similar to our approach they consider the uni-modal scenario, with only speech available for both test and training times. Still the fine-tuning of the model is trained with audio and text pairs, which still implicitly introduces dependencies to phonology and linguistics.

The previously mentioned literature defines emotional input on a discrete scale. Moreover, it is assumed that emotions/sentiment are order agnostic. Although, Mohammadi and Vuilleumier (2019) show that emotions/sentiment intensities are ordinal and a misclassification should not be penalized equally. Thus, we focus on speech-only opinion-level sentiment intensity recognition taking the ordinal and continuous nature of sentiment intensity labels into account.

[2]Multimodal Opinion Sentiment and Emotion Intensity, `http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/`, last accessed 04.05.2022

3. Architecture

The architecture of our sentiment analysis system is shown in Figure 1. Its components are explained in the following.

As a backbone model, we use a ResNet34-like(He et al., 2016) architecture. We chose a ResNet backbone as it is well established, shows a state-of-the-art performance for many tasks and is well-suited for speech data. Furthermore in comparison with X-vectors, EfficientNets and ECA-ResNets, ResNet achieved the best results. The first two blocks of the ResNet backbone, referred to as Speaker Recognition network-based Extractor (SRExt), are taken from the pre-trained speaker verification model. The configurations of the last two blocks vary and are referred to as StExt. The backbone features are passed to the two parallel decoders which are either a Combination of multiple Global Descriptors (CGD) (Gusev et al., 2020a) block or a CGD block with a speaker embedding fusion.

Then the first output is processed by the regression stream predicting a number indicating the intensity value on a continuous scale. The second output is passed to the COnsistent RAnk Logits (CORAL) (Cao et al., 2020) stream which predicts the vector of probabilities for each of the 6 sentiment intensity bins.

3.1. SRExt

To extract rich latent features out of Mel-scaled filter bank (MFB) acoustic features, we took two lower SE[3] (Hu et al., 2018) ResNet blocks from the SR-SEResNet34 system trained for the speaker verification task on a database of human voices with the much bigger hours of speech compared to CMU-MOSEI. In the following, these 2 lower ResNet blocks are referred to as Speaker Recognition network-based Extractor (SRExt). We use those two blocks in evaluation mode, with no further fine-tuning for **2/2SRExt**, and with fine-tuning of the second block for **1/2SRExt**. These first convolutional blocks are supposed to extract speech-related low-level features to simplify the task for the weaker network trained for sentiment prediction on a smaller dataset of voices. For the detailed architecture and training strategy description of the SR-SEResNet34 we refer to Gusev et al. (2021). We adopted their design of DG-SE-ResNet34 but excluded the domain generalization head.

3.2. StExt

Blocks 3 and 4 maintain the original design of the blocks from SEResNet. Their widths are both set to 64 and the number of layers being fixed to 2 and 1, respectively (for the SEResNet-16 version). Regarding the SEResNet-34 version 128 and 256 convolutional filters were used in 6-layered 3rd and 3-layered 4th blocks, respectively.

Decoder: CGD The output frame-level features are

[3]Squeeze-and-Excitation

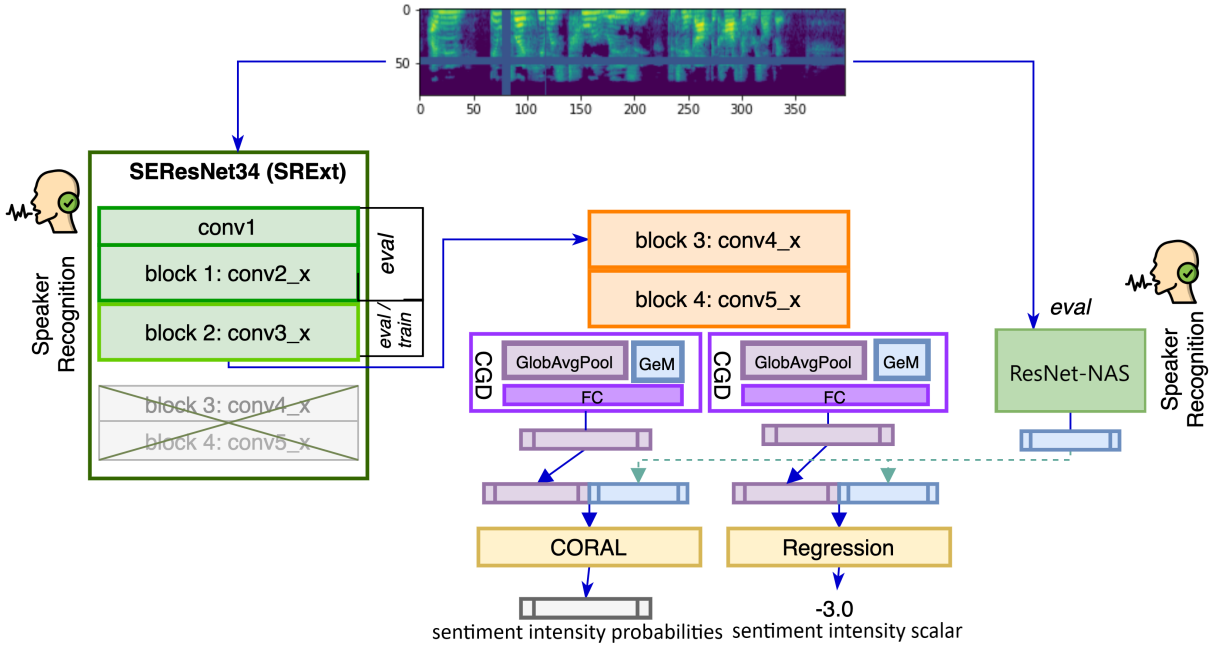

Figure 1: Architecture of the sentiment analysis system.

passed to the two parallel pooling + segment-level network (sln) blocks. We use a CGD-inspired parallel Generalized-Mean (GeM) (Radenović et al., 2018) and global average pooling, as segment-level network a linear layer with a Rectified Linear Unit (ReLU) and applied non-learnable batch normalization. The outputs of the two pooling and segment level blocks are then concatenated to a single vector.

3.3. Decoder: Speaker Verification Feedback (SpE)

To reduce the speaker bias the model suffers from, we fused the pre-extracted speaker embedding with the output of CGD block. The embeddings were extracted by another pre-trained speaker verification ResNet-NAS-based network, which was previously introduced by Gusev et al. (2021). It is larger than the SR-SEResNet34 used for the acoustic features preprocessing. During the training time of the sentiment recognition network, we retrieve the corresponding speaker embedding for the input samples and concatenate the speaker embedding with the output of the segment-level part of the sentiment recognition system. The resulting 128+512 or 512+512 vector is passed to the final classification head.

3.4. Regression Head

Afterward, the two sentiment embeddings extracted by two parallel CGD blocks are passed to two streams. The first regression stream is composed of a single linear layer.

3.5. CORAL Head

The second stream is the Ordinal regression managed by the CORAL head. Let K be the number of discretized sentiment classes, then the second rank consistent ordinal regression stream is composed of a linear layer predicting a single probability value and a bias vector of $K-1$ values that are used to construct a vector of sentiment class probabilities. This vector indicates the probability of each sentiment being present in a given speech segment in a multi-label classification. The $K-1$ binary tasks are designed to share the same weight parameters but have independent bias units which enables to achieve rank-monotony and guarantees binary classifier consistency (*Theorem 1* pointed out by Cao et al. (2020)). To get the final prediction from the CORAL head we transform each class probability to binary classes, such that it is set to 0 for no sentiment being present and 1 for sentiment being present. Then, the final sentiment value is the sum over all $K-1$ binary classes.

4. Objectives

4.1. Regression Loss

We use the Mean Absolute Error (MAE) as the first option to model the loss with varying penalization for the misclassifications depending on the absolute difference between ground truth and predicted classes. Thus,

$$MAE = N^{-1} \sum_{i=1}^{N} |y_i - g_r(f(x_i))|, \qquad (1)$$

where N denotes the number of training samples, g_r the regression head and f the output of the decoder.

4.2. Rank Consistent Ordinal Regression Loss

The prediction of sentiment labels on a discrete scale is performed by applying CORAL (Cao et al., 2020). Let K denote the number of discretized sentiment intensity labels, namely {-3, -2, -1, 0, 1, 2, 3}. Furthermore, f shall denote the stack of layers up to the regression/CORAL heads with the sentiment embedding as its output (the output of the decoder). Then $g_c(f(x_i), W)$ describes the output of the penultimate layer, that shares a single weight with all $K-1$ nodes in the final layer. Next, $K-1$ independent bias units are added to $g_c(f(x_i), W)$, resulting in $K-1$ binary classifier outputs, undergoing sigmoid, to result in the binary class probabilities, denoted as p_k. Thus, we obtain the final target function:

$$\sigma(z) = \frac{1}{(1+\exp(-z))}, \tag{2}$$

$$p_k = \sigma(g_c(f(x_i), W) + b_k), \tag{3}$$

$$L = -\sum_{i=1}^{N}\sum_{k=1}^{K-1}[\log(p_k)y_i^k + \log(1-p_k)(1-y_i^k)], \tag{4}$$

where N is the number of training samples, and y_i denotes the labels one-hot-encoded in a cumulative manner, such that

$$y_i = \cap_{j=1}^{K-1}\mathbb{1}\{y_i > j\}. \tag{5}$$

Hence, the final prediction is obtained by

$$\sum_{k=1}^{K-1}\mathbb{1}\{p_k > 0.5\}. \tag{6}$$

5. Experimental Setup

In the following the experimental setup is explained, i.e. the data processing, acoustic feature extraction and configuration optimization. The described setup and design choices were adjusted during the experiments and tailored to tackle encountered challenges and increase the system's performance.

5.1. Data

The following subsections give a deeper insight data processing in our pipeline.

5.1.1. Speaker Recognition

Both the SR-SEResNet34 as well as the ResNet-NAS-based SpE extractor were trained using the SdSVC 2021 Challenge train dataset (Zeinali et al., 2020). It consists of VoxCeleb 1 and VoxCeleb 2 (SLR47), LibriSpeech, Mozilla Common Voice Farsi and DeepMine (Task 2 Train Partition). The overall number of speakers in the resulting set is 11939. VoxCeleb 1 and the whole Voxceleb 2 together include approx. 2k hours of speechs, LibriSpeech about 2k hours, Mozilla Common Voice Farsi 70 hours and DeepMine 480 hours. Thus adding up to approximately 4550 hours of speech in total.

5.1.2. Sentiment Recognition

CMU-MOSEI (Zadeh et al., 2018) is one of the largest speech emotion recognition datasets from 1000 online YouTube speakers denoting the intensity of sentiment with continuous labels in the range [-3, 3] (from highly negative, to highly positive). It contains more than 23,500 single-speaker sentence utterance videos (more than 65 hours of annotated videos), containing mostly feedback and reviews. Hence, it is suited for predicting user sentiment in argumentative dialogues. The audios are segmented into non-opinion and opinion frames. Emotional and sentiment tagging is performed only for the latter ones.

5.2. Acoustic Features Extraction

5.2.1. Mel-filter Bank Energies 80

The input features for all models in this paper are 80-dimensional log Mel-filter bank energies extracted from a $16\,\mathrm{kHz}$ raw signal with $25\,\mathrm{ms}$ frame-length and $15\,\mathrm{ms}$ overlap. Additionally, we use per-utterance Cepstral Mean Normalization (CMN) over a $3\,\mathrm{s}$ sliding window over the stack of MFBs to compensate for the channel effects and noise by transforming data to have zero mean (Furui, 1981). In order to remove silence frames from the utterances we make use of a U-net-based Voice Activity Detector (VAD) (Gusev et al., 2020b; Lavrentyeva et al., 2020).

5.2.2. SpecAugment

To increase the data variability in training we use SpecAugment as a widely used data augmentation technique in speech recognition. It applies a random number of random-width masks on the time and spectral dimensions of the 80-dim MFB features.

5.2.3. Weighted Data Sampling

The training dataset is highly imbalanced regarding its classes. Even though this imbalanced labels distribution is close to the test case scenario, it forces the model to be biased favoring the most frequent samples. Thus, due to the huge class imbalance convergence is not achieved and some of the minority classes are ignored by the network. To resolve this imbalance we randomly oversample values of each class with the probability of the inverse class frequency. As a result, the Gaussian-like distribution of classes is almost transformed into a uniform one, as shown in Figure 2. The blue bars denote the original distribution of continuous sentiment intensity labels for the training data. The oversampling applied to the original distribution of training labels is given in orange. In green, the distribution of training labels rounded to the nearest integer prior to oversampling is shown. Finally, the red bars mark the original distribution of labels in the test data.

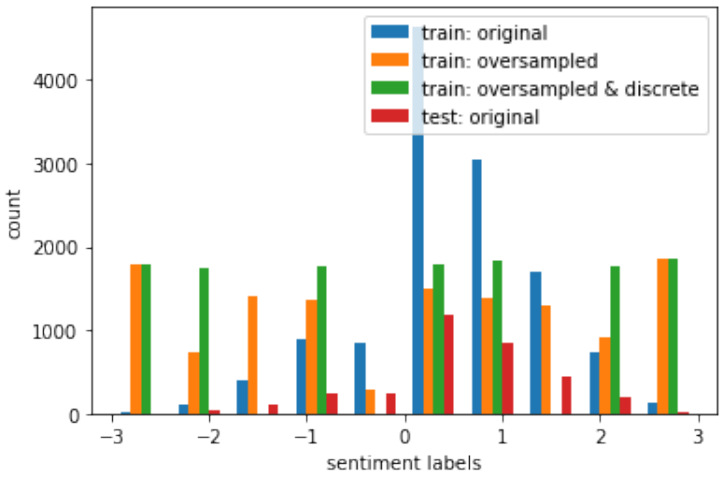

Figure 2: The distribution of sentiment intensity labels over the training and test sets.

5.3. Configurations Optimization

All system variants are trained on random 4 s chunks with a random spatio-temporal masking SpecAugment with maximum number of masks set to 2, maximum time and frequency mask widths equal to 10 and 5. The Adam optimizer is utilized with a multistep learning rate scheduler which decreases the learning rate of 1e-3 10 times as soon as the training converges. In order to transform labels to fit the **CORAL** approach, they are rounded to the closest integer and afterwards the level transformation described in Section 3 is applied. It is essential that all labels are present in both, the train and valid sets of the data[4]. Regarding the **regression**, the labels remain unchanged (no rounding). The usage of the pre-trained low-level features extractor is necessary to assure convergence of the model optimized by MAE loss. CORAL and MAE are used separately and also jointly, where both are added up with equal weights of 1 and optimized together. Thus, during training the sum of both loss functions is optimized. Note, that the outputs of both heads are not fused, but either one of them is used to compute the metric as described in the following.

6. Results and Discussion

To evaluate our approach against previous baselines, four performance metrics[5] based on Sun et al. (2019) are reported: 1) mean absolute error (MAE), 2) F1-score for 7 sentiment intensity levels, 3) binary accuracy (Acc. 2-class), where labels in $[-3, 0)$ considered as negative sentiment and in $(0, 3]$ as positive and 4) 7-class sentiment intensity accuracy (Acc. 7-class). In the case of the two-headed models, we report the metrics for both of the heads either the regression head ("reg") and the CORAL head ("class"). The results of our systems compared to the ones of current state-of-the-art literature are shown in Table 1. Our results show that the regression loss is stable in the optimum and beats unimodal baselines (MULT Audio (Tsai et al., 2019) and ICCN Audio (Sun et al., 2019)). The fusion of pre-trained speaker and sentiment embeddings together with the minimization of the rank consistent ordinal regression loss in our system "1-34 + SpE + C" enables us to reduce the overfitting effect and increase the complexity of the StExt. Before the utilization of SpE the heaviest model to converge with no overfitting was SRExt-SEResNet-16 and after became SRExt-SEResNet-34. This increase in complexity led to a better performance with regard to MAE and both accuracies.

The application of CORAL was successfully shown. Thus, the task of sentiment recognition in speech utterances can be considered an ordinal regression task with cumulative relations between sentiment intensity values. Moreover, we showed that low-level features extracted from a pre-trained speaker recognition extractor can be successfully applied for speech sentiment analysis. By fusing a speaker embedding to the sentiment one we could get even closer to the multimodal baselines which only use sentiment embeddings. Still, we perceive a gap between the multimodal baselines ICCN (Sun et al., 2019) and pre-trained BERT+RoBERTa (Siriwardhana et al., 2020) and our results. But to the best of our knowledge, we are significantly closer to those than any other unimodal approach. Training on bigger datasets as well as reusing deep neural feature extractors pre-trained on large-scale speech datasets with a huge inter- and intra-speaker variability are very likely to improve our results further. Another possible enhancement is to optimize the VAD for the emotion/sentiment recogni-

[4]Please note that the model is trained either with two heads (CORAL and Regression MAE) or with one of them. Whereas CORAL uses discrete labels (rounded integers), Regression (MAE) uses continuous labels.

[5]Regarding Accuracy and F1-score higher is better, in contrast to the mean absolute error (MAE), where lower is better.

System	MAE		F1 7-class		Acc. 7-class		Acc. 2-class	
	reg	class	reg	class	reg	class	reg	class
2/2SREx-SEResNet-16 C + R	0.715	–	42.55	41.68	42.56	41.68	87.94	87.97
1/2SREx-SEResNet-16 + C + R	0.737	41.865	42.316	41.865	42.32	41.86	87.91	86.59
1/2SREx-SEResNet-16 + R	0.851	-	36.360	-	36.360	-	86.50	-
1/2SREx-SEResNet-16 + C	-	0.888	-	34.796	-	34.80	-	86.74
1/2SREx-SEResNet-34 + SpE + C	-	**0.712**	-	**43.10**	-	**43.16**	-	**88.15**
$TFN_{acoustic}$ (Zadeh et al., 2017)	1.23		–		–		65.1	
Audio (Tsai et al., 2019)	0.764		–		41.4		65.6	
Audio (Sun et al., 2019)	0.785		–		38.59		58.75	
ICCN (Sun et al., 2019)	0.565		–		51.58		84.18	
Pre-trained BERT+RoBERTa: Text + Audio (Siriwardhana et al., 2020)	**0.491**		–		**55.971**		**88.04**	

Table 1: Results of our ResNet-like systems, audio baselines and fused multimodal results. R denotes the usage of Regression, C of CORAL and SpE of Speaker Embedding.

tion application. By retraining with emotion/sentiment specific VAD would add complementary information of non-verbal vocal expressions and reduce the uncertainty of predictions.

7. Conclusions and Future Work

In this paper, we have described a speech sentiment intensity level prediction system that can be utilized as a preference detection module in (argumentative) spoken dialogue systems. As visual features (video) are not always accessible and quoting might lead to misinterpretations of the user's stance and sentiment regarding argumentative discourses, we focus on the single modality speech. Consequently, we aim to narrow the gap for speech-only to multimodal approaches by enhancing the robustness of the former to noisy labels. In the future, we aim to investigate how to fuse text and speech, by ensuring quoting and citations of system utterances by the user are not interpreted as indicators for user sentiment.

As far as we know this paper is one of the first to use speech-only sentiment intensity recognition benchmarked on the CMU-MOSEI database and to outperform unimodal baselines and thus, gets significantly closer to the ones of state-of-the-art literature in multimodal scenarios. Furthermore, to the best of our knowledge, we are the first to apply the CORAL approach to preserve the ordinal nature of the sentiment intensity labels when performing speech sentiment analysis. Moreover, we showed that rich low-level acoustic features can be extracted from the SR model for speech sentiment analysis tasks.

In future work, we will test and train the herein introduced systems with other datasets, (like CMU-MOSI, ICT-MMMO, MOUD, etc.) to further investigate its performance. Moreover, we aim to integrate our presented approach to create a sentiment-aware, intelligent spoken argumentative dialogue system which is able to recognize and adapt its behavior in real-time, taking the user's sentiment and preferences into account.

8. Acknowledgements

This work has been funded by the DFG within the project "How to Win Arguments – Empowering Virtual Agents to Improve their Persuasiveness", Grant no. 376696351, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999) and the DAAD Leonhard-Euler-Programm 2020/2021.

9. Bibliographical References

Aicher, A., Rach, N., Minker, W., and Ultes, S. (2021). Opinion building based on the argumentative dialogue system bea. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 307–318. Springer Singapore.

Amgoud, L. and Ben-Naim, J. (2016). Evaluation of arguments from support relations: Axioms and semantics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16*, pages 900—-906. International Joint Conferences on Artificial Intelligence Organization, July.

Amgoud, L. and Ben-Naim, J. (2018). Weighted bipolar argumentation graphs: Axioms and semantics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5194–5198. International Joint Conferences on Artificial Intelligence Organization, July.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. pages 1042–1047, 01.

Byeon, Y.-H. and Kwak, K.-C. (2014). Facial expression recognition using 3d convolutional neural network. *IJACSA*, 5(12).

Cambria, E., Hazarika, D., Poria, S., Hussain, A., and Subramanyam, R. (2017). Benchmarking multimodal sentiment analysis. *arXiv*.

Cao, W., Mirjalili, V., and Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *arXiv*, 140:325–331, Dec.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). Recurrent neural networks for emotion recognition in video. ICMI '15, page 467–474, New York, NY, USA. Association for Computing Machinery.

Ekman, P. and Keltner, D. (1997). Universal facial expressions of emotion. *Nonverbal communication: Where nature meets culture*, pages 27–46.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272.

Gao, C., Lei, W., He, X., Rijke, M. D., and Chua, T. (2021). Advances and challenges in conversational recommender systems: A survey. *ArXiv*, abs/2101.09459.

Gusev, A., Volokhov, V., Andzhukaev, T., Novoselov, S., Lavrentyeva, G., Volkova, M., Gazizullina, A., Shulipa, A., Gorlanov, A., Avdeeva, A., et al. (2020a). Combination of multiple global descriptors for image retrieval.

Gusev, A., Volokhov, V., Andzhukaev, T., Novoselov, S., Lavrentyeva, G., Volkova, M., Gazizullina, A., Shulipa, A., Gorlanov, A., Avdeeva, A., et al. (2020b). Deep speaker embeddings for far-field speaker recognition on short utterances. In *Odyssey 2020 Proceedings*.

Gusev, A., Vinogradova, A., Novoselov, S., and Astapov, S. (2021). SdSVC Challenge 2021: Tips and Tricks to Boost the Short-Duration Speaker Verification System Performance. In *Proc. Interspeech 2021*, pages 2307–2311.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heo, H. S., Lee, B.-J., Huh, J., and Chung, J. S. (2020). Clova baseline system for the voxceleb speaker recognition challenge 2020, 09.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

Kumar, A. and Vepa, J. (2020). Gated mechanism for attention based multimodal sentiment analysis, 02.

Lavrentyeva, G., Volkova, M., Avdeeva, A., Novoselov, S., Gorlanov, A., Andzhukaev, T., Ivanov, A., and Kozlov, A. (2020). Blind speech signal quality estimation for speaker verification systems. In *INTERSPEECH*, pages 1535–1539.

Li, Q., Gkoumas, D., Lioma, C., and Melucci, M. (2021). Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, 65:58–71.

Liu, B. and Mazumder, S. (2021). Lifelong and continual learning dialogue systems: Learning during conversation. pages 1042–1047. AAAI Conference on Artificial Intelligence (AAAI).

Lu, Z., Cao, L., Zhang, Y., Chiu, C.-C., and Fan, J. (2020). Speech sentiment analysis via pre-trained features from end-to-end asr models. *arXiv*.

Mohammadi, G. and Vuilleumier, P. (2019). Towards understanding emotional experience in a componential framework. pages 123–129, 09.

Rach, N., Weber, K., Aicher, A., Lingenfelser, F., André, E., and Minker, W. (2019). Emotion recognition based preference modelling in argumentative dialogue systems. In *In 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*.

Radenović, F., Tolias, G., and Chum, O. (2018). Fine-tuning cnn image retrieval with no human annotation.

Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. (2020). Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv*.

Sun, Z., Sarma, P., Sethares, W., and Liang, Y. (2019). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *arXiv*.

Tian, L., Moore, J., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. 09.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. volume 2019, page 6558.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv*.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*.

Zeinali, H., Lee, K. A., Alam, J., and Burget, L. (2020). Short-duration speaker verification (sdsv) challenge 2021:the challenge evaluation plan. Technical report, arXiv.

Zeng, J., Nakano, Y. I., Morita, T., Kobayashi, I., and Yamaguchi, T. (2018). Eliciting user food preferences in terms of taste and texture in spoken dialogue systems. In *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*, MHFI'18. Association for Computing Machinery.