Detecting Escalation Level from Speech with Transfer Learning and Acoustic-Linguistic Information Fusion

Ziang Zhou[1], Yanze Xu[1], and Ming Li[1,2]

[1] Data Science Research Center, Duke Kunshan University, Kunshan, China
{ziang.zhou372,ming.li369}@dukekunshan.edu.cn
[2] School of Computer Science, Wuhan University, Wuhan, China

Abstract. Textual escalation detection has been widely applied to e-commerce companies' customer service systems to pre-alert and prevent potential conflicts. Similarly, acoustic-based escalation detection systems are also helpful in enhancing passengers' safety and maintaining public order in public areas such as airports and train stations, where many impersonal conversations frequently occur. To this end, we introduce a multimodal system based on acoustic-linguistic features to detect escalation levels from human speech. Voice Activity Detection (VAD) and Label Smoothing are adopted to enhance the performance of this task further. Given the difficulty and high cost of data collection in open scenarios, the datasets we used in this task are subject to severe low resource constraints. To address this problem, we introduce transfer learning using a multi-corpus framework involving emotion detection datasets such as RAVDESS and CREMA-D to integrate emotion features into escalation signals representation learning. On the development set, our proposed system achieves 81.5% unweighted average recall (UAR), which significantly outperforms the baseline of 72.2%.

Keywords: escalation detection · transfer learning · emotion recognition · multimodal conflict detection

1 Introduction

Escalation level detection system has been applied in a wide range of applications, including human-computer interaction and computer-based human-to-human conversation [36]. For instance, there are e-commerce companies [23] that have been equipped with textual conversational escalation detectors. Once an increasing escalation level of the customers is detected, the special agents will take over and settle their dissatisfaction, effectively preventing the conflict from worsening and protecting the employees' feelings. In public areas like transportation centers, and information desks, where many impersonal interactions occur, it is also essential to detect the potential risk of escalations from conversation to guarantee public security. Therefore, audio escalation level analysis is instrumental and crucial.

We adopted two escalation datasets, Aggression in Trains (TR) [22] and Stress at Service Desks (SD) [21] from a previous escalation challenge [36]. These two datasets provide conversation audio recorded on the train and at the information desk respectively. About four hundred training audios from the SD dataset, with an average length of 5 seconds, are used for training. Five hundred audios from the TR dataset are used for testing. Given datasets with limited scales, learning effective escalation signals from scratch would be challenging. Thus, supervised domain adaptation [2] became a better option, for we can adapt a more general feature distribution backed by sufficient data to adapt to the escalation signal domain using limited resources. Since emotion is an obvious indicator in potential conversational conflicts, we have good reasons to assume that the encoding ability of emotion features would be a good starting point to support the modeling of escalation signals from conversations. Through pretraining on large-scale emotion recognition datasets, our model will be more capable of capturing emotion features and knowledge to support fine-tuning using the escalation datasets. Recent research [30] on small sample set classification tasks also showed promising results on pattern recognition via transfer learning.

2 Related Works

2.1 Conflict Escalation Detection

Several conflict escalation research has been done in recent years, focusing on the count of overlaps and interruptions in speeches. In [13], the number of overlaps is recorded in the hand-labeled dataset and used in conflict prediction. And [8,16] uses a support vector machine (SVM) to detect overlap based on acoustic and prosodic features. Kim et al. in [17] analyzed the level of conversation escalation based on a corpus of French TV debates. They proposed an automatic overlap detection approach to extract features and obtained 62.3% unweighted accuracy on the corpus. Effective as they seem, these methods are considered impractical in this Escalation Sub-task. First, the length of audio files in [8] ranges from 3 to 30 minutes, and the length of conversation audio in [17] is 30 seconds. While in our escalation detection task, the average length of the corpus is 5 seconds. Most of the time, an audio piece only contains a single person's voice. Thus, focusing on overlap detection seems to be ineffective. Besides, we did not spot a significant difference in overlap frequency among different escalation classes based on conversation script analysis. Second, with a total training corpus duration of fewer than 30 minutes, the model built on overlap counts will easily suffer from a high bias and low variance [4,28].

2.2 Transfer Learning

In [12], Gideon et al. demonstrate that emotion recognition tasks can benefit from advanced representations learned from paralinguistic tasks. This suggests that emotion representation and paralinguistic representation are correlated in nature. Also, in [39], supervised transfer learning has brought improvements to music classification tasks as a pre-training step. Thus it occurs to us that utilizing transfer learning to gain emotion feature encodability from large-scale emotion recognition datasets might as well benefit the escalation detection task. Research [5] on discrete class emotion recognition mainly focuses on emotions, including happiness, anger, sadness, frustration and neutral.

2.3 Textual Embeddings

Emotions are expressive in multiple modalities. As shown in [6,33], multimodal determination has become increasingly important in emotion recognition. In the TR [22] and SD [21] datasets, manual transcriptions for the conversations are also provided besides the audio signals. Given various lengths of transcriptions, we look for textual embeddings that agree in size with each other. In [34,35], Reimers et al. proposed Sentence-BERT (SBERT) to extract sentimentally meaningful textual embeddings in sentence level. Using conversation transcriptions as input, we encoded them into length-invariant textual embeddings, utilizing the pre-trained multilingual model.

3 Datasets and Methods

An overview of our solution pipeline is shown in Figure 1. We apply librosa toolkit [27] to extract Mel Frequency Cepstral Coefficient (MFCC), which is then fed to the residual network backbone [14] to pretrain the emotion encoder. The embedding extractor is pre-trained on an aggregated dataset of four emotion recognition datasets, learning emotion representations from a broader source.

3.1 Datasets

Escalation Datasets For this escalation level detection task, we employed two datasets: Dataset of Aggression in Trains (TR) [22] and Stress at Service Desk Dataset (SD) [21]. The TR dataset monitors the misbehaviors in trains and train stations, and the SD dataset consists of problematic conversations that emerged at the service desk [36]. The escalation level has been classified into three stages: low, mid, and high. The higher escalation level suggests that the conflict will likely grow more severe. Moreover, the Dutch datasets have an average of 5 seconds for each conversation clip. The SD dataset is used for training, and the TR dataset is used for testing. More details regarding these datasets can be found in the overview of the previous challenge [36].

Emotion Recognition Datasets Previous work [12] has highlighted the correlation between Emotion Recognition Tasks and Paralinguistics tasks. Hence we assume that the escalation level detection and emotion recognition tasks share certain distributions in their feature space. Thus, we aggregated four well-known audio emotion datasets for joint sentimental analysis: **RAVDESS** [26] is

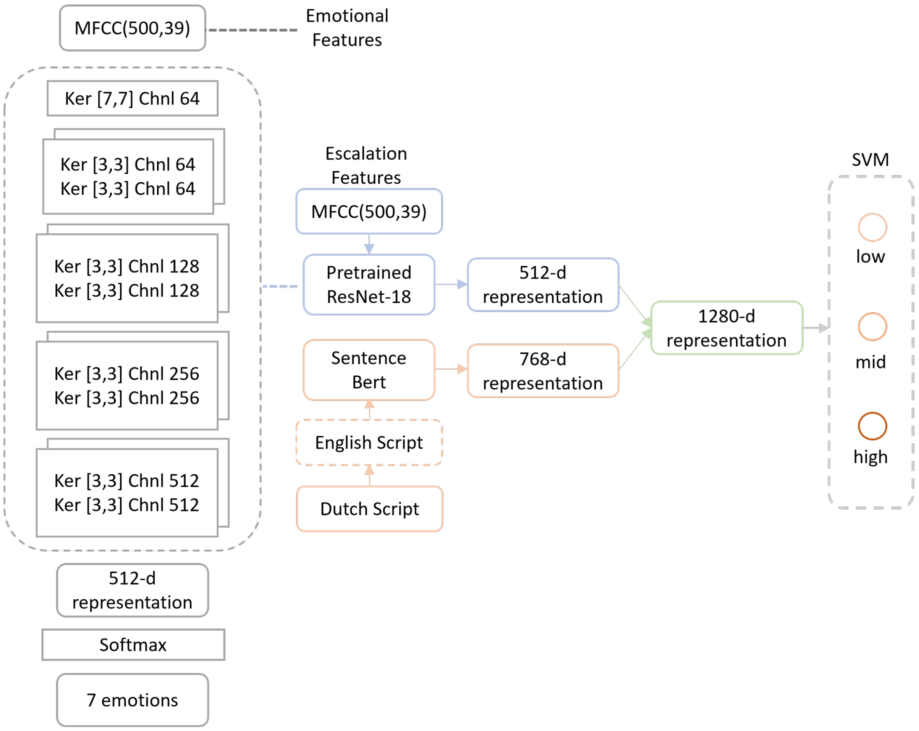

Fig. 1. Pipeline of the Escalation Detection System

a gender-balanced multimodal dataset. Over 7000 pieces of audio are carefully and repetitively labeled, containing emotions like calm, happy, sad, angry, fearful, surprise, and disgust. **CREMA-D** [7] is a high quality visual-vocal dataset, containing 7442 clips from 91 actors. **SAVEE** [11] is a male-only audio dataset with same emotion categories with RAVDESS Dataset. **TESS** [9] is a female-only audio dataset collected from two actresses whose mother languages are both English. TESS contains 2800 audios covering the same emotion categories mentioned above.

With four audio emotion recognition datasets incorporated together, we have 2167 samples for each of Angry, Happy and Sad emotions, 1795 samples for Neutral, 2047 samples for Fearful, 1863 samples for Disgusted and 593 samples for Surprised emotions.

3.2 Methods

Voice Activity Detection Voice activity detection is a process of identifying the presence of human speech in an audio clip [19]. The SD dataset [21] is collected at a service desk and therefore contains background noises. In case the background noise undermines the paralinguistic representations, we implement the WebRTC-VAD [1] tool to label non-speech segments from the audio before feeding the whole pieces for feature extraction.

Transfer Learning Transfer learning has proved effective in boosting performance on low-resource classification tasks [42]. Under our previous assumption, emotion features are essential indicators in escalation-level detection; hence we expect our model to include the capability to involve emotion patterns in the representation learning process. The emotion datasets mentioned in 3.1 are combined to train the ResNet-18 backbone as the emotion recognition model. Note that the emotion classifier and escalation level classifier share the same configuration for a reason. Although linear probing is much cheaper computationally, fine-tuning all parameters of the source model achieves state-of-the-art performance more often [10]. Therefore, we adopt full fine-tuning where all layers of the pre-trained model are updated during the fine-tuning stage on the escalation datasets.

Features Automatic emotion and escalation detection has been a challenging task for the fact that emotion can be expressed in multiple modalities [38]. In multimodal emotion recognition, visual, audio and textual features are the most commonly studied channels. In our task, raw Dutch audio and manually transcribed texts are provided. Thus acoustic and linguistic features can be fused to determine the escalation predictions jointly. Although more recent clustering-based feature extraction approaches have proven instrumental in human speech emotion recognition, e.g., adaptive normalization [15] and pitch-conditioned text-independent feature construction [40], they may further bias our training process due to the extremely low resource limitation. Since MFCC has been widely used in speech emotion recognition tasks [18,20,24], we adopt MFCC as our acoustic feature in this task.

To label the silent intervals, we first apply the WebRTC-VAD [1] to filter out the low-energy segments. Next, MFCCs are calculated from the filtered audio fragments. After that, the MFCCs are applied to fine-tune the emotion classifier, which was previously pre-trained on the aggregated large-scale emotion datasets. On top of the ResNet backbone, we also adopted the Global Average Pooling (GAP) layer structure [25], granting us the compatibility to process the input of variant length during the evaluation phase. According to Tang et al. [37], simply replacing fully connected layers with linear SVMs can improve classification performance on multiple image classification tasks. Therefore, our work does not construct an end-to-end detection system. Instead, we employ Support Vector Machine (SVM) [25] to conduct the backend classification task.

For the textual embeddings extraction, we adopt the pre-trained multilingual model `distiluse-base-multilingual-cased-v1` from Universal Sentence Encoder (USE) [41] from Sentence-BERT(SBERT) [34,35] to extract the sentence-level embeddings. We also compared the Unweighted Average Recall (UAR) metric by extracting Dutch embeddings directly and by extracting embeddings from Dutch-to-English translation. The experiment result shows that the sentence embeddings from the English translation outperformed the original Dutch transcriptions; thus, we adopted the former for textual embedding extraction.

4 Experimental Results

For the Escalation Sub-challenge, we aim to build a multimodal model to determine whether the escalation level in a given conversation is low, medium or high. UAR has been a reliable metric in evaluating emotion recognition tasks under data imbalance constraints. So we followed the metric choice of UAR by previous work on similar settings [31,32]. This section will introduce our experiment setup, results and several implementation details.

4.1 Feature Configuration

In the audio preprocessing stage, we first applied the open-source tool WebRTC-VAD [1] to filter out the silent segments in the audio from the temporal domain. The noise reduction mode of WebRTC-VAD is set to 2. Next, we extract MFCCs from the filtered audio segments. The window length of each frame is set to 0.025s, the window step is initialized to 0.01s, and the window function is hamming function. The number of mel filters is set to 256. Also, the frequency range is from 50Hz to 8,000 Hz. The pre-emphasize parameter is set to 0.97. The representation dimension is set to 512.

4.2 Model Setup

As for the emotion classification task, both the architecture and the configuration of the representation extractor are the same with the escalation model, except that the former is followed by a fully connected layer that maps a 128 dimension representation to a seven dimension softmax probability vector and the latter is followed by a linear SVM classifier of three levels. Weighted Cross-Entropy Loss is known as capable of offsetting the negative impact of imbalanced data distribution [3] and is set as the loss function for that reason. The optimizer is Stochastic Gradient Descent (SGD), with the learning rate set to 0.001, weight decay set to 1e-4, and momentum set to 0.8. The maximum training epochs is 50, with an early stop of 5 non-improving epochs. In the fine-tuning stage, the system configurations remain unchanged, except that the training epochs are extended to 300, and no momentum is applied to the optimizer to reduce overfitting.

The dimension of textual embeddings extracted from the multilingual pretrained model `distiluse-base-multilingual-cased-v1` is 768-d [35]. In the fusion stage of our experiment, textual embeddings will be concatenated with the 512-d acoustic representations, forming 1280-d embeddings at the utterance level.

4.3 Results

Prior to fine-tuning the emotion recognition model on the escalation datasets, we first train the ResNet-18 architecture on emotion recognition datasets to learn emotion representations from audio. The highest UAR achieved by our emotion recognition model is 65.01%. The model is selected as the pre-trained model to be fine-tuned on the Escalation dataset.

To learn better escalation signals in the fine-tuning process, we introduced three factors that may impact model performance positively. Besides fine-tuning the pretrained emotion recognition model, VAD and acoustic-linguistic information fusion are also tested to boost the escalation level detection ability further. We start by analyzing whether voice activity detection will leverage the performance of the development set. Then, we fine-tune the emotion recognition model pre-trained on the four audio emotion datasets to analyze any notable improvement. Finally, we examine whether the fusion between textual embeddings extracted from SBERT [34,35] and acoustic embeddings can improve the performance.

To evaluate the effect of VAD on the prediction result, we did several controlled experiments on the development set. Table 1 demonstrates the effect of VAD on various metrics. First, we calculated features from unprocessed audio and fed them into the embedding extractor. With acoustic embeddings alone, we scored 0.675 on the UAR metric using the Support Vector Machines (SVM) as the backend classifier. With all conditions and procedures remaining the same, we added VAD to the audio pre-processing stage, filtering out non-speech voice segments. The result on the UAR metric has increased from 0.675 to 0.710. This shows that VAD is critical to the escalation representation learning process.

We believe that the escalation detection tasks, to some degree, share certain advanced representations with emotion recognition tasks. [12] Thus, we also experimented with fine-tuning parameters on the escalation dataset with the

Table 1. Effects of Voice Activity Detection (VAD) **TE**: Textual Embeddings fused.

Model Name	Precision	UAR	F1-Score
MFCC	0.640	0.675	0.647
MFCC+VAD	0.675	0.710	0.688
MFCC+TE	0.652	0.690	0.664
MFCC+VAD+TE	0.676	0.721	**0.691**
Baseline Fusion [36]	-	**0.722**	-

model pre-trained on the emotion datasets. Table 2 shows the experiment results after implementing transfer learning to our system. We have witnessed a positive impact of VAD on our experiment results on the development set. Thus the base experiment has been implemented with VAD applied. We can see that, after applying the pre-trained model to the MFCC+VAD system, with acoustic embeddings alone, the score reached 0.810 on the metric UAR, which turns out to be a significant improvement. This also proves that emotion features can benefit paralinguistic tasks by transfer learning. Additionally, the MFCC+VAD+PR system may have already been stable enough that the textual embedding fusion brings no noticeable improvement.

Besides the experiments recorded above, we also implemented various trials involving different features, networks, and techniques. As listed in Table 3, we recorded other meaningful experiments with convincing performance on the devel set that could be applied to final model fusion. Other standard acoustic features like the Log filterbank are also under experiment. Label Smoothing technique [29] is applied on the MFCC+VAD+PR model but brings a slightly negative impact. According to the experiment results, Voice Activity Detection has again been proven effective in enhancing model performance on the development set. ResNet-9 without being pre-trained is also implemented, whose classification result is 74.9% UAR on the devel set.

To further improve our model performance, we proposed model fusion on three models of the best performance on the development set. The fusion is conducted in two ways, early fusion and late fusion. We also proposed two approaches to deal with the embeddings in the early fusion stage. The first approach is concatenating the embeddings, and the second is simply taking the mean value of the embeddings. Both scenarios employ SVM as the classifier. As for the late fusion, we proposed a voting mechanism among the three models' decisions. The fusion result is shown in Table 4. By implementing late fusion, we got the best system performed on the devel set with a UAR score of 81.5%.

5 Discussion

Our proposed best fusion model exceeded the devel set baseline by 12.8%. It is worth mentioning that the WebRTC-VAD system is not able to tease out every non-speech segment. Instead, its value cast more light on removing the blank

Table 2. Effects of fine-tuning **PR**: Pre-trained Emotion Recognition Model applied.

Model Name	Precision	UAR	F1-Score
MFCC+VAD	0.675	0.710	0.688
MFCC+VAD+PR	0.807	**0.810**	0.788
MFCC+VAD+PR+TE	0.807	**0.810**	0.788
Baseline Fusion [36]	-	0.722	-

Table 3. Extra Experiments. **LS**: Label Smoothing.

Model Name	Precision	UAR	F1-Score
Logfbank	0.670	0.743	0.684
Logfbank+VAD	0.711	0.778	0.733
MFCC+VAD+PR+LS	0.781	**0.781**	**0.761**
MFCC+VAD+ResNet-9	0.727	0.749	0.725
Baseline Fusion [36]	-	0.722	-

or noisy segments at the beginning and end of an audio clip. This is reasonable since the unsounded segments in a conversation are also meaningful information to determine the escalation level and emotion. Dialogues would be more likely labeled as high escalation level if the speaker is rushing through the conversation and vice versa. Thus we agreed that a more complicated Neural-Network-based voice activity detector may be unnecessary in this task.

The significant improvement of our system on the devel set has again proved that emotion recognition features and paralinguistic features share certain advanced representations. Just as we mentioned in the related work part, they benefit from each other in the transfer learning tasks. However, due to the small scale of the training set, the overfitting problem is highly concerned. Thus we chose ResNet-18 to train the model on a combined emotion dataset, containing 12,000+ labeled emotion clips. This architecture has also been proved to be effective in the Escalation detection task.

For this task, we adopted Sentence-BERT [34] as the textual embeddings extractor. We utilized the pre-trained multilingual BERT model which is capable of handling Dutch, German, English, etc. We chose to translate the raw Dutch text to English text before feeding them into the embedding extractor, for we did a comparative experiment, which showed that the English textual embeddings alone significantly outperformed the Dutch textual embeddings. The UAR on the devel set achieved by English Embeddings alone is around 45%, whose detection ability is very likely to be limited by the dataset scale and occasional errors in translation. Had we have richer textual data, the linguistic embeddings should be of more help.

An unsuccessful attempt is adding denoising into the preprocessing attempt. Denoising should be part of the preprocessing stage since most of the collected audios contain background noise from public areas. According to [23], they first

Table 4. Model Fusion. Selected models: MFCC+VAD+PR, MFCC+VAD+PR+LS, Logfbank+VAD

Fusion Approach	Precision	UAR	F1-Score
Concatenate	0.783	0.800	0.779
Mean	0.789	0.805	0.789
Voting	0.810	**0.815**	**0.803**

denoise the police body-worn audio before feature extraction, which turns out rewarding for them in detecting conflicts from the audios. However, our attempt does not improve the performance. Our denoised audio is agreed to be clearer in human perception and contains weaker background noises. However, the performance on the devel set is significantly degraded. Our assumption is that, unlike the police body-warn audio, which mostly contains criminality-related scenarios, the conversation audios in TR and SD datasets happen with richer contextual environments. The speech enhancement system might also affect the signal of speech which might be the reason for the performance degradation.

6 Conclusions

In this paper, we proposed a multimodal solution to tackle the task of escalation level detection under extremely low resource contraints. We applied Voice Activity Detection to pre-process the escalation datasets. We also pre-trained an emotion recognition model with ResNet backbone and fine-tune the parameters with the escalation dataset. We also validated that the learning process of escalation signals can benefit from emotion representations learning. By integrating linguistic information in the classification process, the model can become more stable and robust. The single best model can achieve 81.0% UAR, compared to 72.2% UAR basline. By doing the late fusion the models after fusion are able to achieve the 81.5% UAR. Future efforts will be focusing on addressing the over-fitting problem.

Acknowledgements This research is funded in part by the Synear and Wang-Cai donation lab at Duke Kunshan University. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

References

1. Webrtc-vad (2017), https://webrtc.org/
2. Abdelwahab, M., Busso, C.: Supervised domain adaptation for emotion recognition from speech. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5058–5062. IEEE (2015)
3. Aurelio, Y.S., de Almeida, G.M., de Castro, C.L., Braga, A.P.: Learning from imbalanced data sets with weighted cross-entropy function. Neural processing letters **50**(2), 1937–1949 (2019)
4. Brain, D., Webb, G.I.: On the effect of data set size on bias and variance in classification learning. In: Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales. pp. 117–128 (1999)
5. 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 **42**(4), 335–359 (2008)
6. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. pp. 205–211 (2004)
7. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
8. Caraty, M.J., Montacié, C.: Detecting speech interruptions for automatic conflict detection. In: Conflict and Multimodal Communication, pp. 377–401. Springer (2015)
9. Dupuis, K., Pichora-Fuller, M.K.: Toronto emotional speech set (tess)-younger talker_happy (2010)
10. Evci, U., Dumoulin, V., Larochelle, H., Mozer, M.C.: Head2toe: Utilizing intermediate representations for better transfer learning. In: International Conference on Machine Learning. pp. 6009–6033. PMLR (2022)
11. Fayek, H.M., Lech, M., Cavedon, L.: Towards real-time speech emotion recognition using deep neural networks. In: 2015 9th international conference on signal processing and communication systems (ICSPCS). pp. 1–5. IEEE (2015)
12. Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., Provost, E.M.: Progressive Neural Networks for Transfer Learning in Emotion Recognition. In: Proc. Interspeech 2017. pp. 1098–1102 (2017). https://doi.org/10.21437/Interspeech.2017-1637
13. Grèzes, F., Richards, J., Rosenberg, A.: Let me finish: automatic conflict detection using speaker overlap. In: Proc. Interspeech 2013. pp. 200–204 (2013). https://doi.org/10.21437/Interspeech.2013-67
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Huang, C., Song, B., Zhao, L.: Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering. International Journal of Speech Technology **19**(4), 805–816 (2016)
16. Kim, S., Valente, F., Vinciarelli, A.: Annotation and detection of conflict escalation in Political debates. In: Proc. Interspeech 2013. pp. 1409–1413 (2013). https://doi.org/10.21437/Interspeech.2013-369

17. Kim, S., Yella, S.H., Valente, F.: Automatic detection of conflict escalation in spoken conversations. pp. 1167–1170 (2012). https://doi.org/10.21437/Interspeech.2012-121
18. Kishore, K.K., Satish, P.K.: Emotion recognition in speech using mfcc and wavelet features. In: 2013 3rd IEEE International Advance Computing Conference (IACC). pp. 842–847. IEEE (2013)
19. Ko, J.H., Fromm, J., Philipose, M., Tashev, I., Zarar, S.: Limiting numerical precision of neural networks to achieve real-time voice activity detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2236–2240. IEEE (2018)
20. Lalitha, S., Geyasruti, D., Narayanan, R., M, S.: Emotion detection using mfcc and cepstrum features. Procedia Computer Science **70**, 29–35 (2015). https://doi.org/https://doi.org/10.1016/j.procs.2015.10.020, https://www.sciencedirect.com/science/article/pii/S1877050915031841, proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems
21. Lefter, I., Burghouts, G.J., Rothkrantz, L.J.: An audio-visual dataset of human–human interactions in stressful situations. Journal on Multimodal User Interfaces **8**(1), 29–41 (2014)
22. Lefter, I., Rothkrantz, L.J., Burghouts, G.J.: A comparative study on automatic audio–visual fusion for aggression detection using meta-information. Pattern Recognition Letters **34**(15), 1953–1963 (2013)
23. Letcher, A., Trišović, J., Cademartori, C., Chen, X., Xu, J.: Automatic conflict detection in police body-worn audio. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2636–2640. IEEE (2018)
24. Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U.: Speech based human emotion recognition using mfcc. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). pp. 2257–2260 (2017). https://doi.org/10.1109/WiSPNET.2017.8300161
25. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
26. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one **13**(5), e0196391 (2018)
27. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. vol. 8, pp. 18–25. Citeseer (2015)
28. Mehta, P., Bukov, M., Wang, C.H., Day, A.G., Richardson, C., Fisher, C.K., Schwab, D.J.: A high-bias, low-variance introduction to machine learning for physicists. Physics reports **810**, 1–124 (2019)
29. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf
30. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 443–449 (2015)
31. Peng, M., Wu, Z., Zhang, Z., Chen, T.: From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In: 2018 13th IEEE

International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 657–661. IEEE (2018)

32. Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F.: Emotion classification in children's speech using fusion of acoustic and linguistic features. In: Proc. Interspeech 2009. pp. 340–343 (2009). https://doi.org/10.21437/Interspeech.2009-110
33. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536 (2019)
34. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
35. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4512–4525 (2020)
36. Schuller, B.W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Leon J. M.R., Zwerts, J., Treep, J., Kaandorp, C.: The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In: Proceedings INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association. ISCA, Brno, Czechia (September 2021), to appear
37. Tang, Y.: Deep learning using linear support vector machines (2013). https://doi.org/10.48550/ARXIV.1306.0239, https://arxiv.org/abs/1306.0239
38. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing **11**(8), 1301–1309 (2017). https://doi.org/10.1109/JSTSP.2017.2764438
39. van den Oord, Aäron and Dieleman, Sander and Schrauwen, Benjamin: Transfer learning by supervised pre-training for audio-based music classification. In: Conference of the International Society for Music Information Retrieval, Proceedings. p. 6 (2014)
40. Wu, C., Huang, C., Chen, H.: Text-independent speech emotion recognition using frequency adaptive features. Multimedia Tools and Applications **77**(18), 24353–24363 (2018)
41. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., et al.: Multilingual universal sentence encoder for semantic retrieval. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 87–94 (2020)
42. Zhao, W.: Research on the deep learning of the small sample data based on transfer learning. In: AIP Conference Proceedings. vol. 1864, p. 020018. AIP Publishing LLC (2017)