TU Delft

Delft University of Technology

A Cross-Corpus Speech-Based Analysis of Escalating Negative Interactions

Lefter, I.; Baird, Alice; Stappen, Lukas; Schuller, Björn W.

DOI
10.3389/fcomp.2022.749804

Publication date
2022

Document Version
Final published version

Published in
Frontiers in Computer Science

Citation (APA)
Lefter, I., Baird, A., Stappen, L., & Schuller, B. W. (2022). A Cross-Corpus Speech-Based Analysis of Escalating Negative Interactions. *Frontiers in Computer Science*, *4*, Article 749804. https://doi.org/10.3389/fcomp.2022.749804

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

frontiers in Computer Science

METHODS
published: 07 March 2022
doi: 10.3389/fcomp.2022.749804

A Cross-Corpus Speech-Based Analysis of Escalating Negative Interactions

Iulia Lefter[1]*, *Alice Baird*[2], *Lukas Stappen*[2] and *Björn W. Schuller*[2,3]

[1] Department of Multi-Actor Systems, Delft University of Technology, Delft, Netherlands, [2] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [3] Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

The monitoring of an escalating negative interaction has several benefits, particularly in security, (mental) health, and group management. The speech signal is particularly suited to this, as aspects of escalation, including emotional arousal, are proven to easily be captured by the audio signal. A challenge of applying trained systems in real-life applications is their strong dependence on the training material and limited generalization abilities. For this reason, in this contribution, we perform an extensive analysis of three corpora in the Dutch language. All three corpora are high in escalation behavior content and are annotated on alternative dimensions related to escalation. A process of label mapping resulted in two possible ground truth estimations for the three datasets as low, medium, and high escalation levels. To observe class behavior and inter-corpus differences more closely, we perform acoustic analysis of the audio samples, finding that derived labels perform similarly across each corpus, with escalation interaction increasing in pitch (F0) and intensity (dB). We explore the suitability of different speech features, data augmentation, merging corpora for training, and testing on actor and non-actor speech through our experiments. We find that the extent to which merging corpora is successful depends greatly on the similarities between label definitions before label mapping. Finally, we see that the escalation recognition task can be performed in a cross-corpus setup with hand-crafted speech features, obtaining up to 63.8% unweighted average recall (UAR) at best for a cross-corpus analysis, an increase from the inter-corpus results of 59.4% UAR.

Keywords: affective computing, negative interactions, cross-corpora analysis, conflict escalation, speech paralinguistics, emotion recognition

OPEN ACCESS

Edited by:
Joseph Andrew Allen,
The University of Utah, United States

Reviewed by:
Hiroki Mori,
Utsunomiya University, Japan
Alexey Karpov,
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia
David Doukhan,
Institut National de l'Audiovisuel, France

***Correspondence:**
Iulia Lefter
i.lefter@tudelft.nl

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *29 July 2021*
Accepted: *31 January 2022*
Published: *07 March 2022*

Citation:
Lefter I, Baird A, Stappen L and Schuller BW (2022) A Cross-Corpus Speech-Based Analysis of Escalating Negative Interactions. Front. Comput. Sci. 4:749804. doi: 10.3389/fcomp.2022.749804

1. INTRODUCTION

Automatic recognition of escalating interpersonal interactions has many real-world use-cases, including in health care (e.g., various Virtual Reality-based therapies), monitoring conflicts during business meetings, and surveillance, e.g., to observe the need for support during customer service roles. While emotion recognition has become a well-established field of research, there has been relatively little attention on automatically recognizing when an interpersonal interaction may be escalating into a potentially aggressive or dangerous situation. Escalation of negative interactions can be accompanied by a range of negative emotions, stress, and corresponding verbal and non-verbal behaviors. Among them, speech plays an important role, not only by being the main

source of communicating semantic content (linguistic), but primarily through patterns of inflections in the voice that appear when people are discontent or arguing (paralinguistic): the tone that makes the music.

The majority of emotion recognition studies focus on finding the best performing cocktails of features and classifiers given a specific corpus (Vinola and Vimaladevi, 2015). Despite substantial advancements obtained on their target corpus, the question of how well those trained systems would perform "in-the-wild", when exposed to totally new situations, is still unanswered. Aside from learning characteristics of emotion or alternative labels provided to the datasets, classifiers will also be influenced by corpus-specific characteristics, such as recording conditions, language, or speaker-related characteristics such as age and gender (Kaya and Karpov, 2018). These are incorporated in the trained models and will therefore affect performance in a previously unseen setup.

In this paper, we present an audio-based escalation analysis study, where we investigate the various cross-corpus performance of a range of experimental designs. The core task is to classify speech instances for their state of escalation as either low, medium, or high levels. The experiments are performed using three corpora containing a large variety of escalating situations: The Dataset of Aggression in Trains (TR) (Lefter et al., 2013), The Stress at Service Desk Dataset (SD) (Lefter et al., 2014), and The Dataset of Negative Affect and Aggression (NAA) (Lefter et al., 2017).

The elicitation method in all three datasets is based on role-plays, where the participants receive short role and scenario descriptions and the interaction builds up as they react to each other. While in TR and SD all participants are actors, NAA contains interactions between actors and non-actors, bringing it a step closer to real-life. A set of differences between topics, the number of participants in the scene (dyads in NAA, dyads and larger groups in TR and SD), the amount of overlapping speech, recording quality, and label distribution of the three datasets, together with the escalation-rich content create an interesting setup for a cross-corpus study.

Within this contribution, we experiment with two acoustic representations for our approach, including hand-crafted speech methods and deep representations derived from spectrograms. We detail the process of mapping labels between corpora and explore the effects of label mapping choices on recognition performance. Additionally, we perform an acoustic analysis to explore and understand speech patterns related to different escalation levels, and whether escalation is represented similarly in the three corpora. Given unbalanced representations of the different classes, we additionally explore the advantage of class-based data upsampling. Furthermore, we evaluate the effects of merging corpora for training which has shown contradicting results in previous work (Schuller et al., 2010; Zhang et al., 2019) and explore how it links to different label mapping approaches.

Additional challenges of deploying recognition systems in real-world scenarios are related to the fact that much of the training material may result from acted interactions. Particularly in negative interactions, the preponderance of genuine expressions is very limited due to ethical and privacy concerns of exposing participants to negativity. Therefore, we perform independent testing on partitions of the NAA dataset consisting of only actor and only non-actor material to explore such effects. With the introduced data sets, features, and cross-corpus experiments, we provide insights into supporting their generalization ability to unseen situations.

While utilizing already established features and classifiers, our contributions lie in the careful exploration of how escalation can be recognized across corpora, and are summarized as follows: (i) exploring the impacts of different label mapping procedures which will almost always be needed in cross-corpus experiments, (ii) exploring whether escalation was expressed in similar acoustic ways in the three corpora, regardless of the different scenarios, (iii) observe differences between acted and spontaneous speech in the context of escalation, and (iv) performing a range of experiments to evaluate which setup proved most stable across corpora. The paper is a continuation of efforts part of the Interspeech Compare Paralinguistic Challenge, where two of the datasets (SD and TR) were considered as train and test set respectively. In addition to that, the NAA dataset has been added and we further explored the efficacy of speech-based features for understanding escalating negative interactions.

This paper is organized as follows; firstly, in section 2, we outline related work, particularly concerning cross-corpus analysis in the affective computing domain. We then proceed to introduce the three corpora that are used in our experiments in section 3, as well as the label mapping which was applied to derive the degree of negative escalation (section 3.2). Following this, we perform an initial acoustic analysis on each of the corpora in section 4, in this way beginning to understand the vocal behaviors that are present as escalation increases. We then outline the experimental setting which is applied for the classification task as well as a detailed description of the extract acoustic features and the experimental paradigms applied in section 5. The results obtained are discussed in detail in section 6, which is then followed by concluding remarks and future work outlook in section 7.

2. RELATED WORK

While the research community less explores the escalation of negative interactions compared to emotion recognition in general, the recognition of affective states such as anger, frustration, high arousal, and interpersonal states such as conflicts are related to our work. One special category is related to the analysis of call-center recordings, which often contain complaints and include manifestations of anger. Polzehl et al. (2011) found that loudness and MFCC features stood out for anger vs. non-anger recognition in three datasests two out of which contained call-center data. State-of-the-art deep learning algorithms outperformed classical approaches for call-center anger recognition (Deng et al., 2017). Wizzard-of-Oz setups are a successful method for emotion elicitation, and in Song et al. (2021) this method was used to elicit frustration in participants playing a game and recognizing frustration using a range of speech features and deep learning architectures. While

in research related to negative emotions the focus and annotation is often perform per person, in our work the focus is on the interaction between subjects and how it evolves.

A step closer to our work are several studies focusing on conflicts in political debates. In Kim et al. (2012), the authors show the general feasibility for a statically significant automatic conflict escalation prediction using TV debates. By using SVMs, they achieved an unweighted accuracy of 49.5%. Doing so, the minimum pitch and the overlap of speakers between turns were particularly predictive features. Georgakis et al. (2017) focused on audio-visual escalation resolution analysis, highlighting, in particular, the paucity of available labeled data in this area. For this reason, they present a new dataset with more than 2 h of Greek political television debates. Furthermore, they use a linear SVM to predict the binary escalation classes. The visual pose features outperformed the expressions ones, but fell behind audio features. Likewise, Kim et al. (2014)'s work focuses on televised political debates aiming to use dimensional annotations instead of classes. The models based on the Bayesian Gaussian process reached a correlation value of above 0.7 between the actual and predicted conflict levels on the SSPNet Conflict Corpus. The probabilistic model employed also indicates a high relevance of audio intensity (the minimum speech behavior is louder) and pitch for conflict prediction. In contrast to them, the types of interactions included in the datasets we explore focus on situations that might occur in daily life situations.

Several studies focus on cross-corpus prediction, as this offers a deeper understanding of a given phenomenon and combats the lack of appropriate data. In the following, we highlight a few of them related to general emotion recognition (especially arousal) using speech-driven approaches. In an extensive series of experiments, Schuller et al. (2010) utilized six databases for emotion recognition experiments using SVMs and audio normalization. The results vary considerably both between the corpora and between emotions. However, the generalisability of the results is limited, and they advocate for more consistency for cross-corpus learning in terms of, for example, emotion types and languages. Similarly, in Kaya and Karpov (2018), the impact of normalization strategies for cross-corpus acoustic emotion recognition is investigated across five languages captured in different datasets. In particular, they focus on the speaker-level features rather than the acoustic level and conclude that this form of normalization minimizes speaker- and corpus-related effects. In Schuller et al. (2011), two new voting strategies for the fusion of multiple corpora are investigated. Involving multiple corpora into the experimental setting it turned out that, especially when using emotion recognition in real applications, these models yields more stable prediction results. The best results for arousal are achieved with 63.4% UAR using all available data sets. The implications of non-emotional factors, such as domain and gender, in cross-corpus emotion recognition systems, are addressed in the work of Zhang et al. (2019). However, instead of using classical fusion, they use a multi-task learning approach. Here, each task considers one of these components as a training target. In terms of arousal, prediction performs better when the task is defined by corpus than by gender and the predictive power improves with an increasing number of corpora.

A similar work to our research which targets negative interaction utilizing two of the corpora included in this work, is from Lefter and Rothkrantz (2017), where several datasets are fused for the recognition of negative interactions. To investigate the importance of context in escalation scenarios, they vary it in the experimental setting by selecting and separating the context within a corpora and between corpora in distinct training and test sets. This study showed that different contexts lead to significant differences in performance—even within a single data set. In addition, further experiments compare the influence of varying feature origins in terms of modality so that audio features and video features and audio-visual feature-level fusion are considered. Among the modalities, audio appeared to be the most robust in terms of prediction results due to context-independent paralinguistic features. The further multi-corpus experiments did not improve the uni-corpus outcomes.

The research applied within this contribution is a continuation of experiments, which were included as part of the "Escalation" sub-challenge for the INTERSPEECH 2021 ComParE Challenge (Schuller et al., 2021). For the experiments that were part of the challenge, two out of the three datasets utilized in this study were applied in a similar cross-corpus manner, where one was used for training and the other for testing. From the experiments of this challenge, we see that hand-crafted features were performing competitively over deeper representation of the audio signal. Further to this, the end-to-end deep learning approach applied in the ComParE challenge did not perform as strongly as the conventional Support Vector Machine approach, which obtained a baseline score of at best 59.8% UAR, 5 percent points stronger than the closest end-to-end results. The results obtained proved promising and within this study, we explore these phenomena further and include additional related corpora.

3. DATA

This section describes the three corpora used in the experiments of this paper. Section 3.1 addressed the datasets content-wise, with respect to elicitation methods and their annotation. Section 3.2 elaborates on the process of finding a common ground truth concerning escalation levels between their annotations, and the resulting label mapping used in experimentation. **Table 1** gives an overview of the datasets in terms of numbers of scenarios, session, use of actors, whether the corpora contain interactions in dyads or larger groups, as well as the total number of samples (utterance-based), while a visual impression of the datasets can be found in **Figure 1**. **Table 2** analyses the datasets in terms of annotations provided and inter-rater agreements. Overall, the three datasets last approximately 3 h.

3.1. Datasets Description

3.1.1. The Dataset of Aggression in Trains (TR)

The Dataset of Aggression in Trains (TR) (Lefter et al., 2013) is an audio-visual corpus consisting of 21 scenarios of unwanted behaviors in trains and train stations (e.g., harassment, hooligans, theft, begging, traveling without a ticket, rude or

TABLE 1 | Description of the considered datasets in terms of numbers of scenarios, total played sessions, type and number of speaker in total and interacting per session, and total samples.

Database	Scenarios	Sessions	Participants	Speakers	Speakers per session	Samples
TR	21	21	Actors	12+	2/2+	534
SD	4	8	Actors	8	2/2+	413
NAA	3	36	Actors and non-actors	16	2	2,417

FIGURE 1 | From (left) to (right) TR, SD, and NAA corpora stills taken from the video modality.

irritating behavior toward other passengers). These scenarios were constructed from a surveillance perspective based on what should or should not happen in trains. The scenarios were presented as short descriptions (no scripts) to a group of actors, who then improvised based on these descriptions. The dataset has a high amount of escalating behavior content, including instances of verbal and physical aggression.

The TR dataset was annotated on aggression levels on a 3 point scale, from a surveillance perspective indicating moments where an operator would increase attention to a scene (medium aggression), or decide to act (high aggression). It contains annotation based on solely the audio content, solely video, and multimodal content. In this study, the audio-only annotation was used, which proved to be more suitable given that there would be instances of aggression with no audible content (e.g., physical offences).

3.1.2. The Stress at Service Desk Dataset (SD)

The Stress at Service Desk Dataset (SD) (Lefter et al., 2014) contains scenarios of problematic interactions situated at a service desk (e.g., , a slow and incompetent employee while a customer has an urgent request, employee not being helpful because of lunch break, or visitor in phone conversation blocking the service desk) performed by eight subjects. Just like TR, it was also designed for surveillance purposes, to detect when a situation gets out of hand and support is needed. It consists of mostly dyads, but occasionally multiple persons are interacting in the scene. Four scenarios are played two times each by a group of actors who received short scenario descriptions. The dataset is not scripted and again interaction is build up in response to each other (improvisations).

SD was annotated based on levels of stress on a three-point scale. Stress, in this case, results from the inability to achieve one's desires due to barriers posed by a third party, which manifests itself in escalating behavior. The annotation by four raters was performed based on the audio-visual recordings.

TABLE 2 | Available annotations per dataset, the number of raters and inter-rater agreement expressed as Krippendorf's alpha for the original labels of the three datasets prior to label mapping, according to Lefter et al. (2013, 2014, 2017).

Dataset	Available annotation	Scale	Raters	Agreement	α Actors	α Non-actors
TR	Aggression	3 pt.	7	0.77		
SD	Stress	5 pt.	4	0.75		
NAA	Aggression	5 pt.	3	0.78	0.8	0.33
	Valence	5 pt.	3	0.74	0.75	0.60
	Arousal	5 pt.	3	0.72	0.72	0.34
	Dominance	5 pt.	3	0.62	0.63	0.49

In the case of NAA, the agreement is provided for speech samples, and also separately per actors and non-actors speech.

3.1.3. The Dataset of Negative Affect and Aggression (NAA)

The Dataset of Negative Affect and Aggression (NAA) (Lefter et al., 2017) was originally designed in the context of VR aggression regulation therapy to support the behavior recognition component to which avatars in the VR would respond (role-plays are an important part of such therapies). Aside from audio-visual recordings, it also contains motion capture and a range of bio-signals (ECG, GSR, and EMG). It has a similar role-play setup, but as opposed to the other two datasets, it consists of interactions between aggression regulation training actors and non-actors (students)—who did not know upfront the nature of their opponents. In this case, both participants receive short role and goal description (opposing in nature), and the actors elicited the escalating behavior. A separate study (Lefter and Fitrianie, 2018) on validating the emotional experience of the non-actors involved, found evidence of significant changes

in the heart rate variability of the non-actors corresponding to degrees of the aggressive behavior of the actors. Example scenarios include a person being late for an interview needing to convince the bus driver to let them travel without a ticket, or convincing an educational advisor to register for an exam after the deadline.

In contrast to TR and SD, where the annotation was performed based on the entire scene, in NAA, the annotation was performed per person, based on utterance level segmentation, irrespective of whether the participants were speaking or not. The annotated dimensions are levels of aggression, valence, arousal, and dominance, on a five-point scale. The dataset was annotated by three raters, and **Table 2** outlines the inter-rater agreement for segments where either the actors or non-actors were speaking, overall, but also separately.

One aspect that can be noticed by examining the inter-rater agreement for NAA, is that the agreement on the actor data was higher for each dimension than the one for non-actors. In particular, low agreement values are observed for non-actor aggression and arousal. By further investigating confusion matrices between raters for these two dimension, we noticed that most often confusions took place between neighboring classes. We computed weighted Krippendorff's alpha with quadratic weights (Antoine et al., 2014) for these two dimensions. This agreement measure for ordinal data penalizes confusions between neighboring classes less, and confusions between class labels far apart more. The weighted alpha values are 0.48 for non-actor aggression, and 0.40 for non-actor arousal, which indicate more reasonable agreement of these annotations.

3.2. Label Mapping

The task of cross-corpus training and testing often requires that labels between corpora are mapped. Therefore, the system's performance will be influenced by how the concepts to be recognized (e.g., emotions, or escalation level) were defined and how the annotators were instructed. Often, the choice is made to reduce the considered datasets to subsets (e.g., the emotion categories that jointly exist), or mapping sets of emotions to negative and positive valence, or finding an equivalence between labels.

TR was annotated based on aggression levels on a three-point scale, SD was annotated for stress levels on a five-point scale, and NAA was annotated for aggression, valence, arousal, and dominance by three raters. Given these different annotation schemes, a process of label mapping to none, medium, and high escalation levels had to be performed. Further, it is essential to note that escalation is considered a characteristic of the interaction in the scene and not of a specific person.

While for TR and SD the data was annotated from an overall scene perspective, for NAA, the annotation was performed per individual participating in the interaction, for every utterance, irrespective of whether they were speaking or not. To create a scene-based label from these, the maximum levels the subject's aggression or arousal were considered. In addition, NAA contains annotations of whether the speaker is an actor, a non-actor, or whether there is speech overlap. As speech overlap has proven to be an important characteristic of escalating interactions (Caraty and Montacié, 2015); these segments are included, and their label is the maximum escalation level between the participants.

TABLE 3 | Procedure for mapping the original levels of the three datasets to low, medium, and high escalation.

		Escalation			
Dataset	**Original label**	**Low**	**Medium**	**High**	**Name**
TR	Aggression	1	2	3	
SD	Stress	1 & 2	3	4 & 5	
NAA	Aggression	1 & 2	3	4 & 5	NAA Agg.
NAA	Arousal	1 & 2	3	4 & 5	NAA N-A

Two options are provided for NAA, to which we will refer to as NAA Agg (constructed from the aggression label) and NAA N-A (Negative Arousal). Important: NAA N-A is constructed based on arousal level, only given that valence is negative.

Table 3 describes the label mapping procedure. While for TR and SD a relatively straightforward approach of maintaining or downsampling was adopted, the availability of multiple labels in NAA allowed us to explore two alternative mappings. NAA Agg is a label resulting from downsampling the aggression labels on a three-point scale. NAA N-A stands for Negative-Arousal and has the down-sampled level of arousal on a three-point scale given that valence is negative. Instances of positive valence are labeled as low escalation irrespective of the arousal level.

Figure 2 (left) shows the resulting distribution of escalation levels resulting from the mapping process. On the right side, the plot depicts how resulting escalation levels are distributed for the actor and the non-actors parts of NAA. Note that in the case of NAA Agg, there are almost no instances of increased escalation for non-actors (third bar). The situation changes with the other definition of escalation (first bar), meaning that non-actors were often perceived as negative arousal, but they were not aggressive.

4. ACOUSTIC ANALYSIS

To analyse the audio samples for each of the escalation derived classes, we perform an acoustic analysis on all the corpora. The fundamental frequency (F0) and volume intensity in the voice are known to vary during aggressive interactions (Pisanski et al., 2020). Further to this, there is a relationship between the two features. For example, typically, when individuals are portraying dominance, their F0 may be less influential than loudness (Idemaru et al., 2020). For these reasons, we extract both of these features from the audio signals of each class. For the F0 extraction, we only extract F0 for voiced parts of the samples to not skew the result based on the substantial segments of silence in the audio files.

In **Figure 3**, we report the standard deviation, median, and range as a violin plot for F0 and Intensity (dB) each of each corpus. When observing F0 in **Figure 3**, we can see for all corpora there is an increase in pitch along with the increase in escalation. Most prominently, the TR corpus shows the most dynamic change across classes for F0, with the higher class appearing to represent much more variety. There is little difference between the N-A label and Aggression for the NAA corpus—although, perhaps, slightly lower-end frequency dominance. However, as

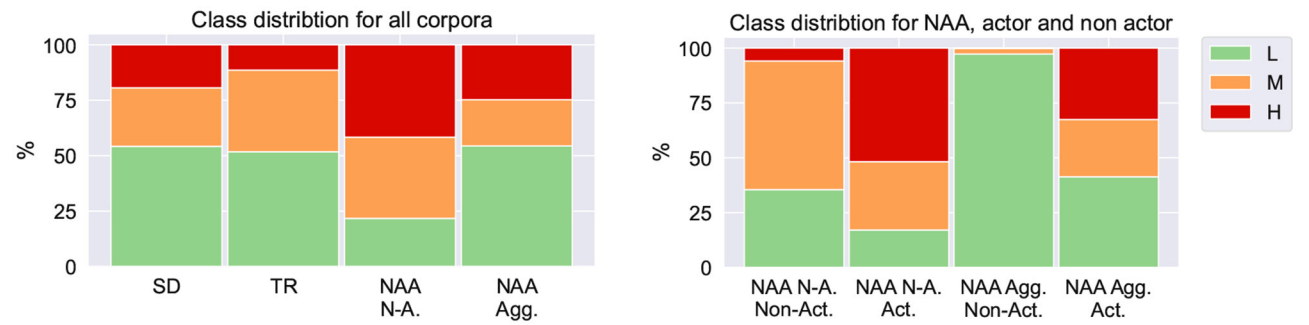

FIGURE 2 | Class distribution (as % of the total samples) across the corpora utilized in this study (left), as well as separation between actor and non-actor in the NAA corpus (right).

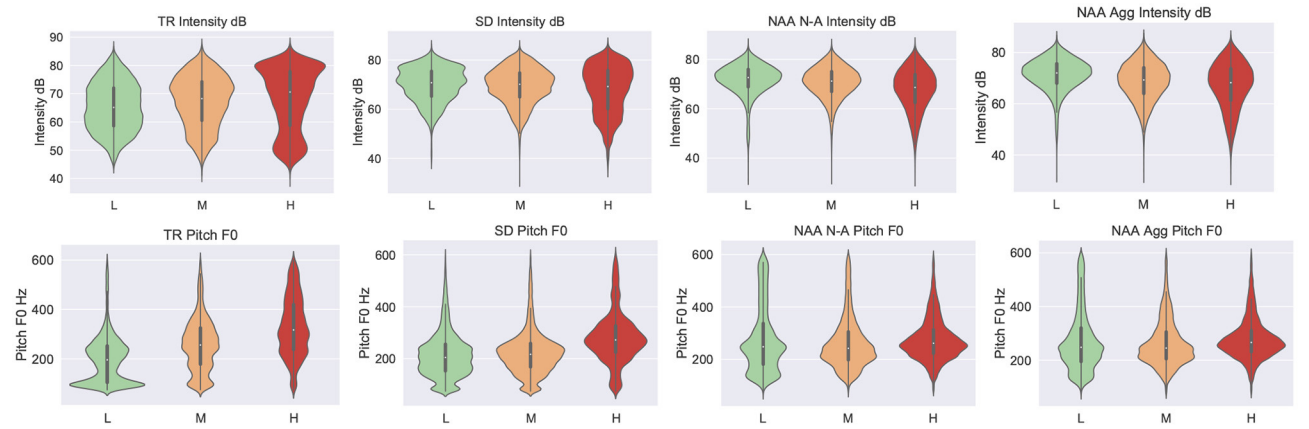

FIGURE 3 | Violin plot representation for the pitch (F0) and intensity (dB) vocal behavior for each of the corpora in utilized in this study. Summarized across all speakers. cf. **Figure 2** for legend.

with other corpora, there is a slight increase for pitch as escalation level increases.

When observing dB intensity level, we also see as expected that there is an increase in volume as escalation increases. This is most prominent for the TR and SD corpora, and in fact, there is a slight decrease for the NAA corpus. We consider that this slight decrease in intensity for NAA may be due to the nature of recording for this data set, which is mainly distant from the microphone. As can be seen in **Figure 1**, the NAA corpus is recorded in a larger room which may loose vocal instensity, as the acosutic space is known to effect loudness in general (Rychtáriková et al., 2013). This is also seen in the intensity results for SD which was also captured in a larger space. This is further supported by the increase in dynamics range for the higher class, which can be seen across each of the corpora; it appears that speakers in this class show a more comprehensive range of spoken volume. As we discussed previously, the NAA corpus has two labels which we consider to be related to the degree of escalation, aggression (Agg), and the combined negative-valence and arousal (N-A) signal. From the acoustic analysis, we can see that the classes for these two labels behave very similarly.

We also extract a selection of spectrograms in **Figure 4** from a 2.5 s segment of various audio examples. We do this to observe the vocal behavior qualitatively within the class.

We have plotted spectrograms for both male and female speakers in all escalation levels for all three datasets. We noticed for both female and male subjects a consistent increase in pitch and intensity with increasing escalation levels. To not generate confusion between separate female and male characteristics, and for compactness of the figure, in **Figure 4** we present spectrograms for male participants only.

Of most interest from these spectrogram representations, we affirm that there is an increase in pitch. Furthermore, we also affirm this broader divergence in the pitch as the degree of escalation increases. This is shown by the consistent increase in both the mean and standard deviation for F0. To conclude, we observe similarities in escalating behavior over the three datasets.

5. EXPERIMENTAL SETTINGS

For our main experiments, we create numerous combinations for the three corpora to explore (i) the phenomena of escalating emotion in speech, (ii) the effect of label definition and label

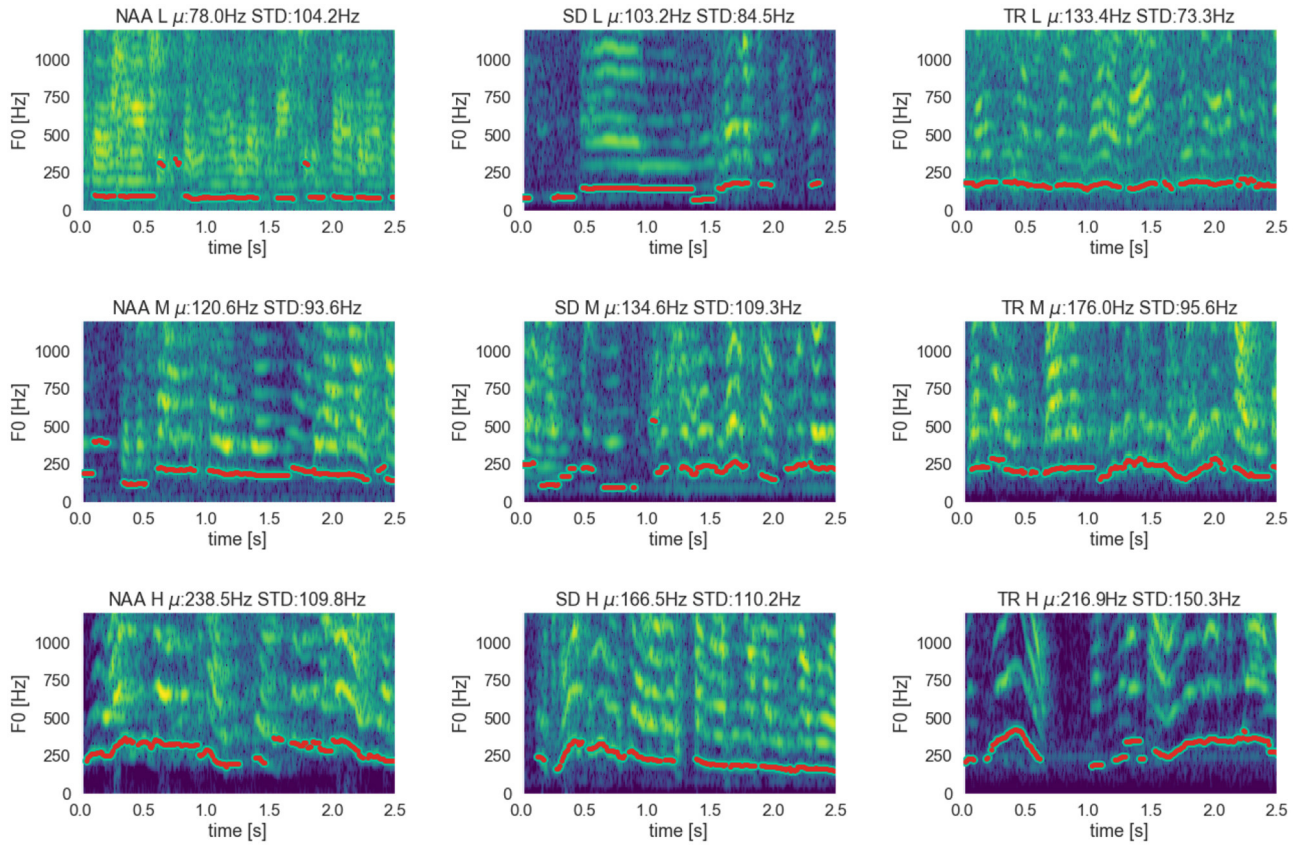

FIGURE 4 | Spectrogram representation of three males samples (one for each escalation class) for the three corpora being analyzed in this study. Showing for each escalation class (L)ow, (M)edium, and (H)igh, and reporting mean (μ), and standard deviation (STD) for Pitch (F0).

mapping in the context of escalating emotion (aggression vs. stress vs. "negative arousal"), and (iii) the effect of spontaneous and acted speech.

5.1. Labels

The experiments are performed using the escalation labels resulting from the mapping procedure described in section 3.2. In the case of NAA, to evaluate the effects of alternative label definitions, both NAA N-A and NAA Agg will be explored. In the case of NAA, the complete dataset (including overlapping speech) is used in the cross-corpus experiment.

For testing on actor vs. non non-actor samples only the NAA N-A label is considered, given the unbalanced representation of non-actor samples in NAA Agg. In the case of actor and non-actor testing, the overlapping speech samples from NAA were removed to ensure clean testing conditions per target group.

5.2. Features

We first extract features from each of the three corpora. Namely, the ComParE hand-crafted set, utilizing the openSMILE toolkit, and additionally we utilize the DeepSpectrum toolkit[1] as a state-of-the-art spectrogram-based deep learning approach.

The ComParE acoustic feature set is a well established set which has shown to give consistent insights for related domains of speech analysis (Stappen et al., 2019), including states of stress (Baird et al., 2019; Stappen et al., 2021), and anxiety (Baird et al., 2020). The ComParE feature set is also used as the baseline feature for the INTERSPEECH ComParE challenges since 2013 (Schuller et al., 2013), and further extended in 2016 (Schuller et al., 2016). As with the 2021 ComParE challenge (Schuller et al., 2021), we extract the features from the entire audio samples, resulting in feature sets of 6,373 static features, which are derived from the calculation of static-functionals obtained from low-level descriptor (LLD) contours (Eyben et al., 2013; Schuller et al., 2013).

Much like the ComParE set, DeepSpectrum has also shown to be applicable for a variety of speech driven tasks, particularly in emotion (Baird et al., 2021). The toolkit applies pre-trained convolutional neural networks (CNNs) to extract deep representations from the input audio data (Amiriparian et al., 2017). For this method, first, the audio signal is transformed into a Mel-spectrogram applying a Hanning window of width

[1]https://github.com/DeepSpectrum/DeepSpectrum.

TABLE 4 | Results obtained for all corpora combinations for both ComParE and DeepSpectrum feature sets using the NAA N-A label.

			No upsampling				Upsampling			
			Dev (inner)		Test (cross)		Dev (inner)		Test (cross)	
Training set	**Test set**	**Feature set**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**
TR	NAA	ComParE	54.5	45.0	**57.4**	**57.1**	59.4	53.2	**55.3**	**55.4**
TR		DeepSpectrum	50.9	56.8	43.1	38.0	52.0	50.4	51.6	42.7
SD		ComParE	70.5	64.4	53.1	47.2	71.8	67.8	45.1	36.8
SD		DeepSpectrum	64.2	63.6	54.9	52.1	61.4	64.4	54.6	50.6
SD + TR		ComParE	65.0	62.6	51.1	51.3	63.4	65.1	52.3	53.7
SD + TR		DeepSpectrum	58.0	60.1	49.1	54.3	55.1	53.0	50.5	53.4
TR	SD	ComParE	54.5	45.0	**62.8**	59.6	59.4	53.2	**63.8**	62.5
TR		DeepSpectrum	50.9	56.8	54.6	62.2	52.0	50.4	55.1	59.1
NAA		ComParE	68.5	72.4	51.0	42.1	67.6	73.2	51.7	45.3
NAA		DeepSpectrum	65.6	71.3	52.4	**63.7**	64.5	69.9	47.2	**66.1**
TR + NAA		ComParE	66.5	66.4	52.2	49.9	67.4	69.2	53.3	52.3
TR + NAA		DeepSpectrum	64.6	68.2	51.8	62.0	62.6	66.1	50.2	57.9
NAA	TR	ComParE	68.5	72.4	54.3	44.4	67.6	73.2	58.6	50.8
NAA		DeepSpectrum	65.6	71.3	56.4	58.4	64.5	69.9	54.4	59.2
SD		ComParE	70.5	64.4	60.0	**61.2**	71.8	67.8	50.1	57.6
SD		DeepSpectrum	64.2	63.6	57.7	55.8	61.4	64.4	57.3	57.9
SD + NAA*		ComParE	68.2	69.6	**61.7***	56.6	68.7	72.3	**62.7***	60.0*
SD + NAA*		DeepSpectrum	63.8	68.4	59.2*	58.4*	65.8	70.7	59.5*	**61.4***

To observe the effect of potential class imbalance, we are reporting UAR and WAR (%), as well as results obtained with and with and without data upsampling. Emphasized in bold are the strongest results for a given paradigm's test set UAR and WAR and are discussed in section 6. Results marked with a "" represent improvements of merging corpora for training over using a single training corpus.*

32 ms and an overlap of 16 ms—128 Mel frequency bands are then computed from this. These extracted spectrograms are then forwarded to a pre-trained DenseNet121 network (Huang et al., 2017), and as with the challenge baseline, the activations of the networks "avg_pool" layer are extracted. This process results in a feature set of 2,048-dimensional feature vectors, which can be seen as a lower-level representation of the initial Mel-spectrograms (Amiriparian, 2019).

Within the machine learning pipeline, all the feature representations were scaled to zero mean and unit standard deviation, utilizing the (StandardScaler of Scikit-learn). The scaling was applied using the parameters from the respective training set—for the final test, we combine the training and development set, and in this case, the scaling parameters are calculated again.

5.3. Data Partitioning and Training Procedure

Each of the three datasets was split in two subsets, one to serve as training and one as development sets for the cross-corpus experiments. Special care has been paid to ensuring speaker independence and a balancing of the labels in each subset. For TR and SD the same split as in the Interspeech ComParE Challenge was used, to ensure that the results generated here are comparable to the ones from the challenge. For NAA a similar approach was used, where in addition to speaker independence, we also accounted for a balance between the actor and non-actor speech samples between partitions.

Proceeding the feature extraction, we prepare a series of partition combinations to analyse across-corpora the conflict escalation label across multiple scenarios. In the *cross*-corpus analysis, we utilize one corpus as training and development and then another as testing. In this way, from the development result, we can observe the *inner*-corpus result, and from the test set, we can observe how conflict escalation is recognized across-corpora. To further evaluate the cross-corpus effect, we additionally perform a *combined* analysis where two of the corpora are combined for training and development: here both training sets are merged and both development sets are merged for form a merged train and development set respectively. Lastly, as the NAA corpus contains both spontaneous (non-actor) and acted speech, we also evaluate this by splitting the NAA test in this way.

5.3.1. Classifier

We choose to utilize a Support Vector Machine (SVM) with a linear kernel to classify "escalating negativity" for several reasons. Applying this system allows for more transparent reproducibility and aligns with the ComParE 2021 Challenge approach, thus making it slightly more comparable to others utilizing this dataset. As well as this, the smaller size of the datasets being utilized means that SVMs may produce more

TABLE 5 | Results with the NAA Agg labels.

			No upsampling		Upsampling	
Training set	**Test set**	**Feature set**	**Dev (Inner)**	**Test (Cross)**	**Dev (Inner)**	**Test (Cross)**
TR	NAA	ComParE	54.4	**59.3**	57.9	58.5
		DeepSpectrum	50.9	47.4	52.0	49.2
SD		ComParE	70.4	55.9	72.1	47.7
		DeepSpectrum	64.2	**57.9**	61.4	57.4
SD + TR		ComParE	64.9	**57.6**	60.7	56.4
		DeepSpectrum	58.0	50.4	55.7	56.8
NAA	TR	ComParE	70.9	**56.1**	71.9	**56.1**
		DeepSpectrum	67.5	58.4	66.2	58.5
SD + NAA		ComParE	71.0	59.1	70.2	**60.4**
		DeepSpectrum	67.4	60.2	66.7	58.8
NAA	SD	ComParE	70.9	56.7	71.9	55.3
		DeepSpectrum	67.5	**60.6**	66.2	59.6
TR + NAA*		ComParE	68.0	65.0*	68.7	**66.1***
		DeepSpectrum	66.3	60.8*	66.4	58.8*

*Reporting Unweighted Average Recall (%). We include results which stem from training data with and without upsampling of the minority classes. For results involving only the SD and TR datasets and not influenced by the NAA labeling, please refer to **Table 4**. Emphasized in bold are the strongest results for a given paradigm's test set UAR and WAR and are discussed in section 6. Results marked with a "*" represent improvements of merging corpora for training over using a single training corpus.*

robust and more easily interpretable results than deep neural networks. As mentioned for optimization, the development set is utilized, and the complexity parameter C is altered ($C \in 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1}). The training and development set is fused for testing, and we utilize the best value for C.

Given the strong class imbalance in specific corpus combinations, we also upsample the training set (utilizing the SCIKIT-LEARN RANDOMOVERSAMPLER function, which picks samples at random and duplicates them to increase the minority class and ensure there is balance between all class. We report results for both with and without upsampling, as this was not always beneficial. Furthermore, to observe the possible effect of class imbalance, we report both unweighted average recall (UAR) and weighted average recall (WAR) as an evaluation measure for all experiments.

6. RESULTS AND DISCUSSION

An overview of all classification results for each corpus paradigm is given in **Tables 4**, **6**. Of promise, when observing all results from **Table 4**, we see that in all cross-corpus testing for both feature sets, the results are above chance-level (33.3%) for the reported UAR. These robust results show us that escalating negativity is present in a similar manifestation for all the corpora. We see solid results for testing on the SD and TR corpora, which may be due to both similar class imbalance and the derived label type being stress and aggression, respectively. In other words, there could be more highly aroused examples in this data as compared to the NAA results.

In general, for these results, we see that COMPARE features are performing more robustly than DEEPSPECTRUM. This performance may be due to the sizeable dimensional space for COMPARE feature; however, given its more specific focus on classical speech features, we expect that in the context of our analysis, this was the most meaningful addition from the COMPARE features as compared to the image-based DEEPSPECTRUM features.

Due to the imbalance in classes, we applied upsampling to the training data. From **Tables 4**, **5**, we notice that often the overall test performances with and without upsampling are not very different. While for a number of cases upsampling results in an improvement, it can also decrease the overall performance. We therefore can't report of finding a consistent trend generated by upsampling. Furthermore, the efficiency of upsampling can be influenced by the initial distribution of the classes in the training set. Hence, the inconsistent results could have been influenced by the fact that in our case the distribution in the training set is changing with the dataset(s) chosen as training material (recall **Figure 2**-left).

Both **Figures 5**, **6** depict a selection of confusion matrices for training on TR and testing on either the non-actor part of NAA, or the actor part of NAA respectively. This allows us to inspect the performance for the high escalation class, which may be very important not to miss from an application perspective. From both figures we can observe a higher recognition of the high escalation class when the COMPARE features are applied, as opposed to when the DEEPSPECTRUM features are applied. An important thing to note is that for the non-actors, the high escalation class is highly unrepresented (see **Figure 2**-right), consisting of 41 samples as opposed to 512 samples for actors. Another aspect that can be noted consistently from the two figures is that upsampling results in increasing recall of the high escalation class when utilizing the DEEPSPECTRUM features, while for COMPARE the opposite effect is observed. In all cases displayed in the two figures, the highest confusion of high escalation is with medium escalation.

In **Table 6** the results show that, in general, escalation can be recognized above chance-level for both acted and spontaneous (non-actor) speech. The major limitation for the field, in general, is obtaining enough spontaneous samples with highly negative emotional expression. However, when observing the WAR results, which exclude the class imbalance, we see more robust results, which leads us to consider more promise for this approach upon obtaining representative data.

There are additional factors that may have affected the drop in performance for the non-actor segments, in addition to authenticity of expression. It is acknowledged in Lefter et al. (2017) that the behavior of non-actors was more difficult to interpret and annotate, which also resulted in lower inter-rater agreement for all annotated dimensions (cf. **Table 2**). Furthermore, the annotation was performed from an audio-visual perspective, and it is possible (for both NAA and SD) that triggers of higher escalation may have resulted from the video modality only.

TABLE 6 | Results obtained for all corpora combinations where NAA is used as the testing set using the NAA N-A label.

			No upsampling				Upsampling			
			Dev (Inner)		Test (Cross)		Dev (Inner)		Test (Cross)	
Training set	**Test set**	**Feature set**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**
TR	NAA (actor)	ComParE	54.5	45.0	**62.9**	**63.2**	59.4	53.2	58.9	**59.9**
TR		DeepSpectrum	50.9	56.8	45.3	33.8	52.0	50.4	45.7	39.3
SD		ComParE	70.5	64.4	56.6	52.7	71.8	67.8	49.4	41.4
SD		DeepSpectrum	64.2	63.6	54.8	54.9	61.4	64.4	**54.2**	52.6
SD + TR		ComParE	65.0	62.6	50.9	51.9	63.4	65.1	52.4	53.2
SD + TR		DeepSpectrum	58.0	60.1	48.9	50.6	55.1	53.0	49.6	51.9
TR	NAA (non-actor)	ComParE	54.5	45.0	50.3	50.5	59.4	53.2	48.2	50.8
TR		DeepSpectrum	50.9	56.8	37.9	51.2	52.0	50.4	37.0	47.4
SD		ComParE	70.5	64.4	40.6	39.5	71.2	67.8	37.7	38.8
SD		DeepSpectrum	64.2	63.6	**53.9**	40.8	61.4	64.4	**49.9**	41.3
SD + TR		ComParE	65.0	62.6	38.2	53.5	63.4	65.1	40.6	60.5
SD + TR		DeepSpectrum	58.0	60.1	40.4	**70.5**	55.1	53.0	28.2	**61.2**

Splitting the data into non-actor and actor samples for the NAA corpus. To observe the effect of potential class imbalance, we are reporting UAR and WAR (%), as well as results obtained with and with and without data upsampling. Emphasized in bold are the strongest results for a given paradigm's test set UAR and WAR and are discussed in section 6.

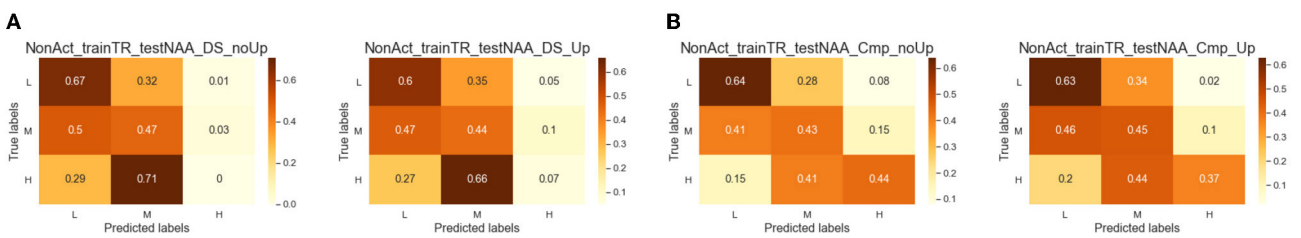

FIGURE 5 | Confusion matrices (in percentage) for training on TR and testing on the non-actor part of NAA. **(A)** Experiments using the DEEPSPECTRUM features with no upsampling (left) and upsampling (right). Note a slight increase in recall of the high class for upsampling. **(B)** Experiments using the COMPARE features with no upsampling (left) and upsampling (right). Note a slight decrease in recall of the high class for upsampling.

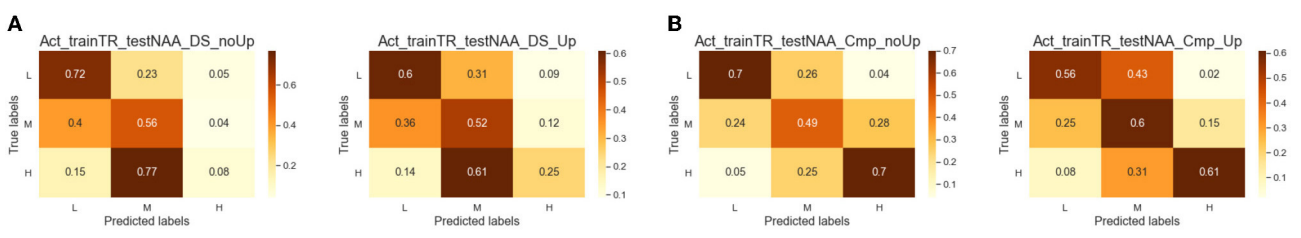

FIGURE 6 | Confusion matrices (in percentage) for training on TR and testing on the actor part of NAA. **(A)** Experiments using the DEEPSPECTRUM features with no upsampling (left) and upsampling (right). Note a slight increase in recall of the high class for upsampling. **(B)** Experiments using the COMPARE features with no upsampling (left) and upsampling (right). Note a slight decrease in recall of the high class for upsampling.

The approach of merging datasets for training is based on the fact that intuitively, more training data would lead to better recognition. Inconsistent results are shown in the literature, where merging improves or performs below the best single training corpus. In our study, we explore the effects of merging datasets for the two escalation definitions created for the NAA dataset. Results for NAA using the negative arousal equivalent of escalation (NAA N-A) are presented in **Table 4**, while **Table 5** outlines results for NAA using the aggression-based equivalent of escalation (NAA Agg).

On the one hand, when using the NAA annotation focusing on aggression (NAA Agg), the only cross-corpus experiment resulting in better results for merged dataset training instead of single dataset training is when NAA and TR are merged. Note that NAA Agg and the annotation of TR are both focused on aggression. On the other hand, when the NAA annotation

focusing one negative arousal was used (NAA N-A), the only merging dataset combination that proved successful was NAA with SD. Both of these labels are related to negative arousal. Based on these observations, we conclude that the only situations when merging datasets actually worked, was when the label definitions of the mapped original labels were mostly in line.

A fraction of our results are compsarable with the results from the Interspeech ComParE Challenge (Schuller et al., 2021), namely the part where training is performed on the SD dataset, and TR is used for testing. The best performing combination for the Escalation subchallenge baseline was of an SVM classifier with a bag of audio words features, resulting in 59.0 UAR. The winning team employed a combination of the ComParE feature set and Fisher Vectors to increase UAR to 63.2 with a linear SVM classifier (Egas-López et al., 2021).

There are a number of limitations of our work, which we discuss in this paragraph. First, the three datasets used in this study contain predominantly footage of actors, which may have an impact on the realism of the data. However, the interactions were not scripted and evolved naturally as subjects reacted to each other, which has contributed to degree of naturalness of the recordings. Second, for the NAA dataset which featured non-actor participants, the inter-rater agreement for the non-actor segment is lower than for the actor segments. By investigating the confusion matrices between annotators we noticed that much of the confusions were between neighboring classes, and part of these confusions were diminished by the label mapping procedure. Third, the datasets have a relatively small number of samples, which may impact performance, generalization and the applicability and success of deep learning methods. Fourth, as the datasets came with different labeling schemes, the label mapping process—as actually shown in the experiments—has an important weight in the results.

7. CONCLUSION AND FUTURE WORK

In this contribution, we performed a series of cross-corpus experiments to analyse the generalization ability of escalation recognition systems. Three corpora with different recording conditions but rich in a variety of escalation behaviors were employed. Each of these datasets was annotated based on a different dimension and at different granularities. All cross-corpus classification tasks resulted in above chance-level recognition rates. Among the two feature sets utilized, higher overall performance was obtained using ComParE, while DeepSpectrum provided more robust representations of the highest escalation class, which is highly relevant from an applications' perspective.

After label mapping was conducted, an acoustic analysis indicated that similar escalation vocal patterns are present in the three datasets. Increases in pitch and intensity accompanied an increase in escalation.

Merging datasets for training has shown contradicting results in the literature. By examining results using two alternative escalation definitions for the NAA dataset, we conclude that the degree to which a set of corpora will merge well as a training set is highly dependent on the similarity of the labeling definitions. While this is not surprising, to the best of our knowledge, this is the first study that explores alternative label mapping strategies and their consequences. These effects of label mapping are presents even though both labels considered for the NAA dataset had similar acoustic patterns.

Concerning evaluating recognition when stepping away from actor data, we observe a decrease in recognition rates but still stable behavior above chance performance. The drop in performance can be attributed to factors other than the degree of spontaneity, including the fact that the behavior of the non-actors was much more challenging to interpret also for humans, resulting in lower inter-rater agreement and consequently in less robust labels.

From the various findings of this study, there are many areas that can be targeted in future work. As a priority, the machine learning research community should work collaboratively with researchers in psychology to collect more spontaneous interactions for instances of escalating speech, especially as the higher escalation level class may be more fruitful for security and safety-driven applications. Further to this, with additional data, the somewhat rudimentary methodologies applied in this article for machine classification can be extended to state-of-the-art deeper frameworks, which may show more robust results.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study contains personally identifiable information and, as such, are not available in public domain. Requests to access TR and SD should be directed to i.lefter@tudelft.nl. All requests will be reviewed and granted if the necessary privacy and ethical requirements can be met.

ETHICS STATEMENT

The current study did not directly involve human participants. The data analyzed in this study is part of previous contributions which have been cited accordingly. Data reuse was performed in accordance with the conditions imposed during the initial data collection.

AUTHOR CONTRIBUTIONS

IL: literature analysis, experimental design, manuscript preparation, and drafting manuscript. AB: experimental design, computational analysis, manuscript preparation, and drafting manuscript. LS: literature analysis, experimental design, and manuscript preparation. BS: manuscript editing. All authors revised, developed, read, and approved the final manuscript.

FUNDING

This project received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

REFERENCES

Amiriparian, S. (2019). *Deep representation learning techniques for audio signal processing* (Ph.D. thesis). Technische Universität München, Munich, Germany.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., et al. (2017). "Snore sound classification using image-based deep spectrum features," in *Proc. Interspeech 2017* (Stockholm), 3512–3516. doi: 10.21437/Interspeech.2017-434

Antoine, J.-Y., Villaneau, J., and Lefeuvre, A. (2014). "Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation," in *EACL 2014* (Gotenborg). doi: 10.3115/v1/E14-1058

Baird, A., Amiriparian, S., Cummins, N., Sturmbauer, S., Janson, J., Messner, E.-M., et al. (2019). "Using speech to predict sequentially measured cortisol levels during a trier social stress test," in *Proc. INTERSPEECH 2019* (Graz: ISCA), 534–538. doi: 10.21437/Interspeech.2019-1352

Baird, A., Amiriparian, S., Milling, M., and Schuller, B. W. (2021). "Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention," in *Proc. 2021 Spoken Language Technology Workshop (SLT)* (Virtual: IEEE), 397–402. doi: 10.1109/SLT48900.2021.9383542

Baird, A., Cummins, N., Schnieder, S., and Schuller, B. W. (2020). "An evaluation of the effect of anxiety on speech-computational prediction of anxiety from sustained vowels," in *Proc. INTERSPEECH 2020* (Shanghai: ISCA), 4951–4955. doi: 10.21437/Interspeech.2020-1801

Caraty, M.-J., and Montacié, C. (2015). "Detecting speech interruptions for automatic conflict detection," in *Conflict and Multimodal Communication* (Springer), 377–401. doi: 10.1007/978-3-319-14081-0_18

Deng, J., Eyben, F., Schuller, B., and Burkhardt, F. (2017). "Deep neural networks for anger detection from real life speech data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (San Antonio, TX: IEEE), 1–6. doi: 10.1109/ACIIW.2017.8272614

Egas-López, J. V., Vetráb, M., Tóth, L., and Gosztolya, G. (2021). Identifying conflict escalation and primates by using ensemble x-vectors and fisher vector features. *Proc. Interspeech* 2021, 476–480. doi: 10.21437/Interspeech.2021-1173

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM Multimedia* (Barcelona), 835–838. doi: 10.1145/2502081.2502224

Georgakis, C., Panagakis, Y., Zafeiriou, S., and Pantic, M. (2017). The conflict escalation resolution (confer) database. *Image Vis. Comput.* 65, 37–48. doi: 10.1016/j.imavis.2016.12.001

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of Computer Vision and Pattern Recognition* (Honolulu, HI), 4700–4708. doi: 10.1109/CVPR.2017.243

Idemaru, K., Winter, B., Brown, L., and Oh, G. E. (2020). Loudness trumps pitch in politeness judgments: evidence from korean deferential speech. *Lang. Speech* 63, 123–148. doi: 10.1177/0023830918824344

Kaya, H., and Karpov, A. A. (2018). Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275, 1028–1034. doi: 10.1016/j.neucom.2017.09.049

Kim, S., Valente, F., Filippone, M., and Vinciarelli, A. (2014). Predicting continuous conflict perception with bayesian gaussian processes. *IEEE Trans. Affect. Comput.* 5, 187–200. doi: 10.1109/TAFFC.2014.2324564

Kim, S., Yella, S. H., and Valente, F. (2012). "Automatic detection of conflict escalation in spoken conversations," in *Thirteenth Annual Conference of the International Speech Communication Association* (Portland, OR). doi: 10.21437/Interspeech.2012-121

Lefter, I., Burghouts, G. J., and Rothkrantz, L. J. (2014). An audio-visual dataset of human-human interactions in stressful situations. *J. Multimodal User Interfaces* 8, 29–41. doi: 10.1007/s12193-014-0150-7

Lefter, I., and Fitrianie, S. (2018). "The multimodal dataset of negative affect and aggression: a validation study," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO: ACM), 376–383. doi: 10.1145/3242969.3243013

Lefter, I., Jonker, C. M., Tuente, S. K., Veling, W., and Bogaerts, S. (2017). "NAA: a multimodal database of negative affect and aggression," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (San Antonio, TX), 21–27. doi: 10.1109/ACII.2017.8273574

Lefter, I., and Rothkrantz, L. J. (2017). "Multimodal cross-context recognition of negative interactions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (San Antonio, TX: IEEE), 56–61. doi: 10.1109/ACIIW.2017.8272586

Lefter, I., Rothkrantz, L. J., and Burghouts, G. J. (2013). A comparative study on automatic audio-visual fusion for aggression detection using meta-information. *Pattern Recogn. Lett.* 34, 1953–1963. doi: 10.1016/j.patrec.2013.01.002

Pisanski, K., Raine, J., and Reby, D. (2020). Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries. *R. Soc. Open Sci.* 7, 191642. doi: 10.1098/rsos.191642

Polzehl, T., Schmitt, A., Metze, F., and Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. *Speech Commun.* 53, 1198–1209. doi: 10.1016/j.specom.2011.05.002

Rychtáriková, M., Müllner, H., Urbán, D., Chmelik, V., Roozen, B., and Glorieux, C. (2013). "Influence of temporal and spectral features of neighbour's noise on perception of its loudness," in *42nd International Congress and Exposition on Noise Control Engineering 2013, INTER-NOISE 2013: Noise Control for Quality of Life* (Innsbruck), 1408–1415.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., et al. (2016). "The Interspeech 2016 computational paralinguistics challenge: deception, sincerity & native language," in *Proc. Interspeech* (San Francisco, CA), 2001–2005. doi: 10.21437/Interspeech.2016-129

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The Interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. Interspeech* (Lyon), 148–152. doi: 10.21437/Interspeech.2013-56

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., et al. (2010). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Trans. Affect. Comput.* 1, 119–131. doi: 10.1109/T-AFFC.2010.8

Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G. (2011). "Using multiple databases for training in emotion recognition: to unite or to vote?" in *Twelfth Annual Conference of the International Speech Communication Association* (Florence). doi: 10.21437/Interspeech.2011-468

Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., et al. (2021). The interspeech 2021 computational paralinguistics challenge: covid-19 cough, covid-19 speech, escalation & primates. *arXiv[Preprint].arXiv:2102.13468*. doi: 10.21437/Interspeech.2021-19

Song, M., Mallol-Ragolta, A., Parada-Cabaleiro, E., Yang, Z., Liu, S., Ren, Z., et al. (2021). Frustration recognition from speech during game interaction using wide residual networks. *Virtual Real. Intell. Hardw.* 3, 76–86. doi: 10.1016/j.vrih.2020.10.004

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E., et al. (2021). "The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress," in *Proc. 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop* (Chengdu: ACM). doi: 10.1145/3475957.3484450

Stappen, L., Karas, V., Cummins, N., Ringeval, F., Scherer, K., and Schuller, B. (2019). "From speech to facial activity: towards cross-modal sequence-to-sequence attention networks," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)* (Kuala Lumpur: IEEE), 1–6. doi: 10.1109/MMSP.2019.8901779

Vinola, C., and Vimaladevi, K. (2015). A survey on human emotion recognition approaches, databases and applications. *Electron. Lett. Comput. Vis. Image Anal.* 14, 24–44. doi: 10.5565/rev/elcvia.795

Zhang, B., Provost, E. M., and Essl, G. (2019). Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground

while preserving differences. *IEEE Trans. Affect. Comput*. 10, 85–99. doi: 10.1109/TAFFC.2017.2684799

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Lefter, Baird, Stappen and Schuller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.