Entheos: A Multimodal Dataset for Studying Enthusiasm

Carla Viegas
Carnegie Mellon University
Pittsburgh, USA
cviegas@cs.cmu.edu

Malihe Alikhani
University of Pittsburgh
Pittsburgh, USA
malihe@pitt.edu

Abstract

Enthusiasm plays an important role in engaging communication. It enables speakers to be distinguished and remembered, creating an emotional bond that inspires and motivates their addressees to act, listen, and coordinate (Bettencourt et al., 1983). Although people can easily identify enthusiasm, this is a rather difficult task for machines due to the lack of resources and models that can help them understand or generate enthusiastic behavior. We introduce Entheos, the first multimodal dataset for studying enthusiasm composed of video, audio, and text. We present several baseline models and an ablation study using different features, showing the importance of pitch, loudness, and discourse relation parsing in distinguishing enthusiastic communication.

1 Overview

Although different emotional constructs such as *anger* and *happiness* have been studied extensively in the field of natural language processing (NLP), more fine-grained emotional expressions such as enthusiasm or charisma are relatively unexplored. Such models and datasets can benefit different areas of NLP and AI. Multimodal human-machine interaction can be more effective if systems can find a deeper understanding of more complex emotional responses or generate appropriate emotionally-aware communicative presentations. Given the importance of enthusiasm in teaching (Bettencourt et al., 1983; Zhang, 2014), for instance, researchers are studying the effect of virtual agents and robots that can behave in an enthusiastic manner (Liew et al., 2017, 2020; Saad et al., 2019). The current research is far from generating natural enthusiastic behavior.

Although previous research results in psychology, education, and business have studied the importance of enthusiasm in communication (Bettencourt et al., 1983; Sandberg, 2007; Keating, 2011; Antonakis et al., 2019), it is relatively unexplored in the NLP and dialogue literature. We take a step to bridge this gap by introducing the first multimodal dataset labeled with levels of enthusiasm following the definition that Keller et al. (2016) provided.

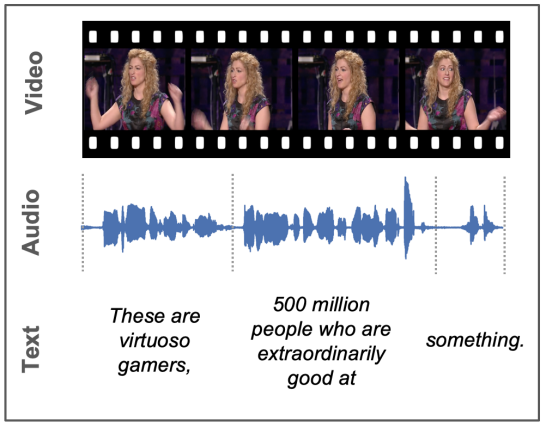

Figure 1: An enthusiastic sample from the Entheos dataset, showing aligned video frames, audio, and text.

Our contributions are as follows: First, we present Entheos (ἔνθεος: being possessed by a god, root for enthusiasm), the first multimodal dataset of TED talk speeches with annotated enthusiasm level[1] (Section 3). It contains sentence segments, labeled as either monotonous, normal, or enthusiastic. Figure 1 shows an example of an enthusiastic sample. Second, in search of finding multimodal signals for understating enthusiasm, we present an analysis of our data to identify attributes present in enthusiastic speech in different modalities (Section 3.5 and 5). Finally, we also provide several baseline models using different kinds of features extracted from text, speech, and video. In

[1] https://github.com/clviegas/Entheos-Dataset

addition, we show the importance of identifying discourse relations in predicting enthusiasm (Section 5).

2 Related Work

In this paper, we focus on investigating resources and models that can help us gain insights into ways by which computers can understand and predict enthusiasm. This topic is relatively unexplored in the computer science field although it has been extensively studied in psychology (Bettencourt et al., 1983; Sandberg, 2007; Keating, 2011; Antonakis et al., 2019).

Enthusiasm Limited work exists on the automatic detection of enthusiasm and has been mainly done in the text domain. Inaba et al. (2011) worked on the detection of enthusiasm in human text-based dialogues, using lexical features and word co-occurrences with conditional random fields in order to distinguish enthusiastic utterances from non-enthusiastic ones. They defined enthusiasm as "the strength of each participant's desire to continue the dialogue each time he/she makes an utterance". In our work, we instead combine different modalities and features to detect enthusiasm and we define an enthusiastic speaker as "stimulating, energetic, and motivating" (Keller et al., 2016). Tokuhisa and Terashima (2006) also worked with human-to-human conversational dialogues and annotated dialogue acts (DAs) and rhetorical relations (RRs) on a sentence-level. An enthusiasm score in the range of 10-90 was given without providing examples to the annotators. The relationship between DAs, RRs, and enthusiasm was analyzed based on the frequencies. They found that affective and cooperative utterances are significant in an enthusiastic dialogue. We detected RRs automatically and trained a feed forward network to classify enthusiasm in three levels: monotonous, normal, and enthusiastic. During data annotation, examples for each category were available as references. Twitter data have also been used to detect enthusiasm. Mishra and Diesner (2019) created a dataset with enthusiastic and passive labels. Enthusiastic tweets had to include personal expression of emotion or call to action, whereas passive tweets lacked clear emotive content or call to action. They trained logistic regression models using salient terms. We evaluate emotional expressions in several modalities. We use acoustic features that relate to emotion such as pitch and voice quality, and also Facial Action Units extracted from videos which measure the intensity of different facial expressions.

Charisma Enthusiasm is also a trait that can be displayed by charismatic speakers (Spencer, 1973), which in addition are perceived as competent, passionate, and self-confident (Niebuhr, 2020). Charisma is a desired trait for leaders in economy and politics (Antonakis et al., 2019; De Jong and Den Hartog, 2007) because it can influence followers to undertake personally costly yet socially beneficial actions. Niebuhr et al. (2016) have investigated the prosodic attributes of charismatic speakers. They analyzed pitch level, pitch variation, loudness, duration of silence intervals, etc and concluded that charisma can be trained as far as melodic features are concerned. In addition to analyzing the relationship of different attributes with enthusiasm, we also trained a model that can distinguish between different levels of enthusiasm.

Although sentiment analysis and emotion detection have been studied extensively in unimodal and multimodal frameworks as shown in several surveys (Marechal et al., 2019; Garcia-Garcia et al., 2017; Seyeditabari et al., 2018; Sudhakar and Anil, 2015) there is a gap in the analysis, detection and generation of enthusiastic behavior. Our dataset will allow to extend the work in understanding human behavior and also generate more natural virtual agents (Zhang, 2014; Keller et al., 2014; Liew et al., 2020; Viegas et al., 2020).

3 Entheos Dataset

In this section we present the Entheos dataset. We describe our domain choice and label selection, the annotation process, extracted features, as well as statistics of the dataset.

3.1 Data Acquisition

Enthusiastic speakers are passionate about their message, wanting to gain their audience for their purpose and persuading them to change their perspective or take action. Given that TED is well-known for spreading powerful messages that can change attitudes and behavior, we use TED talk speeches as our domain for creating a multimodal enthusiasm dataset. We randomly selected 52 male and female speakers from the TEDLIUM corpus release 3 (Hernandez et al., 2018), which contains audio of 2351 talks. Transcripts were obtained

Rating	Description
4: Advanced	Excellent use of vocal variation, intensity and pacing; vocal expression natural and enthusiastic; avoids fillers
3: Proficient	Good vocal variation and pace; vocal expression suited to assignment; few if any fillers
2: Basic	Demonstrates some vocal variation; enunciates clearly and speaks audibly; generally avoids fillers (e.g. um, uh, like)
1: Minimal	Sometimes uses a voice too soft or articulation too indistinct for listeners to comfortably hear; often uses fillers
0: Deficient	Speaks inaudibly; enunciates poorly; speaks in monotone; poor pacing; distracts listeners with fillers

Table 1: Description of the Public Speaking Competence Rubric (PSCR) (Schreiber et al., 2012) evaluated as potential label to describe the use of vocal expressions and paralanguage during a talk.

Vocal Attributes	Description	Rating
Variation	Vocal variety is the spice of speech. Tone, pace, and volume should all be varied over the course of a presentation.	4: excellent, 3: good, 2: some, 1: almost no vocal variation, 0: speaks in monotone
Intensity	Speaks loudly and clearly enough for listeners to hear and understand what is being said.	4: excellent use, 3: good, 2: enunciates clearly and speaks audibly , 1: sometimes voice too soft or articulation too indistinct for listeners to comfortably hear, 0: inaudibly, enunciates poorly
Pacing	Speaks in an understandable rate and places pauses for emphasis.	4: excellent use including well placed pauses, 3: good, 2: pace is appropriate but could have more/less pauses, 1: poor pacing, 0: poor pacing with no/too many pauses
Expression	Emotion delivered by the voice.	4: natural and enthusiastic, 3: suited to assignment, 2: some expressions, 1: few expressions, 0: no expressions)

Table 2: Fine-grained description of vocal attributes derived from PSCR, evaluated as potential label categories on sentence-level.

through the Google cloud transcription service[2]. The talks were segmented into sentences, based on punctuation. We extend the samples from the TEDLIUM corpus with aligned video segments downloaded from the official TED website.

3.2 Label Selection and Temporal Granularity

In order to define the temporal granularity for annotation and what labels to use, we performed preliminary annotation experiments with three annotators.

Three audio recordings of talks were chosen from speakers with different proficiency level. One recording was a TED talk by Al Gore[3], and the remaining were recordings of participants in a pilot study with our institution in which they introduce themselves and describe their skills.

We evaluated two different temporal granularities: sentence-level and entire talk. In addition, we explored the use of three different sets of labels, which will be described in the following.

PSCR (Public Speaking Competence Rubric) PSCR (Schreiber et al., 2012) was developed to effectively assess students' skills in public speaking. It is composed of eleven skills that are assessed during speaking with a 0-4 scale. We focused on the seventh, which evaluates the effective use of vocal expression and paralanguage to engage the audience. During annotation, annotators had Table 1 available for a detailed description on how the speaker articulates for the corresponding rating.

Vocal Attributes Based on the PSCR descriptions we crystallized four main components of the effective use of the voice: vocal variation, intensity, pacing, and expression. Each one was evaluated with a score of 0-4 and described as depicted in Table 2.

Enthusiasm and Emphasis As a final set of labels, we decided to use intuitive categories, namely enthusiasm and emphasis. For enthusiasm, we chose the definition provided by Keller et al. (2016) as they study enthusiasm in context of spoken

[2] https://cloud.google.com/speech-to-text

[3] https://www.ted.com/talks/al_gore_averting_the_climate_crisis

Category	Description	Rating
Enthusiasm	Speaker is passionate, energetic, stimulating, and motivating.	0: monotonous, 1: normal, 2: enthusiastic
Emphasis	One or more words are emphasized by speaking louder or pronouncing them slowly.	0: no emphasis, 1: emphasis existent

Table 3: Intuitive labels used to evaluate as potential categories to annotate sentence-level samples.

Label	Fleiss' κ	Agreement
PSCR	0.31	fair
Variation	0.56	moderate
Intensity	0.81	almost perfect
Pacing	0.55	moderate
Expression	0.63	substantial
Enthusiasm	**0.82**	**almost perfect**
Emphasis	**0.87**	**almost perfect**

Table 4: Inter-rater agreement using different labels computed with Fleiss' kappa with interpretations based on Landis and Koch (1977). Enthusiasm, emphasis, and vocal intensity achieved almost perfect agreement.

monologues (similar to our data) while Inaba et al. (2011) studied written dialogues. We also asked annotators to label enthusiasm in three levels: monotonous, normal, and enthusiastic. As Table 3 shows, annotators were asked to label emphasis as existent or not, depending on whether words were emphasized by speaking louder or pronouncing words slowly.

Experiment Description The experiment was composed of two parts. First the entire audio recordings were played and the annotators were asked to use only the PSCR annotation scheme, rating each talk with a single score. Afterwards, seven sentences of each talk were played with pauses in between to allow annotation using vocal attributes, enthusiasm and emphasis labels. Each sentence was annotated with six scores. For both parts, the annotators had access to the description of the labels during annotation as shown in Tables 1,2,3. Once all annotators finished labeling a sample, the next one was played.

Results and Conclusion In Table 4 the inter-rater agreement for the different annotation schemes is shown in terms of Fleiss' kappa Landis and Koch (1977). We can see that PSCR, which rated the entire talk, has the lowest agreement. Vocal variation and pacing have moderate agreement, while vocal intensity, enthusiasm, and emphasis show almost perfect agreement.

Given these results, we annotated audio recordings on a sentence-level using enthusiasm and emphasis labels.

3.3 Data Annotation Protocol

Our study was approved by our institution's human subject board and annotators were paid $20/h. Seventeen subjects participated in data annotation and signed the consent form before the study. For data annotation, an internal tool was created that enabled annotators to listen to audio samples and annotate them through their web browser at their time of convenience. As labeling availability fluctuated, instead of randomly choosing samples from the entire dataset, we decided to release small batches of data to obtain as many annotations per sample as possible. In a bi-weekly rhythm, small batches of 200 samples were available to annotate in a randomly chosen order for each annotator. As our definition for enthusiasm (Table 3) allows subjective interpretations, we included three reference audio files for each enthusiasm level in the web interface of our annotation tool as depicted in Figure 2. Annotators were indicated to listen to the reference files after every 10 labeled samples and when insecure on how to label a sample. In addition, annotators were given the definition of enthusiasm and emphasis shown at Table 3. Besides enthusiasm and emphasis, also the corresponding perceived gender was annotated. We limited the options for perceived gender to female and male, based on prior work which used these two genders to improve the performance in emotion detection (Li et al., 2019). Samples with laughter or clapping were asked to be labeled as noisy files.

Annotator Quality Assessment: Annotation was performed by 17 different annotators. As noisy annotations are common when crowdsourcing and not using expert annotators due to spammers and malicious workers (Burmania et al., 2015), we com-

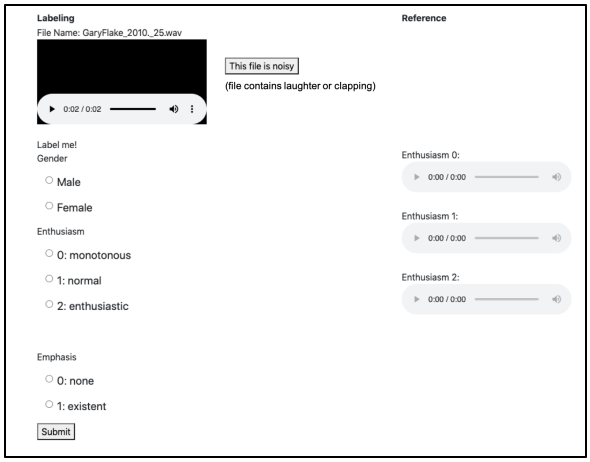

Figure 2: Layout of the annotation interface. On the top left is the sample to be annotated and below are the different labels: perceived gender, enthusiasm, and emphasis. On the top center is the option to mark the sample as noisy if laughter or clapping is present. On the right side are reference samples for the three different levels of enthusiasm.

pared the percentage agreement of each individual's annotations with a preliminary majority vote. The analysis showed that 12 annotators had lower agreement than 30%. The same annotators had also labeled less than 17% of the data. To ensure high quality of annotation we used the remaining five annotators who labeled more than 50% of the data. The remaining annotators identify themselves as latino, asian, and white. We removed all samples that had only one or two different annotations and computed the final majority vote for the remaining 1,126 samples. To confirm high inter-rater agreement, we computed Cohen's kappa (McHugh, 2012) in a pairwise manner for the five annotators and obtained an average agreement of 0.66.

3.4 Final Data Selection

Out of 1,819 labeled samples, we kept 1,126 which had more than one annotation. The selected samples are from 113 different TED talk speeches, being 60 from male and 53 from female speakers. We created a test split with 108 samples from five speakers of each perceived gender. The training set, composed by 55 male and 48 female speakers, has a total of 1,018 samples. There is no overlap of speakers between training and test set. In Figure 3 (top) we can see the label distribution in our train-test split.

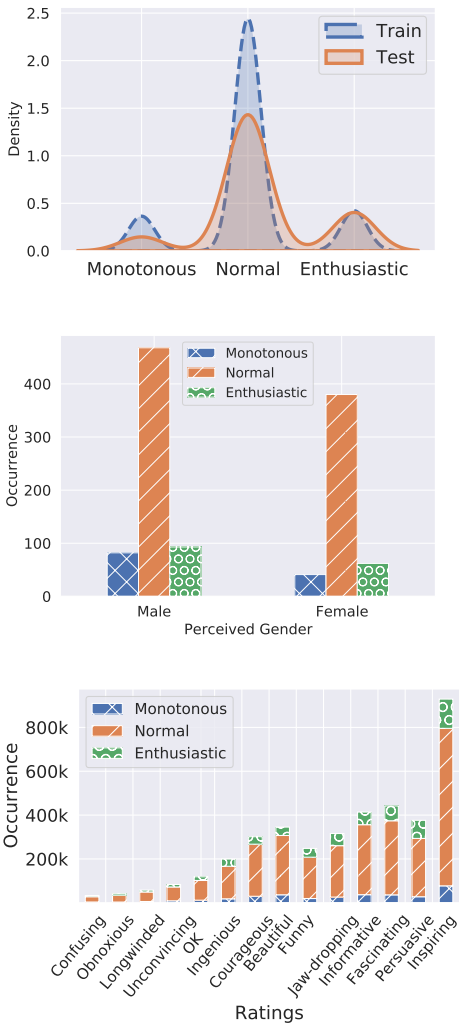

Figure 3: From top to bottom: Label distribution in our train-test split, among perceived gender, and ratings given by TED viewers. Top: Training set and testing set reflect the same imbalance of class labels. Center: Female speakers have proportionally fewer monotonous samples and more normal samples than male, but the same proportion of enthusiastic samples. Bottom: Samples labeled as enthusiastic have been mainly rated as fascinating, persuasive, and inspiring. They have rarely been rated negatively.

3.5 Data Statistics

In the following we will describe the relationship between the different enthusiasm levels and other attributes of the talks such as viewer ratings, number of views and comments, and perceived gender of the speakers. This metadata was obtained from a Kaggle competition[4] that collected data about TED talks until September 21st, 2017.

In Figure 3 (center), we can see that the enthusi-

[4] https://www.kaggle.com/rounakbanik/ted-talks

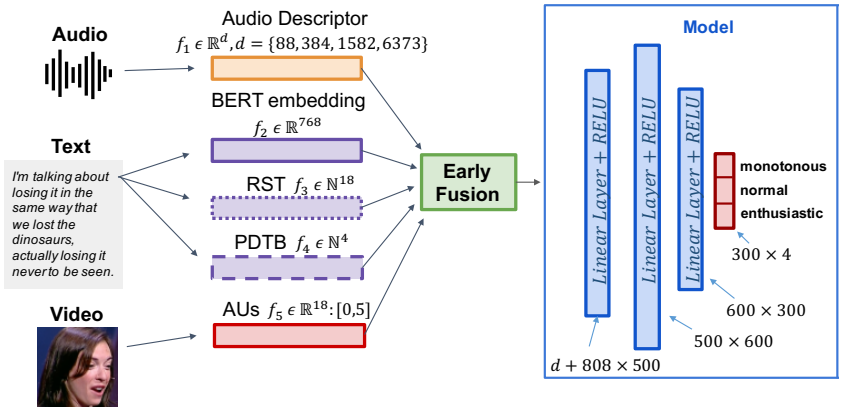

Figure 4: An overview of our proposed multimodal dataset and model for predicting levels of enthusiasm using different features extracted from video, audio, and text.

asm levels are similarly distributed for both gender labels. We computed the Pearson's chi-squared test for independence to evaluate if there is a significant difference in enthusiasm level between gender. With a significance level of 5%, we obtained $p = 0.04$, meaning that gender of the speaker and enthusiasm level are dependent of each other. In Figure 3 (bottom), the label distribution among the different ratings that were given by viewers is shown. There are nine positive ratings (funny, beautiful, ingenious, courageous, informative, fascinating, inspiring, persuasive, jaw-dropping) and five negative ratings (longwinded, confusing, unconvincing, ok, obnoxious) which viewers could select. The ratings have been sorted by increasing number of enthusiastic samples. We can see that the negative ratings have the least number of enthusiastic samples. The ratings with the three highest numbers of enthusiastic samples are fascinating, persuasive and inspiring. We also performed two one-way ANOVAs to evaluate if the number of views and comments depend on the enthusiasm level. The resulting p-values were correspondingly $p = 0.3844$ and $p = 0.6892$ which means that views and comments are not influenced by the enthusiasm level of the speaker.

4 Computational Experiments

In the experiments of this paper, we aim to establish a performance baseline for the Entheos dataset using only the enthusiasm annotations. We train our model with different feature combinations to understand the role of different modalities in enthusiasm detection (see Figure 4). In the following we describe different features that were extracted and the model architecture that we used.

4.1 Features

Given the small number of labeled samples, instead of training an end-to-end model, we extract different features that will serve as input for our model. In the following we will describe the features used per modality.

Video: As enthusiasm is related to emotions, we extracted Facial Action Units (FAUs) which describe the intensity of muscular movements in the face based on the Facial Action Coding System (FACS) (Friesen and Ekman, 1978). We used OpenFace (Baltrusaitis et al., 2018) to obtain the intensity of 18 FAUs in a scale of 0-5. As FAUs vary over time, we computed the average and standard deviation for each AU and concatenated them in a feature of 36 dimensions per sample.

Acoustic: We extracted different audio features using OpenSMILE (Eyben et al., 2010), a toolbox that can extract over 27k features. We extracted four different feature combinations, which have been thoroughly studied in the speech community in affective computing tasks: a) eGEMAPS (88 attributes) (Eyben et al., 2015), b) Interspeech 2009 Emotion Challenge (384 attributes) (Schuller et al., 2009), c) Interspeech 2010 Paralinguistic Challenge (1582 attributes) (Schuller et al., 2010), and d) Interspeech 2013 Compare (6373 attributes) (Schuller et al., 2013). Each feature collection differs in the selection of features, functionals, and statistical measures. Examples of features covered are voice quality (jitter and shimmer), pitch (F0), energy, spectral, cepstral (MFCC) and voicing related low-level features (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and

psychoacoustic spectral sharpness.

Text: As a low-level feature, we used the bert-large-uncased model[5] to obtain word-embeddings on a sentence-level. For each sample we obtained a feature of 768 dimensions. As high-level features, we extracted two types of discourse relations: Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008) relations. We used the RST parser from Wang et al. (2017) and the PDTB parser from Lin et al. (2014) for automated discourse relation annotation. Elementary discourse units (EDU) were obtained by using the method presented by Wang et al. (2018). For both parsers, samples can have more than one relation or none at all. The annotations were converted into a bag-of-words representation, obtaining features of 18 dimensions for RST and 4 for PDTB.

4.2 Model Architecture

Our model is composed by four fully connected layers with ReLU activation functions in between. We use concatenation to combine different features in the multimodal setting. Given our imbalanced dataset, we compute class weights, which represent the relation of samples per label and the total sample number. The class weights are then passed to our loss function (cross entropy loss) to give more weight to samples of the underrepresented classes. We use the Adam optimizer (Kingma and Ba, 2015) and during training, we perform early stopping to avoid overfitting. We train the model for a three class problem using all enthusiasm levels and also in a binary manner, combining "monotonous" and "normal" labels to the category called "non-enthusiastic".

5 Results and Evaluation

In this section, we present the performance results of our model using different combinations of features. We also evaluate the performance of the discourse parsers used and show statistical analysis of visual and acoustic features. All results of our statistical analysis are shown in the Appendix A.

5.1 Predicting Enthusiasm level

For each feature combination, we performed hyperparameter search with 10-fold cross-validation. The best hyperparameter combination was used to train the model with the entire training set. We evaluated the performance of the models on our test set. In Table 5, the weighted average results for precision, recall, and F1-score are shown. We see that in the unimodal case, BERT embeddings perform the best in the binary classification as well as in the three-class problem. Although PDTB has a higher F1-score in the binary case, RST performs better in the multi-class problem. Out of the different audio features, eGEMAPS performs slightly better than the other acoustic features. In the multi-class case, IS09 features are the best performing acoustic features.

When all features except AUs are combined, we reach the highest F1-score for the binary problem, improving the best unimodal performance by 0.08. We also see that combining both discourse relation features with eGEMAPS and BERT improves F1-score by 0.08 compared to using only one of them. In the multi-class problem, the best performing feature combination shows only a slight improvement of 0.04 compared to the unimodal case. Although manually annotating the entire resource was beyond the scope of this paper, we believe that it is necessary to understand the weaknesses and strengths of automatic parsers when used in spoken monologues. With current efforts being made in the field of creating discourse parsers for speech, the role of discourse parsers for enthusiasm detection will be better understood.

5.2 Evaluating the Effect of Discourse Features

We see in Table 5 that discourse relations help the model achieve the highest F1-score. However, we obtained the discourse relations by using discourse parsers that are trained on Wall Street Journal data[6], which is different from monologues.

To evaluate the performance of the parsers, 40 samples of our data were manually annotated with RST and PDTB relations by two annotators. The annotation protocol was approved by our institution's human subject research center. The inter-rater agreement was $\kappa = 0.88$. The accuracy of the RST parser on our data sample was 46.7 and for the PDTB parser 60.0. Although the accuracy of the parsers is low using our data, we have seen that concatenating both discourse relation features

[5] https://huggingface.co/bert-large-uncased

[6] https://catalog.ldc.upenn.edu/LDC93S6A

Features	Precision [B/M]	Recall [B/M]	F1-Score [B/M]
RST	0.67/0.55	0.64/0.47	0.65/0.50
PDTB	0.70/0.68	0.70/0.29	0.70/0.32
BERT	**0.77/0.66**	**0.81/0.56**	**0.75/0.60**
EGEMAPS	0.80/0.59	0.71/0.47	0.74/0.50
IS09	0.70/0.60	0.76/0.57	0.72/0.55
IS10	0.68/0.56	0.70/0.44	0.69/0.48
IS13	0.65/0.68	0.69/0.37	0.67/0.43
AU	0.67/0.77	0.76/0.50	0.70/0.57
BERT + PDTB	0.77/0.66	0.80/0.57	0.77/0.61
BERT + RST	0.79/0.66	0.81/0.56	0.77/0.60
EGEMAPS + BERT	0.81/0.62	0.60/0.58	0.64/0.59
EGEMAPS + PDTB	0.75/0.69	0.75/0.52	0.75/0.54
EGEMAPS + RST	0.77/0.71	0.7/0.61	0.73/0.64
EGEMAPS + BERT + PDTB	0.74/0.72	0.77/0.65	0.75/0.67
EGEMAPS + BERT + RST	0.77/0.71	0.81/0.58	0.75/0.61
EGEMAPS + RST + PDTB + BERT	**0.83/0.63**	**0.84/0.65**	**0.83/0.64**
EGEMAPS + RST + PDTB + BERT + AU	0.81/0.65	0.65/0.58	0.68/0.60

Table 5: Weighted average precision, recall, and F1-score for binary (**B**) and multiclass (**M**) classification. The same model architecture was used to train different feature combinations. BERT embeddings performed best in the unimodal setting. Combining acoustic with text features performed best in the multimodal setting.

to BERT and eGEMAPS improved our model's performance from an F1-score of 0.64 to 0.83 in the binary classification.

In Figure 5(a,b) we evaluated the relative occurrence of each enthusiasm level for RST and PDTB relations in ascending order of enthusiastic samples. In Figure 5a we can see that most samples do not have any discourse relation. However, there is a clear difference in the number of monotonous and enthusiastic samples that show *contingency*, as well as *temporal* relations. In Figure 5b we see that enthusiastic samples compared to monotonous samples use more elaboration, attribution, and joint relations. We performed the Pearson Chi Square test to verify our null hypotheses that discourse relations and enthusiasm level are independent from each other. We obtained a p-value of 0.0001 for PDTB and a p-value of 0.008 for RST, which permits us to reject our null hypothesis, meaning that the discourse relations influence the level of enthusiasm.

5.3 Investigating Visual Features

Given that AUs have not helped our model improve, we evaluated their dependence with our labels. We performed two separate one-way ANOVAs to evaluate the dependence of the mean of the 18 AUs with our labels, as well as the standard deviation of the AUs with our labels. The AUs with p-value < 0.05 are AU 12 (lip corner puller), AU 15 (lip corner depressor), AU 17 (chin raiser), and AU 26 (jaw drop). In Figure 5(c,d) the label distribution for the mean of AU 26 and standard deviation of AU 12 is shown. In both cases, we can observe that monotonous samples have more frequently a mean and standard deviation of zero compare to enthusiastic samples. We can also see in Figure 5d that enthusiastic samples have more frequently a standard deviation of AU 12 > 0.02.

5.4 Investigating Acoustic Features

We have seen that acoustic features are important in improving our model's performance. In this section we want to evaluate if pitch (F0) and loudness are independent from enthusiasm level. We perform a one-way ANOVA for the mean F0 per sample and its enthusiasm level, as well as for the mean loudness. Both p-values are < 0.05, meaning that the enthusiasm labels depend on the acoustic features. In Figure 5e, we can see that monotonous samples have a lower mean F0 than that of enthusiastic samples. We can also see in Figure 5f that monotonous samples have lower mean loudness than that of enthusiasm. These observations agree with the intuition that enthusiastic speakers speak louder and increase their pitch.

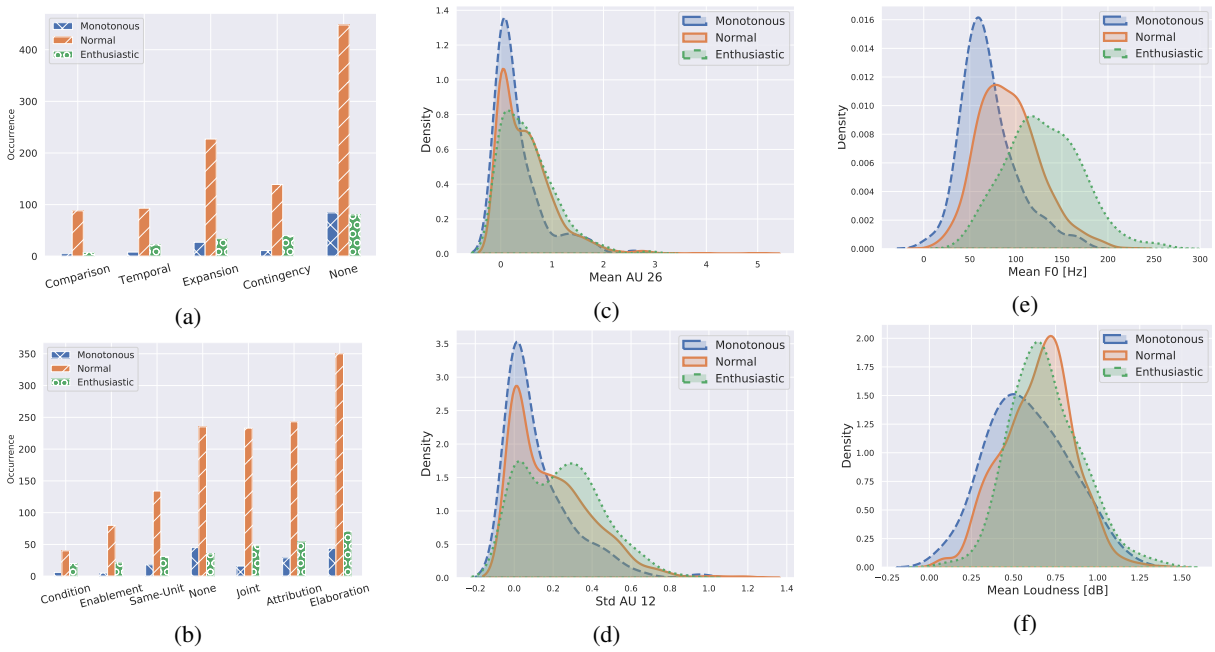

Figure 5: Label distribution of different enthusiasm levels in relation to discourse relations (a,b), acoustic features (c,d), and facial action units (e,f). In (a), most samples had no PDTB relation, however there is a visible difference between monotonous and enthusiastic samples in the occurrence of temporal and contingency relations. In (b), RST relations show that enthusiastic samples compared to monotonous samples use more elaboration, attribution, and joint relations. In (c), we can see that monotonous samples have more often low intensities for AU26 (jaw drop) than enthusiastic samples. (d) shows that monotonous samples have mostly very low standard deviation for AU12 (lip corner puller), but enthusiastic samples have higher standard deviation. In (e), we can see that enthusiastic samples have a higher mean F0 (pitch) compared to monotonous samples. (f) shows that monotonous speech tends to have lower mean loudness compared to enthusiastic speech.

6 Discussion and Conclusion

We present the first multimodal dataset for enthusiasm detection called Entheos[7] and discuss several baseline models. In addition, we present qualitative and quantitative analyses for studying and predicting enthusiasm using the three modalities of text, acoustic, and visual.

Our work has several limitations. TED talks are a very specific form of monologues as they are well-rehearsed and prepared. However, it is more likely that we can find enthusiastic speakers or well-structured sentences in TED talks. To understand enthusiastic behaviors in daily conversations, more data from other domains need to be annotated and studied. We hope that our annotation protocol will help other researchers in the future.

Further theoretical and empirical research is needed for better studying enthusiastic behaviors in general. The signals and definitions that we have worked with are not fine-grained or well-connected when exploring different modalities. Facial expressions and gestures can potentially provide meaningful contributions. Our experiments with facial action units were not successful. Our baseline approach used statistical information of each AU instead of the raw signal, which may dilute useful information. More experiments are needed to evaluate if and how AUs can help predict enthusiasm.

We hope our resources provide opportunities for multidisciplinary research in this area. Given the difficulties of annotating multimodal datasets in this domain, future work needs to investigate weakly supervised approaches for labeling multimodal data.

[7] https://github.com/clviegas/Entheos-Dataset

Acknowledgments

We thank TalkMeUp Inc. for supporting this research.

References

John Antonakis, Giovanna d'Adda, Roberto Weber, and Christian Zehnder. 2019. Just words? just

speeches? on the economic value of charismatic leadership. *NBER Rep. 4*.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Edward M Bettencourt, Maxwell H Gillett, Meredith Damien Gall, and Ray E Hull. 1983. Effects of teacher enthusiasm training on student on-task behavior and achievement. *American educational research journal*, 20(3):435–450.

Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2015. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388.

Jeroen PJ De Jong and Deanne N Den Hartog. 2007. How leaders influence employees' innovative behaviour. *European Journal of innovation management*.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5.

Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. 2017. Emotion detection: a technology review. In *Proceedings of the XVIII international conference on human computer interaction*, pages 1–8.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer.

Michimasa Inaba, Fujio Toriumi, and Kenichiro Ishii. 2011. Automatic detection of "enthusiasm" in non-task-oriented dialogues using word co-occurrence. In *2011 IEEE Workshop on Affective Computational Intelligence (WACI)*, pages 1–7. IEEE.

Caroline F Keating. 2011. Channelling charisma through face and body status cues. *Social psychological dynamics*, pages 93–111.

Melanie M Keller, Thomas Goetz, Eva S Becker, Vinzenz Morger, and Lauren Hensley. 2014. Feeling and showing: A new conceptualization of dispositional teacher enthusiasm and its relation to students' interest. *Learning and Instruction*, 33:29–38.

Melanie M Keller, Anita Woolfolk Hoy, Thomas Goetz, and Anne C Frenzel. 2016. Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review*, 28(4):743–769.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. 2019. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*, pages 2803–2807.

Tze Wei Liew, Su-Mae Tan, Teck Ming Tan, and Si Na Kew. 2020. Does speaker's voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning? *Information and Learning Sciences*.

Tze Wei Liew, Nor Azan Mat Zin, and Noraidah Sahari. 2017. Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment. *Human-centric Computing and Information Sciences*, 7(1):9.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Catherine Marechal, Dariusz Mikołajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Węgrzyn-Wolska. 2019. *Survey on AI-Based Multimodal Methods for Emotion Detection*, pages 307–324. Springer International Publishing, Cham.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. 2004. The Penn Discourse Treebank. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Shubhanshu Mishra and Jana Diesner. 2019. Capturing signals of enthusiasm and support towards social issues from twitter. In *Proceedings of the 5th International Workshop on Social Media World Sensors*, pages 19–24.

Oliver Niebuhr. 2020. Space fighters on stage-how the f1 and f2 vowel-space dimensions contribute to perceived speaker charisma. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pages 265–277.

Oliver Niebuhr, Jana Voße, and Alexander Brem. 2016. What makes a charismatic speaker? a computer-based acoustic-prosodic analysis of steve jobs tone of voice. *Computers in Human Behavior*, 64:366–382.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Elie Saad, Joost Broekens, Mark A Neerincx, and Koen V Hindriks. 2019. Enthusiastic robots make better contact. In *IROS*, pages 1094–1100.

Birgitta Sandberg. 2007. Enthusiasm in the development of radical innovations. *Creativity and Innovation Management*, 16(3):265–273.

Lisa M Schreiber, Gregory D Paul, and Lisa R Shibley. 2012. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Martin E Spencer. 1973. What is charisma? *The British Journal of Sociology*, 24(3):341–354.

Rode Snehal Sudhakar and Manjare Chandraprabha Anil. 2015. Analysis of speech features for emotion detection: a review. In *2015 International Conference on Computing Communication Control and Automation*, pages 661–664. IEEE.

Ryoko Tokuhisa and Ryuta Terashima. 2006. Relationship between utterances and "enthusiasm" in non-task-oriented conversational dialogue. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 161–167.

Carla Viegas, Albert Lu, Annabel Su, Carter Strear, Yi Xu, Albert Topdjian, Daniel Limon, and JJ Xu. 2020. Spark creativity by speaking enthusiastically: Communication training using an e-coach. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 764–765.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Qin Zhang. 2014. Assessing the effects of instructor enthusiasm on classroom engagement, learning goal orientation, and academic self-efficacy. *Communication Teacher*, 28(1):44–56.

A Statistical Tests

In this section we present the results of the statistical tests performed to the facial action units and prosody features extracted from the entire dataset.

A.1 AU Statistical Tests

In order to understand which AU influence the enthusiasm level, we performed two different statistical tests: ANOVA for the three levels of enthusiasm (monotonous, normal, enthusiastic), and T-test for two levels of enthusiasm (enthusiastic, non-enthusiastic). In Table 6 on the left we can see the results of the ANOVA, analyzing the mean value of the different AUs per sample with the three levels of enthusiasm. All mean AUs that show p-value < 0.05 are highlighted. As AU 26 has the lowest p-value, the label distribution is shown in Figure 5c.

In Table 6 on the right we can see the results of the ANOVA, analyzing the standard deviation of the different AUs per sample with the three levels of enthusiasm. As AU 12 has the lowest p-value, the label distribution is shown in Figure 5d.

We also performed T-tests for the binary case using the labels enthusiastic and non-enthusiastic. Table 7 on the left shows that AU 17 (chin raiser)

Mean Action Unit	F-Statistic	P-value
AU 01	0.1393	0.87
AU 02	0.4541	0.6351
AU 04	1.415	0.2434
AU 05	0.385	0.6805
AU 06	1.1288	0.3238
AU 07	0.3578	0.6993
AU 09	2.4968	0.0828
AU 10	2.4397	0.0877
AU 12	**4.7553**	**0.0088**
AU 14	1.1253	0.3249
AU 15	**5.1991**	**0.0057**
AU 17	**4.672**	**0.0095**
AU 20	0.8012	0.449
AU 23	1.1192	0.3269
AU 25	0.9896	0.3721
AU 26	**6.0058**	**0.0025**
AU 45	1.0887	0.337

Std Action Unit	F-statistic	P-value
AU 01	**5.614114**	**0.003749**
AU 02	1.052148	0.349531
AU 04	0.905609	0.404591
AU 05	1.778094	0.169435
AU 06	**5.960989**	**0.002660**
AU 07	1.948772	0.142930
AU 09	**10.337395**	**0.000036**
AU 10	**8.383927**	**0.000243**
AU 12	**12.263390**	**0.000005**
AU 14	**5.483568**	**0.004266**
AU 15	**4.347689**	**0.013155**
AU 17	**12.201065**	**0.000006**
AU 20	**3.262334**	**0.038662**
AU 23	**8.411563**	**0.000237**
AU 25	**6.328837**	**0.001848**
AU 26	**11.989375**	**0.000007**
AU 45	**4.585977**	**0.010385**

Table 6: ANOVA significance test for three levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with lowest p-value are highlighted.

is the only AU with a p-value < 0.05. The distribution of the average values of AU 17 are shown in Figure 6(a). For comparison, the distribution of the average AU 02 (outer brow raiser) with highest p-value is shown in Figure 6(b). For both analysis, ANOVA and T-test, the differences of standard deviations among the enthusiasm levels are statistically significant for almost all AUs. This is not the case when analyzing the average values of AUs.

A.2 Prosody Statistical Tests

We performed statistical significance tests using the mean and standard deviation for F0 (pitch) and loudness. In Table 8(left), the ANOVA analysis results are shown and in Table 8(right), the results of the T-test. In both significance tests all variables have a p-value$<$ 0.05, which means that all of them influence the enthusiasm level. Figure 6(c-f) show the label distribution for different values of the variables used in the significance test.

Mean Action Unit	F-Statistic	P-value
AU 01	-0.3995	0.6896
AU 02	0.0357	0.9715
AU 04	1.5205	0.1287
AU 05	0.4318	0.666
AU 06	-0.0535	0.9573
AU 07	0.7846	0.4328
AU 09	-1.8848	0.0597
AU 10	-0.9503	0.3422
AU 12	-1.1706	0.242
AU 14	0.5841	0.5592
AU 15	-0.9274	0.3539
AU 17	**-2.9922**	**0.0028**
AU 20	-1.2633	0.2067
AU 23	-1.4888	0.1368
AU 25	-0.586	0.558
AU 26	-1.2643	0.2064
AU 45	-0.3449	0.7303

Std Action Unit	F-statistic	P-value
AU 01	**-2.290794**	**0.022160**
AU 02	-1.451265	0.146985
AU 04	-0.909680	0.363186
AU 05	-1.067368	0.286035
AU 06	**-2.203459**	**0.027765**
AU 07	-1.574206	0.115721
AU 09	**-4.374609**	**0.000013**
AU 10	**-3.239400**	**0.001233**
AU 12	**-3.181258**	**0.001507**
AU 14	-1.460543	0.144420
AU 15	**-2.571532**	**0.010253**
AU 17	**-4.600027**	**0.000005**
AU 20	**-2.514758**	**0.012050**
AU 23	**-2.810554**	**0.005031**
AU 25	**-1.972713**	**0.048773**
AU 26	**-2.491479**	**0.012865**
AU 45	-1.378754	0.168245

Table 7: T-test for two levels of enthusiasm and AU mean values on the left and standard deviation of AU on the right. AUs with lowest p-value are highlighted.

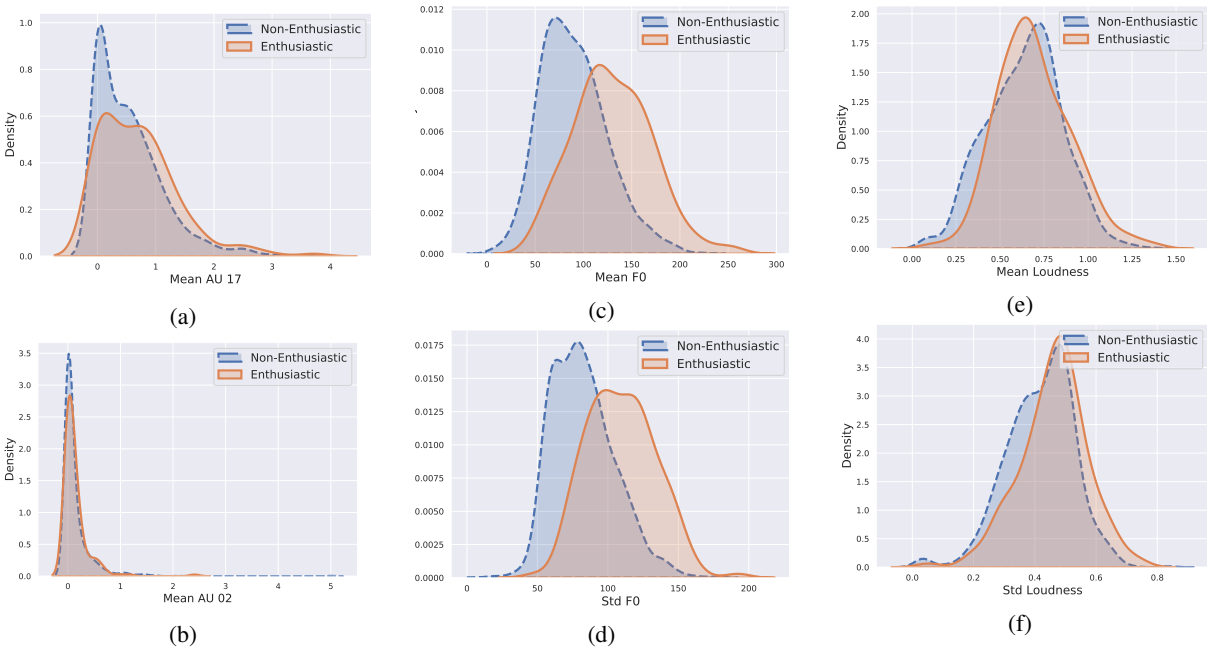

Figure 6: Label distribution of enthusiastic and non-enthusiastic samples in relation to the (a) mean AU 17 (p-value $= 0.0028$), mean AU 17 (p-value $= 0.9715$), (c) mean F0 (p-value $= 0.0$), (d) std F0 (p-value $= 0.0$), (e) mean loudness (p-value $= 0.0034$), (f) std loudness (p-value $= 0.00$).

	F-statistic	P-value
Mean F0	113.4309	0.0000
Mean Loudness	8.2467	0.0003
Std F0	146.9639	0.0000
Std Loudness	16.9411	0.0000

	F-statistic	P-value
Mean F0	-13.1960	0.0000
Mean Loudness	-2.9355	0.0034
Std F0	-13.9376	0.0000
Std Loudness	-4.508	0.0000

Table 8: Significance test for mean and standard deviation of F0 and loudness to evaluate the dependence with the different enthusiasm levels. Left: ANOVA significance test results three enthusiasm levels shows that all p-value< 0.05, which means that all variables influence the enthusiasm level. Right: T-test significance test for two levels of enthusiasm also shows that all variables influence the enthusiasm level.