Automatic Assessment of Speaking Skills Using Aural and Textual Information

Sofia Eleftheriou
Institute of Informatics & Telecommunications,
NCSR "Demokritos"
Athens, Greece

Panagiotis Koromilas
Institute of Informatics & Telecommunications,
NCSR "Demokritos"
Athens, Greece

Theodoros Giannakopoulos
Institute of Informatics & Telecommunications,
NCSR "Demokritos"
Athens, Greece

sofiaeleftheriou13@gmail.com
{pakoromilas, tyianak}@iit.demokritos.gr

Abstract

In this work we propose a multimodal speech analytics framework for automatically assessing the quality of a public speaker's capabilities. For this purpose, we present the Public Speaking Quality (PuSQ) dataset, a new publicly available data collection that contains speeches from various speakers, along with respective annotations of how are these speeches perceived by the audience in terms of two labels namely: "expressiveness" and overall "enjoyment" (i.e. if the listener enjoys the speech as a whole). Towards this end, several annotators have been asked to provide their input for each speech recording and inter-annotator agreement is taken into account in the final ground truth generation. In addition, we present a multimodal classifier that takes into account both audio and text information and predicts the overall recordings' label with regards to its speech quality (in terms of the two aforementioned labels). To this end, we adopt a hierarchical approach according to which we first analyze the speech signal in a segment-basis (50ms of audio and sentences of text) to extract emotions from both text and audio and then aggregate these decisions for the whole recording, while adding some high-level speaking style characteristics to produce the overall representation that is used by the final classifier.

1 Introduction

Public speaking (also called oratory or oration) is the act of giving speech face to face to live audience. However, due to the evolution of public speaking, it is lately viewed as any form of speaking (formally and informally) between an audience and the speaker. Traditionally, public speaking was considered to be a part of the art of persuasion. The act can accomplish particular purposes including information, persuasion, and entertainment. Additionally, differing methods, structures, and rules can be utilized according to the speaking situation.

Currently, technology continues to transform the art of public speaking through newly available techniques such as videoconferencing, multimedia presentations, and other nontraditional forms. Knowing when speech is most effective and how it is done properly are key to understanding the importance of it.

While most current methods for evaluating speech performance attend to both verbal and nonverbal aspects, almost all existing assessments in practice require human rating (Ward, 2013; Schreiber et al., 2012b; Carlson and Smith-Howell, 1995). Due to the obvious need to use our speech in our daily lives, its evaluation and its improvement is also very important. This evaluation becomes easier and faster if it is performed by an automated process that mostly uses machine learning methodologies.

Speech is everywhere and the way we speak is just as important as what we say. Therefore, multimodal speech analytics (using text and audio) is an important process that can be applied not only to assess public speakers speech quality, but also in other speech-related fields of application such as the identification of learning disabilities related to speech (dyslexia, autism), the analytics of call center data and the speech-based assessment of psychological and psychiatric conditions. The proposed pipeline for assessing the public speech quality can be adopted in such applications, as soon as respective ground truth have been made available for training the supervised models.

The related works in assessing public speakers' skills is very limited and usually focuses in two specific tasks, namely learning analytics and persuasive analysis. In particular, (Chen et al., 2016) focuses on the design of a multimodal automated assessment framework for public speaking skills an-

alytics, which is based on the public speaking competence rubric (PSCR)(Schreiber et al., 2012a) for scoring and uses both audiovisual and textual features. With regards to the persuasiveness prediction application domain, a widely used dataset, named Persuasive Opinion Multimedia (POM) (Park et al., 2014) has been created, which contain multiple communication modalities (audio, text and visual). A deep learning approach for this task that is evaluated on POM is presented by (Nojavanasghari et al., 2016), where the authors design a deep multimodal fusion architecture, that has the ability to combine signals from the visual, acoustic, and text modalities effectively.

However, the aforementioned methodologies do not address the task of assessing the public speakers' skills in a generalized manner. This paper proposes an ML framework for classifying long speech recordings in terms of: (a) overall speech "expressiveness", as perceived by the audience and (b) perceived "enjoyment", i.e., how much the listeners enjoyed each speech recording. Towards this end, we demonstrate a Python open-source library that utilizes segment-level (size of 20ms to 50ms) audio and text classifiers related to emotional and speaking style attributes. The final recording-level decision is extracted by a long-term classifier that is based on feature aggregates of the segment-level decisions. Apart from the open-source library, we present an openly available dataset of real-world recordings, annotated in terms of perceived expressiveness and enjoyment. Extensive experimental results prove that the proposed ML framework can discriminate between positive and negative speech samples, despite the simplicity of the baseline segment-level classifiers.

The paper is organized as follows: Section 2 shows the conceptual diagram of the proposed methodology, Section 3 presents the segment-level audio and text classifiers related to emotional attributes, Section 4 introduces the aggregation of class posteriors among with some high-level features that are calculated across the entire recording, Section 5 refers to the newly constructed Public Speaking Quality (PuSQ) dataset, Section 6 is responsible for the reporting of the implemented experiments and Section 7 sets out the final conclusions.

2 System overview

The system architecture developed in the context of speech quality assessment is divided into two parts: segment-level analysis and recording-level analysis. In the first, we break the information (audio or text) into temporal segments and use segment-classifiers related to emotional content. In the second, we aggregate the previously produced class posteriors and combine with high-level features that characterize the overall speaking style. This rationale is followed for both textual and audio modalities and the final decisions can either be used independently or combined in the final recording-level classifiers. The conceptual diagram of the proposed system architecture is shown in Figure 1.

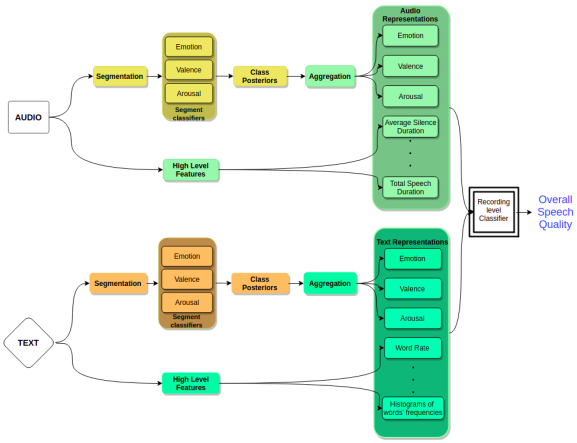

Figure 1: System Architecture

3 Segment-Level Analysis

3.1 Audio Analysis

The audio recording is split into segments and for each segment audio feature extraction is performed and segment classifiers are applied to produce a series of emotion-specific classification decisions, which are then aggregated for the whole recording's classification in terms of overall speech quality. The goal of this section is to describe this segment-level process.

3.1.1 Segmentation and Feature Extraction

For each 3s segment, a short-term window process is followed, i.e. the segment is further split into short-term windows (frames) of 50 ms long with a step of 50 ms (no overlapping). For each short-term window, a series of hand-crafted audio features is extracted, that have been widely used in speech classification tasks. These *low-level audio*

features are: Zero-crossing rate, Energy, Energy entropy, Spectral centroid, Spectral spread, Spectral Entropy, Spectral Flux, Spectral Rolloff, the first 13 MFCCs , the Chroma Vector (12-dimensional) and Chroma Standard Deviation. All these features summing up to 34 in total. We further add the deltas of these features, i.e. the difference between each feature in the current short-term window and the value it had in the immediately preceding short-term window. So we end up with 34 such derivatives (deltas), so 68 features in total for each frame.

Then, for each segment, we extract two feature statistics for each sequence of short-term features described above. The statistics are: the average μ and the standard deviation σ^2 of the respective short-term feature sequences, among the whole 3s segment. Therefore, each segment is now represented by 134 (68×2) feature statistics. To extract this representation, we used the Pyaudioanalysis (Giannakopoulos, 2015) open source Python library.

3.1.2 Speech Segment Classifiers

As described above, each speech segment is represented by 134 audio feature statistics. Then, we have selected to train segment-level classifiers related to the underlying *emotions*, because emotions are strongly associated to the overall speaking style of a public speaker and therefore to the respective speech quality. Towards this end, we have adopted both categorical and dimensional Speech Emotion Recognition annotations (attributes).

The categorical attributes consist of some basic classes of emotions. According to Ortony and Turner (1990), basic emotions are often the primitive building blocks of other non-essential emotions, which are considered variations, or mixtures of basic emotions. In Ekman (1992) six basic emotions are suggested, based on the analysis of facial expressions (anger, disgust, fear, joy, sadness and surprise). We choose to use the 4 basic emotions: anger, sadness, neutral and happiness, as provided by the, widely used in the literature (Koromilas and Giannakopoulos, 2021), processed version of the IEMOCAP dataset.

The main disadvantage of the categorical model is that it has a lower resolution than the, associated with continuous values, dimensional model because it uses categories. The true number of individual emotions and their tones encountered in different types of communication are much richer than the limited number of emotion categories in the model. The smaller the number of classes in the categorical model,the greater the simplification of the description of emotions (Grekow, 2018). That is why dimensional attributes are also widely used in emotion recognition. These representations allocate emotions in dimensional spaces that can mainly capture the similarities and differences between them. Wundt and Judd (1897), proposed the first dimensional model by disassembling the space of emotions along three axes, namely: *valence* (positive-negative), *arousal* (calm-excitement) and intensity (intensity-relaxation). In this work, a usual scheme in the literature (Le et al., 2017) has been adopted with the use of a discretized version of the first two axes (valence and arousal) with the following classes: negative, neutral, positive for valence and high, neutral, low for arousal.

For the training and the evaluation of the three above-mentioned models (emotions, valence, arousal), we use 5 open-source speech emotion datasets, as well as a proprietary dataset that had been created by the authors. The open source datasets are: Emovo (Costantini et al., 2014), EmoDB (Burkhardt et al., 2005), Savee (Jackson and ul haq, 2011), Ravdess (Livingstone and Russo, 2018) and IEMOCAP (Busso et al., 2008). The 6th dataset is named Emotion Speech Movies and contains audio files from movie scenes that are divided into 5 emotional classes.

Some of the aforementioned datasets contain more classes of emotions, therefore we only used the samples corresponding to the classes of interest. Also "excitement" was merged with happiness, since they are quite related expressions. For the valence and arousal tasks, one can observe that the only dataset which contains corresponding labels is IEMOCAP. These labels are continuous values and as we address classification problems, we divide these value ranges into three identical intervals. For valence: the samples with value in the range [1,2.5) are considered to be "negative", the samples with value in the range [2.5,4) are considered to be "neutral" and the samples with values in the range [4,5.5] are considered to be "positive". For arousal: the samples with value in the range [1,2.53) are considered to be "low", the samples with value in the range [2.3,3.6) are considered to be "neutral" and the samples with values in the range [3.6,5] are considered to be "high". For all other data sets that do not contain valence and arousal tags, we distribute the emotion tags in the above 6 valence

Classification Task	Dataset						Average
	IEMOCAP	Savee	Emovo	Emo-db	Ravdess	EmotionSpeechMovies	
Emotion	79.7	71.2	75.5	80.6	67	50.4	**70.7**
Valence	52.9	62.3	59.5	76	63.9	53.2	**61.3**
Arousal	60.3	75.3	79.9	81.9	68.3	64.1	**70**

Table 1: Inner-dataset Evaluation of Audio Models

Classification Task	Test Dataset						Average
	IEMOCAP	Savee	Emovo	Emo-db	Ravdess	EmotionSpeechMovies	
Emotion	39	36	45.5	57.6	29.8	36.6	**40.8**
Valence	39.7	37.9	32.7	37	26.3	42	**35.9**
Arousal	40.1	41.8	40.3	51.1	38.4	38.6	**41.7**

Table 2: Cross-dataset Evaluation of Audio Models

and arousal classes based on the circumplex model shown in Figure 2.

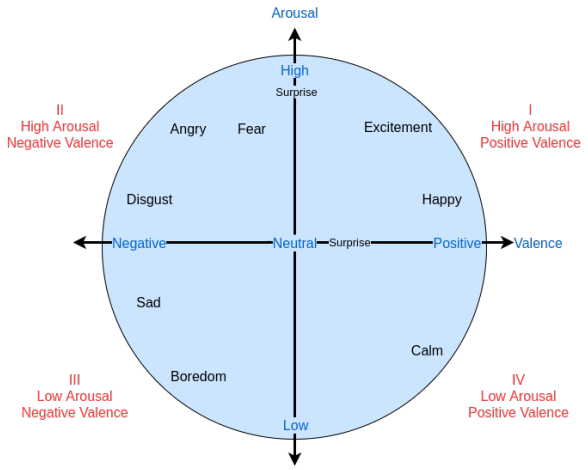

Figure 2: Distribution of Emotions in Circumplex Space

3.1.3 Segment Classifiers Training and Evaluation

As described above, each audio segment is represented by a 134-dimensional feature vector, while three classification tasks have been defined: emotion, valence and arousal, using 6 different datasets. For these classification tasks, we have first performed a separate evaluation pipeline for each different dataset. The results of this *inner-dataset evaluation* procedure is shown in Table 1. The classification algorithm used for this experimentation was the SVM with RBF kernel, which outperformed all traditional classification methods (decision trees, random forests and k-nearest-neighbors).

Apart from these dataset-dependent results we have also conducted experiments using a "leave-one-dataset-out" rationale, in order to perform a *cross-dataset evaluation*. In both evaluations a repeated cross validation approach has been adopted with a 80% - 20% train-test data split and 100 iterations. The cross-dataset evaluation results for the best classifier (again SVM with RBF kernel) are shown in Table 2. All the metrics shown are f1 macro-averaged.

Comparing the results of the above two tables, we can observe that in the cross-dataset evaluation

Classification Task	Merged Dataset		
	Xgboost	CNN	SVM
Emotion	60.4	60.5	64.4 (+/-0.9)
Valence	51.7	52.6	55.2 (+/-1.2)
Arousal	64.2	69.3	66.8 (+/-1.1)

Table 3: Merged-dataset Evaluation of Audio Models

(Table 2), the results are just slightly better than random guess on average (25%). This implies that the problem that these models are called upon to solve is directly dependent on the specific sub-domain. By "sub-domain", we mean the set of context conditions and types of speakers in each dataset. Cross-domain adaptation is one of the most common difficulties in speech emotion recognition. And it is beyond the scope of this paper to handle this issue. The most straightforward way for our scope (which is to create segment classifiers that can be used as feature extractors for the recording-level decisions), is to simply train our segment models on a *merged emotional dataset*. The results of cross-validation on this merged dataset are presented in Table 3. The machine learning algorithm used for this type of experiments is SVM with RBF kernel, as well as xgboost (Chen and Guestrin, 2016). In addition, to manage the imbalance of datasets and to increase the performance, we also used a sampler SMOTE-Tomek (Wang et al., 2019), a StandardScaler, a VarianceThreshold with threshold set equal to zero and a PCA (Kabari and Nwamae, 2019). During the training, a gridsearch was applied which uses RepeatedStratifiedKFold cross validation with 5 folds. Hyperparameter tuning is performed in three hyperparameters: the number of components of the pca that are maintained, the hyperparameters γ and C of the SVM.

For comparison reasons, we also evaluated the performance of a Convolutional Neural Network (CNNs) with melgrams used as audio features, which is a common approach in Deep Learning for audio classification. Towards this end, the open-source Python library *deep_audio_features* (Theodoros Giannakopoulos, 2020) was used. The CNN has 4 convolutional layers with kernels 5×5, single stride and zero padding, while max pooling of size 2 was used. The output channels (i.e. the

third dimension), for the first layer are 32, for the second 64, for the third 128 and for the forth 256. After the convolution layers, we use 3 linear layers, with the first having an output dimension of 1024, the second 256 and the third equal to the number of classes.

The above results show that the best performance is achieved when using the SVM classifier. CNN is outperforming only for the arousal task, which indicates that more data may be needed for this deep approach to outperform. Of course, more sophisticated approaches could be used also capturing temporal dependencies between features (such as LSTMs or Transformers), but this is to be considered for future work. Finally, we have experimented with speaker independent experiments, by evaluating the SVM classifier on a subset of audio segments of unseen speakers (i.e. speakers whose segments were not available in the training data). This speaker-independent evaluation showed results that were on average 3% worst than the ones appearing on Table 3. This indicates that the speaker independence assumption does not significantly affect the performance for this particular model.

3.2 Text Analysis

3.2.1 Segmentation and Feature extraction

The Speech API provided by Google was used in order to extract textual information from the initial audio signal. The text from the whole recording can be segmented using three different approaches: sentence-level splitting, splitting into windows of predefined number of words or splitting in fixed time windows. In order to train the models described in the next paragraphs, the samples are pre-segmented in sentences.

In Natural Language Processing (NLP), word embedding is a term used to represent words in text analysis, usually in the form of a real valued vector that encodes the meaning of the words so that they can be represented in a joint representation space where the closest words are expected to have a similar meaning (Mikolov et al., 2013). To obtain word embeddings, experimentations with two pre-trained natural language models, ie. FastText (Bojanowski et al., 2016) (trained on data of English Wikipedia) and BERT (Devlin et al., 2018) (trained on data of BooksCorpus (Zhu et al., 2015) and English Wikipedia), were conducted. For the BERT architecture, given a text segment/sentence, the embeddings of the last 4 layers were averaged in order to get a more general representation.

	IEMOCAP		
Classification Task	SVM with FastText Embeddings	XGBOOST with BERT Embeddings	SVM with BERT Embeddings
Emotion	66.5 (+/-1)	63.9 (+/-1.7)	**69.5 (+/-1.4)**
Valence	61.5 (+/-1)	59.4 (+/-0.9)	**63.8 (+/-1)**
Arousal	48.8 (+/-1.1)	48.2 (+/-1)	**51 (+/-1.1)**

Table 4: IEMOCAP Evaluation of Text Models

As the IEMOCAP dataset is the only data collection, from the ones presented in section 3.1.2, that contains transcriptions (textual information), this is the one that will be used for the training/evaluation of proposed models (emotions/valence/arousal).

3.2.2 Segment classifiers training and evaluation

Experiments with both Fasttext and BERT embeddings were conducted on all three classification tasks. For training, an appropriate parameter tuning of a pipeline cosnisting of SMOTETomek (to handle imbalance), StandardScaler, VarianceThreshold, PCA and either SVM with RBF kernel or XGBOOST classifier was held. The evaluation procedure was performed using RepeatedStratifiedKFold validation scheme with 5 folds and 3 repetitions. The macro-averaged f1 score metric is shown in Table 4, where the +/- sign indicates the standard deviation of the metrics across different test folds.

From the listed results it can be clearly seen that *(i)* the use of BERT embeddings results in better performance probably due to stronger word representation power, and *(ii)* the SVM classifier is superior to the XGBOOST for all three classification tasks.

4 Recording-Level Analysis

The overall goal of segment-level analysis is to be used in order to extract recording-level information. Towards this end, we will combine segmented information with high-level features in order to perform a recording analysis.

4.1 Aggregation of Class Posteriors

Segment-classifiers result in three labels associated with emotion, valence and arousal respectively. In order to characterize the whole speech signal, an **aggregation** of the class posteriors across the recording length is performed. For example, the average emotion confidence per label may be: $P(emotion = sad) = 0.3$, $P(emotion =$

$neutral) = 0.4$, $P(emotion = happy) = 0.1$, $P(emotion = angry) = 0.2$

4.2 High Level Features

In order to capture long-term dependencies some high-level features, for both audio and text, need to be calculated across the input signal.

For the aural modality, a voice activity detection is firstly performed using features of the pyAudioAnalysis library (Giannakopoulos, 2015), in order to train in a semi-supervised fashion an SVM classifier so as to detect periods of silence. After identifying the parts of voice and silence in speech, the following high-level features are calculated: average silence duration, silence segment per minute, standard deviation of silence duration, speech ratio and word rate in speech. In order to extract different kinds of silence (ie. inter-word and intra-word), the aforementioned features are calculated for 2 different short-term windows: windows of 0.25 step and 0.5 length and windows of 0.25 step and 1 length respectively, resulting in 10 high-level features for each audio file.

As for the textual modality, the following high-level features are extracted: word rate, unique word rate and 10-bin histogram of word frequencies (frequency must range between 0 to 0.1 in order to filter out non-informative words), resulting in 12 high-level features.

5 Public Speaking Quality Dataset

Speech Quality Assessment is a task of interest in psychology, public speaking, rhetoric and a variety of other related sciences. However, this problem is quite difficult to track with the use of computational approaches. In order to address this need we introduce the Public Speaking Quality (PuSQ) Dataset, a data collection that contains speech audio and text files annotated from human listeners.

5.1 Data Acquisition

For the needs of the data collection process, a web application[1], named RecSurvey was created and the participants could use it through their personal recording set-up (ie. headset or PC microphone). The participants had to firstly fill out their demographic information (age, ethnicity, gender, English fluency etc) and then record themselves either reading some of the 40 predefined English texts (4-5 lines each) of different topics (politics, books, machine learning, etc) or answering some of the 20 general questions such as "What do you like most about your current job?". Here it has to be noted that, although our purpose is to evaluate the quality of free speech, a variety of predefined texts were used in order to have a quantitative control of the result.

The process of data acquisition resulted in a total of 695 recordings/speeches from 42 different individuals, of which 26 were female. In addition, people of different nationalities were considered so as to have a variety of pronunciation and speaking styles.

5.2 Annotation

The annotation process is mandatory in order to create a labeled dataset. Towards this end another web application[2] was created and used so as to annotate the collected speeches based on three indicators:

- **expressiveness:** how active, emotional or passionate the speech is, regardless of its content.
- **ease of following:** the evaluation of verbal clarity, fluency and rate of speech, for the specific content described. It is noted that fluency, clarity and rate can be correlated. For example one speaker, despite speaking fast, may deliver an easy-to-follow speech, while the opposite may hold for another speaker.
- **enjoyment:** defines the listener's / annotator's personal view of whether the speech was exciting, entertaining or motivating.

The marking/annotation in each of the above classes was done using 5 staggered labels, ie. the annotator had to rate each recording in the range from 1 to 5, with 1 being the worst and 5 the best measure.

In total, 14 annotators made 2687 markings by labeling 689 out of the 695 recordings for each of the 3 tasks. Further details on the number of annotations per user are illustrated in Figure 3.

5.3 Annotations Aggregation

Although the labels are distinct (5 labels / values per task), they have a continuous, scalable form as the smaller label (1) corresponds to the worst performance, while the larger label (5) corresponds to

[1] https://github.com/lobracost/RecSurvey

[2] https://github.com/sofiaele/audio_annotator

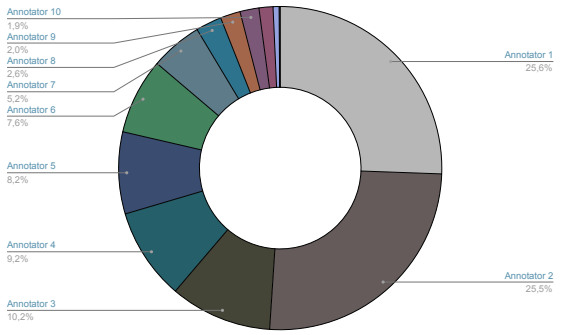

Figure 3: Annotations per User

	Expressiveness			
	Female		Male	
	Positive	Negative	Positve	Negative
Mean Thresholding	$\mu >= 4$	$\mu <= 2$	$\mu >= 3.1$	$\mu <= 2$
Number of samples after Mean Thresholding	72	80	70	67
Deviation Thresholding	$\sigma < 0.75$	$\sigma < 0.75$	$\sigma < 0.75$	$\sigma < 0.75$
Number of samples after Deviation Thresholding	63	71	48	60
Number of samples after Minimum Annotators	52	53	41	50
Average Disagreement	**0.52**		**0.53**	

Table 5: Definition of Expressiveness Dataset

the best one. Therefore, instead of aggregating the annotations of a recording based on majority voting, averaging have been used. More specifically, for each task (expressiveness, ease of following, enjoyment) and for each gender (female, male), 2 classes / labels were exported: one negative and one positive. Here it has to be noted that the gender separation was required since, the annotated quality of speech was biased to favor female speakers and adding the fact that the two genders are easily distinguishable due to different speech sound frequencies, the problem of over-fitting arose.

To aggregate the annotations and produce the binary datasets, the following three filterings have been applied on each sample:

- samples with less than three annotators are excluded
- the mean value of the labels given by different annotators, must either be under a lower or above an upper threshold. Thus, the instance is either labeled as negative/positive or excluded from the dataset (when $> lower$ and $< upper$).
- the median absolute deviation of the annotations is required to be less than or equal to a predetermined threshold. Practically, this value indicates the average deviation that results from the deviations of each annotation from the average.

Furthermore, some agreement metrics have been calculated in order to get an idea of how well defined the final labels are. Firstly, the average disagreement of annotators is defined as the average value of the median absolute deviations from all samples. A second metric that indicates the annotators validity, is the average disagreement for each participant. That is, for each sample that the user annotated, the deviation of the label she/he has set from the average value of all the annotations of this sample is calculated and then averaged across all the user's annotations in order to get the average user disagreement.

The results of the filtering procedure and the calculated agreement metrics for the expressiveness task are listed in Table 5. The corresponding tables for the remaining tasks can be found on the dataset's repository[3].

Here, it has to be noted that the average disagreement of the task "ease of following" is high enough (ie. 0.57 female - 0.58 male) which indicates that this task is ill-defined and thus it will not be accounted for the experiments.

5.4 Data Availability

PuSQ is publicly available in https://github.com/sofiaele/PuSQ in the form of extracted audio and text features and ASR text files for the two valid tasks (expressiveness and enjoyment).

6 Experiments

For each of the two tasks (expressiveness, enjoyment), 8 different types of experiments, in terms of the features used, were conducted. More specifically, the below features, together with early or late fusion among them, were used:

1. **Meta Audio (MA):** Audio features derived from the segment-level classifiers together with the high-level audio features described in section 4.2, resulting in a 20d feature vector.
2. **Text (T)**: Text features derived from segment-level classifiers together with the high-level text features described in previous sections, resulting in a 22d feature vector.

[3] https://github.com/sofiaele/PuSQ

	Individual Modalities			Fusion Methods				
	Meta Audio MA	Text T	Low Level Audio LLA	MA + T	MA and LLA Late Fusion	MA and LLA Early Fusion	MA + T and LLA Late Fusion	MA + T and LLA Early Fusion
Female Expressiveness	71	37	**77**	66	**77**	76	75	75
Male Expressiveness	69	41	**71**	71	75	-	**79**	-
Female Enjoyment	44	**65**	57	51	44	57	51	**62**
Male Enjoyment	57	48	**70**	60	64	70	**74**	66
Free Text Expressiveness	75	65	**87**	91	87	84	**93**	86
Free Text Enjoyment	**71**	57	56	**86**	66	62	75	67

Table 6: Evaluation of recording-level classification (AUC metric)

3. **Low Level Audio (LLA)**: Low level audio features which are a long-term average of the features that were presented in section 3.1.1 and were used for segment classifiers, resulting in a 136d feature vector.

During the conducted experimentation three types of classifiers were tested: *(i)* SVM with RBF kernel, *(ii)* Gaussian Naive Bayes, and *(iii)* Logistic Regression. After the appropriate data pre-processing and parameter tuning techniques, a Leave-One-Speaker-Out (LOSO) validation was used in order to evaluate the performance of the classifiers in a speaker-independent manner. The metric of interest was the Area Under the ROC Curve (ROC-AUC) calculated on the aggregated probabilities of the LOSO validation. This metric was chosen instead of other widely used classification metrics, such as the f1-score, since the data are minimal and thus f1-score is prone to small changes.

In Table 6, the final results are summed up. The last two rows include the outcome of the evaluation of free-text only samples, ie. only the answers to questions and not predefined texts. It has to be noted that the Gaussian Naive Bayes was the best performing algorithm for all tasks, except from Male Expressiveness/Meta Audio where Logistic Regression was chosen.

From the presented evaluations, it can be easily seen that in all cases, the best fusion method has either increase or keep equivalent performance compared to the best individual method. The only exception is Female Enjoyment, where there is a slight deterioration of an absolute 3%, which however can be considered negligible.

Another important observation is associated with the tasks of Male Expressiveness and Male Enjoyment, where the combination of Meta Audio with Text features (MA + T), seems to result in increased performance compared to MA and T individually. This fact shows that the textual information can significantly help in distinguishing the two classes (negative, positive), mostly when involving free text, where the recording differs among participants and contains different semantical information.

In addition, it is observed that in most cases (4 out of 6), the combination of information from all feature spaces results in the best performance. Also, most of the times, late fusion marks better results than early fusion, which indicates that late fusion introduces a normalization factor in that dataset, since the models are not directly exposed to the low level features that may result in over-fitting.

The code of the experimentations is open sourced and can be accessed in `https://github.com/tyiannak/readys`.

7 Conclusion

In this work we presented PuSQ, a public speaking quality dataset, that introduces the tasks of speech expressiveness and enjoyment in public speech data. In order to address these speech quality assessment tasks, we designed a hierarchical classifier that is based on both segment-level emotion analysis and recording-level analysis where the aforementioned information is aggregated along with some high-level speech features. It is noteworthy that the presented pipeline can be used for any multimodal (audio and text) speech analytics process, and that both the dataset and the proposed ML framework are openly provided.

In a future work, several issues can be addressed, such as the extension of the dataset, the integration of learning methods that take into account the annotation confidence (eg. Sharmanska et al., 2016; Fornaciari et al., 2021), the use of more robust segment-level classifiers (CNNs,LSTMs,Transformers etc) (eg. Fayek et al., 2017; Jiang et al., 2020) as well as the inclusion of domain adaptation techniques (eg. Mao et al., 2016, 2017; Ocquaye et al., 2019; Huang et al., 2017) and the application of transfer learning from unsupervised temporal models (eg. Wang and Zheng, 2015; Feng et al., 2019).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. volume 5, pages 1517–1520.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Robert E. Carlson and Deborah Smith-Howell. 1995. Classroom public speaking assessment: Reliability and validity of selected evaluation instruments. *Communication Education*, 44(2):87–97.

Lei Chen, Gary Feng, Chee Wee Leong, Jilliam Joe, Christopher Kitchen, and Chong Min Lee. 2016. Designing an automated assessment of public speaking skills using multimodal cues. *Journal of Learning Analytics*, 3:261–281.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an Italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3501–3504, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

P. Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68. Advances in Cognitive Engineering Using Neural Networks.

Kexin Feng, Megha Yadav, Md Nazmus Sakib, Amir Behzadan, and Theodora Chaspari. 2019. Estimating public speaking anxiety from speech signals using unsupervised transfer learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PLOS ONE*, 10:e0144610.

Jacek Grekow. 2018. *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*. Springer, Cham.

Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. 2017. Unsupervised domain adaptation for speech emotion recognition using pcanet. *Multimedia Tools and Applications*, 76.

Philip Jackson and Sana ul haq. 2011. Surrey audio-visual expressed emotion (savee) database.

Changjiang Jiang, Junliang Liu, Rong Mao, and Sifan Sun. 2020. Speech emotion recognition based on dcnn bigru self-attention model. In *2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pages 46–51.

Ledisi Kabari and Believe Nwamae. 2019. Principal component analysis (pca) -an effective tool in machine learning.

Panagiotis Koromilas and Theodoros Giannakopoulos. 2021. Deep multimodal emotion recognition on human speech: A review. *Applied Sciences*, 11(17):7962.

Duc Le, Zakaria Aldeneh, and Emily Mower Provost. 2017. Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. In *Interspeech*, pages 1108–1112.

S. R. Livingstone and F. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13.

Qirong Mao, Guopeng Xu, Wentao Xue, Jianping Gou, and Yongzhao Zhan. 2017. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication*, 93:1–10.

Qirong Mao, Wentao Xue, Qiru Rao, Feifei Zhang, and Yongzhao Zhan. 2016. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2608–2612.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013*

conference of the north american chapter of the association for computational linguistics: Human language technologies, pages 746–751.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

Elias Nii Noi Ocquaye, Qirong Mao, Heping Song, Guopeng Xu, and Yanfei Xue. 2019. Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition. *IEEE Access*, 7:93847–93857.

Andrew Ortony and Terence Turner. 1990. What's basic about basic emotions? *Psychological review*, 97:315–31.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57.

Lisa Schreiber, Gregory Paul, and Lisa Shibley. 2012a. The development and test of the public speaking competence rubric. *Communication Education*, 61.

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

Viktoriia Sharmanska, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Theodoros Giannakopoulos. 2020. *Pytorch implementation of deep audio embedding calculation Resources*.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237.

Zhe Wang, Chunhua Wu, Kangfeng Zheng, Xinxin Niu, and Xiujuan Wang. 2019. Smotetomek-based resampling for personality recognition (july 2019). *IEEE Access*, PP:1–1.

A. E. Ward. 2013. The assessment of public speaking: A pan-european view. In *2013 12th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 1–5.

Wilhelm Max Wundt and Charles Hubbard Judd. 1897. *Outlines of psychology*. Leipzig, W. Engelmann, New York, G.E. Stechert.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.