MSP-Face Corpus: A Natural Audiovisual Emotional Database

Andrea Vidal
University of Texas at Dallas
Richardson, Texas
axv170003@utdallas.edu

Ali Salman
University of Texas at Dallas
Richardson, Texas
ans180000@utdallas.edu

Wei-Cheng Lin
University of Texas at Dallas
Richardson, Texas
wei-cheng.lin@utdallas.edu

Carlos Busso
University of Texas at Dallas
Richardson, Texas
busso@utdallas.edu

ABSTRACT

Expressive behaviors conveyed during daily interactions are difficult to determine, because they often consist of a blend of different emotions. The complexity in expressive human communication is an important challenge to build and evaluate automatic systems that can reliably predict emotions. Emotion recognition systems are often trained with limited databases, where the emotions are either elicited or recorded by actors. These approaches do not necessarily reflect real emotions, creating a mismatch when the same emotion recognition systems are applied to practical applications. Developing rich emotional databases that reflect the complexity in the externalization of emotion is an important step to build better models to recognize emotions. This study presents the MSP-Face database, a natural audiovisual database obtained from video-sharing websites, where multiple individuals discuss various topics expressing their opinions and experiences. The natural recordings convey a broad range of emotions that are difficult to obtain with other alternative data collection protocols. A feature of the corpus is the addition of two sets. The first set includes videos that have been annotated with emotional labels using a crowd-sourcing protocol (9,370 recordings – 24 hrs, 41 m). The second set includes similar videos without emotional labels (17,955 recordings – 45 hrs, 57 m), offering the perfect infrastructure to explore semi-supervised and unsupervised machine-learning algorithms on natural emotional videos. This study describes the process of collecting and annotating the corpus. It also provides baselines over this new database using unimodal (audio, video) and multimodal emotional recognition systems.

CCS CONCEPTS

• **Information systems** → **Multimedia databases**; • **Human-centered computing** → *Social content sharing*; Social media.

KEYWORDS

Multimodal emotional database, emotion recognition

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418872

ACM Reference Format:
Andrea Vidal, Ali Salman, Wei-Cheng Lin, and Carlos Busso. 2020. MSP-Face Corpus: A Natural Audiovisual Emotional Database. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418872

1 INTRODUCTION

Recent advances in machine-learning have led to impressive performance for several tasks, including *automatic speech recognition* (ASR) [43], object recognition [33, 34], and face recognition [11, 44]. However, there are important tasks where achieving high performance is still a challenge. One of these tasks is emotion recognition. Emotion recognition is a challenging task due to its subjective nature [6, 28, 40]. The underlying emotional state in a video or audio is often approximated with subjective evaluations, which do not always agree in their judgment, creating noisy labels to train emotional classifiers [20, 40]. Furthermore, existing databases are often limited in the number of participants, emotional content and naturalness [10, 22]. There is also a need for multimodal databases, where complementary information can be derived across modalities [7]. In fact, the addition of multiple modalities for emotion recognition including audio, images, and text, can help improve the performance of a system. The research community has created several multimodal emotional databases to build emotion recognition systems [5, 8, 18, 29, 41, 45]. However, several of these databases are acted [8, 10, 19, 23]. Other databases rely on different emotional elicitation techniques [24, 35]. Both of these approaches have limitations in representing the emotional content observed in real daily interactions.

This paper introduces the MSP-Face database, which is a natural emotional multimodal corpus collected from video-sharing websites. The recordings include people in front of a camera speaking about different situations from their daily life, giving their opinions. The videos consist of natural and spontaneous recordings, where the emotions are not acted or artificially elicited. The data collection protocol is flexible and scalable, and addresses key limitations of existing multimodal databases. For example, the corpus is collected from multiple participants, expressing a broad range of emotions, which is not easily achieved with other data collection protocols. It provides an ideal resource to study the expression of emotions across audiovisual modalities. The database has two sets, which is a novel feature of the corpus. The first set is the labeled data, which consists of 9,370 videos (24hrs, 41m) annotated with

Table 1: Overview of some of the existing multimodal emotional databases. The symbol '-' indicates that the information is not available.

Database	Type	Size [hrs]	Speakers
eNTERFACE'05 [23]	Acted	-	42
CREMA-D [10]	Acted	-	91
RAVDESS [19]	Acted	-	24
IEMOCAP [5]	Acted	≈ 12	10
MSP-IMPROV [8]	Acted	≈ 9.5	12
CreativeIT [25, 26]	Acted	≈ 8	16
SEMAINE [24]	Natural	≈ 75	150
MAHNOB-HCI [41]	Natural	-	27
RECOLA [35]	Natural	≈ 3.75	46
SEWA [18]	Natural	44	398
CMU-MOSEI [45]	Natural	≈ 65	1,000
MSP-Face	Natural	≈ 24.7 (+46)	302

emotional labels. The emotional content is determined with perceptual evaluations conducted with crowdsourcing. We annotate the emotional content with categorical and attribute-based descriptors. The categorical-based descriptors include primary emotion (i.e., the most dominant emotion perceived in the video), and secondary emotions (i.e., all other emotional traits also perceived in the video). The attribute-based descriptors include valence, arousal, and dominance. Each video is annotated by at least five workers. The second set is the unlabeled data, which consists of 17,955 videos (45hrs, 57m). This set is intentionally left without annotations of emotions to explore semi-supervised and unsupervised machine-learning algorithms for emotion recognition, with videos that are similar to the recordings in the labeled set. With over 70 hrs of labeled and unlabeled data, this corpus provides an important resource that complements and extends existing emotional databases. This corpus can be used not only for emotion recognition problems, but also in other applications such as generating visual agents with expressive behaviors [38]. We will share this corpus with the research community to increase the infrastructure available to address these important research areas.

This paper is organized as follows. Section 2 describes the main differences between existing multimodal databases and our corpus. Section 3 presents the MSP-Face database, describing in details the data collection process, and the annotation of the emotional content using crowdsourcing-based perceptual evaluations. Section 4 presents our analysis of the emotional content of the corpus. Section 5 provides baselines for mono-modal (audio, video) and multimodal emotion recognition systems. Finally, Section 6 concludes our paper, summarizing the key features of the corpus and discussing future directions.

2 RELATED WORKS

A key requirement to facilitate the study of emotion is access to large natural databases that capture the complexity of the externalization of emotion. This is particularly important with new deep learning formulations that require large training sets. This section describes some of the existing multimodal datasets, focusing our discussion on audiovisual corpora.

Table 1 lists some of the audiovisual databases used in the community, highlighting some of their key features. The first type of databases corresponds to acted databases, where speakers are asked to read sentences expressing emotions. This technique was used in several corpora, including the RAVDESS [19], eNTERFACE'05 [23], and CREMA-D [10] databases. Other emotional databases were also collected from actors, but the protocol was designed to have more realistic interactions. The MSP-IMPROV [8], IEMOCAP [5] and the CreativeIT [25, 26] databases are examples of this approach, where actors improvise specific scenarios carefully selected to elicit certain emotions. This approach is effective, leading to recordings that are perceived as more natural than recordings from actors pretending to express emotions while reading a script [8]. Another technique used to collect multimodal databases is to elicit emotions during a conversation. An example of this approach is the SEMAINE database [24], where one of the participants portrays a given personality, aiming to elicit a target emotional reaction in the other participant. Another approach to elicit more natural emotional interactions is to induce emotions by showing videos intended to affect the mood of the participants. This approach was used in the MAHNOB-HCI database [41]. Unlike previous databases, they used self-report to annotate the emotions (categorical and attribute annotations). This emotion induction process can also be used before the subjects engage in a conversation. Examples of the use of this emotion induction process are the collections of the RECOLA [35] and SEWA [18] databases, where people interacted with each other online after watching emotional stimuli. More natural multimodal databases include the CMU-MOSEI [45] corpus, where the recordings did not include acted renditions, and the speakers were not emotionally induced before the recordings. This corpus includes YouTube videos, where people talk in front of the camera.

This paper presents the MSP-Face corpus, a natural multimodal emotional database that consists of videos of people talking about their experiences, feelings, and opinions. The experienced emotions in those videos are natural, providing a better representation of the externalization of emotion. The size, number of speakers and emotional diversity in the recordings make this corpus a unique resource in the study of multimodal emotion recognition. The most similar database to our corpus is the CMU-MOSEI database [45]. Three important advantages of our corpus are the emotional descriptors, the emotional distribution, and the unlabeled sets included in the corpus. First, the CMU-MOSEI corpus was annotated with six categorical emotions and one attributed-based descriptor (i.e., sentiment, which is equivalent to valence). The labels were obtained by only three annotators. Our description of emotion is more complete, obtaining annotations by five or more workers. We describe the emotions in the videos with primary and secondary emotions, as described in Section 3.2. These annotations are complemented with three attribute-based annotations (arousal, valence and dominance) providing a rich characterization of the emotional content of the videos. Second, the distribution of the emotional categories in the CMU-MOSEI corpus was heavily biased towards one emotion ("happiness"), representing approximately 40% of the corpus. In contrast, our selection of videos aimed to have enough samples per emotional class, without a dominant class. The predominant emotion

in our database corresponds to only 23% of the annotated videos ("sadness"). Finally, our corpus provides additional videos without labels to facilitate semi-supervised and unsupervised solutions for multimodal emotion recognition.

3 THE MSP-FACE DATABASE

The MSP-Face corpus is a database collected by the *Multimodal Signal Processing* (MSP) Laboratory at The University of Texas at Dallas. The MSP-Face corpus is a collection of online videos of people talking in front of a camera about diverse topics, expressing a broad range of emotions.

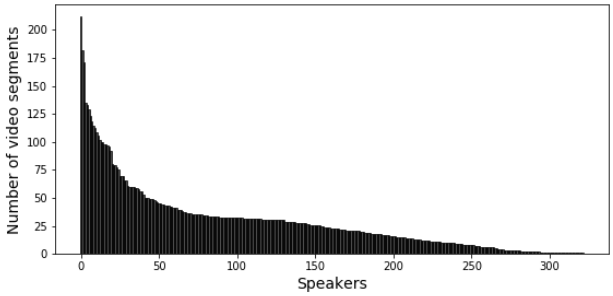
3.1 Protocol for the Selection of the Recordings

The data collection process of the MSP-Face corpus is flexible and scalable. The first step consists of selecting candidate recordings from video-sharing websites. The videos were selected to have a single speaker, frontal face, clear speech, and no background music. We select videos where one person is talking in front of the camera, where his/her face is visible. Using single-person videos removes the need for speaker diarization to know who is speaking. It also eliminates speech overlap, which is an important problem in speech processing during multiparty interactions. We require the face to be mostly frontal so we can reliably extract facial features. We also select videos without background music, where the speech quality is acceptable. Our goal is to have fair conditions for speech-based emotion recognition systems.

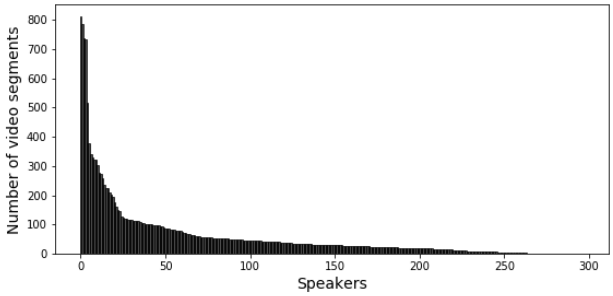
The second step is to manually segment the videos into short recordings. Our emotion annotation process relies on sentence-level annotations, where evaluators provide a global descriptor after watching the entire video segment. This data annotation process imposes constraints on the duration of the video segment. We avoid having recordings that are too short, where evaluators do not have enough contextual information to reliably assess the emotional content. Also, we avoid having recordings that are too long, where emotions may fluctuate within the segment. To balance this tradeoff, we segment the videos into recordings with a duration between three and ten seconds. In total, we have collected 27,325 video segments, which correspond to 70hrs and 38m of audiovisual recordings.

The final step in this process is to manually annotate the identity of the participants in the video. The identity of the participants is provided as a unique identification number. In total, the corpus has videos from 491 people. Figure 1 shows the number, in decreasing order, of segmented videos for each participant in the MSP-Face database. The corpus has at least 50 videos from 140 participants. This database includes speakers from the black, latino, and LGBTQ communities, which provides the speaker diversity that most multimodal databases lack.

The recording are split into labeled (9,370 videos –24hrs, 41m) and unlabeled (17,955 videos –45hrs, 57m) sets. The labeled set is annotated with emotional labels using perceptual evaluations (Sec. 3.2). The recordings from this set include videos from 322 people. The partitions of the labeled set were created by selecting videos of 222 participants for the train set (70% of the labeled data), videos of 27 participants for the development set (10% of the labeled data), and videos of 73 participants for the test set (20% of the labeled data). This partition creates participant-independent sets, where the recordings of one subject are exclusively included in one of these sets. The unlabeled set includes recordings from 297 participants. There is an overlap between the participants in the labeled and unlabeled sets. From 297 participants, 128 of them have segments in the labeled set, which corresponds to 17,291 segments (12,049 from the unlabeled and 5,242 from the labeled set). The unlabeled set is intentionally left without emotional annotations to facilitate research on semi-supervised and unsupervised machine-learning algorithms for multimodal emotion recognition.



(a) Labeled set



(b) Unlabeled set

Figure 1: Number of videos per speaker in the MSP-Face database (labeled and unlabeled sets). The speakers are listed in descending order.

3.2 Annotation of Emotional Content

The emotional content of the labeled set is annotated with perceptual evaluations using a crowdsourcing protocol implemented in *Amazon Mechanical Turk* (AMT). The database is annotated using workers, who complete *human intelligence tasks* (HITs). The reliability of the evaluations in our annotation process is important. For that reason, the workers must meet certain requirements to qualify for our HITs. The workers must live in The United States, have at least 100 accepted HITs, and have a 95% acceptance rate. In addition to qualification criteria, we replicate the online approach proposed in Lotfian and Busso [22], which is a modified version of the crowdsourcing protocol presented by Burmania *et al.* [4]. This crowdsourcing protocol tracks the performance of the workers in real-time, stopping the evaluation when their quality drops below

a certain threshold. We assess the performance of the workers by placing videos with annotations, which are used as references. During the annotation process, we create blocks of four videos, where each block consists of three videos to be annotated and one reference video. The placement of the reference video is randomized so the workers cannot determine the videos that are used to track their performance. For each HIT, a worker initially annotates 12 video segments, including three reference videos. If their performance is considered adequate, they can evaluate four more videos. The quality assessment always considers the last three reference videos. The workers are allowed to annotate up to 100 video segments per HIT. The study of Lotfian and Busso [22] provides more details. In addition, we frequently measure the inter-evaluator agreement of each worker, creating a list with the ones that provide consistently low agreement across HITs. These workers are blocked for future HITs.

Figure 2 shows the emotional questionnaire used for the data collection, which is similar to the method used in previous data collection efforts [8, 22]. We evaluate the corpus with at least five annotators per video, since studies have shown that five evaluations is a reasonable number to evaluate emotional databases [3]. Some of the videos have more than five annotations, since they are used in the reference set in the crowdsourcing protocol. The workers evaluate the video segments with attribute-based and categorical-based emotional representation.

Motivated by the core affect theory [36], we annotate the emotional content with emotional attributes. We use three of the most common attributes used in previous studies: arousal, valence, and dominance. These attributes are evaluated using a seven-point Likert scale, which characterizes each of the emotional attributes as follows: valence (from very negative to very positive), arousal (from very calm to very active), and dominance (from very weak to very strong). These evaluations are implemented using *self-assessment manikins* (SAMs) [15, 16], as shown in Figures 2(a)-2(c). These pictorial representations help the workers visualize the description and meaning of these emotional attributes, helping to calibrate their emotional judgments. A sample video is shown to the workers before the annotation process to explain examples of videos with clear emotional attribute scores. This video is short since we do not want to bias the annotation. The consensus label for each emotional attribute is the average of the scores assigned by the workers.

We also annotate the emotional content in terms of categorical emotions, which has some benefits from an application perspective. The workers have to choose primary and secondary emotions. The primary emotion is defined as the dominant emotion perceived in the video. We restrict the options to anger, sadness, happiness, surprise, fear, disgust, contempt, neutral state, and "other." Figure 2(d) shows this part of the questionnaire. The selected emotional classes include all basic emotions [13]. Furthermore, the same set of eight emotions has been used by not only commercial tools such as the Face API from Microsoft, but also existing databases such as the MSP-Podcast [22] and AffectNet corpora [27]. The class "other" is used when none of the emotional states listed in the questionnaire are appropriate. Including the class "other" is important to avoid artifacts associated with forced choices [37]. The class with the most selections across the workers is used as the consensus label for a video.

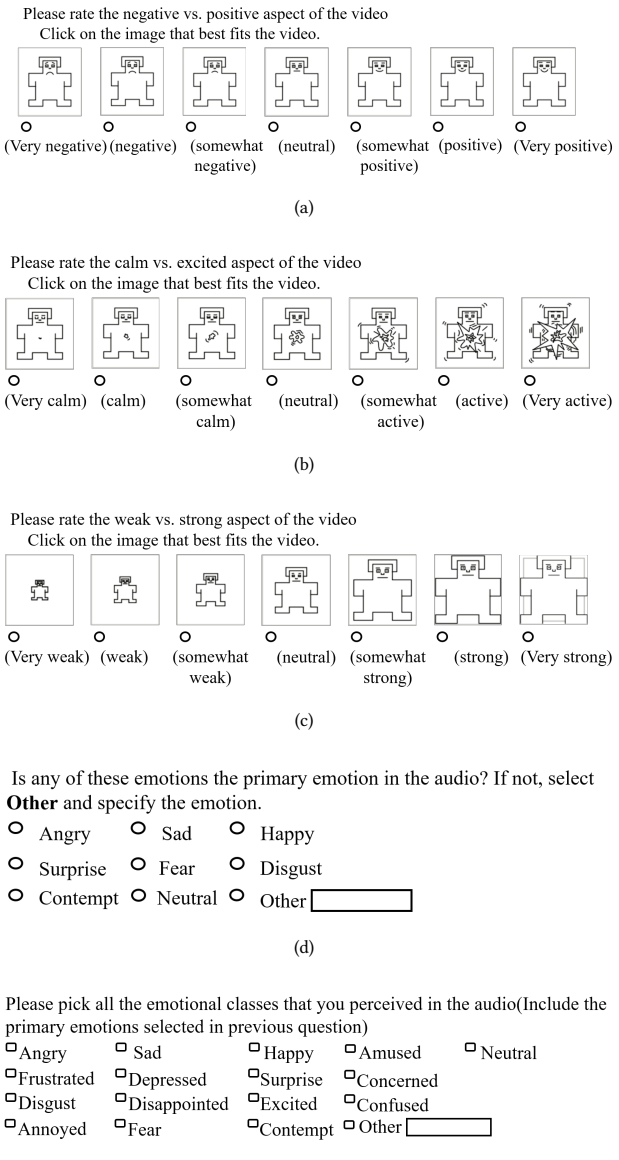


Figure 2: Emotion questionnaire used for the perceptual evaluation. The figure shows the seven-point Likert scale for the emotional attributes valence (a), arousal (b), and dominance (c). It also shows the lists of classes for the primary (d) and secondary (e) emotions.

In addition to the primary emotion, we also request the worker to annotate the secondary emotions, which helps complement the main emotion perceived from the video. Figure 2(e) shows this part of the questionnaire. The workers can select as many emotions as they want from an extended list: anger, frustrated, disgust, annoyed, sadness, depressed, disappointed, fear, happiness, surprise, excited, contempt, amused, concerned, confused, neutral and other. As Figure 2(e) shows, these emotions are grouped in broader classes with

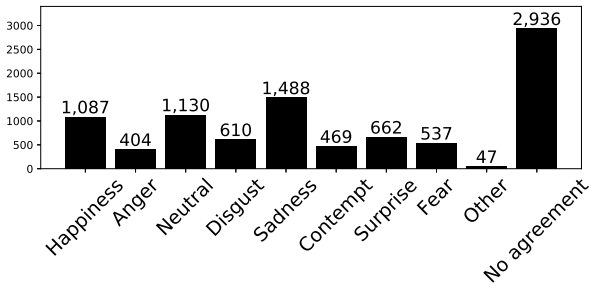


Figure 3: Histogram of the consensus labels for the primary emotions on the labeled part of the MSP-Face corpus.

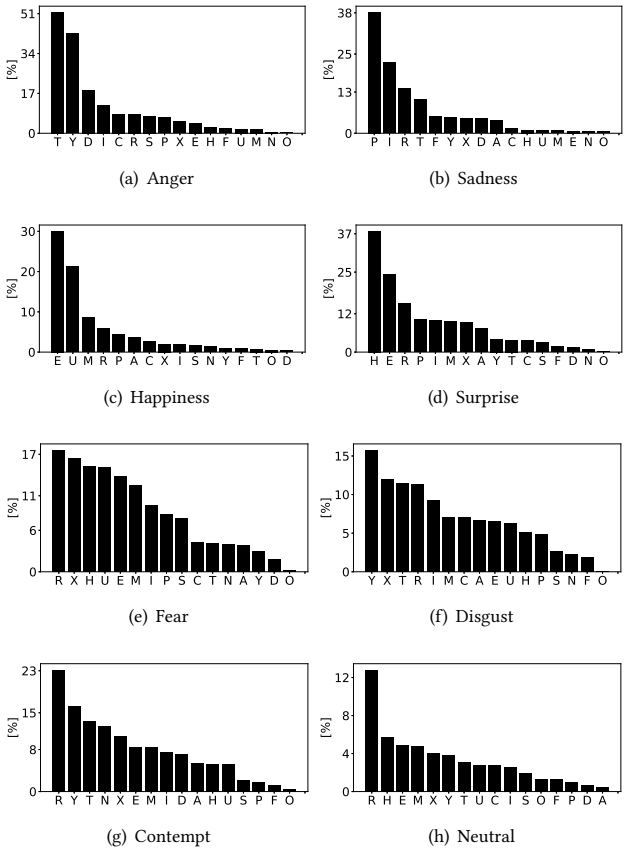
similar emotions to reduce the cognitive load of finding the right emotion within the list.

The assessment of the worker's performance in the crowdsourcing protocol is implemented with primary emotions and attribute-based descriptors.

4 ANALYSIS OF THE MSP-FACE CORPUS

This section describes the emotional content of the labeled portion of the corpus by analyzing the annotations provided by the workers. First, we consider the distribution of the consensus labels for the primary emotions, which is displayed in Figure 3. The figure shows that happiness, neutral, and sadness are well-represented in the corpus with over 1,000 videos. All the other emotional classes have at least 400 videos. There are several videos for which consensus agreement was not reached, given the challenging task of assigning emotional categories to spontaneous, natural emotions.

Secondary emotions are very useful to provide a more complete description of the emotional content of a video. They are also useful to understand the relationship between emotional classes. In computational models, studies have effectively used secondary emotions to retrieve speech samples with similar emotions to an anchor sentence [17], and as a secondary task in multitask learning for speech emotion recognition problems [21]. The workers selected an average of 1.12 secondary emotions per video, without considering the selected primary emotion. The most popular classes selected as secondary emotions are "concerned," "frustrated," and "annoyed." Figure 4 shows the histogram of secondary emotions selected for each primary emotion. For example, Figure 4(a) shows that 51.6% of the videos selected with the primary emotion "anger," received the secondary emotion "frustrated." The figure shows important relations between emotional classes. For example, the classes "happiness," "excited" and "surprise" are mutually related. Figure 4(c) shows that "surprise" and "excited" are the main secondary emotions of "happiness." Figure 4(d) shows that "happiness" and "excited" are the main secondary emotions of "surprise." Furthermore, there are primary emotions that are often selected with the same secondary emotions. For example, for the primary emotion "anger" and "disgust," the emotion "annoyed" is the top-two classes selected as secondary emotions (Figs. 4(a) and 4(f)). We observe similar patterns for the primary emotions "neutral" and "fear," where the main secondary emotion is "concerned" (Figs. 4(h) and 4(e)).



(a) Anger (b) Sadness

(c) Happiness (d) Surprise

(e) Fear (f) Disgust

(g) Contempt (h) Neutral

Figure 4: Secondary emotions selected for each primary emotion in descending order. Each figure considers all the individual evaluations where a worker selected a given primary emotion, reporting the selection rate for each secondary emotions. [A: anger; C: contempt; D: disgust; E: excited; F: fear; H: happiness; I: disappointed; M: amused; N: neutral; O: other; P: depressed; R: concerned; S: sadness; T: frustrated; U: surprise; X: confused; Y: annoyed].

Figure 5 shows the distributions of the consensus scores assigned to valence, arousal, and dominance for the videos of the MSP-Face corpus. This result shows that the labeled part of the database has a broad range of emotional content. Furthermore, we observe that the distribution of valence and arousal are balanced. The distribution for dominance is biased towards the right, which indicates that the emotions are perceived to be more dominant. Figure 6 shows the distribution of the samples in the arousal and valence space, where each video is represented with a circle. For visualization, the scores are normalized between -1 and 1. The figure shows that the emotional content of the corpus covers most of the arousal-valence space, with the exception of the lower right quadrant (e.g., videos with low arousal, positive valence). Figure 6 also highlights with a different color the consensus label for primary emotions assigned to each video. The clusters of the categorical classes are in the expected regions in the arousal-valence space. An interesting observation is

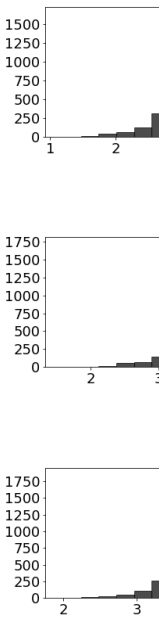


Figure 5: Histograms fo
the videos for valence, arousal, and dominance. This figure considers the labeled part of MSP-Face corpus.

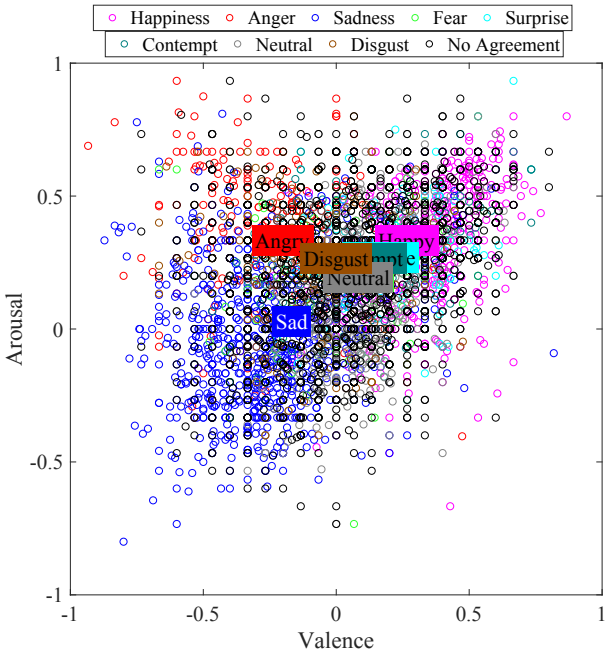


Figure 6: Coverage of the emotional content in the arousal and valence space, where each video is represented with a circle. The color of the circle represents the primary emotion assigned to the video. The name of the categorical emotions are placed at the center of their respective clusters.

the spread of each emotion in this space, which suggests variability within each class (i.e., different degree of happiness). Describing emotion with categorical and attribute-based descriptors provides a more complete representation of the emotional content of the corpus.

4.1 Inter-evaluators Agreement

Each of the 9,370 utterances in the labeled set has at least five annotations from different workers using the process described in Section 3.2. To estimate the inter-evaluator agreement, we consider the first five annotations to keep a fixed number of annotations per video. We use the Fleiss's Kappa (κ) statistic for categorical emotions. The value of the Fleiss's Kappa among our evaluators is $\kappa = 0.091$ for the primary emotions. Most databases use a reduced list of emotional classes so the reported agreements tend to be higher. We also measure the inter-agreement of the evaluators for the emotional attributes using the Krippendorff's alpha coefficient (α). The values for valence, arousal, and dominance are $\alpha_{val} = 0.214$, $\alpha_{aro} = 0.143$, and $\alpha_{dom} = 0.106$, respectively. Annotating the emotional content of naturalistic recordings is a subjective and challenging task due to differences in emotional perception. As a result, the inter-evaluator agreements tend to be low.

5 EMOTION RECOGNITION EVALUATIONS

This section provides baselines for emotion prediction systems for speech-only (Sec. 5.1), face-only (Sec. 5.2) and audiovisual (Sec. 5.3) systems. This section only considers the labeled portion of the corpus.

For attribute-based descriptors, the task is a regression problem that aims to predict the emotional value for valence, arousal, and dominance. We implement the baseline using a *multitask learning* (MTL) formulation with three output nodes, predicting the three attributes at the same time. MTL formulations have been successfully used in emotion recognition [12, 31, 32]. The loss function for the attribute-based model is based on the *concordance correlation coefficient* (CCC), which corresponds to a simple linear combination of the losses for valence, arousal and dominance: $\mathcal{L} = \frac{1}{3}(\mathcal{L}_{val} + \mathcal{L}_{aro} + \mathcal{L}_{dom})$. We also assess the performance of the system with CCC. For emotional categories, we explore two problems. First, we implement a five-class problem by considering the emotional classes happiness, anger, neutral, disgust, and sadness. Second, we implement an eight-class problem with all the emotional classes. We use cross-entropy for both of these classification problems. We measure performance with both macro-average F1-score (i.e., treating all the classes as equally important regardless of the number of samples), and micro-average F1-score (i.e., accounting for the number of samples per class). We use the Adam optimizer for all the baseline models.

Table 2: Architectures of the baseline emotion recognition models. The output layer depends on the target task. For the attributed dimensions, the MTL output layer consists of linear layer without an activation function. For the categorical emotions, the output layer consists of a softmax layer. The number of nodes is determined by the number of output classes.

SER model			FER model		
Layer	Dimension	Activation	Layer	Dimension	Activation
Input	6373	N/A	Input	1024	N/A
Linear	256	N/A	Linear	512	ReLU
Dropout	rate=0.3	N/A	Dropout	rate=0.5	N/A
Linear	256	ReLU	Linear	256	ReLU
Dropout	rate=0.3	N/A	Dropout	rate=0.5	N/A
Linear	**256**	**ReLU**	Linear	128	ReLU
Output	depends	depends	Dropout	rate=0.5	N/A
			LSTM	**512**	**Tanh**
			Output	depends	depends

Audiovisual model		
Layer	Dimension	Activation
Concatenate	256+512	N/A
Linear	256	ReLU
Linear	256	ReLU
Output	depends	depends

5.1 Speech-only Emotion Recognition

This section describes the emotion recognition evaluation using acoustic features. The goal of this section is to provide baselines for multi-class emotion recognition and attribute-based descriptors using acoustic features. The *speech emotion recognition* (SER) system relies on the high dimensional feature representation provided by the OpenSmile toolkit [14]. We use the feature set proposed for the computational paralinguistics challenge at Interspeech 2013 [39], which consists of a 6,373 dimensional vector per sentence. These features are statistical values such as mean, range and moments of *low-level descriptors* (LLDs) (e.g., energy and fundamental frequency).

Table 2 shows the model architecture for our speech-based baseline, which is constructed with several linear layers implemented with dropout to impose regularization. The input of the model is the 6,373 dimensions vector, which is standardized using the mean and standard deviation values estimated over the training samples.

The first column of Table 3 shows our baseline speech emotion recognition results. The table shows that our attributed-based models for arousal and dominance have similar performance. The performance is lower for valence since detecting this attribute from acoustic features is a difficult task [1, 9, 42]. For the emotional classes, the five-class problem achieves reasonable recognition rates given the spontaneous recordings considered in this corpus. The recognition results for the eight-class problem is more challenging, in part, because of the unbalanced distribution of the classes. These recognition results are provided as a reference for future research using this database.

Table 3: Baseline results for the speech-only, face-only and audiovisual models.

	Speech-only	Face-only	Audiovisual
Aro-CCC	0.3794	0.2065	0.3961
Val-CCC	0.2924	0.2677	0.3453
Dom-CCC	0.3390	0.2085	0.3430
5 class F1-score (macro)	0.2835	0.3027	0.3010
5 class F1-score (micro)	0.3599	0.3494	0.3641
8 class F1-score (macro)	0.1629	0.1308	0.1690
8 class F1-score (micro)	0.2637	0.3161	0.2710

5.2 Face-only Emotion Recognition

For our *facial emotion recognition* (FER) baseline, we use face images as input. We extract the faces using the OpenCV DNN library [2], which gives areas of the image detected as faces with their corresponding confidence. We select the predicted faces, choosing the one with the highest confidence that is correctly classified as a face. The baseline system has two modules: feature representation and temporal modeling. The feature representation is obtained with the VGG-16 model [46]. We initialize the weights with the VGG-FACE model provided by Parkhi *et al.* [30]. Then, we train the system using the AffectNet corpus [27]. This process is separately done for the emotional attributes, five-class, and eight-class problems. The feature representation model for the emotional attributes is a multi-task model predicting arousal and valence at the same time (dominance is not included in the annotations of AffectNet). For the multi-class problem, the output of the feature representation module is a softmax layer with one node per target emotion. Next, we use the first linear layer (1,024 neurons) of the VGG model as the feature representation. The temporal model is implemented with *long short-term memory* (LSTM) to predict the emotion in the video from the sequence of images. Table 2 describes the structure of the temporal modeling module. For the attribute-based model, the last output of the LSTM is passed to three linear neurons to jointly predict valence, arousal, and dominance. For the five and eight-class models, the output layer is a softmax layer of five and eight neurons, respectively.

Table 3 shows our baseline FER results. The CCC of the face-only model achieves lower performance compared to speech-only for arousal, valence and dominance. For the categorically based emotion recognition, the speech-only and face-only models achieve comparable results for the five-class task.

5.3 Audiovisual Emotion Recognition

We build our audiovisual model by simply concatenating the feature representations obtained just before the output layers in the speech-only and face-only models (these layers are highlighted with bold font in Table 2). Note that while training the audiovisual model, we freeze the parameters of the speech-only and face-only models. Therefore, the error is not back-propagated to those models. While this approach may result in lower performance than a multimodal system trained over the entire network, we choose this setting to simplify the training process. Furthermore, this approach allows us to generate fixed hidden representation from different modalities,

which facilitate the analysis of the effectiveness of cross modality fusion.

The last column of Table 3 shows the results for our multimodal baseline. The attribute-based model benefits from fusing the modalities, reaching the best performance for arousal, valence and dominance. An interesting case is valence. While both unimodal models achieve similar performance, the clear improvement in the multimodal system indicates that facial and acoustic features are complementary in describing valence. The fusion of categorical-based models also increases the recognition performance by relying on acoustic and facial features. We expect that future studies on this corpus will improve the results reported in Table 3.

6 CONCLUSIONS

This paper presented the MSP-Face corpus, which is a new emotional audiovisual database with naturalistic recordings. The corpus is obtained from video-sharing websites, and consist of good quality videos with clear audio and without background music. Each recording contains one individual talking to the camera about a broad range of topics. The variety of speakers (491), its size and the diversity in the emotional content make this corpus a perfect resource to study speech-only, face-only, or audiovisual emotion recognition. A portion of the corpus is annotated with emotional labels (24hrs, 41m), including attribute-based descriptors and categorical descriptors. The corpus was annotated with perceptual evaluations conducted using a crowdsourcing protocol. A key feature of this corpus is the addition of recordings without emotional labels (45hrs, 57m) that were selected and segmented using the same protocol used for the labeled set. We expect that the labeled and unlabeled sets will facilitate research on supervised, semi-supervised and unsupervised algorithms for emotion recognition.

Our goal is that the MSP-Face corpus becomes a valuable resource for the study of multimodal emotion recognition. Therefore, we will make this corpus and the source code for the baselines available to the research community [1].

7 ACKNOWLEDGMENTS

This work was funded by National Science Foundation grants IIS:1718944.

REFERENCES

[1] M. Abdelwahab and C. Busso. 2018. Study Of Dense Network Approaches For Speech Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada, 5084–5088. https://doi.org/10.1109/ICASSP.2018.8461866

[2] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[3] A. Burmania and C. Busso. 2017. A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors. In *Interspeech 2017*. Stockholm, Sweden, 152–157. https://doi.org/10.21437/Interspeech.2017-1278

[4] A. Burmania, S. Parthasarathy, and C. Busso. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing* 7, 4 (October-December 2016), 374–388. https://doi.org/10.1109/TAFFC.2015.2493525

[5] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42, 4 (December 2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[6] C. Busso, M. Bulut, and S.S. Narayanan. 2013. Toward effective automatic recognition systems of emotion in speech. In *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella (Eds.). Oxford University Press, New York, NY, USA, 110–127. https://doi.org/10.1093/acprof:oso/9780195387643.003.0008

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. 2004. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004*. ACM Press, State College, PA, 205–211. https://doi.org/10.1145/1027933.1027968

[8] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing* 8, 1 (January-March 2017), 67–80. https://doi.org/10.1109/TAFFC.2016.2515617

[9] C. Busso and T. Rahman. 2012. Unveiling the Acoustic Properties that Describe the Valence Dimension. In *Interspeech 2012*. Portland, OR, USA, 1179–1182.

[10] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma. 2014. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing* 5, 4 (October-December 2014), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. Long Beach, CA, USA, 4685–4694. https://doi.org/10.1109/CVPR.2019.00482

[12] T. Devries, K. Biswaranjan, and G. W. Taylor. 2014. Multi-task Learning of Facial Landmarks and Expression. In *Canadian Conference on Computer and Robot Vision*. Montreal, QC, Canada, 98–103. https://doi.org/10.1109/CRV.2014.21

[13] P. Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (1992), 169–200. https://doi.org/10.1080/02699939208411068

[14] F. Eyben, M. Wöllmer, and B. Schuller. 2010. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *ACM International conference on Multimedia (MM 2010)*. Florence, Italy, 1459–1462.

[15] L. Fischer, D. Brauns, and F. Belschak. 2002. *Zur Messung von Emotionen in der angewandten Forschung*. Pabst Science Publishers, Lengerich.

[16] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication* 49, 10-11 (October-November 2007), 787–800.

[17] J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso. 2019. Retrieving Speech Samples with Similar Emotional Content Using a Triplet Loss Function. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. Brighton, UK, 7400–7404. https://doi.org/10.1109/ICASSP.2019.8683273

[18] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic. 2020. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). https://doi.org/10.1109/TPAMI.2019.2944808

[19] S.R. Livingstone and F.A. 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, 5 (May 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[20] R. Lotfian and C. Busso. 2017. Formulating Emotion Perception as a Probabilistic Model with Application to Categorical Emotion Classification. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*. San Antonio, TX, USA, 415–420. https://doi.org/10.1109/ACII.2017.8273633

[21] R. Lotfian and C. Busso. 2018. Predicting Categorical Emotions by Jointly Learning Primary and Secondary Emotions Through Multitask Learning. In *Interspeech 2018*. Hyderabad, India, 951–955. https://doi.org/10.21437/Interspeech.2018-2464

[22] R. Lotfian and C. Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *IEEE Transactions on Affective Computing* 10, 4 (October-December 2019), 471–483. https://doi.org/10.1109/TAFFC.2017.2736999

[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW 2006)*. Atlanta, GA, USA.

[24] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3, 1 (January-March 2012), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

[25] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan. 2010. The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*. Valletta, Malta.

[26] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan. 2016. The USC CreativeIT Database of Multimodal Dyadic Interactions: From Speech and Full Body Motion Capture to Continuous Emotional Annotations. *Journal of Language Resources and Evaluation* 50, 3 (September 2016), 497–521. https://doi.org/10.1007/s10579-015-9300-0

[27] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions*

[1] https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Face.html

on Affective Computing 10, 1 (January-March 2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[28] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S.S. Narayanan. 2009. Interpreting Ambiguous Emotional Expressions. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*. Amsterdam, The Netherlands, 1–8. https://doi.org/10.1109/ACII.2009.5349500

[29] B. Nojavanasghari, T. Baltrušaitis, C.E. Hughes, and L.-P. Morency. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *ACM International Conference on Multimodal Interaction*. Tokyo, Japan, 137–144. https://doi.org/10.1145/2993148.2993168

[30] O.M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference (BMVC 2015)*. Swansea,UK, 1–12. https://doi.org/10.5244/c.29.41

[31] S. Parthasarathy and C. Busso. 2017. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In *Interspeech 2017*. Stockholm, Sweden, 1103–1107. https://doi.org/10.21437/Interspeech.2017-1494

[32] S. Parthasarathy and C. Busso. 2018. Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes. In *Interspeech 2018*. Hyderabad, India, 3698–3702. https://doi.org/10.21437/Interspeech.2018-1391

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. Las Vegas, NV, USA, 779–788. https://doi.org/10.1109/CVPR.2016.91

[34] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in neural information processing systems (NIPS 2015)*. Montreal, Canada, 91–99.

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. 2013. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*. Shanghai, China, 1–8. https://doi.org/10.1109/FG.2013.6553805

[36] J.A. Russell and L.F. Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology* 76, 5 (May 1999), 805–819. https://doi.org/10.1037/0022-3514.76.5.805

[37] J. A. Russell. 1993. Forced-Choice Response Format in the Study of Facial Expression. *Motivation and Emotion* 17, 1 (March 1993), 41–51. https://doi.org/10.1007/BF00995206

[38] N. Sadoughi and C. Busso. 2020. Speech-Driven Expressive Talking Lips with Conditional Sequential Generative Adversarial Networks. *IEEE Transactions on Affective Computing* To appear (2020). https://doi.org/10.1109/TAFFC.2019.2916031

[39] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Interspeech 2013*. Lyon, France, 148–152.

[40] V. Sethu, E. Mower Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan. 2019. The ambiguous world of emotion representation. *ArXiv e-prints (arXiv:1909.00360)* (May 2019), 1–19. arXiv:cs.HC/1909.00360

[41] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (January-March 2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

[42] K. Sridhar, S. Parthasarathy, and C. Busso. 2018. Role of Regularization in the Prediction of Valence from Speech. In *Interspeech 2018*. Hyderabad, India, 941–945. https://doi.org/10.21437/Interspeech.2018-2508

[43] S. Thomas, M. Suzuki, Y. Huang, G. Kurata, Z. Tuske, G. Saon, B. Kingsbury, M. Picheny, T. Dibert, A. Kaiser-Schatzlein, and B. Samko. 2019. English Broadcast News Speech Recognition by Humans and Machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. Brighton, United Kingdom, 6455–6459. https://doi.org/10.1109/ICASSP.2019.8683211

[44] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision (ECCV 2016)*, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.). Lecture Notes in Computer Science, Vol. 9911. Springer Berlin Heidelberg, Amsterdam, The Netherlands, 499–515. https://doi.org/10.1007/978-3-319-46478-7_31

[45] A. Zadeh, P.P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.-P. Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACM Association for Computational Linguistics (ACL 2004)*, Vol. 1. Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[46] Y. Zhang, W. Chan, and N. Jaitly. 2017. Very deep convolutional networks for end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*. New Orleans, LA, USA, 4845–4849. https://doi.org/10.1109/ICASSP.2017.7953077