The MSP-Podcast Corpus

Carlos Busso, *Fellow, IEEE*, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, *Member, IEEE*, Lucas Goncalves, *Member, IEEE*, Srinivas Parthasarathy, *Member, IEEE*, Abinay Reddy Naini, *Student Member, IEEE*, Seong-Gyun Leem, Luz Martinez-Lucas, *Graduate Student Member, IEEE*, Huang-Cheng Chou, *Member, IEEE*, and Pravin Mote, *Student Member, IEEE*

***Abstract*—The availability of large, high-quality emotional speech databases is essential for advancing *speech emotion recognition* (SER) in real-world scenarios. However, many existing databases face limitations in size, emotional balance, and speaker diversity. This study describes the MSP-Podcast corpus, summarizing our ten-year effort. The corpus consists of over 400 hours of diverse audio samples from various audio-sharing websites, all of which have Common Licenses that permit the distribution of the corpus. We annotate the corpus with rich emotional labels, including primary (single dominant emotion) and secondary (multiple emotions perceived in the audio) emotional categories, as well as emotional attributes for valence, arousal, and dominance. At least five raters annotate these emotional labels. The corpus also has speaker identification for most samples, and human transcriptions of the lexical content of the sentences for the entire corpus. The data collection protocol includes a machine learning-driven pipeline for selecting emotionally diverse recordings, ensuring a balanced and varied representation of emotions across speakers and environments. The resulting database provides a comprehensive, high-quality resource, better suited for advancing SER systems in practical, real-world scenarios.**

***Index Terms*—Affective computing, speech emotional database, speech emotion recognition.**

Received 10 September 2025; revised 24 March 2026; accepted 24 March 2026. This work was supported by the National Science Foundation (NSF) under Grant CNS-1823166, Grant CNS-2016719, and Grant CAREER IIS-1453781. Recommended for acceptance by P. Lopez-Otero. *(Corresponding author: Carlos Busso.)*

Carlos Busso is with Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: busso@cmu.edu).

Reza Lotfian is with Athenahealth, Boston, MA 02135 USA (e-mail: rlotfian@athenahealth.com).

Kusha Sridhar is with Accenture LLP, Mountain View, CA 94041 USA (e-mail: k.sridhara.murthy@accenture.com).

Ali N. Salman is with ARRAY Innovation, Manama, Bahrain (e-mail: ali.salman@array.world).

Wei-Cheng Lin is with the Bosch Center for Artificial Intelligence, Bosch Research, Pittsburgh, PA 15222 USA (e-mail: wei-cheng.lin@us.bosch.com).

Lucas Goncalves was with Amazon, Sunnyvale, CA 94089 USA. He is now with Oracle AI (e-mail: lucas.g.goncalves@oracle.com).

Srinivas Parthasarathy is with Amazon, Sunnyvale, CA 94089 USA (e-mail: parsrini@amazon.com).

Abinay Reddy Naini, Luz Martinez-Lucas, and Pravin Mote are with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: AbinayReddy.Naini@utdallas.edu; luz.martinez-lucas@utdallas.edu; Pravin.Mote@UTDallas.edu).

Seong-Gyun Leem is with Reality Labs, Meta Platforms, Inc., Burlingame, CA 94010 USA (e-mail: sgleem@meta.com).

Huang-Cheng Chou is with the Signal Analysis and Interpretation Laboratory (SAIL), Ming Hsieh Department of Electrical and Computer Engineering,, University of Southern California (USC), Los Angeles, CA 90089 USA (e-mail: huangchengchou@gmail.com).

Digital Object Identifier 10.1109/TAFFC.2026.3678489

I. Introduction

AFFECTIVE computing is a prominent research field focused on understanding, analyzing, recognizing, and synthesizing human emotions. Enriching interfaces with emotional awareness has the potential to enable significant applications across diverse domains, including *human-computer interaction* (HCI), mental health, security and defense, education, and entertainment. Among the various modalities, speech plays a critical role in these interfaces by conveying information beyond the literal meaning of words. However, recognizing emotions from speech in realistic settings poses considerable challenges, largely due to the subtle and complex expressive behaviors inherent in human interactions [1]. To effectively develop and evaluate methods that address naturalistic scenarios, it is crucial to have access to datasets that accurately represent these real-world conditions. A common issue in building *speech emotion recognition* (SER) systems is the limited availability of datasets that provide sufficient data, diversity, and representativeness of naturalistic interactions. This scarcity impedes further advancements in the field of speech affective computing and related research areas.

Over the years, numerous studies have focused on developing diverse methods for collecting emotionally rich databases. These approaches include using actors delivering predefined sentences with specific emotional states [2], [3], [4], [5], employing speakers in semi-structured scenarios designed to evoke natural emotional responses [6], [7], [8], recording colloquial conversation between participants [9], [10], utilizing acted TV shows as source for emotional content [11], [12], [13], and collecting data from audio and video sharing platforms [14], [15], [16], [17], [18]. However, utilizing some of these aforementioned methods comes with issues. Using actors with predefined sentences often results in exaggerated or stereotypical emotional expressions that may not reflect natural human behavior. The scripted nature also limits variability and spontaneity, potentially biasing models trained on such datasets. Semi-structured scenarios aim for more spontaneity, but may still fail to capture authentic emotional experiences. For example, the participants' awareness of being observed can influence their behavior, leading to unnatural responses. Acted TV shows, while providing large amounts of emotional material, face challenges such as exaggerated externalizations of emotions for dramatic effect and a lack of authenticity. Additionally, the context in TV shows may not generalize well to real-world scenarios, and ethical and copyright issues can complicate the use of such data in research.

These limitations highlight the need for the MSP-Podcast corpus, which contains naturalistic, diverse, and well-annotated data to advance the study of emotion in speech. Collecting authentic emotional data in real-world settings without scripts or actors can provide more genuine samples. The diversity in the data is crucial for developing models that generalize across various scenarios and contexts. Moreover, including a variety of annotator opinions ensures that the dataset can more accurately capture the complexity of human emotions.

This paper presents the MSP-Podcast corpus, summarizing our 10-year effort to collect this corpus. Mariooryad et al. [19] presented the initial idea for a scalable data collection protocol that inspired our effort for the MSP-Podcast corpus. Lotfian and Busso [17] formulated a protocol for using machine-learning methods to retrieve emotional recordings that are carefully annotated with emotional labels. The focus of this paper is to describe the resulting database, detailing the changes made to the protocol to enhance the quality of the data. The final release of the MSP-Podcast corpus comprises 409 hours of annotated data collected from more than 3,641 speakers, incorporating diverse audio samples from various sources with diverse emotional content. The continuous growth of multimedia content on the Internet offers an abundant resource for audio data, particularly podcasts that cover a wide array of topics and emotions. Our primary challenge was to select audio segments that provide a balanced representation across the emotional spectrum. We carefully selected and downloaded podcasts featuring natural conversations among various speakers on diverse subjects, including both positive and negative topics, such as personal stories, debates, and cultural discussions. To ensure the database can be shared widely within the research community, we focused on recordings available under Creative Commons licenses with minimal restrictions. The audio was processed to extract clean, single-speaker segments by removing silence, background noise, music, and overlapping speech, utilizing advanced algorithms for voice activity detection, speaker diarization, and noise estimation. We employed enhanced machine learning models trained on larger corpora to identify segments exhibiting specific emotional categories and values for the attributes of valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). This refined approach enables greater control over the emotional content, increases speaker diversity, and preserves the spontaneous nature of the recordings.

This paper presents our methods for curating a more diverse and emotionally rich set of naturalistic speech samples from podcasts available on audio-sharing platforms. We describe the emotional annotation process, which began with crowdsourcing evaluations and continued with a carefully controlled annotation process involving trained students from our institution. At least five raters annotated each speaking turn, providing rich labels for primary (single dominant emotion) and secondary (all emotions perceived in the speech) emotional categories, and emotional attributes for valence, arousal, and dominance. We describe our strategy to enhance the quality of annotations, which includes tracking the performance of annotators on a weekly basis, providing detailed feedback, and implementing a training strategy to improve their annotations if their quality falls below a given threshold. We also describe other annotations included in the corpus, including speaker identification for most of the corpus and human transcriptions, with a focus on the quality control methods we implemented. The contribution of this study is not only the resulting database but also the lessons learned from this multi-year effort, which can guide future data collections.

The remainder of this paper is organized as follows. Section II provides a brief overview of existing emotional databases. Section III outlines the protocol used for data collection, including the selection of podcasts, segmentation into short turns, post-processing and filtering steps, and procedures for emotional annotation. Section IV describes the annotations of the corpus, including emotions, speaker information, and lexical content. Section V provides the partitions of the corpus and a brief recollection of early releases of this corpus. Section VI presents SER baselines for classifying primary emotions and predicting emotional attributes. Section VII highlights new research opportunities opened by key features of this corpus. Finally, Section VIII concludes the paper with a summary and final remarks.

II. Related Work

A. Emotional Databases

Table I presents some emotional databases. Although the research community has access to numerous emotional databases, they come with certain limitations that restrict their effectiveness in tackling ongoing research problems. These limitations include the lack of naturalness in the emotional expressions, unbalanced emotional content, and constraints in size and speaker diversity.

Traditional emotional corpora designed for emotion recognition largely depended on actors who were directed to vocalize sentences with intended emotions. This practice was used to create several well-known emotional databases, such as the Emo-DB [4], RAVDESS [5], TESS [37], CREMA-D [2], and the Chen Bimodal [34] databases. While these datasets have played an essential role in early research efforts, the use of acted emotions presents challenges in truly mirroring the complex and spontaneous nature of genuine human emotions, as discussed by Devillers et al. [38] and Batliner et al. [39]. Some databases have been designed to address this limitation. The DUSHA corpus [20] was constructed using a hybrid data collection methodology, combining elicited speech from non-professional actors with spontaneous speech extracted from podcasts. This approach aims to balance the experimental control inherent in acted performances with the ecological validity of naturalistic recordings. Other databases, such as the USC-IEMOCAP [40], MSP-IMPROV [8], and THAI-SER [25] corpora, aimed to bridge this gap by incorporating more naturally occurring emotional expressions within dyadic interactions, thereby deviating from the more scripted monologues of previous databases. These endeavors made significant strides in producing dialogue that closely mimics the nuances of real-world emotional exchanges. Yet, the usage of professional actors remained a barrier to capturing naturalistic emotional responses.

TABLE I
SELECTED EXAMPLES OF SPEECH EMOTION DATABASES

Corpus	Size	#spk	Avail	Size	# Speakers	Type	Lang.
MSP-PODCAST 2.0 (this paper)	✓	✓	✓	409h	3,641+	Spontaneous	English
Dusha [20]	✓	✓	✓	346h36m	8,308	Acted, Spontaneous	Russian
Crowdsourcing Emotional Speech [21]	✓	✓	✓	187h	2,965	Spontaneous	English
BIIC-Podcast [15]	✓	✗	✓	147h26m	Unknown	Spontaneous	Taiwanese Mandarin
MIKU-EmoBench [22]	✓	✗	✓	131h12m	Unknown	Spontaneous	Multiple
CMU-MOSEAS [23]	✓	✓	✓	68h49m	1,645	Spontaneous	Multiple
CMU-MOSEI [24]	✓	✓	✓	65h53m	1,000	Spontaneous	English
THAI-SER [25]	✗	✓	✓	41h36m	200	Acted	Thai
CEMO [26]	✗	✓	✓	20h	688	Spontaneous	French
IEMOCAP [6]	✗	✗	✓	12h26m	10	Acted	English
MELD [13]	✗	✓	✓	30h45m	407	Acted	English
TUM AVIC [27]	✗	✗	✓	10h23m	21	Spontaneous	English
MSP-IMPROV [8]	✗	✗	✓	9h35m	12	Acted	English
FAU-AIBO [28]	✗	✗	✓	9h12m	51	Spontaneous	German
CHEAVD 2.0 [29]	✗	✓	✓	7h54m	527	Acted	Mandarin
DEMoS [30]	✗	✗	✓	7h40m	68	Induced	Italian
Emozionalmente v1.1 [31]	✗	✓	✓	7h18m	431	Acted	Italian
WHiSER [32]	✗	✗	✓	6h21m	Unknown	Spontaneous	English
SEMAINE [33]	✗	✗	✓	6h30m	20	Induced	English
Chen Bimodal [34]	✗	✓	✗	5h36m	100	Acted	English
CREMA-D [2]	✗	✗	✓	5h16m	91	Acted	English
NNIME [10]	✗	✓	✓	11h	43	Acted	Taiwanese Mandarin
UrduSER [35]	✗	✗	✓	3h2m	10	Acted	Urdu
RECOLA [9]	✗	✗	✓	3h50m	46	Spontaneous	French
CMU-MOSI [36]	✗	✗	✓	2h34m	98	Spontaneous	English
VAM-Audio [12]	✗	✗	✓	48m	47	Spontaneous	German
Emo-DB [4]	✗	✗	✓	3h	10	Acted	German
RAVDESS [5]	✗	✗	✓	1h28m	24	Acted	English

Check mark (✓) indicates the dataset has at least 50 hrs (Size), has at least 100 speakers (#spk), and is publicly available (Avail).

In the pursuit of authenticity, other datasets have relied on spontaneous interactions derived from sources such as colloquial conversations (SEMAINE [33], RECOLA [9], TUM-AVIC [27]), television programs (VAM [12], MELD [13], CHEAVD [41], UrduSER [35]), the Internet (BIIC-Podcast [15], WHiSER [32], CMU-MOSI [36], CMU-MOSEI [24], CMU-MOSEAS [23]), and customer service calls (CEMO [26]). This shift towards spontaneity was critical in capturing genuine emotional displays, but these databases faced the obstacle of skewed emotional representations, constrained by the contexts from which they were sourced. For instance, television programs broadcasting relationship issues might lean towards negative emotions [12], while casual conversations might predominantly exhibit positive emotions [9]. The emotional imbalance also poses a challenge for SER models, which require diverse and evenly distributed emotional examples to learn effectively. For example, Naini et al. [42] demonstrated SER improvements by just undersampling the training set to match the emotional distribution of the target domain.

A prominent trend in emotion corpus development involves leveraging crowdsourcing to acquire data from a large pool of participants using their personal, consumer-grade devices. In this paradigm, exemplified by corpora such as Emozionalmente [31] and the dataset by Smith et al. [21], annotation is also frequently crowdsourced to enhance scalability and cost-effectiveness. A direct consequence of this methodology is significant acoustic variability due to differences in microphone types and recording environments. More recent approaches automate this process; for instance, MIKU-EmoBench [22] is constructed by applying an automated pipeline to extract and label content from large-scale, user-generated video platforms. Although the acquisition is automated, this strategy retains the core benefit of crowdsourcing by capturing a wide spectrum of speech from the varied settings and diverse speaker demographics present in the original online content. While crowdsourcing and automated retrieval have expanded the scale and diversity of emotional databases, these approaches often struggle with annotation consistency, emotional ambiguity, and quality control. As a result, many large-scale corpora exhibit high variability in recording conditions and occasional inaccuracies in emotional labeling. These limitations highlight the need for frameworks that not only scale to large datasets but also maintain annotation reliability and emotion authenticity.

B. Relation to Prior Work

The effort to collect the MSP-Podcast corpus was motivated by retrieval-based strategies explored by Mariooryad et al. [19]. The core idea was to identify emotional segments with machine learning models. We noticed this approach can scale if we design an emotion perceptual evaluation using crowdsourcing [43]. Lotfian and Busso [17] formally introduced the original protocol, describing early results, showing the effectiveness of our strategy in retrieving emotional speech with the intended emotional content (e.g., finding positive speech with high valence values). Since then, we have released early versions of the corpus over the years, from version 1.0 in November 2017 to version 1.12 in June 2024. With this study, we release version 2.0 of the MSP-Podcast corpus, the final release.

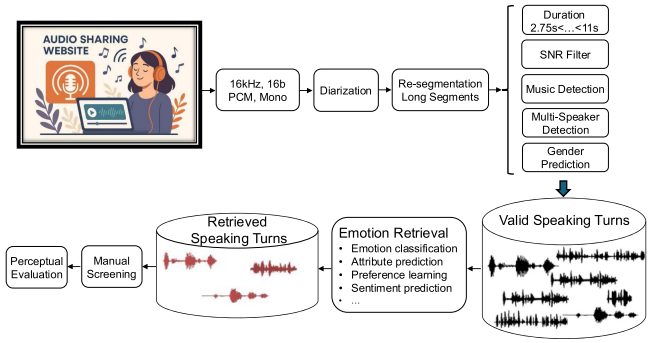


Fig. 1. Protocol for the data collection of the MSP-Podcast corpus. Section III-A presents the selection of podcasts. Section III-B discusses the data segmentation process. Section III-C describes the selection of speaking turns. Section III-D explains the perceptual evaluation.

We have prepared this paper minimizing the overlap with the protocol described in Lotfian and Busso [17]. Instead, we have focused on describing the final release of the corpus and the modifications that we implemented to improve the quality of the annotations. The resulting corpus consists of 409 hours of speech, offering large-scale emotional and speaker diversity, while providing natural, conversational speech collected over a decade of continuous acquisition. The enhancements make the final version of the MSP-Podcast corpus a far more comprehensive and robust resource for SER research, positioning it as a superior dataset for real-world emotion recognition tasks.

III. Protocol for the MSP-Podcast Corpus

The protocol for data collection in the MSP-Podcast corpus is explained in Lotfian and Busso [17]. This section summarizes the protocol, with a focus on the changes implemented to enhance the quality of the data. Fig. 1 shows a diagram of the data collection protocol.

A. Selection of Podcasts

We source our speech data from online sources that host publicly available audio. Our goal is to have an emotionally diverse and gender-balanced corpus. We also want speaker diversity. Therefore, we collect podcasts, talk shows, and lectures about sports, popular media, politics, personal struggles, societal issues, public health, crime, technology, and daily life. We use five criteria when searching for podcasts: (1) clean audio, no background music or speech, and not too much noise, (2) English speech, (3) emotional speech, prioritizing queries likely to convey target emotions, (4) diverse speaker demographic, and (5) appropriate license. The podcasts were identified primarily through manual searches, where researchers selected search terms that could elicit emotional topics and chose podcasts that met the aforementioned criteria. 4,742 (78.9%) podcasts in the corpus were found in this way (manually). Eventually, we wrote a script to automatically find podcasts. A researcher can input a list of search terms, and the script will find podcasts that meet the criteria and download them. The script first downloads the metadata of some of the search results, then filters them by language (if available) and license. The script then downloads the audio of the chosen podcasts. We implement automatic steps to filter podcasts based on a music detector [44] and a noise detector [45]. Finally, a researcher briefly listens to each podcast selected by the script, verifying whether the chosen recordings meet the target criteria. 1,265 (21.1%) podcasts in the corpus were found this way (automatically). In total, the MSP-Podcast corpus includes recordings from 6,007 unique podcasts.

TABLE II
Percentage of Podcasts With a Specific License in the Corpus

License	Perc. of Podcasts	# of Podcasts	# of Turns
Public Domain	2.88%	173	5,872
CC-BY	90.86%	5,458	242,699
CC-BY-SA	5.59%	336	18,910
Unknown	0.67%	40	424
Total	–	6,007	267,905

We select podcasts that are shared with licenses that allow us to distribute and modify them freely. We mainly focus on podcasts with Public Domain licenses or Creative Commons licenses with minimal restrictions (https://creativecommons.org/). Table II shows the number and percentage of podcasts in the corpus that were selected with specific licenses. Our practice was to save a screenshot of the website to document the license of the podcasts. There are 40 podcasts whose license information was not saved when initially collected, despite being selected with the target Creative Commons license. When we searched for the license information at a later date, the podcasts had been removed from the online website. Therefore, we do not have precise license information for these 40 podcasts in the corpus, which we denote as having an "Unknown" license in Table II.

After choosing and downloading the podcasts, we convert all of them to the same audio format as described in Lotfian and Busso [17]. This conversion is done due to the inconsistent nature of the source audio (i..e., data in a variety of audio and video formats). We convert the podcasts to wave audio format with a mono channel, a sample rate of 16 kHz, and 16-bit *pulse code modulation* (PCM) with the Librosa toolbox [46].

B. Data Segmentation

The next step in the pipeline is to split the podcasts into *speaking turns*. We define a speaking turn as a segment spoken by a speaker, which may comprise one or more sentences or phrases. We started the project by manually conducting this step. Researchers manually split the first 279 (4.64%) podcasts. However, this process was very time-consuming considering the final size of the corpus. We decided to use an automated tool to split the remaining podcasts. Since podcasts can contain music or noisy segments and often feature multiple speakers, we need a tool that can segment the audio into speaking turns while also keeping speakers and noise separate. The diarization of the podcasts into sentences was mostly done using the Microsoft Azure Video Indexer[1]. 3,667 (61.0%) podcasts in the corpus

[1]https://azure.microsoft.com/en-us/products/ai-video-indexer

were segmented using this tool. We eventually switched to using Pyannote [47], [48], which we used to segment the remainder podcasts. We also obtain automatic transcriptions using either Microsoft Azure Video Indexer or Whisper [49]. For Whisper, we use the pre-trained large-v2 model in the HuggingFace library [50]. Our data collection process lasted for 10 years. Therefore, we used different versions of these toolkits.

C. Automatic Filtering & Selection of Speaking Turns

After the podcasts are split into speaking turns, the next step involves employing multiple filters designed to aid our system in selecting only the highest-quality recordings to proceed with our annotation process (single speaker, no music, clean recording, with target duration, and target emotion). During this stage, we conduct several key operations: speaking turn duration estimation, resegmentation of long segments using word alignment, music detection, noise estimation, multiple speaker detection, gender prediction, automatic emotion retrieval, and final inspection by a trained human worker. This section explains each of these filters used to select the speaking turns to be included in the corpus.

The initial step involves verifying the timings and word content of the speech segments. Our goal is to have speaking turns with a duration between 2.75 and 11 seconds. The lower threshold is justified by the need to have enough context for a rater to reliably infer an emotional label during the perceptual evaluation. Cowie & McKeown [51] reported that emotion judgments become more reliable when speech segments span on the order of a few seconds, and recommended temporal windows in the range of approximately 1-3 seconds as a practical lower bound for reliable emotion coding. Our lower threshold of 2.75 seconds is consistent with these recommendations. The higher threshold was imposed because emotions can vary during a speaking turn, so having a single label may not accurately reflect the emotional content of the speaking turn. Audios shorter than 2.75 seconds are automatically excluded, while those exceeding 11 seconds undergo a re-segmentation process. This step involves utilizing the automatic transcriptions from Section III-B and executing an automatic word-level alignment with the audio segments using a Python module [52]. This module facilitates interaction with Praat's TextGrid [53] to align transcripts with audio. We then evaluate the alignments to identify pauses in speech lasting at least 0.3 seconds, at which point we crop the audio to create smaller segments within the target range of 2.75 to 11.0 seconds. The 0.3-second threshold is applied to identify pauses indicative of a potential sentence completion by the speaker. Following this resegmentation, we aggregate all audio segments within the 2.75 to 11.0-second duration and automatically review their transcriptions to exclude any speaking turns with fewer than five words, thus eliminating segments lacking substantial spoken content. Fig. 2 shows the distribution of the durations of the selected speaking turns included in the corpus.

The audio segments are then evaluated with music detection and noise estimation algorithms. In particular, we employ a pre-trained audio tagging model [44] to identify segments where music is present. Segments where music constitutes more than

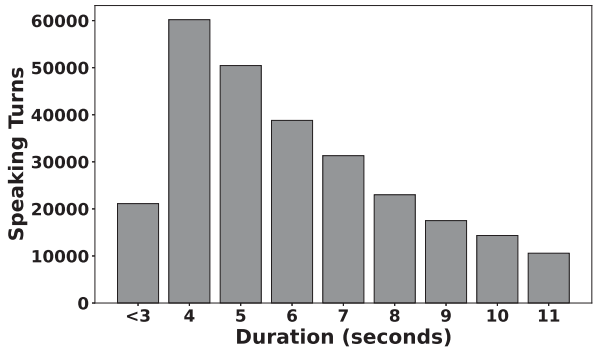


Fig. 2. Histogram showing the distribution of speaking turn durations in the MSP-Podcast corpus. The x-axis shows duration in seconds.

50% of the duration are filtered out. Following this step, we estimate the *signal-to-noise ratio* (SNR) using the WADA-SNR algorithm [54], based on *waveform amplitude distribution analysis* (WADA). Audio segments with an SNR below 15 dB are subsequently rejected. The remaining audio segments are further processed using the pyannote.audio speaker diarization toolkit [55], [56] to ensure that each audio segment contains speech from only a single speaker. The use of this toolkit enables the automatic exclusion of samples containing multiple speakers.

All audio segments that meet the aforementioned filters are then subjected to a series of predictive models to automatically identify speaker and recording characteristics. One of the traits is gender. Gender prediction is achieved through a pre-trained speech *long short-term memory* (LSTM)-based model, capable of distinguishing between "Female" and "Male" [57]. This process is done to gender balance the selected speaking turns. As discussed in Section IV-D (Table IV), the segments with speaker ID include 1,598 female speakers (159.58 hrs) and 2,043 male speakers (175.36 hrs).

We have millions of valid speaking turns obtained from the 6,007 podcasts that passed our criteria. Most of these segments are expected to be emotionally neutral. As explained in Lotfian and Busso [17], we can prioritize the annotation of emotional recordings by selecting speaking turns predicted to have target emotions. Therefore, we implement an automatic emotional retrieval step. We mitigate the potential problem of biasing the selected speaking turns towards specific SER systems by employing multiple models and formulations. The SER models encompass multiple versions of emotion classification [58], [59], emotion attribute prediction [60], [61], ranking-based preference learning prediction [62], and textual sentiment analysis [63]. We consider opensource implementations [58], [60], [61], [63], [64], [65], [66] and internally trained variants. The final retrieval system relies on over 48 criteria dictated by emotion models. It employs various pre-trained models developed from extensive emotional corpora, including CREMA-D [2], MSP-IMPROV [8], IEMOCAP [67], earlier versions of MSP-Podcast [17], and Twitter sentiment data [68]. These models also utilize a comprehensive range of inputs, including *low-level descriptors* (LLDs), *high-level descriptors* (HLDs), raw audio

for foundational *self-supervised learning* (SSL) models, and textual data derived from audio transcriptions. The models were updated and retrained multiple times during the project. This emotion retrieval step is crucial for assembling an emotionally diverse and naturalistic corpus that spans a broad spectrum of emotional states.

After running all these models on the audios, we compile a set of master lists with predictions retrieved for each task using each model, and rank these predictions from high to low accordingly for each model. We ensure that the lists are set up to dynamically change as new data is processed and entered into our master lists. Such a ranking system is instrumental in our methodology, helping us select high-emotional content and minority emotional states for annotation. Additionally, we created separate master lists for each gender. We fine-tune our selection using dynamic thresholds to maintain a balanced representation of genders and emotional states, adapting our approach as new data enters the annotation pipeline. This strategy ensures the creation of a more inclusive and precise annotated dataset, effectively minimizing bias. Updates to our master lists ensure that each sample is selected only once, avoiding redundancy in future selections. Moreover, we document the rationale behind each selection (e.g., a sample *A* is chosen due to its high emotional rating by model *B*), facilitating an evaluation of our models' effectiveness in identifying emotionally relevant samples for subsequent selection rounds and threshold adjustments or model removals. We weekly monitored the performance of these SER models during the project.

Selected samples are then forwarded to a trained evaluator who conducts a thorough review, listening to each audio to confirm its suitability for annotation. This final check aims to identify any samples that, despite passing through our filters, might still present issues such as background music, low signal-to-noise ratios, unintelligible speech, foreign language usage, extremely brief sentences, profanity, multiple speakers, or excessive background noise. The evaluator's task is to identify and exclude samples based on these criteria, compiling a final list to be used for annotation. Notice that the evaluator listens only to the selected samples, instead of the millions of speaking turns in the entire pool considered for the corpus. Approximately, we extract around 3 million speaker segments from the 6,007 podcasts. Out of them, around 1.7 million are valid speaking turns. We only annotated 267,905 out of them, using our retrieval-based strategy.

D. Perceptual Evaluation

The last step in the protocol is to annotate the selected speaking turns. We annotate emotional categories (e.g., anger, happiness, etc.) and emotional attributes (valence, arousal, and dominance). Sections IV-A and IV-B describe the instrument used to annotate the corpus. The original protocol employed a slightly modified crowdsourcing strategy introduced in Burmania et al. [43]. The approach tracks the quality of annotations provided by a worker in real-time during a session, stopping the session if the quality drops below a given threshold. We can measure *quality* by including reference sentences that we have already annotated so that we can estimate inter-evaluator agreements. Lotfian and Busso [17] introduced specific changes to the original protocol, aiming to increase the frequency of checkpoints and incorporate primary emotional annotations and attribute-based annotations into the quality estimation. We followed this approach for the first part of the project.

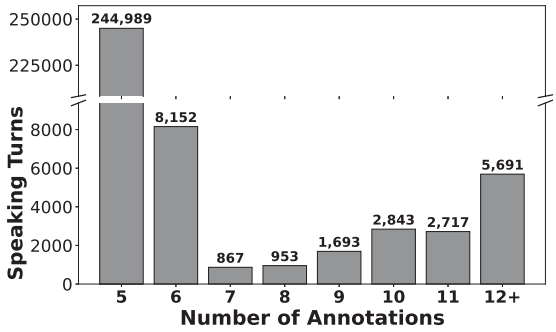


Fig. 3. Histogram showing the number of files in the MSP-Podcast 2.0 corpus by the number of valid annotations. Each file has at least five annotations.

Around September 2021, we noticed important issues with our crowdsourcing platform. We noticed that *human intelligent tasks* (HITs) were immediately taken when we uploaded them, suggesting the presence of bots. Several HITs returned with random annotations (e.g., all the sentences in the batch were labeled as "happy"). Our first step was to suspend every worker found to be showing this behavior. Next, we audited and hardened the perceptual evaluation code, adding safeguards to thwart automated bot submissions and improve overall robustness. While refining our code, we developed an alternative approach to prevent delays in the annotations. We decided to hire student workers from the *University of Texas at Dallas* (UT Dallas) to annotate the corpus. Because emotionrecognition skill varies across individuals, we created and administered a screening test to ensure we could retain only highperforming candidates. The resulting student annotations proved consistently higher in quality than those obtained through traditional crowdsourcing. This new process enabled us to provide regular feedback to our student workers, which was not possible with crowdsourcing workers. As a result, we decided to discontinue our crowdsourcing effort and transition entirely to perceptual evaluations conducted by our student workers. Regularly, we had between 14 and 20 student workers annotating the corpus. We developed a website that connected to the server used for the perceptual evaluation, displaying the number of annotations provided by each student worker in real-time, thereby providing a powerful tool to track our progress. It was easy to identify student workers who were not actively involved in the evaluation.

We collect five or more annotations from different workers for the crowdsourcing evaluation and the perceptual evaluation conducted by our student workers. Some of the speaking turns have more than five evaluations, since they were used as reference sentences in our crowdsourcing protocol. Fig. 3 shows the distribution of the number of annotations per sentence in the corpus. By providing multiple annotations per speaking turn,

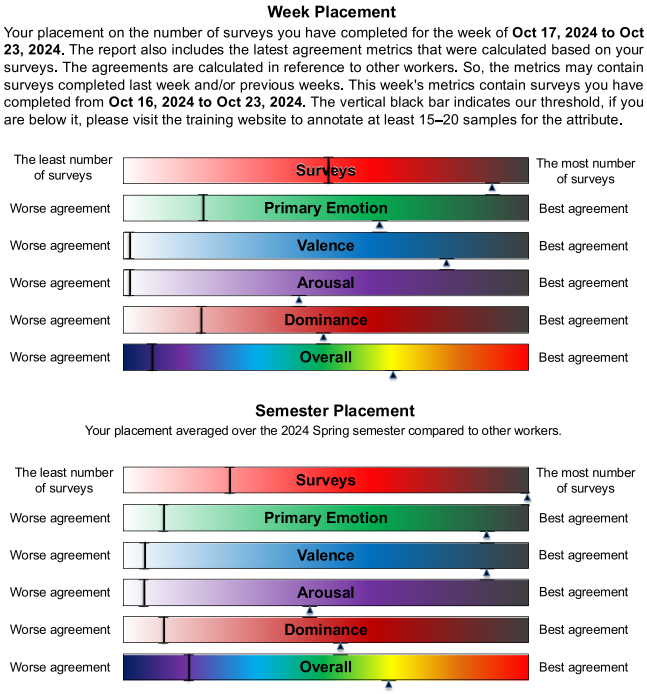
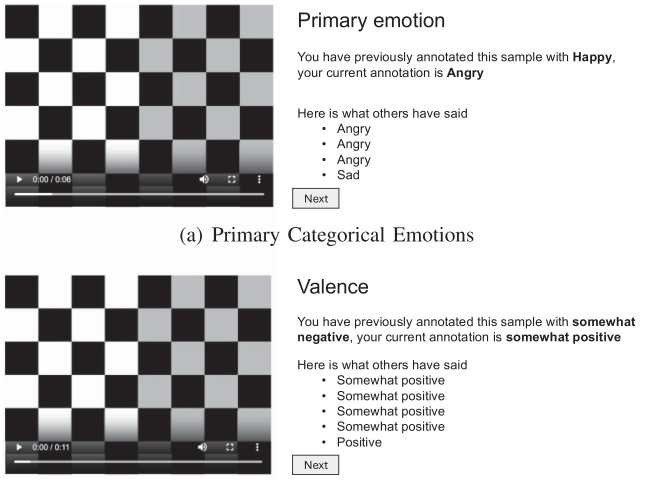


Fig. 4. An example of the weekly email report sent to student workers. The email shows the number of surveys completed, their inter-agreement levels with fellow student workers, and the performance threshold (represented by the thin vertical black line) for both weekly and semester-based placements.



(a) Primary Categorical Emotions

(b) Valence

Fig. 5. Example of training interface for primary emotions and valence. Not shown on the image are the instructions that explain the target emotional descriptor. These screens are shown after the student worker re-annotated the carefully selected speaking turns, showing the original annotation by the target worker and the labels provided by the other student workers.

we enable the exploration of multiple research problems related to utilizing the subjectivity of human emotional perception, such as curriculum learning training strategies [69], exploring co-occurrence of emotion to improve the cost function [70], training with soft labels [71], [72], [73], [74], [75], implementing oversampling strategies for minority classes [76], and finding trends across annotations [77], [78].

With our student workers, we did not implement the crowdsourcing strategy to track the quality in real-time. Instead, we focused on providing weekly feedback. A research assistant trained the student workers before they began annotating data, describing emotional descriptors, particularly the concepts of valence, arousal, and dominance. The student worker completed the first session with the research assistant, who answered any questions raised during the perceptual evaluation. In addition, we wanted to provide frequent feedback to the student workers, so they were aware if we were satisfied with their annotations. We implemented a weekly report that provides their relative ranking with respect to other student workers. Fig. 4 shows an example of the document shared with our student workers. The report presents weekly-based performance (top part of the report) and semester-based reports (bottom part of the report). Instead of providing the actual values of the metrics used to estimate inter-evaluator agreements, we provide a relative ranking comparing the worker with the rest of the workers. For each indicator, we denote the performance with an arrow placed between two extremes. The closer to the right extreme, the better (see Fig. 4). The bars also include a black vertical line that indicates the lower threshold we tolerate. This threshold is set to Fleiss $\kappa = 0.25$ for primary emotions and Krippendorff's $\alpha = 0.25$ for emotional attributes. Given that some emotional attributes are inherently easier to annotate than others, most annotators rank above the threshold for dimensions such as valence, whereas dominance is more challenging, and many workers fall closer to or below the threshold. The first indicator includes the number of annotations completed by the student workers. Then, the report includes the agreement for primary emotions and attribute-based annotations (arousal, valence, and dominance). It also includes the overall score, which is the average of all the emotional descriptors. In the example in Fig. 4, the student worker was very good at annotating primary emotions and valence (both for the current week and the entire semester). However, the annotations for arousal and dominance were average. In all cases, the quality of the worker was above our minimum threshold. The reports were automatically generated, so this process did not require much continuous effort from our team.

We also implemented a targeted training to re-train our student workers with lower inter-evaluator agreements. We created a training website that focuses on a single emotional descriptor (primary emotions, valence, arousal, or dominance). Therefore, the student workers only work on the emotional descriptor that they are struggling with. For example, if a student worker has low inter-evaluator agreement on dominance, the application only includes samples to improve this emotional attribute. We automatically identify speaking turns where the annotations from the target student workers differ from consistent annotations obtained from other student workers. The application asks the student workers to re-annotate these carefully selected samples. Then, it lists their original annotations and the annotations made by the other student workers. These annotations are only revealed after the student worker re-annotates the speaking turn. Fig. 5 shows an example for primary emotions (Fig. 5(a)) and for valence (Fig. 5(b)). Not shown on the figures are the precise

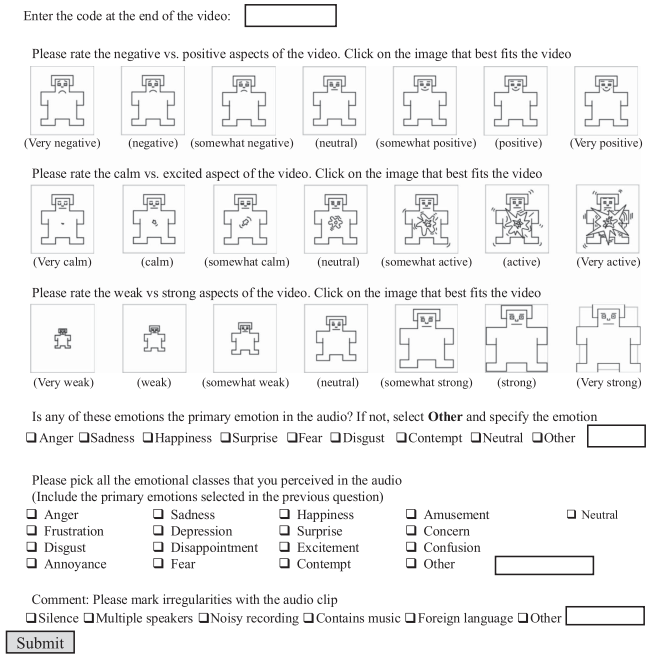
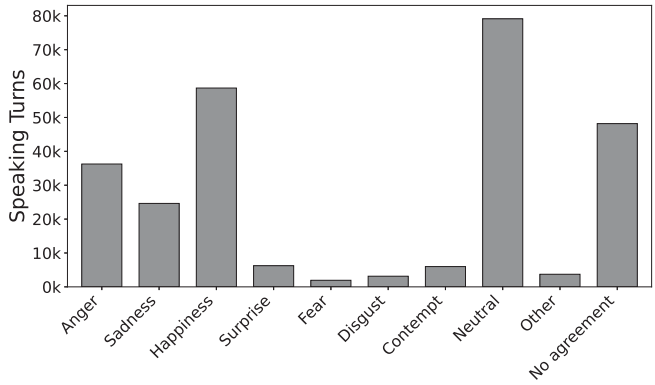
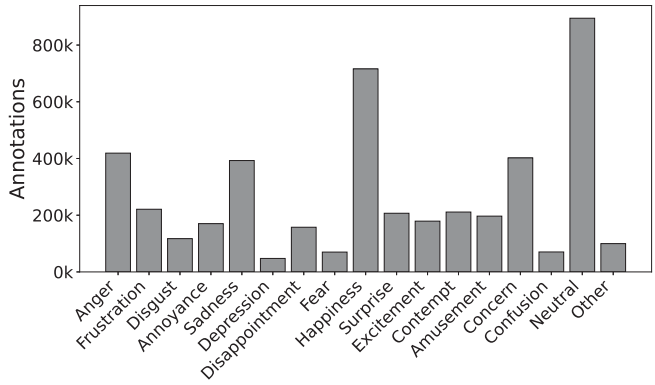


Fig. 6. shows the survey for annotating the MSP-Podcast audios.



(a) Primary Emotions (consensus labels)



(b) Secondary Emotions (individual annotations)

Fig. 7. Histogram of the emotional classes selected by the workers for (a) primary and (b) secondary emotions. For primary emotions, we present the consensus labels, showing the histogram of the consensus emotions assigned to the speaking turns (plurality rule). For secondary emotions, we present a histogram of the secondary emotions selected in the individual evaluations.

instructions given to the student workers to understand the corresponding emotional descriptors. This training was mandatory for student workers with quality below our minimum thresholds, and optional for all others who may want to practice to solidify their understanding of the emotional descriptors used in this corpus.

A later addition to the perceptual evaluation website was an optional field where a student worker could indicate that a speaking turn still had issues, despite our efforts to filter out overlapped speech, silence, noisy recordings, foreign language, or speech with background music (see bottom part of the questionnaire in Fig. 6). When a file was flagged, it was immediately separated from the perceptual evaluation until we manually checked if the speaking turn should be removed entirely from the database. This step was very important to avoid annotating data that we would later discard. The effectiveness of this annotation protocol is reflected in the inter-annotator agreement results reported in Section IV-C.

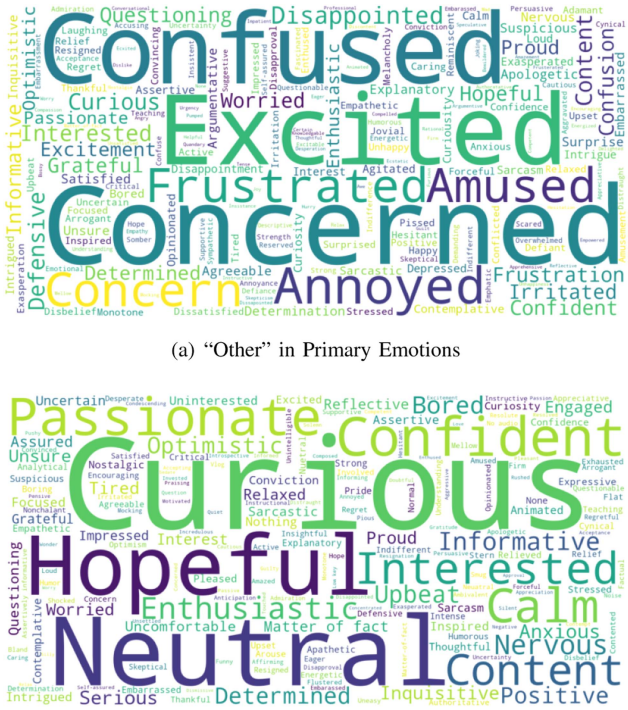
IV. Annotation of the Corpus

A key feature of the corpus is the annotations of the speaking turns. This section describes the annotations for emotions, speaker identification, human transcription, and phonetic alignment. For emotions, we utilize both categorical and dimensional attributes to describe emotions adequately.

A. Annotation of Categorical Emotions

The MSP-Podcast corpus offers a rich set of emotion content from natural conversational speech. Fig. 6 shows the questionnaire used for the perceptual evaluation for the evaluations using crowdsourcing and student workers. The categorical annotation (bottom part in Fig. 6) was inspired by the work of Devillers et al. [38], which includes dominant (Major) and secondary (Minor) labels to capture mixtures of emotions. The primary emotions in the perceptual evaluation include anger, sadness, happiness, surprise, fear, disgust, contempt, and neutral speech. The workers can also select "other" and add their label to add flexibility and avoid the forced-choice response bias discussed by Russell [79]. The workers select only one primary emotion.

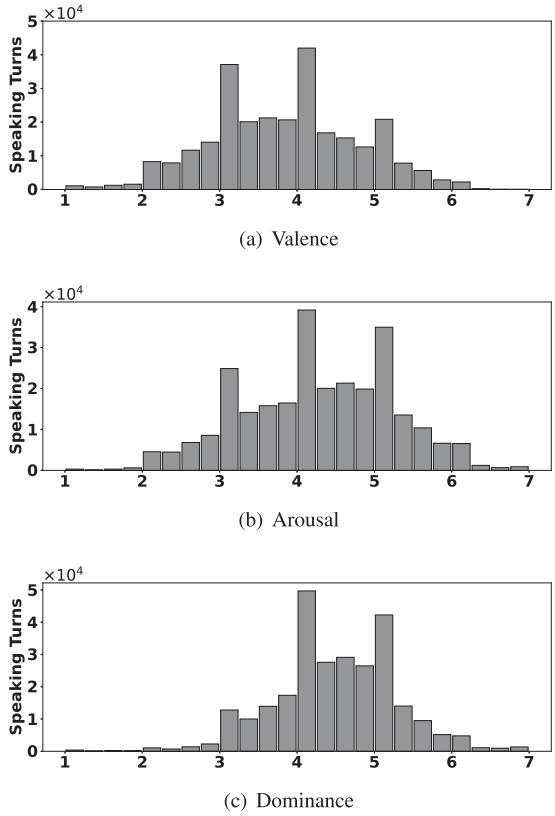
Fig. 7(a) shows the number of speaking turns assigned to each primary emotion category using the plurality rule. We include the class "no agreement" for speaking turns that do not reach agreement under the plurality rule. The histogram reflects the frequency at which emotions appear in natural conversation, with many samples for classes such as happiness, anger, sadness, and neutral speech, and few samples for surprise, fear, disgust, and contempt. Neutral speech is the most dominant class in regular conversation. However, we only have 28% of the speaking turns labeled as neutral, demonstrating the effectiveness of our retrieval-based strategy (Section III-C). Fig. 8(a) shows the word cloud of the labels provided when workers selected "other" as the primary emotions. The figure identifies the emotions "confused,"



(a) "Other" in Primary Emotions

(b) "Other" in Secondary Emotions

Fig. 8. Word cloud representing the manually typed emotions by the annotators when selecting the option "other." (a) Other in primary emotion annotations, and (b) other in secondary emotion annotations.



(a) Valence

(b) Arousal

(c) Dominance

Fig. 9. Histogram distributions of valence, arousal, and dominance attributes in the MSP-Podcast corpus.

"excited," and "concerned" as the most common terms. These emotions are potential candidates for inclusion in the primary emotions for future evaluations.

The secondary emotions extend the list of eight primary emotions by adding frustration, annoyance, depression, disappointment, excitement, amusement, concern, and confusion (16 emotions). We also include the "other" option, allowing them to add their own labels. The workers are asked to select all the secondary emotions that they perceived in the speaking turn. We explicitly requested that the primary class be included as one of the secondary emotions, but the workers did not always follow this instruction. Secondary emotions can play a crucial role in understanding the complex blend of emotions expressed in the speaking turns. Fig. 7(b) shows the histogram of secondary labels selected in the individual annotations. We did not aim to obtain consensus labels like the case with primary emotions. For consistency, we added the primary emotion to the secondary emotion list when the worker did not include it. Neutral, happiness, and anger are the most commonly selected classes. If we do not include the primary emotions, the most popular selections were concern, amusement, and frustration. The classes confusion and depression were the least frequent selections. Fig. 8(b) shows the word cloud with the emotional labels provided by the workers when they selected the option "other." The classes "curious," "hopeful," "neutral," are the most frequent labels, followed by "passionate," "confident," "interested," "calm and "content." The word cloud figures highlight the nuanced and co-occurring nature of emotions that need richer expressive descriptors to represent affective states.

B. Annotation of Emotional Attributes

Emotional attributes are an alternative, powerful strategy to characterize emotions. We include the emotional attributes of valence (negative to positive), arousal (calm to active), and dominance (weak to strong). The top part of Fig. 6 shows the questionnaire for these attributes. We rely on *self-assessment manikin* (SAM) [80] to visually capture the essence of each emotional attribute. We use a Likert scale from 1 to 7, with 1 indicating the lower extreme (e.g., very negative, very calm, or very weak) and 7 indicating the higher extreme (e.g., very positive, very active, or very strong). The consensus label for an attribute is the average score assigned across workers for each speaking turn.

Fig. 9 illustrates the emotional attribute histograms of the speaking turns. Each distribution resembles a unimodal Gaussian distribution. For valence, the center of the distribution is around 4, which corresponds to the neutral range in this emotional dimension. For arousal and dominance, the distributions are slightly shifted to the right, indicating more active and dominant speech recordings.

Fig. 10 displays each speaking turn in the arousal-valence space, with colors indicating the consensus primary emotion assigned to them. The name of each emotional class is

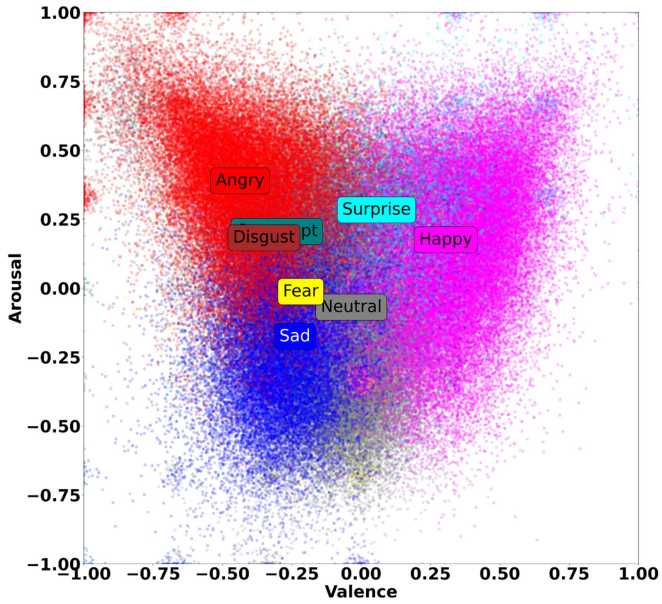


Fig. 10. Illustration of the emotional distribution of the MSP-Podcast corpus in the arousal and valence space, where each point is a speaking turn. The color of the points corresponds to the consensus primary class to which they are assigned. Each emotional class label is placed at the average arousal and valence values associated with that emotion. The class behind "disgust" is "contempt".

positioned at the mean arousal and valence coordinates for that emotion. The figure shows that we have speaking turns with expressive content covering most of the arousal-valence space. The emotional classes are located in the expected quadrant of the arousal-valence space. The figure also reinforces the importance of having categorical and attribute-based annotations. We observe important intra-class variability for primary emotions, indicating that speaking turns assigned to the same class can exhibit a wide range of emotional variability (e.g., cold anger versus hot anger). By having both emotional descriptors, we can effectively capture the emotional content of the speaking turns, opening research directions that are not possible if only one of these descriptors is provided.

C. Inter-Evaluator Agreement

Having quality emotional annotations has been a key goal of our effort. Given the struggles we experienced with crowdsourcing evaluations, we decided to estimate the inter-evaluator agreement for each worker, especially those recruited in our crowdsourcing perceptual evaluation. Based on the agreements, we removed 430 crowdsourcing workers and their 44,968 annotations. These speaking turns were reannotated by our student workers. After these corrections, we have 1,446,270 emotional annotations from 13,280 workers. Out of them, we have 13,205 crowdsourcing workers who completed 494,340 annotations (34.18% of the annotations), and 75 student workers who completed 951,930 annotations (65.82% of the annotations). The release of the corpus includes the age and gender of the annotators. The inter-evaluator agreement significantly increased after re-annotating labels provided by unreliable crowdsourcing workers. The weekly feedback and the training procedure also helped improve the reliability of the labels.

TABLE III
INTER-EVALUATOR AGREEMENT IN THE MSP-PODCAST CORPUS. WE ESTIMATE AGREEMENT FOR PRIMARY EMOTIONS USING FLEISS κ, AND FOR EMOTIONAL ATTRIBUTES USING KRIPPENDORFF'S α.

Descriptor	All	Train	Dev.	Test1	Test2	Test3
Primary [κ]	0.411	0.391	0.410	0.412	0.294	0.510
Valence [α]	0.508	0.461	0.598	0.573	0.228	0.593
Arousal [α]	0.441	0.412	0.515	0.471	0.205	0.610
Dominance [α]	0.386	0.358	0.498	0.378	0.212	0.584

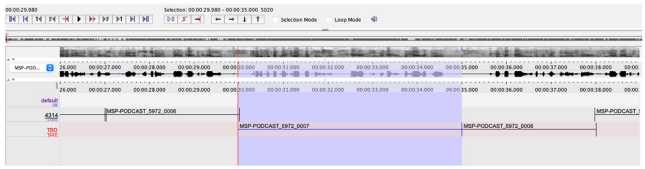


Fig. 11. Annotation process for speaking information using Elan. An audio track contains a previously annotated tier related to speaker 4314, providing contextual information for new annotations. A tier named 'TBD' contains the speaking turns, without speaker information, to be annotated.

Table III presents the inter-evaluator agreement for the entire database and individual partitions (as described in Section V-A). For primary emotions, the Fleiss *kappa* statistic is 0.411 for the entire data. This agreement is high with respect to annotations from previous datasets [6], [8], [81], considering the naturalness of the recordings and the inclusion of eight classes. For emotional attributes, the value for Krippendorff's α for valence is better than the value for arousal. Dominance is the dimension with lower agreement, although its score is above $\alpha > 0.38$. Across the entire corpus, student workers have higher agreement, achieving a kappa of 0.438 for primary emotions, and an alpha of 0.536 for valence, 0.469 for arousal, and 0.407 for dominance. In comparison, crowdsourcing workers obtain a kappa of 0.362 for primary emotions, and an alpha of 0.472 for valence, 0.410 for arousal, and 0.351 for dominance.

D. Speaker Information

It is essential to ensure that data splits for train, validation, and test are speaker-independent for effective SER performance that replicates the expected results on unseen data. This step requires speaker information. Knowing the identity of the speakers is also helpful to explore the role of emotions in other speech tasks such as speaker verification and identification [82], [83], [84], [85], [86] and speech synthesis [87], [88]. Therefore, we manually annotate the speaker information of most of the corpus.

As an initial step in the manual annotation process, we identify all speakers participating in a podcast session using the available information on the source webpage. Then, we listen to each speaking turn selected from that podcast and assign it to its respective speaker. Fig. 11 shows the Elan interface used for this annotation. For each audio track, we create tiers for existing speaker annotations and a new tier

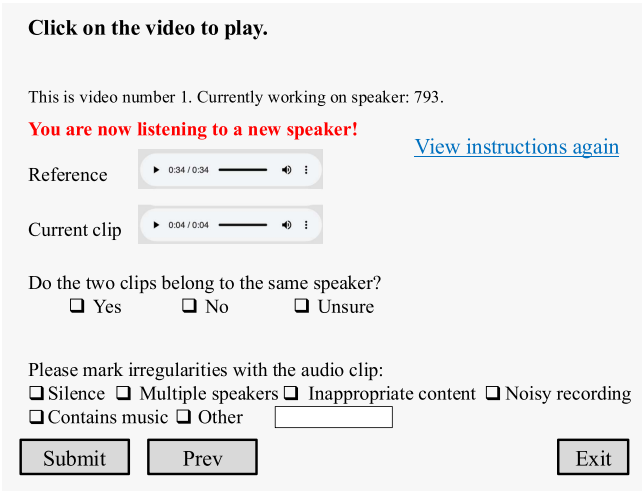


Fig. 12. Interface of the verification website to correct the speaker information. The annotator listens to both the reference audio and a clip that is supposed to belong to the same speaker (speaker 793 in the example). Sequentially listening all speaking turns associated with a given speaker facilitates identifying potential errors in speaker annotations.

for speaking turns without speaking information that we aim to annotate. As illustrated in Fig. 11, an audio track contains two annotation tiers named '4314' and 'TBD', which indicate the previously annotated speaking turns associated with speaker 4314 and the one to be annotated. We then listen to the audio around the segments, assigning speaker information to each. To maintain anonymity, each speaker is assigned a unique identification number. Some speaking turns are very hard to assign to a speaker in the conversation, even after listening to the context from nearby segments. The instruction was to mark these speakers as "unknown," prioritizing precision in the annotations.

We conducted a manual speaker verification process to correct potential mistakes made in the speaker annotations. During this process, all speaking turns associated with an individual speaker are reviewed sequentially using the user interface shown in Fig. 12. For each individual speaker, a 30-second reference audio is created by concatenating manually selected, error-free audio segments. Each speaking turn is then evaluated against this reference audio and marked to indicate whether the current clip belongs to the reference speaker. The annotators can directly compare the voice of the reference speaker with the voice of each speaking turn associated with that speaker. This method facilitates filtering outliers and inconsistencies in speaker annotations. The speaking turns flagged with wrong speaker information by this verification step are manually re-annotated to refine the speaker identities. In total, we have 3,641 unique speakers, where 2,043 are females and 1,598 are males. Table IV provides the number of speakers for the entire corpus and for the partitions described in Section V-A. If it is too hard to identify the speaker, we preferred to omit adding a potentially erroneous speaker label. These segments are labeled as *unknown*. Since we do not know the number of distinct speakers for these sentences, Table IV has question marks in the corresponding subsets.

TABLE IV
SPEAKER AND GENDER INFORMATION FOR THE MSP-PODCAST CORPUS. THERE IS AN OVERLAP IN THE SPEAKERS INCLUDED IN THE TEST SETS

	Train	Dev.	Test1	Test2	Test3	All
Female	1,013	298	184	53	171	1,598
Male	1,207	406	281	59	257	2,043
Unknown	?	0	0	?	0	?
All	2,220	704	465	112	428	3,641

TABLE V
TYPE OF NON-VERBAL INDICATOR

Name	Count	Description
[inaudible]	7,813	Unclear or unintelligible sound
[crosstalk]	1,970	Short overlapping speech in conversation
(affirmative)	380	A sound indicating agreement or acknowledgment (e.g., mm-hmm, uh-huh)
(negative)	12	A sound indicating disagreement or negation (e.g., uh-uh, mmm-mmm)
(laughing)	78	A general laughing sound, range from soft to loud laughter
(beep)	49	A beep sound, often indicating a censored word or alert tone
(singing)	24	Singing voice, such as humming or melodic singing
(breathing)	1	A breathing sound, such as sighs or heavy breathing
(cheering)	2	A cheering sound from crowds

E. Transcription

Linguistic content can provide rich information for predicting an emotion. Including text, for example, was key in recent emotion recognition challenges [64], [89]. Therefore, we provide transcription for the collected speech samples. We ask human annotators to transcribe the speaking turns in the corpus. For this purpose, we provided the collected audio files to *REV.com*, which generated transcripts. Transcribers provide several indicators to describe non-verbal sounds that do not include spoken words, such as laughter or affirmative sounds. We remove indicators irrelevant to spoken information, such as *(music)* or *(sound)*. For consistency, we also cluster indicators that denote similar sounds, leaving eight non-verbal indicators in our transcript shown in Table V.

We evaluate the quality of the annotated transcript by comparing the prediction result of robust *automatic speech recognition* (ASR) systems with the collected transcript. We use OpenAI WhisperX [49] and NVIDIA NeMo Canary [90] ASR systems for this process. We downloaded the following pre-trained checkpoints: *whisper-medium.en* for OpenAI WhisperX and *canary-1B* for NVIDIA NeMo Canary. These ASR systems were at the top of the rank in the Open ASR Leaderboard [91] (observed on Oct/23/2024). With these checkpoints, we got the ASR prediction for each speaking turn. We modified the configuration of the ASR model to make it only predicts alphabet characters without having any digits or special characters. We then compute the *word error rate* (WER) between the prediction and the annotated transcript, resulting in two WERs for each of the annotated speaking turns. We ignore non-verbal indicators while computing the WER. We re-annotate transcripts for the speaking turns when both WERs are above 70%.

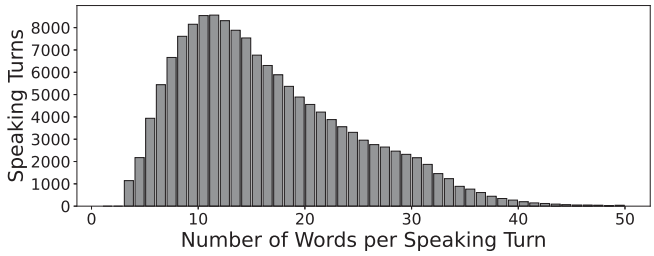


Fig. 13. Histogram of number of words in the speaking turns.

The corpus contains 4.3 million tokens and 50,677 unique words, reflecting a high degree of lexical diversity. The average length of the speaking turns is 15.89 words, capturing the natural variability and spontaneity of conversational speech. Fig. 13 shows a histogram of the number of words per speaking turn, with a peak at 11 words. This distribution is consistent with conversational speech, where speakers tend to produce short but semantically rich segments.

F. Phonetic Alignment

We provide time-aligned phonetic information for each speech segment in the corpus. This level of granularity enables fine-grained analysis of how phonetic structure interacts with emotions, which can support both acoustic modeling and prosody-aware emotion recognition. Importantly, these alignments facilitate cross-lingual and cross-corpus comparisons for emotion recognition, where phone-level correspondences often provide a more robust basis for knowledge transfer than lexical content alone [92], [93], [94]. To generate these alignments, we use the *Montreal Forced Aligner* (MFA) [95], a widely-used tool that performs state-of-the-art alignment of phonetic units for speech given its corresponding transcript. MFA utilizes an acoustic mode implemented with *Gaussian mixture models* (GMM) and *hidden Markov models* (HMM). The GMM-HMM model utilizes a pronunciation dictionary to align sequences of phonemes with audio, resulting in precise timestamps for each individual phone. We used the English pretrained model and default settings provided by MFA. The resulting alignments are released in TextGrid format.

V. Organization and Sharing of the Corpus

A. Partitions

The entire dataset is divided into multiple splits for training, development, and evaluation purposes. Table VI shows the distribution of primary emotions across splits. The class imbalance observed with each split is proportionally consistent with the class distribution across the whole dataset, except for *test 2* and *test 3*, as explained later in this section. A key distinction of our database is the addition of three test sets, which have different characteristics. The *test 1* set has approximately 17.2% of the corpus collected from 465 speakers (Table VI). Table III shows inter-evaluator agreements very similar to the values observed for the entire corpus.

TABLE VI
Emotional Class Distribution for Each Partition. The MSP-Podcast Corpus Has 267,905 Speaking Turns.

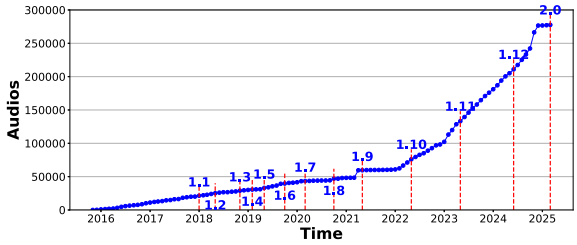
Emotion	Train	Dev.	Test1	Test2	Test3	All
Anger	22,609	5,728	6,985	538	400	36,260
Contempt	2,765	1,476	1,040	304	400	5,985
Disgust	1,324	534	744	141	400	3,143
Fear	794	285	348	116	400	1,943
Happiness	37,048	7,487	10,948	2,801	400	58,684
Neutral	51,149	8,318	12,457	6,793	400	79,117
Sadness	18,256	2,351	3,041	581	400	24,629
Surprise	3,220	1,025	1,206	394	400	6,245
Other	1,746	677	1,019	277	0	3,719
No agreement	30,279	6,518	8,506	2,877	0	48,180
Total	169,190	34,399	46,294	14,822	3,200	267,905

The *test 2* set was collected without the retrieval-based protocol presented in Section III-C. An early feedback we received was that machine learning models may bias the selection of speaking turns. We mitigate this issue by utilizing over 48 criteria based on different SER formulations, trained on different databases, features, and modalities, as explained in Section III-C. In response to this problem, we also created the *test 2* set. We selected 117 podcasts for this set, annotating all the speaking turns that satisfy our requirements, except the emotion retrieval step (Fig. 1). A consequence of this distinction is the higher proportion of speaking turns labeled as neutral (around 45.8% – Table VI). This test set includes recordings from 112 known speakers. An observation from this set in Table III is the lower inter-evaluator agreement compared to other partitions since neutral speech tends to be more uncertain [96].

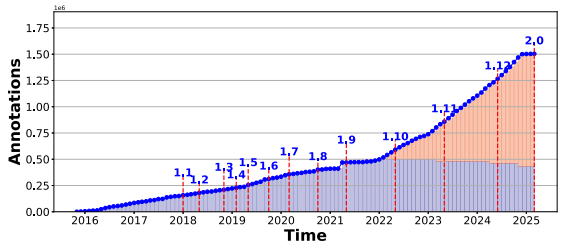
The *test 3* set comprises 3,200 speaking turns, with a balanced representation based on primary categorical emotions (Table VI). These speaking turns come from 428 speakers. We are not releasing the emotional labels, transcriptions, or speaker information for this set, as it aims to provide an unbiased test set where different groups can evaluate their models and compare their results. Early versions of this test set were successfully used for SER challenges (Odyssey 2024 [64] and Interspeech 2025 [89]). We have developed a website-based interface for research groups to submit their results for classification of primary emotions and prediction of emotional attributes[2]. The website displays a leaderboard for each of these two SER formulations, which are automatically updated with the results of new submissions. Notice that the balance of emotional classes resulted in higher inter-evaluator agreements (Table III).

The development set has 12.9% of the corpus (Table VI), and its purpose is to allow research teams to optimize the performance of their SER models on this set during training, including hyperparameters. This practice avoids using the test set(s) during training. The set includes recordings from 704 speakers, which are not included in either the test sets or the training set. The training set includes recordings of the remaining 2,220 speakers and the speaking turns with unknown speakers. The partitions aim to be speaker-independent, although some unknown speakers in the training set may overlap with those

[2]https://lab-msp.com/MSP-Podcast_Competition/SERB/



(a) Number of Speaking Turns Over Time



(b) Number of Annotations Over Time

Fig. 14. Development of the MSP-Podcast corpus over time. The figure shows the number of (a) speaking turns and (b) annotations over time. The vertical lines indicate a released version of the corpus. For Fig. (b), the blue lines correspond to crowdsourcing evaluations and the red lines correspond to student worker evaluations.

in the development or test partitions. The test sets should never be used for training SER models, since there is speaker overlap between test sets (e.g., data from some speakers are included in both *test 1* and *test 2* sets).

B. Sharing Early Versions of the Corpus

The effort to collect the MSP-Podcast corpus started in 2015. Instead of waiting for the full corpus to be ready, we have provided partial releases so the community can benefit from this resource.[3] Fig. 14(a) shows the number of speaking turns completed over time. The vertical lines indicate the different releases of the corpus. After transitioning to perceptual evaluations with student workers, the size of the corpus began to grow more rapidly (from 2022 to 2024). For example, in 2024 the median number of fully annotated speaking turns per week was 1,588 (up from 403 in 2020, the last year we fully relied on crowdsourcing). Fig. 14(b) shows the number of annotations over time, indicating in blue the crowdsourcing worker annotations and in red the student worker annotations. The plot also shows an increased rate in the number of annotations from the time we fully transitioned to perceptual evaluation conducted by student workers. By the end of the project, 65.82% of the annotations were provided by our student workers.

At the time of writing this paper, we have signed data transfer agreements with 363 academic research groups worldwide: Africa (5), Asia (181), Australia (9), Europe (103), North America (58), and South America (7). The corpus is widely used today, playing a key role in advancing the area of speech emotion recognition.

[3]All the differences between the MSP-Podcast releases, including partition adjustments, emotional labels and speaker annotations, are documented in the dataset release package. The release has all the information and files needed to recreate previous releases.

TABLE VII
BASELINE PERFORMANCE ON CATEGORICAL EMOTION RECOGNITION AND EMOTIONAL ATTRIBUTES RECOGNITION

Categorical Emotions

	Test 1		Test 2		Test 3	
Model	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi
WavLM	0.297	0.394	0.206	0.280	0.356	0.373
Wav2vec 2.0	0.238	0.325	0.156	0.166	0.289	0.316
HuBERT	0.285	0.390	0.192	0.264	0.344	0.361

Emotional Attributes

	Model	Valence	Arousal	Dominance
Test 1	WavLM	0.722	0.724	0.645
	Wav2vec 2.0	0.692	0.718	0.639
	HuBERT	0.720	0.708	0.648
Test 2	WavLM	0.549	0.547	0.467
	Wav2vec 2.0	0.479	0.553	0.467
	HuBERT	0.541	0.533	0.465
Test 3	WavLM	0.632	0.632	0.479
	Wav2vec 2.0	0.625	0.634	0.476
	HuBERT	0.641	0.630	0.489

VI. BASELINE

This section presents SER results that can serve as a baseline for other researchers using this corpus. We use pre-trained SSL models built on WavLM [97], Wav2vec 2.0 [98], or HuBERT [99]. These models contain 24 transformer layers and are comprised of ∼310 M parameters. We utilized the pre-trained off-the-shelf models from Hugging Face [50]. As evidenced in previous studies [58], [59], [61], [64], [100], [101], fine-tuning pre-trained SSL models for SER can lead to a significant performance boost. For categorical emotion recognition, we fine-tuned the models on eight emotion classes using focal loss, with a simple two-layer fully connected head. For attribute prediction, we adopted a staged fine-tuning strategy: first, adapting SSL models using *concordance correlation coefficient* (CCC) loss to predict valence, arousal, and dominance, and then jointly training with categorical classification using focal loss. After the fine-tuning stage, for attribute-based predictions, we employ a single-task setup, where we train a separate regression model for each of the three emotion attributes, while keeping the SSL encoder frozen and updating only the head. We fine-tuned both models for 20 epochs, with a learning rate of 1e-5, a batch size of 32, and the Adam optimizer.

Table VII summarizes baseline results for categorical emotion recognition and emotional attribute prediction. Overall, we observed consistent improvements across all test partitions compared to the previous MSP-Podcast v1.12 release, highlighting the benefit of expanding the training set and removing low-agreement labels. On the *speech emotion recognition benchmark* (SERB) [89], these refinements translated into ∼8% relative gains over the earlier baselines. WavLM generally outperformed both wav2vec2 and HuBERT in both categorical and attribute tasks. The large gap between F1-macro and F1-micro scores in Test 1 reflects the severe imbalance across the eight emotion classes, where frequent categories (e.g., neutral, happiness) dominate the micro-average. These results provide a stronger

and more reliable baseline for future work in categorical and dimensional SER.

VII. Discussion

The MSP-Podcast corpus opens new research possibilities due to its unique features, including its diversity in speakers, emotions, and environments. Wagner et al. [61] and Naini et al. [101] demonstrated that finetuning SSL models such as WavLM with emotional data is beneficial for SER tasks. This corpus is sufficiently large to support effective finetuning, providing a stronger starting point for models tailored to a specific domain where less annotated data may be available. This database unlocks a range of novel opportunities. We focus here on highlighting a few notable ones.

A. Perception of Emotions

With 1,446,224 annotations from 13,278 workers, this corpus is well-suited for studying human emotion perception. We are releasing all individual annotations, along with the timestamps indicating when each annotation was completed. This information enables research that incorporates contextual factors into emotion perception. For instance, it allows investigation of the priming effect – how previously annotated sentences influence the perception of subsequent speaking turns [102], [103]. The sequential order of the annotations can also support preference learning strategies, where direct comparisons are used to establish relative labels (e.g., one speaking turn is more positive than another) [104].

A related resource is the MSP-Conversation corpus [18], which includes time-continuous annotations of 10–20 minute segments from the same podcasts used in the MSP-Podcast corpus. These annotations provide continuous traces of perceived changes in valence, arousal, and dominance over time. There are 12,561 segments in the MSP-Podcast that overlap with the recordings in the MSP-Conversation corpus. This overlapping set offers an opportunity to study the relationship between continuous-time annotations (MSP-Conversation) and sentence-level annotations (MSP-Podcast) [105].

B. Robustness to Environments

The variety of podcasts used in this corpus provides a perfect resource for evaluating speech models that are robust to multiple environments. We highlight two prominent efforts in this area. Leem et al. [65] recorded an early version of the MSP-Podcast corpus by playing the speaking turns and radio noise in a single-walled sound booth (release 1.8). The microphone and the speaker were strategically placed at different locations to achieve target SNRs. This noisy version of the corpus has been extremely useful to explore robust SER models [106]. The second effort is the work of Grageda et al. [107], [108], which recorded a noisy version of the MSP-Podcast corpus in the context of *human robot interaction* (HRI) (*test1* of release 1.9). The microphone was mounted on a robot, which moved, changing the relative distance between the noise source, the speech source, and the microphone. This effort has led to improvements in distant SER models [109].

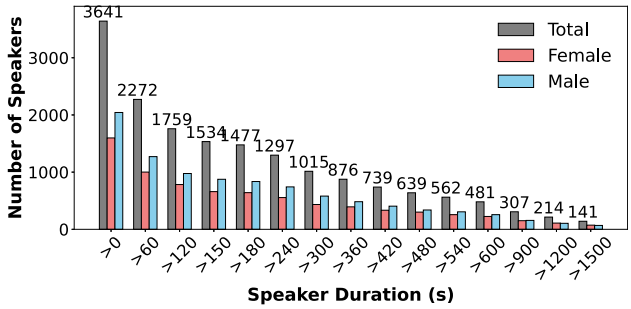


Fig. 15. Cumulative distribution of speakers with increasing recording duration. The bars show the number of male, female, and total speakers who have more than a given duration of data in seconds.

C. Emotions and Other Speech Tasks

The size of the corpus and the speaker information make this resource ideal for exploring how emotion affects other speech tasks, such as speaker verification and speaker recognition tasks [82], [83], [84], [85], [86]. To support these tasks, we made a key decision to collect multiple podcast episodes from the same speakers whenever possible. Speaker verification evaluations are often conducted across sessions collected on different days under different conditions. Different episodes are often collected on different days, which approaches this evaluation setting where several speakers appear in multiple podcasts. Likewise, many applications and experimental settings require sufficient recordings from individual speakers, which we ensured by including multiple episodes per speaker. For example, speaker verification tasks require an enrollment set to build the models. Also, *text-to-speech* (TTS) requires enough data to build a speaker model. Fig. 15 shows an accumulative plot with the number of speakers having a given amount of data. For example, there are 1,015 speakers with 300 seconds (5 minutes), and 141 speakers with 1,500 seconds (25 minutes) of data. These features make this corpus ideal for *voice conversion* (VC) and TTS tasks [87], [88].

D. Rich Emotional Descriptors

Most emotional corpora provide either categorical or attribute-based annotations. In contrast, the MSP-Podcast offers both, along with secondary emotion labels, where annotators select all emotions they perceive in a recording. We have shown the value of secondary emotions by using them as auxiliary tasks in classification problems [110], and in retrieval tasks aimed at finding recordings with emotions similar to a reference (anchor) sample [111], [112]. As described in Section IV-A, the annotation protocol allows evaluators to provide their own labels for both primary and secondary emotions when none of the predefined options are appropriate. This information is also valuable, as demonstrated by Chou et al. [113], who transformed the free-text labels into polarity vectors (negative, positive, ambiguous) using LIWC [114]. These examples showcase the potential of the rich emotional descriptors provided in the corpus.

E. Support for Other Data Collections

The focus of this project is on speech recordings in English. There is a need to collect similar databases in other languages. We created the *affective naturalistic database consortium* (AndC)[4]. This initiative aims to provide all the tools used to collect the MSP-Podcast corpus to other researchers, enabling them to create new databases and expand the infrastructure for affective computing. We have partnered with collaborators from the National Tsing Hua University in Taiwan to test this initiative. They followed the code and protocol used for our corpus. The result of this effort is the BIIC-Podcast corpus [15], with recordings in Taiwanese Mandarin. Another example is the collection of the *White House tapes speech emotion recognition* (WHiSER) corpus [32]. Using a variation of the proposed protocol, we annotated the emotions of ambient recordings from the Oval Office during the presidency of Richard Nixon. This set provides a perfect test set for SER models in challenging recording conditions (distant speech, low-quality microphones, noisy environment). We expect that this consortium will encourage the creation of new resources.

Another collaboration that started from this effort is the NaturalVoices corpus [115], [116]. The MSP-Podcast corpus derives its 409 hours of selected material from a larger pool of 6,007 recordings totaling 5,046 hours. The NaturalVoices corpus uses the entire recordings from the the 6,007 podcasts (5,046 hours). While MSP-Podcast was originally developed for SER, NaturalVoices is tailored for speech generation tasks, particularly *voice conversion* (VC) [115] and *emotional voice conversion* (EVC) [116]. Its annotations and data processing pipeline are specifically designed to support these tasks, although the corpus is also suitable for other speech synthesis applications such as *text-to-speech* (TTS). The original podcast recordings are freely available[5]. The MSP-Conversation corpus [18] also beneficed from the collection of the MSP-Podcast corpus.

VIII. Conclusion

This paper presented the results of a 10-year effort to develop the MSP-Podcast corpus – a large, naturalistic emotional speech database containing diverse recordings from multiple speakers across various environments. The database reflects the emotions observed in daily human interactions. The corpus includes a rich set of emotional descriptors, enabling new research in emotion analysis, recognition, and synthesis. To ensure high-quality annotations, we implemented several strategies, including a screening test for student workers prior to hiring, weekly feedback, and targeted training to improve consistency in labeling. In addition to releasing the final version of the corpus, we also provide the code used in the protocol (Section VII-E), with the intention of supporting replication efforts that will expand affective computing resources in other languages.

MSP-Podcast 2.0 introduces several methodological advances that distinguish it from previous SER datasets. It leverages a retrieval-based selection of samples and a protocol that scales to large, continuously growing collections of naturalistic speech. The corpus provides unique features such as naturalness, size, and diversity in both speakers and conversational topics. It also includes various test sets that measure different goals, enabling targeted evaluation across conditions. Finally, the MSP-Podcast corpus offers a comprehensive annotation protocol that facilitates multiple formulations to be explored, including categorical, dimensional, and multi-label perspectives. These characteristics collectively position the dataset as a robust resource for advancing emotion recognition research.

While MSP-Podcast offers extensive coverage of emotional speech, it also presents several areas for improvement. The corpus does not capture longer conversational contexts, limiting the temporal modeling of emotion compared to resources such as the MSP-Conversation corpus [18]. As with any dataset relying on human annotations, labeling errors may occur (e.g., the speaker ID, transcription, emotional label). Additionally, the corpus does not include visual modality, which restricts multimodal research including facial expressions [117], [118], [119]. These factors highlight opportunities for future extensions of the resource.

[4]http://andc.ai/

[5]https://github.com/3loi/NaturalVoices

Acknowledgment

The authors are grateful to the more than 13,720 individuals who contributed to this effort. They use AI systems for editing and grammar enhancement.

References

[1] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social Emotions in Nature and Artifact: Emotions in Hum. and Hum.-Comput. Interact.*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford Univ. Press, Nov. 2013, pp. 110–127.

[2] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.

[3] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, *Emotional Prosody Speech Transcripts*. Philadelphia, PA, USA: Linguistic Data Consortium, 2002.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[5] S. Livingstone and F. Russo, "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*, vol. 13, no. 5, pp. 1–35, May 2018.

[6] C. Busso et al., "IEMOCAP: Interact. emotional dyadic motion capture database," *J. Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[7] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA Tut. Res. Workshop Speech Emotion*, Newcastle, Northern Ireland, U.K., Sep. 2000, pp. 39–44.

[8] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.

[9] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and Affect. interactions," in *Proc. 2nd Int. Workshop Emotion Representation, Anal. Synth. Continuous Time Space*, Shanghai, China, 2013, Apr. pp. 1–8.

[10] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "NNIME: The NTHU-NTUA chinese interactive multimodal emotion corpus," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 292–298.

[11] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "Memor: A dataset for multimodal emotion reasoning in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, 2020, pp. 493–502, doi: 10.1145/3394171.3413909.

[12] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, Hannover, Germany, Jun. 2008, pp. 865–868.

[13] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy, Jul. 2019, pp. 527–536.

[14] A. Vidal, A. Salman, W.-C. Lin, and C. Busso, "MSP-face corpus: A natural audiovisual emotional database," in *Proc. ACM Int. Conf. Multimodal Interact.*, Utrecht, The Netherlands, Oct. 2020, pp. 397–405.

[15] S. Upadhyay et al., "An intelligent infrastructure toward large scale naturalistic Affect. speech corpora collection," in *Proc. 11th Int. Conf. Affect. Comput. Intell. Interact.*, Cambridge, MA, USA, Sep. 2023, pp. 1–8.

[16] D. Kollias et al., "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, no. 6/7, pp. 907–929, Jun. 2019, doi: 10.1007/s11263-019-01158-4.

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.

[18] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1823–1827.

[19] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 238–242.

[20] V. Kondratenko, N. Karpov, A. Sokolov, N. Savushkin, O. Kutuzov, and F. Minkin, "Hybrid dataset for speech emotion recognition in Russian language," in *Proc. Interspeech*, 2023, pp. 4548–4552.

[21] J. Smith, A. Tsiartas, V. Wagner, E. Shriberg, and N. Bassiou, "Crowdsourcing emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5139–5143.

[22] Y. Cheng, R. Zhang, and J. Shi, "MIKU-PAL: An automated and standardized multi-modal method for speech paralinguistic and affect labeling," in *Proc. INTERSPEECH*, Rotterdam, The Netherlands, Aug. 2025, pp. 4308–4312, doi: 10.21437/Interspeech.2025-648.

[23] A. Bagher Zadeh et al., "CMU-MOSEAS: A multimodal language dataset for," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 1801–1812. [Online]. Available: https://aclanthology.org/2020.emnlp-main.141/

[24] A. Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. ACM Assoc. Computat. Linguistics*, vol. 1, Melbourne, Australia, Jul. 2018, pp. 2236–2246.

[25] J. Wongpithayadisai et al., "THAI speech emotion recognition (THAI-SER) corpus," 2025, *arXiv:2507.09618*.

[26] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. Interspeech Int. Conf. Spoken Lang.*, Pittsburgh, PA, USA, Sep. 2006, pp. 801–804.

[27] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces*, Nagoya, Aichi, Japan, Nov. 2007, pp. 30–37.

[28] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU aibo emotion corpus," in *Proc. 2nd Int. Conf. Lang. Resour. Eval. Int. Workshop Emotion: Corpora Res. Emotion Affect.*, Philadelphia, PA, USA, May 2008, pp. 28–31.

[29] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "MEC 2017: Multimodal emotion recognition challenge," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 1–5.

[30] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception," *Lang. Resour. Eval.*, vol. 54, no. 2, pp. 341–383, 2020.

[31] F. Catania, J. W. Wilke, and F. Garzotto, "Emozionalmente: A crowdsourced corpus of simulated emotional speech in italian," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 1142–1155, 2025.

[32] A. Reddy Naini, L. Goncalves, M. Kohler, D. Robinson, E. Richerson, and C. Busso, "WHiSER: White House Tapes speech emotion recognition corpus," in *Proc. Interspeech*, Kos Island, Greece, Sep. 2024, pp. 1595–1599.

[33] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan.–Mar. 2012.

[34] L. Chen, "Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction," Ph.D. dissertation, Univ. Illinois Urbana-Champaign, Champaign, IL, USA, 2000.

[35] M. Z. Akhtar, R. Jahangir, Q. Ain, M. A. Nauman, M. Uddin, and S. S. Ullah, "UrduSER: A comprehensive dataset for speech emotion recognition in urdu language," *Data Brief*, vol. 60, 2025, Art. no. 111627. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340925003580

[36] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[37] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," *Scholars Portal Dataverse*, vol. 1, 2020, Art. no. 2020.

[38] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, May 2005.

[39] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: Actors, wizards and human beings," in *Proc. ISCA Tutorial Res. Workshop Speech Emotion*, Newcastle, Northern Ireland, U.K., Sep. 2000, pp. 195–200.

[40] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: A closer look," in *Proc. 2nd Int. Conf. Lang. Resour. Eval. Int. Workshop Emotion: Corpora Res. Emotion Affect*, Marrakech, Morocco, May 2008, pp. 17–22.

[41] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: A Chinese natural emotional audio–visual database," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 6, pp. 913–924, 2017.

[42] A. Reddy Naini, D. Robinson, E. Richerson, and C. Busso, "Domain-specific adaptation in speech emotion recognition using emotional distribution alignment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Hyderabad, India, Apr. 2025, pp. 1–5.

[43] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Oct.–Dec. 2016.

[44] J. Lee, J. Park, K. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proc. 14th Int. Conf. Sound Music Comput. Conf., Sound Music Comput. Netw.*, 2017, pp. 220–226.

[45] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Commun.*, vol. 111, pp. 44–55, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639318304308

[46] B. McFee et al., "librosa: Audio and music signal analysis in python," in *Proc. Python Sci. Conf.*, Austin, TX, USA, Jul. 2015, pp. 18–25.

[47] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. ISCA Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3222–3226.

[48] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe," in *Proc. ISCA Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1983–1987.

[49] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., 202, Jul. 2023, pp. 28492–28518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[50] T. Wolf et al., "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771v5*.

[51] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," in *Proc. SEMAINE Rep. D6b*, Belfast, Northern Ireland, U.K., Sep. 2010. [Online]. Available: http://semaine-project.eu

[52] K. Gorman, "Python classes for Praat TextGrid and TextTier files (and HTK. mlf files)," 2017. [Online]. Available: https://github.com/kylebgorman/textgrid

[53] P. Boersma, "Praat, A system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[54] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 2598–2601.

[55] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. ISCA Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3222–3226, doi: 10.21437/Interspeech.2023-205.

[56] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe," in *Proc. ISCA Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1983–1987, doi: 10.21437/Interspeech.2023-105.

[57] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Appl. Acoust.*, vol. 156, pp. 351–358, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19304281

[58] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 1168–1172.

[59] L. Goncalves and C. Busso, "Harnessing multi-modal unlabeled data for enhanced speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 17, no. 1, pp. 1134–1146, Jan.–Mar. 2026.

[60] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, Sep. 2020.

[61] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.

[62] A. R. Naini, M. A. Kohler, and C. Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[63] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Proc. Findings Assoc. Comput. Linguistics*, Nov. 2020, pp. 1644–1650. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.148

[64] L. Goncalves et al., "Odyssey 2024 - Speech emotion recognition challenge: Dataset, baseline framework, and results," in *Proc. Speaker Lang. Recognit. Workshop*, 2024, pp. 247–254.

[65] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Proc. Interspeech*, Brno, Czech Republic, Aug.–Sep. 2021, pp. 2871–2875.

[66] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3698–3702.

[67] C. Busso and S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Proc. Interspeech*,Brisbane, Australia, Sep. 2008, pp. 1670–1673.

[68] I. Naji, "TSATC: Twitter sentiment analysis training corpus," 2012, Accessed: Apr. 10, 2026. [Online]. Available: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

[69] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 815–826, Apr. 2019.

[70] H.-C. Chou, C.-C. Lee, and C. Busso, "Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier," in *Proc. Interspeech*, Incheon, South Korea, Sep. 2022, pp. 161–165.

[71] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. Int. Joint Conf. Neural Netw.*, Vancouver, BC, Canada, Jul. 2016, pp. 566–570.

[72] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, San Antonio, TX, USA, Oct. 2017, pp. 415–420.

[73] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Nara, Japan, Sep.–Oct. 2021, pp. 1–8.

[74] H.-C. Chou et al., "Embracing ambiguity and subjectivity using the all-inclusive aggregation rule for evaluating multi-label speech emotion recognition systems," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Macao, China, Dec. 2024, pp. 502–509.

[75] H.-C. Chou, L. Goncalves, S.-G. Leem, A. Salman, C.-C. Lee, and C. Busso, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 41–55, Jan.–Mar. 2025.

[76] R. Lotfian and C. Busso, "Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 870–882, Oct.–Dec. 2021.

[77] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 402–416, Apr.–Jun. 2021.

[78] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 252–256.

[79] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation Emotion*, vol. 17, no. 1, pp. 41–51, Mar. 1993.

[80] M. Bradley and P. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[81] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, Oct.–Nov. 2007.

[82] S. Parthasarathy and C. Busso, "Predicting speaker recognition reliability by considering emotional content," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, San Antonio, TX, USA, Oct. 2017, pp. 434–436.

[83] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, May 2020, pp. 7169–7173.

[84] S. Parthasarathy, C. Zhang, J. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 5540–5544.

[85] M. Bancroft, R. Lotfian, J. Hansen, and C. Busso, "Exploring the intersection between speaker verification and emotion recognition," in *Proc. Int. Workshop Social Emotion AI Ind.*, Cambridge, U.K., Sep. 2019, pp. 337–342.

[86] I. Ülgen, Z. Du, C. Busso, and B. Sisman, "Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seoul, Republic of Korea, Apr. 2024, pp. 12081–12085.

[87] A. Mahapatra, I. Ülgen, A. Reddy Naini, C. Busso, and B. Sisman, "Can emotion fool anti-spoofing?," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 5628–5632, doi: 10.21437/Interspeech.2025-1234.

[88] I. R. Ülgen, C. Busso, J. Hansen, and B. Sisman, "We need variations in speech synthesis: Sub-center modelling for speaker embeddings," 2024, *arXiv:2407.04291*.

[89] A. R. Naini et al., "The Interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 4668–4672, doi: 10.21437/Interspeech.2025-1972.

[90] K. C. Puvvada et al., "New standard for speech recognition and translation from the NVIDIA nemo canary model," HuggingFace repository, 2024. [Online]. Available: https://huggingface.co/nvidia/canary-1b

[91] V. Srivastav et al., "Open automatic speech recognition leaderboard," 2023. [Online]. Available: urlhttps://huggingface.co/spaces/huggingface.co/spaces/open-asr-leaderboard/leaderboard

[92] S. Upadhyay, L. Martinez-Lucas, W. Katz, C. Busso, and C.-C. Lee, "Phonetically-anchored domain adaptation for cross-lingual speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 1631–1645, Jul.–Sep. 2025.

[93] S. Upadhyay et al., "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes island, Greece, Jun. 2023, pp. 1–5.

[94] P. Mote, A. Reddy Naini, D. Robinson, E. Richerson, and C. Busso, "Analysis of phonetic level similarities across languages in emotional speech," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 4343–4347, doi: 10.21437/Interspeech.2025-2112.

[95] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 498–502, doi: 10.21437/Interspeech.2017-1386.

[96] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, May 2020, pp. 8384–8388.

[97] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[98] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.
[99] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[100] H. Wu et al., "EMO-SUPERB: An in-depth look at speech emotion recognition," 2024, *arXiv:2402.13018*.
[101] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seoul, Republic of Korea, Apr. 2024, pp. 12031–12035.
[102] L. Martinez-Lucas, A. Salman, S.-G. Leem, S. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of Affect. priming on emotional annotations," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Cambridge, MA, USA, Sep. 2023, pp. 1–8.
[103] L. Martinez-Lucas et al., "Affect. priming in emotional annotations and its effect on speech emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Aug., 2025, doi: 10.1109/TAFFC.2025.3597034.
[104] A. Reddy Naini, A. Salman, and C. Busso, "Preference learning labels by anchoring on consecutive annotations," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1898–1902.
[105] L. Martinez-Lucas, W.-C. Lin, and C. Busso, "Analyzing continuous-time and sentence-level annotations for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1754–1768, Jul.–Sep. 2024.
[106] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 6447–6451.
[107] N. Grágeda, C. Busso, E. Alvarado, R. Mahu, and N. Becerra Yoma, "Distant speech emotion recognition in an indoor human-robot interaction scenario," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3657–3661.
[108] N. Grágeda, C. Busso, E. Alvarado, R. García, R. Mahu, and N. B. Yoma, "Speech emotion recognition in real static and dynamic human-robot interaction scenarios," *Comput. Speech Lang.*, vol. 89, Jan. 2025, Art. no. 101666.
[109] R. Garcia et al., "Speech emotion recognition with deep learning beamforming on a distant human-robot interaction scenario," in *Proc. Interspeech*, Kos Island, Greece, Sep. 2024, pp. 3215–3219.
[110] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 951–955.
[111] J. Harvill, S.-G. Leem, M. AbdelWahab, R. Lotfian, and C. Busso, "Quantifying emotional similarity in speech," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1376–1390, Apr.–Jun. 2023.
[112] J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso, "Retrieving speech samples with similar emotional content using a triplet loss function," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 7400–7404.
[113] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, May 2022, pp. 7717–7721.
[114] J. Pennebaker, R. Booth, R. Boyd, and M. Francis, "Linguistic inquiry and word count: LIWC2015," *Pennebaker Conglomerates*, Austin, TX, USa: Operator's Manual, 2015. [Online]. Available: www.LIWC.net
[115] A. Salman, Z. Du, S. Chandra, I. Ülgen, C. Busso, and B. Sisman, "Towards naturalistic voice conversion: Naturalvoices dataset with an automatic processing pipeline," in *Proc. Interspeech*, Kos Island, Greece, Sep. 2024, pp. 4358–4362.
[116] Z. Du et al., "NaturalVoices: A large-scale, spontaneous and emotional podcast dataset for voice conversion," 2025, *arXiv:2511.00256*.
[117] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 306–318, Jan.–Mar. 2025.
[118] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, Oct.–Dec. 2022.
[119] L. Goncalves, H.-C. Chou, A. N. Salman, C.-C. Lee, and C. Busso, "Contextual attention for robust audio-visual emotion recognition," *IEEE Open J. Signal Process.*, vol. 7, pp. 42–53, 2026.

Carlos Busso (Fellow, IEEE) is currently a professor with Language Technologies Institute, Carnegie Mellon University, where he is also the director with Multimodal Speech Processing (MSP) Laboratory. His research interests include human-centered multimodal machine intelligence and applications, focusing on the broad areas of speech processing, affective computing, multimodal behavior generative models, and foundational models for multimodal processing. He is also an ISCA fellow.

Reza Lotfian received the the BSc degree from Amirkabir University, in 2006, the MSc degree from Sharif University, in 2010, and the PhD degree in electrical engineering from The University of Texas at Dallas (UTD), Richardson, TX, USA, in 2018. From 2013 to 2018, he was with MSP Lab, UTD, where he contributed to the development of the MSP-Podcast corpus. He is currently a senior machine learning engineer with Athenahealth, developing AI solution for healthcare industry. His interests include speech emotion recognition, affective computing, NLP, LLMs, and scalable ML systems.

Kusha Sridhar received the MS degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2015, and the PhD degree in electrical engineering from the University of Texas at Dallas, in 2021. He was a Staff Research scientist with Hippocratic AI Inc. He is currently a Sr. manager with Accenture's Advanced Computational AI Group. His research interests include areas related to affective computing, conversational speech models, and multi-modal signal processing.

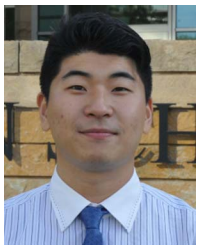
Ali N. Salman received the BS and MS degrees in computer science from Indiana State University, in 2015 and 2017, respectively, and the PhD degree in electrical engineering from the University of Texas at Dallas, in 2024. He is currently a research scientist with ARRAY Innovation. His research interests include affective computing, retrieval-augmented generation (RAG) systems, and facial analysis.

Wei-Cheng Lin (Member, IEEE) received the PhD degree in electrical engineering from the University of Texas at Dallas (UTD), in 2023. He is currently a research scientist with Bosch Research, Bosch Center for Artificial Intelligence, USA. His research interests include multimodal signal processing and deep learning. He was recipient of the Best Dissertation Award from the Association for the Advancement of Affective Computing (AAAC) in 2024.

Lucas Goncalves (Member) received the PhD degree in electrical engineering from The University of Texas at Dallas (UTD), Richardson, TX, USA, in 2024. He is currently an applied scientist with Amazon, USA. His research interests include multimodal signal processing and deep learning, with emphasis on audio–visual learning, speech and language technologies, and vision–language models. He was the recipient of the Erik Jonsson School Excellence in Education Doctoral Fellowship, from 2022 to 2024.

Srinivas Parthasarathy (Member, IEEE) received the PhD degree in electrical engineering from The University of Texas at Dallas (UTD), in 2019. He was a research intern with Amazon, Microsoft Research, and Bosch Research and Training Center. He is currently a senior applied scientist with Amazon. His research interests include computer vision, multimodal large language models, multi-modal signal processing and affective computing. He was the recipient of the Ericsson Graduate Fellowship at UTD, during 2013–2014.

Abinay Reddy Naini (Student Member, IEEE) received the BS degree in electrical engineering from the National Institute of Technology, Warangal, India, and the MS degree in electrical engineering from the Indian Institute of Science (IISc), India. He is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Texas at Dallas (UTD). He is also a visiting researcher with the Language Technologies Institute, Carnegie Mellon University. His research interests include affective computing, speech technology, and machine learning.



Seong-Gyun Leem received the BS and MS degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2018 and 2020, respectively, and the PhD degree in electrical engineering from the University of Texas at Dallas, in 2024. He is currently a research scientist with Reality Labs, Meta Platforms, Inc. His research interests include speech synthesis, speech emotion recognition, noisy speech processing, and machine learning.

Luz Martinez-Lucas (Graduate Student Member, IEEE) received the bachelor's degree in electrical engineering from the University of Texas at Dallas, where she is currently working toward the PhD degree from Electrical and Computer Engineering Department. Her research interests include affective computing, speech technology, and machine learning. She is a student member of AAAC.

Huang-Cheng Chou (Member, IEEE) received the BS and PhD degrees in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016 and 2024, respectively. From 2021 to 2022, he was a visiting scholar with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, TX, USA. He is currently a postdoctoral scholar with the University of Southern California (USC). His research focuses on affective computing.

Pravin Mote (Student Member, IEEE) is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Texas at Dallas. He is also a visiting researcher with the Language Technologies Institute, Carnegie Mellon University. His research interests include speech technology, multimodal affective computing, and machine learning.