The MSP-Conversation Corpus

Luz Martinez-Lucas, Mohammed Abdelwahab, Carlos Busso

Multimodal Signal Processing (MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

luz.martinez-lucas@utdallas.edu, mxa129730@utdallas.edu, busso@utdallas.edu

Abstract

Human-computer interactions can be very effective, especially if computers can automatically recognize the emotional state of the user. A key barrier for effective speech emotion recognition systems is the lack of large corpora annotated with emotional labels that reflect the temporal complexity of expressive behaviors, especially during multiparty interactions. This paper introduces the MSP-Conversation corpus, which contains interactions annotated with time-continuous emotional traces for arousal (calm to active), valence (negative to positive), and dominance (weak to strong). Time-continuous annotations offer the flexibility to explore emotional displays at different temporal resolutions while leveraging contextual information. This is an ongoing effort, where the corpus currently contains more than 15 hours of speech annotated by at least five annotators. The data is sourced from the MSP-Podcast corpus, which contains speech data from online audio-sharing websites annotated with sentence-level emotional scores. This data collection scheme is an easy, affordable, and scalable approach to obtain natural data with diverse emotional content from multiple speakers. This study describes the key features of the corpus. It also compares the time-continuous evaluations from the MSP-Conversation corpus with the sentence-level annotations of the MSP-Podcast corpus for the speech segments that overlap between the two corpora.

Index Terms: Speech emotion recognition, human-computer interaction, time-continuous emotional attributes

1. Introduction

Human-computer interaction (HCI) is becoming increasingly ubiquitous in our lives. An open challenge to improve HCI systems is to develop robust technologies to recognize the emotional state of a user, which can facilitate building HCI systems that are seamless, engaging and effective. An appealing modality is speech, given the increased presence of speech-based interfaces. Therefore, studies have explored different formulations to improve *speech emotion recognition* (SER) systems [1]. For these models to be effective, we need a large amount of naturalistic data annotated with accurate emotional annotations that can describe the complex temporal externalization of expressive behaviors during multiple party interactions.

This paper introduces the MSP-Conversation corpus, which uses naturalistic recordings obtained from online podcasts, conveying a broad range of topics. A key feature of the corpus is that the recordings overlap with the recordings included in the MSP-Podcast database [2], which contains sentence-level annotations of short segments retrieved from podcasts. The MSP-Podcast corpus is not appropriate to study contextual information, as the isolated turns are separately evaluated, missing the temporal relationship between consecutive speaking turns. The MSP-Conversation corpus complements the MSP-Podcast, providing the perfect platform to explore temporal information. The segments in the MSP-Conversation corpus include continuous conversations within the podcasts with duration ranging between 10 and 20 minutes from multiple speakers appearing in multiple podcasts. The corpus has 74 conversation segments from the podcasts (approximately 15h), but the plan is to increase the number of conversational segments in the future. The data is annotated in a time continuous manner with emotional attributes to retain contextual information as well as the dynamic information present in the conversation. The overlap between both corpora provides the perfect resource to explore the relationship between sentence-level annotations and time-continuous annotations.

A key aspect of the MSP-Conversation corpus is the use of time-continuous annotations. Many emotional speech databases use sentence-level annotations, focusing on the emotions of only one speaker. However, emotions are not experienced in a vacuum. They are dynamic and affected by contextual information [3–5]. Time-continuous annotations are particularly effective in capturing dynamic information [6, 7]. They capture the instantaneous emotional perception of evaluators as they listen to a recording, creating emotional traces. These traces can be arbitrarily segmented according to the scope of the analysis, providing the resources to study emotions at various temporal resolutions (e.g., phone, syllable, word, phrase, sentence). The time-continuous annotations in the MSP-Conversation corpus correspond to emotional attributes. Emotional behaviors in natural settings are complex [8–10], so it is difficult to find a finite number of classes that can capture the differences between subtle expressive behaviors. In contrast, using emotional attributes provides appealing descriptors to better characterize emotional events [11], as validated by several studies supporting the core affect theory [12, 13]. The MSP-Conversation corpus is annotated in terms of arousal (active versus calm), valence (positive versus negative), and dominance (strong versus weak), which are the most common attributes previously used.

The MSP-Conversation corpus also has the advantage of containing natural speech that was not recorded for the purpose of creating an emotional database. A substantial amount of emotional speech databases are acted [14–16]. Gathering data in that way has the advantage of substantial control. However, acted corpora tend to over emphasize the emotions resulting in easier problems that does not resemble real world scenarios [17]. By sourcing its speech data from various audio-sharing websites, the recordings of the MSP-Conversation corpus not only allow for an immense amount of data, but also for a more diverse database in terms of speakers and emotions.

This study presents preliminary analysis on the MSP-Conversation corpus. It describes the distribution of the emotional content of the corpus. This paper also presents the inter-evaluator agreement. The average Cronbach's alphas are $\alpha = 0.50$ for arousal, $\alpha = 0.54$ for valence, and $\alpha = 0.41$ for dominance. The analysis shows that certain annotators are more reliable than others, which is useful information for aggregating the traces across evaluators. Finally, we compare the time-continuous evaluations of the MSP-Conversation corpus with the sentence-level annotations of the MSP-Podcast corpus for

Table 1: *Overview of Existing Emotional Databases.*

Database	Type	Duration	Speakers
CreativeIT [15]	Acted	≈8hrs	16
SEMAINE [20]	Natural	15.83hrs	28
RECOLA [21]	Natural	3.83hrs	46
SEWA [22]	Natural	>33hrs	398
MuSe-CaR [23]	Natural	36.87hrs	90
MSP-Conversation	Natural	15.15hrs	197

the speech segments belonging to both corpora. We transform the time-continuous traces into sentence-level descriptors by aggregating the trace values over the duration of the segments in the MSP-Podcast corpus. If the evaluator's reaction lag while annotating the time-continuous labels is considered [18, 19], we observe correlations between $\rho = 0.196$ and $\rho = 0.403$.

2. Related Work

While there are several emotional databases with sentence-level annotations, there are few speech corpora that has time-continuous annotations. Table 1 summarizes the key features of some of these databases. Most of these corpora include naturalistic recordings, where time-continuous annotations are most useful, since emotional traces can capture complex emotional changes during an interaction. The SEMAINE database's Solid SAL approach [20] controlled the emotions of the 'operator' to elicit target emotions from the 'user.' The RECOLA database [21] is another corpus, where participants were asked to solve a problem by communicating through a video conference. The participants were induced by showing neutral, negative or positive stimuli before recordings. A more recent corpus is the SEWA database [22], which takes a similar approach. The data consists of conversations between people, who were emotionally induced by showing emotional media. These databases are able to achieve some level of naturalness, but their data collection involves direct or indirect intervention of the researchers. Also, the emotional contents are often biased to positive interactions given the colloquial interactions [2]. Another recent corpus is the The MuSe-CaR database [23], which uses recording of car reviewers in the wild that are annotated with time-continuous labels for arousal, valence, and trustworthiness. The database includes natural recordings, and, although the domain of the database is limited, it has a wide range of emotional content. We argue that the research community needs more resources to effectively understand the role of temporal information during natural interactions. The proposed approach has the advantage that we can easily balance the emotional content and speaker demography by choosing the right podcasts. The approach does not intentionally manipulate or induce the speakers, resulting in a flexible and scalable approach to collect emotional data

3. The MSP-Conversation Corpus

The motivation for collecting the MSP-Conversation corpus is to address the key limitation of the MSP-Podcast corpus [2], a speech emotional corpus that we are also collecting. The MSP-Podcast corpus does not provide the infrastructure to study temporal information in the externalization of emotions. First, the content of the MSP-Podcast corpus is short sentences with durations less than 11 secs. The speech segments are not necessarily consecutive speech turns in the recordings, since our retrieval criterion is whether our SER models expect these segments to convey emotion. Second, the perceptual evaluation is conducted out-of-order, so full contextual information is not available to the annotators. In contrast, the speech recordings and the perceptual evaluations in the MSP-Conversation corpus are ideal to study temporal information of emotion. The MSP-Conversation corpus introduced in this paper is a perfect complement to the MSP-Podcast corpus.

3.1. Protocol for Selecting Conversations

The audio for the MSP-Conversation corpus is sourced from the same audio recordings used for the MSP-Podcast corpus [2]. We select continuous segments within the podcasts with durations ranging between 10 and 20 minutes. These segments, which we refer to as *conversations*, convey natural emotional content, spanning a broad range of emotions from multiple speakers engaged in natural, spontaneous interactions. Since we have some idea of the emotional content of the dialogues from the speaking turns already annotated on the MSP-Podcast corpus, we select conversation segments that are balanced in terms of gender and emotional content. We also want speaker diversity in the corpus. A final criterion for selection is to have conversations overlapping with segments belonging to the train, test and development sets of the MSP-Podcast corpus, since our aim is to align these partitions with the partitions of the MSP-Conversation corpus. The current version of the MSP-Conversation corpus includes 74 conversations with a total duration of 15hrs, 9min (50.6% female, 49.4% male), but the goal is to increase the size of the corpus in the future.

3.2. Annotations of Emotional Attributes

The MSP-Conversation corpus is annotated using the software CARMA [24], which is designed for time-continuous annotations. As the evaluators listen to the conversations, they use a joystick to place the computer cursor in a *graphical user interface* (GUI), according to their *instantaneous* emotional judgment. The GUI is intended to evaluate one emotional attribute at a time (i.e., one dimensional axis). The extreme values of the interface are *negative* and *positive* for valence, *calm* and *active* for arousal, and *weak* and *strong* for dominance. The central position in the interface corresponds to the neutral values of the attributes. Since each annotator focuses on one specific dimension at a time, we expect more reliable traces. The interface samples the position of the joystick at 60 fps, although the sampling rate is not always uniform. The annotations create emotional traces for each of the attributes.

Annotating long recordings with time-continuous evaluations is an intensive cognitive task. To avoid fatigue, we split the conversations into segments with durations between three to seven minutes. Although the segments were presented at random, annotators were asked to annotate the same attribute for a cycle of 10 consecutive segments before switching to a different emotional attribute. This process was implemented to avoid cognitive mistakes associated with jumping too often from one emotional attribute to another. The color of the interface was also changed depending of the emotional attribute. The duration of the annotation sessions was 1hr long, with at least 30min between sessions.

We have recruited eleven annotators, who were trained before starting the evaluations. First, participants were given an explanation of each emotional attribute, where they could clarify their questions. The attributes were explained, showing examples of their respective extremes. We explained the concept of dominance in the context of the speaker confidence. Then, we asked them to annotate nine dialogues from the SEMAINE database [20]. Their preliminary ratings were compared to previous annotations. We discussed any inconsistencies between their annotations and previous annotations. The training phase

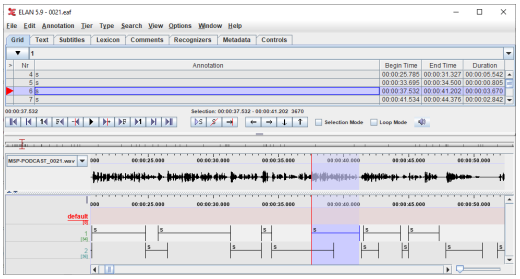


Figure 1: *View of the ELAN software, displaying the annotations of speaker diarization.*

lasted until the annotations were consistent with previously collected traces. After this training phase, they were allowed to start rating the conversations. Our goal is to have at least six different annotators for each conversation and for each emotional attribute. Currently, each conversation was been annotated by at least five annotators. The averages number of annotators per conversation are 6.06 for valence, 6.48 for arousal, and 5.80 for dominance.

3.3. Post-Processing of the Time-Continuous Annotations

We post-process the time-continuous annotations before combining them across annotators. The process aims to (1) smooth the traces, and (2) fix the sampling rate of the traces (the sampling rate provided by CARMA is not uniform, where the average value is approximately 59fps). We achieve these two goals with a moving median filter with an analysis window of 500ms centered at a given point. We shift the target point in jumps of 1/59 seconds, considering all the values of the traces within this 500ms interval. With the sampling rate consistent at 59fps, all the available annotations are averaged across evaluators.

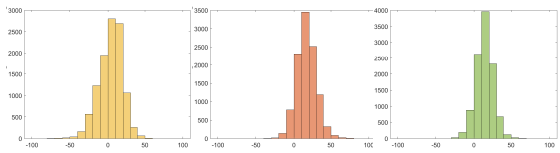
3.4. Annotation of Speaker Diarization

Each conversation was manually diarized, identifying the segments where each participant in the podcast was speaking. Figure 1 shows the annotations in ELAN [25]. Timing labels were created for any speech or emotional sound a speaker made during the conversation, including laughing and crying. Any music or unrelated sound was also annotated. Speaker labels started at any sentence, phrase, or emotional outburst done by a speaker, and ended at the end of the speaker's statements or emotional outburst. We annotate one channel per speaker, so it is easy to identify overlapped speech between speakers. The percentage of overlapped speech is around 4%, although the percentage can increase to as much as 10% depending of the type of conversation. The annotations for speaker diarization also include a unique number for identification of the speakers. In total, the corpus has 197 unique speakers (87 female, 110 male). Three of the conversations (4.1%) have one speaker, 32 conversations (43.3%) have two speakers, 30 conversations (40.5%) have three speakers, four conversations (5.4%) have four speakers, and five conversations (6.8%) have five speakers.

3.5. The MSP-Podcast Corpus

As mentioned, the MSP-Conversation corpus sources all its speech data from the same audio recordings used to collect the MSP-Podcast corpus. The overlap between both corpora is a key strength of our effort. For example, this setting provides the resources to compare sentence-level annotations with time-continuous traces. This section briefly summarizes the MSP-Podcast corpus to better understand the analysis in Section 4.3.

The speech segments in the MSP-Podcast corpus [2] are extracted from recordings available from audio-sharing websites. All the podcasts used for the corpus are available to the public under Creative Commons licenses with the least restrictive conditions (CC-BY or CC-0), so the corpus can be shared with other researchers. Since the selected podcasts contain unscripted conversations that cover a wide range of topics, the corpus has natural, spontaneous interactions from a large number of speakers, expressing a diverse range of emotional behaviors. The podcasts are segmented into short sentences with durations between 2.7s to 11s. We follow the approach proposed by Mariooryad *et al.* [26], where we use machine learning models to identify speech segments with emotional content. These are the segments that are annotated with emotional labels. Notice that these segments are not necessarily consecutive speech turns. The annotation of the corpus is done at the sentence-level (i.e., one label per segment), and in random order without full contextual information. The emotional annotation is conducted with a crowdsourcing protocol that tracks the quality of the workers in real time [27]. The corpus includes categorical and attribute-based labels. The attribute-based labels are also valence, arousal and dominance captured with a seven-point Likert scale. The MSP-Podcast corpus is an ongoing effort. The most current release is version 1.7, which consists of 62,140 speech turns (100hrs). We have annotated the speaker identity of 51,202 sentences (1,163 speakers). The number of speech segments in this version that overlap with the recordings on the MSP-Conversation Corpus is 1,567 turns.



(a) Valence (b) Arousal (c) Dominance

Figure 2: *Histograms of emotional attributes in the corpus. The values correspond to average scores over five second windows.*

3.6. Partition of the MSP-Conversation Corpus

The partition of the conversation takes into account the partitions of the MSP-Podcast corpus. The MSP-Podcast corpus is divided into train, test and development sets, where the goal is to minimize cases where speech recordings from the same speaker are in more than one set. The MSP-Conversation corpus is partitioned so that speakers in the two corpora share the same partition (if a speaker is in the test set for the MSP-Podcast corpus, it will also be in the test set for the MSP-Conversation corpus). With this criterion, 46 conversations are in the train set (62.2%), 17 conversations are in the test set (23.0%), and 11 conversation are in the development set (14.9%).

4. Analysis of the Corpus

4.1. Emotional Diversity

We evaluate the diversity of the emotional content of the corpus. All the aggregated traces were split into five-second segments. Then, we estimate the average value over time across the segments. Figure 2 shows the histograms of the dimensional attributes. Notice that the values of the traces are in the range between -100 and 100. The figure shows that extreme values are uncommon. Most of the annotations are concentrated between -40 to 40 for valence, -20 to 50 for arousal, and -20 to 40 for dominance. Valence has a wider range compared to the other two dimensions, and dominance has the smallest range.

Table 2: *Cronbach's Alpha to assess inter-evaluator agreement.*

	Arousal		Valence		Dominance	
All	0.50		0.54		0.41	
Ann	**Incl**	**Excl**	**Incl**	**Excl**	**Incl**	**Excl**
1	0.50	0.51	0.54	0.53	0.41	0.43
2	0.50	0.46	0.50	0.52	0.39	0.35
3	0.51	0.51	0.47	0.51	0.40	0.37
4	0.58	0.53	0.63	0.57	0.58	0.42
5	0.50	0.45	0.54	0.44	0.41	0.32
6	0.50	0.46	0.54	0.49	0.41	0.37
7	0.50	0.43	0.56	0.51	0.44	0.40
8	0.50	0.41	0.56	0.45	0.44	0.34
9	0.56	0.52	0.57	0.50	0.62	0.54
10	0.56	0.50	0.58	0.52	0.48	0.40
11	0.58	0.50	0.62	0.59	0.54	0.46

Arousal and dominance scores are biased to positive scores (i.e., active and strong, respectively). In the future, we will look for more conversations with low arousal and low dominance scores to balance the corpus (e.g., conversations expressing fear).

4.2. Inter-Evaluator Agreement

This section compares the individual annotations of the MSP-Conversation corpus. We use Cronbach's Alpha [28] to estimate inter-evaluator agreement, which gives a measure of consistency. This metric considers that two raters agree not only if their ratings are similar in value, but also if the ratings increase or decrease at the same rates (i.e., similar trends). We calculate the overall agreement for the emotional attributes and the evaluators. For the attributes, we calculate the agreement for each of the annotated segments, reporting the average scores per emotional attribute. For an annotator, we compare the agreement across traces by including and excluding his/her annotations. The difference between these scores indicates how much his/her annotations improve or worsen the general agreement.

Table 2 lists the agreement results, which shows that the highest agreements are for valence, and the lowest agreements are for dominance. While the agreement values are not high, we would like to highlight that annotating emotion is a complex process, especially with time-continuous annotations. Therefore, inter-evaluator agreements reported in emotional databases are often low [14, 17, 29, 30]. The tables also show different agreements across the annotators, where evaluators 4 through 12 consistently increase the agreements with their traces. The agreement performance can be considered while combining the traces, weighing unreliable workers less. Notice that we have shown that there is value in extra annotations even if they have less reliability [31].

4.3. Time-Continuous Versus Sentence-Level Annotations

The overlap between the MSP-Podcast and MSP-Conversation corpora offers the opportunity to compare sentence-level annotations with time-continuous annotations. This analysis relies on the 1,567 speech turns that overlap the two corpora (Sec 3.5). To compare the annotations, the aggregated time-continuous annotations are averaged in time during the duration of the speaking turns of the MSP-Podcast annotations. Figure 3 illustrates this process for three speech segments, which results in sentence-level scores from the traces. The analysis accounts for the reaction lag of the annotator (i.e., time between the annotator listens the audio, judges the emotional content, and moves the joystick [18, 19]). We consider lags equal to 2.8, 3.0, 3.6, 4.08, 5.44, and 5.6 seconds. The lags that are non-integer correspond to optimal delays found in previous studies for different attributes [18, 32].

We estimate the Pearson correlation coefficient between the

Table 3: *Average correlation of the sentence-level annotations from the MSP-Podcast corpus and the labels derived from the traces of the MSP-Conversation corpus.*

Lag (s)	Arousal	Valence	Dominance
0.00	0.312	0.280	0.222
2.80	**0.373**	0.378	**0.273**
3.00	0.368	0.378	0.271
3.60	0.348	0.403	0.260
4.08	0.324	**0.403**	0.244
5.44	0.266	0.399	0.200
5.60	0.259	0.398	0.196

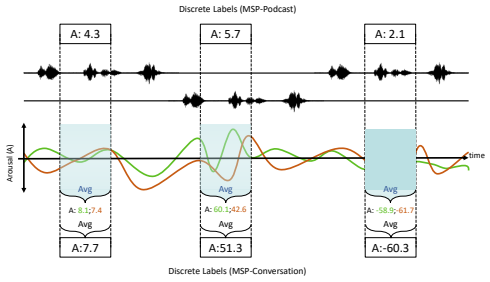


Figure 3: *Illustration of the process to derive sentence-level annotations from the traces. The figure also shows the sentence-level annotations from the MSP-Podcasts for three segments.*

sentence-level labels derived from the MSP-Conversation corpus and the sentence-level labels from the MSP-Podcast corpus. Table 3 shows the averages for each emotional attribute. The table shows higher correlation for valence and arousal. The correlations for dominance are consistently lower. Compensating for the reaction lag leads to higher correlations, where the optimal reaction lag depends on the emotional attributes (4.08s for valence, and 2.8s for arousal and dominance). The correlation coefficients between time-continuous traces and sentence-level labels never exceeded $\rho = 0.403$. While more analysis is needed, these results suggest that lack of contextual information may be the key difference, causing lower correlations.

5. Conclusions

This paper introduced the MSP-Conversation corpus, a speech emotional database annotated with time-continuous traces for the emotional attributes of arousal, valence and dominance. The corpus complements the MSP-Podcast corpus, providing the ideal resource to study temporal information in the externalization of emotions during multiparty interactions. By using recordings available on audio-sharing websites, the corpus provides natural speech interactions, from multiple speakers and with a broad range of emotions. The current version of the corpus has 74 conversations, collected from 197 different speakers. This is an ongoing effort, where our plan is to increase the size of the corpus. We have already identified 52 new conversations that we have started to annotate (28hrs 15min in total).

Our future research directions include the use of this corpus to analyze the role of contextual information in the study of emotion. The annotation of speaker diarization information is critical to evaluate inter-dependencies between speakers [5]. We want to share this corpus with the community. We have a *Federal Demonstration Partnership* (FDP) Data Transfer and Use Agreement for academic institutions interested on this corpus.

6. Acknowledgements

This work was supported by NSF under Grants CNS-1823166 and CNS- 2016719.

7. 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 human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[2] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[3] R. Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 127–131.

[4] M. Jaiswal and et at., "MuSE-ing on the impact of utterance ordering on crowdsourced emotion annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, United Kingdom, May 2019, pp. 7415–7419.

[5] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[6] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, January-June 2012.

[7] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[8] R. Kehrein, "The prosody of authentic emotions," in *Proceedings of the Speech Prosody*, Aix-en-Provence, France, April 2002, pp. 423–426.

[9] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.

[10] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

[11] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[12] J. Russell and L. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *Journal of personality and social psychology*, vol. 76, no. 5, pp. 805–819, May 1999.

[13] J. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, pp. 145–172, January 2003.

[14] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[15] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Journal of Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, September 2016.

[16] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

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

[18] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, April-June 2015, special Issue Best of ACII.

[19] ——, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.

[20] 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 Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[21] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013, pp. 1–8.

[22] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[23] L. Stappen and et al., "MuSe 2020 – the first international multimodal sentiment analysis in real-life media challenge and workshop," in *The Multimodal Sentiment in Real-life Media Challenge (MuSe 2020)*, Seattle, US, October 2020.

[24] J. M. Girard, "CARMA: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, pp. 1–6, July 2014.

[25] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 1556–1559.

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

[27] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[28] L. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, September 1951.

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

[30] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[31] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5190–5194.

[32] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, October 2016, pp. 3–10.