WHiSER: White House Tapes Speech Emotion Recognition Corpus

Abinay Reddy Naini[1], *Lucas Goncalves*[1], *Mary A. Kohler*[2], *Donita Robinson*[2], *Elizabeth Richerson*[2], *Carlos Busso*[1]

[1]Department of Electrical and Computer Engineering, The University of Texas at Dallas, USA
[2]Laboratory for Analytic Sciences, North Carolina State University, USA
{abinayreddy.naini,goncalves,busso}@utdallas.edu,{makohler,drobins7,richer}@ncsu.edu

Abstract

There are several applications for *speech-emotion recognition* (SER) systems in areas such as security and defense and healthcare. SER systems have achieved high performance when they are trained and tested in similar conditions. However, the performance often drops in more realistic and diverse conditions. Most existing SER datasets are too controlled and do not capture complex scenarios relevant to practical applications. This paper presents the *White House tapes speech emotion recognition* (WHiSER) corpus, which includes distant speech with real emotions from conversations in the Oval Office in 1972. This dataset is unique because it combines natural emotional expressions with various background noises, making it a perfect tool to test and improve SER models. Its real-world complexity and authenticity make the WHiSER corpus an excellent corpus for advancing emotion recognition technology, offering insights into how human emotions can be accurately recognized in complex environments.

Index Terms: emotional dataset, speech emotion recognition, expressive speech.

1. Introduction

The development of *speech-emotion recognition* (SER) systems has opened novel opportunities in diverse domains such as healthcare, customer satisfaction, security and defense, education, and entertainment [1–3]. SER systems must accurately detect emotional behaviors regardless of the context, target application, and recording conditions. The effectiveness of these systems relies on the quality and diversity of the datasets used to train the models. It is important to have challenging datasets to test the reliability of SER systems in complex environments that resemble real-world conditions. The existing emotional corpora that capture real human interactions are often recorded using high-quality recording devices with a noiseless environment [4]. However, most of the real-world applications involve different types of background acoustic noise [5, 6]. Therefore, we need more diverse SER datasets that capture the challenges that SER systems are expected to face if deployed in real applications.

Over the years, researchers have created many emotional speech datasets. Initial datasets relied mainly on actors reading a predefined set of sentences portraying targeted categorical emotions (happiness, anger, sadness) [7, 8]. However, research indicates that these emotions tend to be too exaggerated and do not represent the expressions observed in everyday human interactions [9]. A more natural approach is to simulate conversations between individuals to create more natural interactions [4, 10]. However, it is difficult, time-demanding, and expensive to collect a large database with this method. Few studies explored speech datasets generated from uncontrolled settings such as TV shows [11, 12], but these sources often resulted in biased emotional content because of the focus of the show (colloquial conversations in family-oriented programs, negative expressions in conflict-based shows). Recently, large emotional speech corpora have been developed by annotating the data available in media-sharing platforms [4, 13]. Even though these corpora provide adequate resources to develop strong SER systems, the recordings are still clean. We need emotional datasets recorded under real-time conditions including distance speech, reverberation, and background noise.

This paper introduces the *White House tapes speech emotion recognition* (WHiSER) corpus, obtained from the Nixon's Tapes recordings [14]. From 1971 to 1973, President Nixon recorded conversations in the Oval Office, which included authentic emotional interactions under varied recording conditions. The recordings were declassified and archived. They include distant speech, reverberation, and low *signal-to-noise ratio* (SNR), capturing relevant realistic conditions for SER systems. We rely on the pipeline proposed in the *affective naturalistic database consortium* (AndC) to select and annotate emotional recordings. The corpus comprises 6 hours and 21 minutes of data specifically annotated for emotional attributes (arousal, valence, and dominance), and categorical emotions (anger, sadness, happiness, surprise, fear, disgust, contempt, and neutral). The WHiSER database is intended to be used as an independent test set, in the context of the generalization of SER systems in complex environments. Its use can facilitate advances in unsupervised domain adaptation for SER [15–17]. All the data and the emotional labels are released on our GitHub repository. [1]

2. Related Work

Building the right infrastructure to develop computational models for emotions is critical. A common approach in early emotional databases is to record actors instructed to read sentences portraying target emotions. Several emotional datasets are developed using this approach, including the Emo-DB [8], RAVDESS [18], TESS [19], CREMA-D [20], and the Chen Bimodal [21] databases. However, Devillers et al. [22] and Batliner et al. [23] highlighted that such acted emotions do not accurately reflect real-life emotional complexity, suggesting that performances in acted scenarios do not match real-world application accuracies. The IEMOCAP [24] and MSP-IMPROV [25] databases were created to feature conversations in dyadic interactions, moving away from the scripted readings of previous datasets. Although these datasets tackled the problem of exaggerated emotions by presenting more natural dialogues, the recordings still involve actors. Other datasets such as VAM [26], SEMAINE [27], and TUM-AVIC [28] were built, sourcing natural interactions from TV shows and call centers. While

[1] https://github.com/msplabresearch/WHiSER

they moved towards capturing genuine interactions, the emotional content in these datasets became significantly skewed, reflecting the specific contexts of these interactions rather than the range of emotional content observed in everyday human interactions. The MSP-Podcast [4] corpus relied on recordings obtained from audio-sharing websites, enriching the emotional context beyond what acted datasets offer. Podcasts are often recorded with close-talking microphones, providing clean audio. There is a pressing need for emotional datasets that capture naturalistic recordings resembling the conditions expected to exist if SER systems are deployed in real applications. This observation is the motivation for the WHiSER dataset, which provides a valuable resource to comprehensively assess and enhance the capabilities of SER systems.

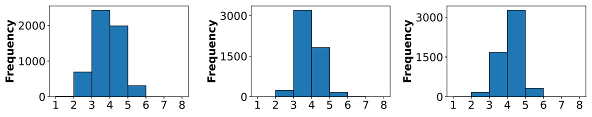
3. The WHiSER Corpus

The WhiSER dataset consists of speech files obtained from the declassified President Nixon's Oval Office recorded conversations from 1971 to 1973, referred to as Nixon's tapes [14]. These recordings cover significant events and meetings in key locations such as the Oval Office, Cabinet Room, the President's Old Executive Office Building, and Aspen Lodge at Camp David. It also includes recordings of the White House telephone calls. The recorded conversations in the dataset provide emotionally rich natural interactions using distant microphones, which makes it an ideal resource for emotional recognition tasks. We follow the pipeline suggested in the *affective naturalistic database consortium* (AndC) to select and annotate emotional recordings. This section details the process.

3.1. Pipeline and Data Annotation

The data collection pipeline's preprocessing stage is divided into two main phases: the audio preparation phase and the filtering phase. In the audio preparation phase, the raw audio collections are transformed into a consistent format using the Librosa toolbox [29] (16kHz, 16-bit, and single-channel). These recordings are then processed by a *voice activity detector* (VAD) to identify speech segments. In the filtering phase, the recordings undergo several filtering steps to enhance speech quality. This step includes automatic algorithms to determine the presence of music, using the approach proposed in Lee et al. [30], and noise, using the noise estimator proposed by Nicolson and Paliwal [31]. We considered 5dB SNR as a lower threshold for removing samples that are too noisy. Only the utterances that satisfy all predefined criteria are considered in the subsequent stages. In total, this set includes 121,482 sentences.

In natural conversations, most of the sentences are emotionally neutral. Therefore, the AndC pipeline involves automatically retrieving samples with machine-learning tools that are likely to convey emotional reactions. We collected emotionally rich samples from the full set of speech files in two stages. Initially, we employed four speech-emotion recognition models based on preference learning on emotional attributes. These models were initialized with the wav2vec2 large model [32] from the HuggingFace [33] library and were fine-tuned with the MSP-Podcast dataset, as detailed in [17]. From these chosen samples, we manually eliminated any unintelligible samples, regardless of their emotional content, ending up with 1,427 speech files to be annotated with emotional labels. In the second stage, we utilized 12 speech-emotion recognition models designed to predict attribute scores, which helped us select emotionally rich samples. These models were based on the state-of-the-art *self-supervised learning* (SSL) models (wavLM [34], wav2vec2 [32], HuBERT [35], Data2vec [36]), fine-tuned and trained with the MSP-Podcast dataset, as explained in [37]. We also removed any unintelligible or duplicate samples from the first stage, resulting in 4,000 speech files to be annotated with emotional labels. The WHiSER corpus incorporates all samples from both stages, totaling 5,427 speech files.



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

Figure 1: *Distributions for the emotional attributes scores for arousal, valence, and dominance.*

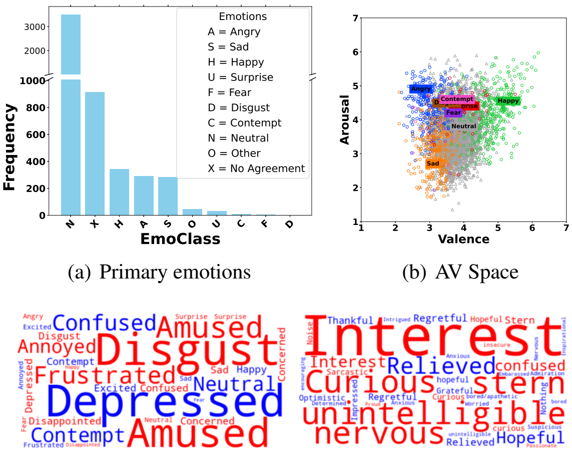
The selected utterances are annotated with emotional labels using a perceptual evaluation. During this phase, annotators label the utterances with the emotional dimensions of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong), using a seven-point Likert scale. We provide *self-assessment manikins* (SAMs) [38] as visual references to help annotators accurately assess these dimensional attributes. The annotations include primary categorical emotion, which is the class that most accurately represents the emotional content of the utterance, choosing from a list of eight primary emotions: anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N). The annotators also provide secondary emotions, including all the emotional classes perceived in the audio. In addition to the eight options, the list for the secondary emotions includes amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed. The corpus was annotated by 33 student workers at <anonymous>. Each speaking turn was annotated by at least five annotators. We use the plurality rule to obtain consensus for primary emotions. For each emotional attribute, we estimate the average scores assigned to each sentence by the annotators, using this value as the consensus score.

3.2. Size of the WHiSER corpus

The WHiSER corpus consists of 5,427 speech files with a total duration of 6 hours 21 minutes. The duration of speech samples within a dataset is a critical factor for the extraction of emotional content and subsequent labeling. Our dataset was developed to include speech files ranging from 3 to 11 seconds, as shown in Figure 3(b). This range was chosen based on previous studies indicating the challenges in perceiving the emotional content from very short audio clips [39]. Likewise, attributing a single emotional label to a long file leads to label noise as emotion can fluctuate within the sentence. We split longer speaking turns into smaller parts to fall within our target duration. Figure 3(b) reveals that the majority of samples are concentrated within the 3 to 6-second window.

3.3. Emotional Content Analysis

The sentences in the corpus were selected based on their predicted content in terms of emotional attributes. Figure 1(a), 1(b), and 1(c) show the distribution of emotional attribute scores for arousal, valence, and dominance, respectively. For each attribute, we observe a *Gaussian* distribution centered around four, which is the central point in the seven-point Likert scale. This distribution captures the expected emotional expressions displayed in daily human interactions, where extreme expressions are less common. The distribution presents a significant number of samples around the scores of three and five. Sridhar and Busso [40] observed a higher level of uncertainty in



(a) Primary emotions (b) AV Space

(c) Secondary Emotions (d) Other Emotions

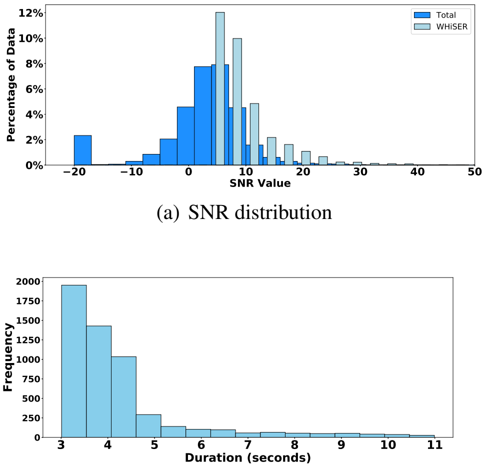
Figure 2: *(a) Distribution of primary emotions, (b) distribution of primary emotions in the arousal-valence space, (c) word cloud representing all secondary emotions, and (d) word cloud representing all the emotions selected for "other".*

models predicting emotional attributes within this range. These sentences on the scale represent emotions that are neither too subdued nor too intense, resulting in the most challenging sentences to accurately recognize. The WHiSER corpus adds a layer of complexity, highlighting the dataset's capacity to challenge SER models and encouraging the development of sophisticated models capable of differentiating across subtle emotions.

The emotional distribution within a dataset can significantly affect its efficacy as a training or testing resource. Figure 2(a) shows the distribution of primary emotions across the dataset. The dataset showcases a substantial portion of 'Neutral' emotional data, with 64.3% instances. This neutrality mirrors the authentic emotional representation of day-to-day conversations, which often skew towards a non-extreme emotional display. Figure 2(b) plots the placement of the sentences assigned to each emotional class in the arousal-valence space. This figure illustrates the variability of the emotional content included within the same classes. We notice that even within the class 'Neutral,' we observe a range of emotional expressions perceived in the speech. Figure 2(a) shows that the dataset exhibits an almost even distribution among 'Happiness', 'Sadness', and 'Anger.' This balanced representation of major emotions makes the dataset an effective tool for testing SER models and provides the ability to distinguish between different emotional states without bias. The frequencies of these categorical emotions validate the dataset's potential to serve as a balanced benchmark for SER systems, ensuring they are well-versed in detecting and interpreting a spectrum of emotions that are commonplace in real-world interactions. Figures 2(c) and 2(d) show word clouds for secondary emotions and terms included when the annotators selected 'other' as the secondary emotions.

3.4. Intended Use of the WHiSER Corpus

The intended use of the WHiSER corpus is to serve as a test set where the generalization of SER models can be assessed. For this purpose, we do not provide train, development, and test sets for this corpus. The WHiSER corpus is ideal for testing *unsupervised domain adaptation* (UDA) strategies aiming to reduce the mismatch between train and test domains.



(a) SNR distribution

(b) Duration

Figure 3: *(a) SNR distribution of the total set before filtering and WHiSER corpus, (b) Distribution of duration (seconds) of all speech files in the dataset.*

4. Speech Emotion Recognition Evaluation

The WHiSER corpus was developed to create a challenging dataset that closely mirrors the real-world testing conditions for SER systems. This section evaluates the effectiveness of SER models using the WHiSER corpus for predicting emotional attributes (arousal, valence, dominance), and classifying emotional categories (happiness, anger, sadness, neutral). We consider models relying on the LLS wavLM, wav2vec2, and HuBERT. These models provide a solid foundation, having been trained on extensive and diverse datasets. We fine-tuned these large representations using the MSP-Podcast dataset using a downstream head consisting of a simple *deep neural network* (DNN). We refer to the fine-tuned models as the *large-robust* (LR) version of the corresponding SSL model. To extend our evaluation, we incorporate UDA techniques to investigate the adaptability of a model trained on a different, sizable corpus. We considered the MSP Podcast corpus as the source dataset and the WHiSER corpus as the target dataset. The first UDA strategy in this study is using the ladder network [15, 41]. It uses an encoder-decoder model with added noise at intermediate representations. The goal is to reconstruct the intermediate representation of the encoder using ladder connections. The cost function has two parts: one for the classification task at the encoder's end, obtained using the fully labeled source data, and another for the reconstruction loss, obtained for both source and unlabeled target data. We refer to this UDA approach as *domain adaptation using ladder network* (DAL). The second strategy is the *adversarial domain adaptation* (ADA), as used by Abdelwahab and Busso [16], which has a domain classifier as an auxiliary task. The approach employs a *gradient reversal layer* (GRL) to make the domain classifier achieve random performance, hence, forcing the shared feature representation layers to generate similar responses for both domains. This method considers task and domain losses for the source domain (labeled data), and only the domain loss for the target domain (i.e., unlabeled target data), alternating between the two data types. These approaches allowed us to explore the extent to which the SER models could adjust to the unique characteristics of the WHiSER dataset without direct supervision, thereby assessing their potential for generalized, cross-domain applications in real-world scenarios. We refer to this UDA method

Table 1: concordance correlation coefficient *(CCC) for arousal, valence, and dominance. The symbol *† indicates that the result is a significant improvement over the result with the symbol *, which is a significant improvement over the results without any symbol.*

CCC	LR	ADA	DAL	Within
		Arousal		
wavLM	0.299	0.396*	0.391*	**0.418***†
wav2vec2	0.301	0.391*	0.384*	**0.412***†
HuBERT	0.292	0.385*	0.377*	**0.408***†
		Valence		
wavLM	0.392	0.441*	0.446*	**0.483***†
wav2vec2	0.395	0.457*	0.461*	**0.493***†
HuBERT	0.383	0.450*	0.453*	**0.476***†
		Dominance		
wavLM	0.338	0.369*	0.379*	**0.387***†
wav2vec2	0.326	0.371*	0.376*	**0.391***†
HuBERT	0.317	0.365*	0.361*	**0.384***†

Table 2: *Comparison of within and cross corpus performance for 4 primary emotions classification. The table reports the average F1 (Macro), UAR (%) values. (* indicates that the model is significantly better than the other three methods)*

	LR		ADA		DAL		Within	
	F1	UAR	F1	UAR	F1	UAR	F1	UAR
wavLM	.613	63.9	.621	64.2	.627	65.3	**.676***	**69.2***
wav2vec2	.592	61.7	.601	61.9	.609	62.6	**.652***	**66.5***
HuBERT	.608	62.8	.604	62.7	.617	63.4	**.661***	**67.4***

as *adversarial domain adaptation* (ADA). To understand the best performance that could be achieved on the WHiSER corpus, we considered within-corpus settings as the fourth method. For the within-corpus settings, we considered a five-fold cross-validation of the corpus, to train the LR models directly on the WHiSER corpus using a downstream head. In each fold, we use data from four partitions as a train set and one as a test set.

For wav2vec2, we fine-tuned the LR model by removing the top 12 transformer layers, which retains performance with fewer parameters [42]. The SER architecture for both fine-tuning and training steps involves two layers with 1,024 nodes each, layer normalization, and ReLU activation. For emotional attributes, the average *concordance correlation coefficient* (CCC) from the three attributes (arousal, valence, and dominance) served as the loss function while fine-tuning the models. For training the SER model, we use three different models, one for each attribute, where the weights are individually set to optimize the performance of the target attribute. For emotional classes, we used the cross-entropy loss for fine-tuning and training stages (a four-class problem). ADA employs a 128-node layer for task and domain classifiers with ReLU activation, while DAL includes two 256-node layers with ReLU and linear activations for the task classifier. We utilize an EC2 g5.4xlarge instance for fine-tuning the models with an NVIDIA A10G GPU with an Adam optimizer at a learning rate of 10e-5. For all other tests, we use an NVIDIA GeForce RTX 3090 GPU.

Table 1 shows the results for the emotional attributes, reporting performance in terms of CCC. We evaluate if the results are statistically significant by dividing the WHiSER corpus data into 20 subsets of similar size. Then, we conducted a two-tailed t-test over the 20 subsets. We defined statistical significance at a p-value<0.05. When we compare the LR results with the ADA and DAL approaches, we observe significant improvements by using UDA strategies. However, the table shows that the within-corpus setting achieves significantly better results than all other cases. There are opportunities to explore better UDA strategies to bridge this gap.

Table 2 shows the classification performance for categorical emotions. We reported the results using F1-score (macro) and *unweighted average recall* (UAR) metrics. Similar to the experiments with emotional attributes, we observed a significantly better result in the within-corpus setting compared to all other cases. In most cases, UDA strategies produced significantly better performance compared to the LR model. Overall, among the three SSL methods considered, wavLM produced better results in most cases with few exceptions in both experiments.

5. Conclusions

The paper presented the *White House tapes speech emotion recognition* (WHiSER) corpus consisting of authentic and diverse emotional content from President Nixon's Oval Office recordings. The corpus closely aligns with real-world challenges faced by SER systems when deployed in practical applications. The WHiSER corpus, with its natural emotional expressions and varied background noises, produces a challenging testing scenario for SER models. Through the detailed analysis of the dataset's emotional content and the evaluation of SER models using this corpus, we have underscored the potential of the WHiSER corpus to serve as an invaluable asset for both testing and improving emotion recognition systems. By incorporating scenarios that feature distant speech, reverberation, and low signal-to-noise ratios, the WHiSER corpus bridges the gap between controlled laboratory conditions and the nuanced, unpredictable nature of real-life settings. Future work will aim to explore innovative approaches to harness the full potential of the WHiSER corpus in advancing deployable SER technology.

6. 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, Nov. 2013, pp. 110–127.

[2] J. Acosta, "Using emotion to gain rapport in a spoken dialog system," Ph.D. dissertation, University of Texas at El Paso, El Paso, TX, USA, December 2009.

[3] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, Nov. 2021.

[4] 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.

[5] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3593–3597.

[6] C. Huang, , G. Chen, H. Yu, Y. Bao, and L. Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

[7] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Philadelphia, PA, USA, 2002, Linguistic Data Consortium.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[9] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, April 2003.

[10] 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, Dec. 2008.

[11] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "Memor: A dataset for multimodal emotion reasoning in videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 493–502.

[12] S. Poria *et al.*, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536.

[13] S. Upadhyay *et al.*, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, Sept. 2023, pp. 1–8.

[14] "Oval 741-2; june 23, 1972; white house tapes; richard nixon presidential library and museum, yorba linda, california." 1972. [Online]. Available: https://www.nixonlibrary.gov/index.php/white-house-tapes

[15] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[16] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[17] A. Reddy Naini, M. Kohler, and C. Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[18] S. R. Livingstone and F. A. 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, p. e0196391, 2018.

[19] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (tess)," *Scholars Portal Dataverse*, vol. 1, p. 2020, 2020.

[20] 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.

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

[22] 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.

[23] 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.

[24] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[25] C. Busso *et al.*, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[26] 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.

[27] 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.

[28] B. Schuller *et al.*, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, Nov. 2009.

[29] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[30] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," 2017.

[31] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, Dec. 2020, pp. 12 449–12 460.

[33] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[34] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, July 2021.

[35] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[36] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning (ICML 2022)*, K. Chaudhuri *et al.*, Eds. Honolulu, HI, USA: Proceedings of Machine Learning Research (PMLR), July 2022, vol. 162, pp. 1298–1312.

[37] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, Seoul, Republic of Korea, April 2024.

[38] L. Fischer, D. Brauns, and F. Belschak, "Zur messung von emotionen in der angewandten forschung," *Beiträge zur Wirtschaftspsychologie*, 2002.

[39] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach," *Advances in Human-Computer Interaction*, vol. 2010, pp. 1–15, January 2010.

[40] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 8384–8388.

[41] A. Rasmusi *et al.*, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[42] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, September 2023.