ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition

Yuanchao Li[†], Zeyu Zhao[†], Ondřej Klejch, Peter Bell, Catherine Lai

Centre for Speech Technology Research, University of Edinburgh, UK

yuanchao.li@ed.ac.uk

Abstract

In Speech Emotion Recognition (SER), textual data is often used alongside audio signals to address their inherent variability. However, the reliance on human annotated text in most research hinders the development of practical SER systems. To overcome this challenge, we investigate how Automatic Speech Recognition (ASR) performs on emotional speech by analyzing the ASR performance on emotion corpora and examining the distribution of word errors and confidence scores in ASR transcripts to gain insight into how emotion affects ASR. We utilize four ASR systems, namely Kaldi ASR, wav2vec2, Conformer, and Whisper, and three corpora: IEMOCAP, MOSI, and MELD to ensure generalizability. Additionally, we conduct text-based SER on ASR transcripts with increasing word error rates to investigate how ASR affects SER. The objective of this study is to uncover the relationship and mutual impact of ASR and SER, in order to facilitate ASR adaptation to emotional speech and the use of SER in real world.

Index Terms: speech recognition, speech emotion recognition, wav2vec2, Conformer, Whisper, confidence measure

1. Introduction

Speech Emotion Recognition (SER) has rapidly developed over the past two decades, and the research on the corpora, features, algorithms, and training models has boomed [1]. Emotion in speech is conveyed both by what is said and how it is said. The former is generally captured as text (i.e., speech transcripts). Recent work on SER consistently shows the benefits of incorporating both textual and acoustic features on benchmark corpora [2, 3]. Although some researchers have also attempted to translate this approach into the wild, e.g., in-car voice systems [4], it is still rare to see SER applications in our daily lives. One of the major reasons is that the majority of SER research uses human annotation, i.e., gold-standard manual transcripts. In contrast, even for the 'lab' emotion corpora (e.g., IEMOCAP), transcripts from a state-of-the-art Automatic Speech Recognition (ASR) system can result in Word Error Rates (WERs) higher than 35%, depending on the emotion label [5]. This means that very few of the findings obtained in the lab can be replicated in the wild.

To bridge this gap, it is necessary to exploit the imperfect textual data generated by ASR for emotion recognition. While previous studies proposed using ASR transcripts for SER [5, 6], the effect of ASR on emotion and vice-versa is not always clear. For example, a high WER on emotional speech is generally assumed to be a result of the distortion relative to neutral speech [7]. However, [5] find neutral speech can in fact have a higher WER than angry speech, depending on utterance length. Thus, even though ASR is a relatively well studied area, key questions about how it can be used in SER remain unresolved, such as 1) *how confident can we be of ASR on different emotions?* and 2) *how is the SER performance affected by ASR errors?* Thus, to make true progress in SER, we need to understand the interrelationship between ASR and SER.

In this work, we look at these fundamental yet longstanding issues by exploring the mutual impact of ASR and emotional speech, with the objective of pushing SER research closer to realistic use scenarios. Specifically, we first take a deep look into the word distribution of emotion corpora and how they are misrecognized in speech recognition. We analyze WER variation to see how ASR performance is affected by different emotions and how this relates to word-level confidence scores. Finally, we produce several ASR transcripts with progressively increased WER to investigate how SER is affected by ASR performance. Our results demonstrate that: **1)** the ratio of words in terms of the Part of Speech (PoS) tag, affective scores, and utterance length can greatly affect ASR performance. **2)** Differences with respect to these characteristics in emotional speech lead to differences in ASR performance bias transcripts, and hence SER. **3)** The relationship between SER and ASR confidence scores requires careful analysis as their interpretation varies between ASR confidence with and without calibration. **4)** WER can greatly affect SER performance in a way that has implications for the use of real-life SER.

2. Related Work

Although the relationship between ASR and SER is an important topic, there is very limited work using ASR transcripts for SER. [8] evaluated the emotion recognition performances with an ASR engine in the loop. Their experiments showed surprisingly low performance degradation ($<3\%$) with real ASR transcripts over ground-truth transcripts based emotion recognition. [6] used two commercial ASR systems to generate transcripts for bimodal SER (audio + text), resulting in a relative loss of 4% and 5.3% in unweighted accuracy compared to ground-truth transcripts, respectively. [5] found emotion classes with low WER typically have a low ratio of short utterances to long utterances, where the shorter the utterance, the higher the WER. They also showed that hierarchically fusing ASR hidden states and transcripts can achieve similar performance as ground-truth text. [9] proposed the use of a confidence measure to adjust the importance weights in ASR transcripts according to the likelihood of a speech recognition error in each word. This mitigated the effects of ASR error on SER performance, achieving about 4% accuracy increase compared to a baseline work without confidence measure.

[†]Equal contribution.

These findings suggest it is possible to utilize ASR for SER, but they still have limitations. [5] only investigated utterance length but did not take other factors into consideration. [6, 8, 9] did not investigate individual emotions, though ASR may behave differently on them. Furthermore, none of these studies have been conducted thoroughly at the word level, and their findings cannot be generalized to other ASR models and corpora. Thus, the interrelationship between ASR and SER is still unclear, and gaps remain on foundational issues such as how ASR confidence varies with different emotions and how word distribution affects ASR performance. Therefore, we perform a word-level investigation on emotional corpora using multiple ASR models, conducting new analyses on ASR performance, confidence, and SER performance, towards understanding the mutual impact of speech and emotion recognition to best bring ASR into SER research.

3. Emotion Corpora and ASR Models

We use three corpora: IEMOCAP [10], MOSI [11] and MELD [12] to cover as many speech conditions as possible for generalizability. IEMOCAP (IEM) consists of five sessions of scripted and improvised dialogues in the research lab. We use four emotion classes: angry, happy (+excited), neutral, and sad, following previous research [13]. MOSI has 2,199 monologue video clips from YouTube, annotated with sentiment scores in the range of [-3, 3]. MELD contains acted dialogue from the TV series *Friends*. Each utterance has been labeled with one of seven emotion classes: anger, disgust, fear, joy, neutral, sadness, and surprise. For IEMOCAP and MELD, we removed utterances whose transcript is blank or whose audio file is too long due to mis-processing in corpora construction, bringing the total utterance numbers to 5,500 and 13,689, respectively.

For the ASR models, we use the Kaldi Librispeech ASR model[1] (KLA), a self-supervised model: wav2vec2-base-960h[2] (W2V2), a Conformer model[3] (CONF) from ESPnet [14], and the medium model of Whisper[4] (WHIS). KLA, W2V2, and CONF are pre-trained on Librispeech960, and WHIS is on 680,000 hours of multilingual and multitask data collected from the web. We use these four models for generalizability.

4. Experiments and Analyses

In this section, we analyze the relationship between ASR and emotion by conducting a series of experiments to identify sources of errors across corpora and word classes and quantifying the effect of WER on SER performance.

4.1. WER Across Corpora

Table 1: *WER (%) of the ASR models on the emotion corpora.*

	IEM	MOSI	MELD
KLA	36.8	40.9	58.5
W2V2	32.7	35.4	57.8
CONF	27.1	30.1	52.1
WHIS	12.3	17.3	34.8

We compute the WERs of the four ASR models on the three emotion corpora as the basis for analysis, as shown in Table 1.

Among the four models, WHIS shows the best performance on all the corpora, significantly outperforming the other three. We see large performance differences among corpora, even though they are all US or UK English emotional speech. This could be due to the different recording settings, the noise that occurred, or the vocabulary that was said.

4.2. PoS and Affective Score Effects on ASR

In addition to the well-known factors that affect ASR (e.g., recording quality, noise), we would like to know if word distribution also affects WER. To investigate this, we analyze the words in each corpus based on two aspects: Part-of-Speech (PoS) tag and affective score. In particular, we explore whether words of a particular class or with strong affective coloring have a strong impact on ASR performance.

PoS. We select seven classes of PoS tags: *Noun, Verb, Adjective, Adverb, Wh-word, Function-word* from the Penn Treebank [15] and *Stop-word* from the stopwords corpus of NLTK [16]. The tagging was also conducted using NLTK.

Affective score. We use an affective words database that has nearly 14 thousand English words [17] to refer the affective scores. Each word is rated in three affective dimensions: Valence (V), Arousal (A), and Dominance (D), ranging from 1 (weak) to 9 (strong). We divide the scores into three classes: low (1-3), medium (4-6), and high (7-9) according to their overall mean value, respectively.

We calculate three metrics to see how different classes of words affect the ASR performance. The Word Ratio (WR) is used to see the distribution of different classes of words in a corpus. The Error Ratio (ER) indicates if misrecognized words largely come from a specific class. The Class Error Rate (CER) measures how difficult words in a class are to recognize. We define them as follows.

$$\text{word ratio} = \frac{\text{word count per class}}{\text{total word count}} \quad (1)$$

$$\text{error ratio} = \frac{\text{word error count per class}}{\text{total word error count}} \quad (2)$$

$$\text{class error rate} = \frac{\text{word error count per class}}{\text{total word count per class}} \quad (3)$$

The WR[5], ER, and CER based on PoS tag and affective score are shown in Table 2 and 3, respectively. For brevity, we averaged the values over the four models. As the ER values from the four ASR models are close and the CER values are proportionally distributed, taking the average does not affect the general findings.

Table 2: *WR, ER, and CER (all in %) based on PoS tag.*

	IEM			MOSI			MELD		
	WR	ER	CER	WR	ER	CER	WR	ER	CER
Noun	19.1	34.8	33.3	21.7	36.8	45.3	25.3	44.0	58.0
Verb	22.5	10.6	26.0	19.3	8.4	25.9	20.3	8.2	34.5
Adj	5.7	2.0	19.4	10.6	3.5	27.5	5.7	2.0	32.1
Adv	8.6	7.0	25.6	8.6	5.1	23.5	7.6	5.9	38.8
Wh	2.1	1.4	26.0	1.7	0.7	21.0	2.4	1.8	44.1
Func	19.2	12.1	20.1	25.1	15.6	24.1	17.5	10.2	30.3
Stop	50.7	33.7	25.7	50.0	30.0	24.8	46.8	29.5	33.7

Table 2 shows that ER generally increases with WR. This is reasonable, as the more words in one class, the more errors will be generated. However, this does not hold for all PoS classes:

---

the WRs for *Noun*, *Verb* and *Func* are similar, but the ER is much lower for *Verbs*. That is, ASR performs better on verbs than expected based on verb frequency. On IEM, *Verb* has a higher WR than *Noun* and *Func* but much lower ER, and similarly for *Adj* and *Adv* on MOSI. This indicates that in ASR models, some classes of words are more difficult to recognize than others. The *Noun* class is the most difficult to recognize, showing the highest CERs across the corpora. However, other word classes do not have uniform patterns across the corpora. For instance, *Wh* words are the easiest to recognize on MOSI yet the second hardest on MELD.

The CERs of most word classes increase from IEM to MELD, which is in line with the change of WERs in Table 1. This is plausible as the word error count increases with the WER. Nevertheless, the rate of growth of CER varies from class to class. For example, *Noun* saw a 74% increase in CER (33.3 to 58.0) compared to 33% for *Verb* (26.0 to 34.5). Again, this reflects the fact that word classes have different ERs, perhaps because they have different tolerances for speech distortion.

Table 3: *WR, ER, and CER (all in %) based on affective score.*

	IEM			MOSI			MELD		
	WR	ER	CER	WR	ER	CER	WR	ER	CER
V_{low}	2.5	3.6	36.1	2.3	3.6	43.2	2.0	2.9	43.0
V_{mid}	10.8	13.8	29.6	9.3	11.7	33.9	9.8	14.2	37.7
V_{high}	13.9	16.2	26.3	15.6	18.9	34.3	12.2	16.2	35.5
A_{low}	15.0	17.4	24.8	12.3	13.8	30.9	13.2	17.3	35.0
A_{mid}	11.1	14.7	30.7	13.6	17.5	36.5	10.1	15.1	39.4
A_{high}	1.0	1.4	32.8	1.3	2.0	43.6	0.7	1.0	40.0
D_{low}	1.0	1.4	28.1	1.1	1.8	47.2	0.8	1.2	38.9
D_{mid}	16.7	20.5	27.7	15.6	20.0	36.2	15.0	20.9	36.9
D_{high}	9.4	11.5	27.4	10.6	11.5	29.2	8.1	11.2	36.7

Looking at the affective score-based distribution in Table 3 and again see that, in general, the higher the WR, the higher the ER. However, in this case, we don't see classes with similar WRs but very different ERs. Nevertheless, differences still remain: CER decreases as valence and dominance scores increase, i.e., high valence and dominance words are better recognized. In contrast, we see that higher-arousal words are more difficult to recognize. Even so, unlike the classes based on PoS tags, the CERs have a relatively balanced distribution here, and their rates of increase with the WER are at the same level.

We see similar but not identical patterns when we look at the results for individual ASR models. We omit detailed results for brevity but note per-class differences in performance, e.g., *Nouns* are better recognized by KLA (32% ER) compared to CONF (35% ER) even though the latter has a lower overall WER. In general, no one model performed the best across all PoS or affective score classes.

4.3. Utterance Length and ASR Performance

Another factor that is likely to have an effect on ASR performance is utterance length. Generally, longer utterances are more likely to have more word errors compared to shorter ones because there are more opportunities for the ASR model to make errors [6]. In fact, [5] showed that longer utterances in the transcripts generated by W2V2 have lower WERs compared to shorter ones, hypothesizing that longer utterances contain more contextual information that W2V2 can use to compensate for the speech distortion. However, it is unknown whether other ASR models have the same behavior, so the general impact of utterance length on ASR performance is not yet clear.

To investigate this, we analyze how WER and WR change with utterance length. The four ASR models displayed the same

trends, so Table 4 shows results for W2V2 for brevity. We see that the shorter the utterances, the higher the WER, regardless of corpus. These support the findings of [5] and additionally show that if a corpus contains many short utterances, ASR may not work well.

Table 4: *WER and word ratio (both in %) according to different utterance lengths (number of words N).*

	IEM		MOSI		MELD	
N	Ratio	WER	Ratio	WER	Ratio	WER
≤ 10	59.6	46.4	53.9	40.6	71.5	73.3
11-20	22.9	32.2	32.2	34.5	23.5	48.6
21-30	10.8	26.3	9.5	31.9	4.4	42.1
≥ 30	6.7	25.0	4.4	32.6	0.6	38.8

4.4. ASR Performance on Different Emotions

It has long been argued that the acoustic characteristics of emotional speech, e.g., prosody variation, can deteriorate ASR performance [18]. Word-level factors, however, are rarely studied. Based on the findings from the above experiments, we conducted a word-level analysis based on different emotions. Specifically, we look at the WER alongside the ratio of *Noun* and short utterance (word count less than 10), which we found to be major factors affecting ASR performance above.

From Table 5, we can see the WER on Neutral is not the best but instead worse than most of the other emotions in both corpora, which supports the finding in [5]. As neutral speech is the least emotionally 'distorted', the reason could be that the ratio of short utterances in Neutral is very high, which offsets the benefit of less emotional distortion. We did not see a large discrepancy in the ratio of *Noun* between emotions, although we found that a high ratio harms ASR performance (Sec 4.2).

Table 5: *WER, ratio of Noun words and short utterances (all in %) according to different emotions on IEM and MELD.*

IEM	Ang	Hap	Neu	Sad
WER	22.8	38.9	36.3	29.5
Noun	20.0	21.3	20.4	19.5
≤ 10	44.8	55.7	60.5	52.7

MELD	Ang	Dis	Fea	Joy	Neu	Sad	Sur
WER	52.7	53.4	58.6	59.9	58.3	52.9	65.3
Noun	26.2	27.6	26.4	29.7	26.9	25.2	24.5
≤ 10	64.1	60.1	62.1	72.4	72.8	62.0	82.8

4.5. Emotion and ASR Confidence

Confidence scores can help identify words of high recognition quality for better use of ASR transcripts. [9] and [19] adjusted the attention weights for words using their confidence scores. [20] proposed removing words with low confidence and selecting words with the highest confidence from multiple ASR hypotheses [21]. However, the measurement and use of confidence scores is a complicated task due to out-of-domain words, overconfidence, and so on [22, 23]. Hence, we analyze how confidence varies with emotion by exploring the word confidence scores. We compared two ASR models–CONF and KLA, and also compared the confidence scores generated by KLA with and without calibration. The calibration was applied by using the official procedure in Kaldi[6]. Since we noticed similar patterns of the confidence in all three corpora, we only show the results of IEM for brevity.

---

[6] https://github.com/kaldi-asr/kaldi/blob/master/egs/tedlium/s5/local/confidence_calibration.sh

Table 6: *Distribution of confidence scores of the ASR models.*

	N	CONF	KLA	KLA (calibrated)
Correct		0.97	0.95	0.62
Incorrect		0.87	0.70	0.53
Ang	≤10	0.96	0.91	0.62
	11-20	0.96	0.91	0.61
	21-30	0.96	0.90	0.61
	≥30	0.96	0.90	0.61
Hap	≤10	0.92	0.80	0.57
	11-20	0.94	0.83	0.58
	21-30	0.95	0.86	0.59
	≥30	0.94	0.85	0.58
Neu	≤10	0.91	0.81	0.58
	11-20	0.94	0.86	0.59
	21-30	0.95	0.88	0.59
	≥30	0.96	0.89	0.60
Sad	≤10	0.92	0.81	0.58
	11-20	0.95	0.87	0.59
	21-30	0.96	0.89	0.60
	≥30	0.97	0.92	0.61

In Table 6, we noticed that there is a big difference in confidence scores between ASR models. The scores of correctly recognized words in both CONF and KLA are close, but the score for incorrectly recognized words is 0.70 in KLA and 0.87 in CONF. The scores of KLA are smaller than those of CONF in general. Besides, the score increases with utterance length in Ang, Neu, and Sad. The longer the utterance, the higher the confidence, which is in line with the change in WER in Table 4. But this trend does not occur in Ang. Note that the score of every emotion increases with utterance length in MELD but barely changes in MOSI, which means using raw confidence scores may not be appropriate. By comparing the confidence with and without the calibration for KLA, we can see that the difference in confidence between correctly and incorrectly recognized words becomes smaller. Similarly, the difference in utterance length almost disappears, which means emotions are more equally treated after calibration. Calibration also adjusts the overconfident behavior of end-to-end ASR models [24]. The overall confidence of each emotion is in line with their respective WER in Table 5 that ASR performance is the best on Ang but worst on Hap. These findings demonstrate that confidence scores vary according to different emotions, utterance lengths, and ASR models. This needs to be taken into account if scores are used to adjust word contributions in SER.

4.6. The Impact of ASR on SER Performance

So far, we have investigated emotion corpora to understand the impact of emotion and related word distributions on ASR. Here, we explore the impact of ASR on emotion recognition by computing the SER performance of using ASR transcripts with progressively increased WER (focusing on IEM for brevity).

First, we used all four models to generate four ASR transcripts with their respective WER shown in Table 1 and divided every transcript into five parts according to the five sessions required by the 5-fold cross-validation training standard of SER on IEM [13]. Thus, we obtained four sub-transcripts of each session. Next, for each session, we produced ten sub-transcripts with WER from 5% to 50%, with an increase step of 5%, by randomly replacing a portion of sentences in a sub-transcript (considering different ASR model performances and the speech distortion in real use, we set the highest WER to 50%). The replacement was done as follows: sentences in transcripts with higher WER replaced those of the same IDs in transcripts with lower WER. If the required WER is outside the range of the

WERs of the sub-transcripts, we used ground truths (for low WERs such as 5%) or worse ASR hypotheses from the CONF model (for high WERs such as 50%) as replacements to obtain the desired overall WER (we generated 30-best ASR hypotheses using CONF). The proportion of replaced sentences was kept as low as possible in order to preserve the transcripts in their original form as real ASR output.

We trained the SER model on ground truth transcripts from four sessions and tested it on the ASR transcripts from the remaining session (5-fold cross-validation). We used a BERT model [25] to extract textual features. The emotion model comprises two bi-directional LSTM layers (hidden state of 32), a self-attention layer (hidden state of 64 and head number of 16), a dense layer (hidden state of 64) with *ReLU* activation and one output layer with *Softmax* activation. The learning rate is set as 5e-4 with *AdamW* optimizer, the weight decay is set as 1e-5, and the batch size is 64. We trained the models using 150 epochs and reported the best unweighted accuracies.
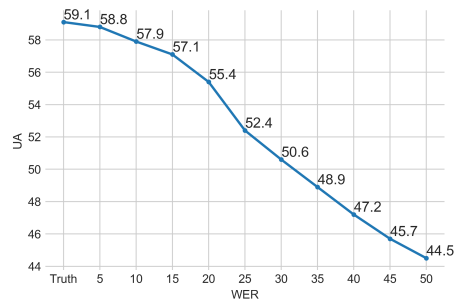

Figure 1: *SER accuracy with increasing WER (both in %).*

Figure 1 shows how SER performance varies with WER. We see that WER does affect the SER accuracy. The higher the WER, the lower the accuracy. The accuracy drop is small at first and gradually becomes larger, with a steep drop between 15% and 25%. Unlike the finding in [8] that WER over 30% only results in an accuracy drop smaller than 3% on the FAU Aibo Emotion Corpus [26], we found that WER can greatly affect SER performance. This may be because we used a different corpus or a different language model (BERT vs. Bag-of-Words). In the future, we will explore how different language models function on imperfect ASR transcripts.

5. Conclusions

In this paper, we address a long-existing yet understudied issue: the relationship between ASR and emotion. We demonstrate that word distribution in a corpus is a key factor that affects ASR performance and that differences exist between emotions. The frequencies of different PoS tags, affective scores, and utterance lengths may have a large impact on ASR performance. For example, the more the *Noun* words and short utterances, the higher the WER, regardless of the speaking style (speech collected in the lab, from monologues or TV shows). Moreover, different ASR models behave similarly in these trends, but with minor differences. Some classes of words are easier to recognize for one model but maybe harder for another. Moreover, the confidence scores generated by different models may show large differences, so the use of confidence scores should be carefully examined. Finally, we show that SER accuracy can greatly degrade with high WER. We expect our findings to pave the way for future SER research using imperfect ASR transcripts.

6. References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] J. Sebastian, P. Pierucci *et al.*, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," in *Proceedings of Interspeech 2019*, 2019, pp. 51–55.

[3] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[4] Y. Li, "Feeling estimation device, feeling estimation method, and storage medium," Aug. 31 2021, US Patent 11,107,464.

[5] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

[6] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription," in *Proceedings of Interspeech 2019*, 2019, pp. 3302–3306.

[7] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[8] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: putting asr in the loop," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4585–4588.

[9] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, and T. Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure." in *Interspeech*, 2021, pp. 1947–1951.

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

[11] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[12] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[13] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proceedings of Interspeech 2019*, 2019, pp. 2803–2807.

[14] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *Proc. Interspeech 2018*, pp. 2207–2211, 2018.

[15] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[16] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.

[17] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[18] Y. Li, Y. Mohamied, P. Bell, and C. Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 868–875.

[19] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Speech emotion recognition based on self-attention weight correction for acoustic and text features," *IEEE Access*, vol. 10, pp. 115 732–115 743, 2022.

[20] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 4961–4965.

[21] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech." in *Interspeech*, 2021, pp. 3810–3814.

[22] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu *et al.*, "Learning word-level confidence for subword end-to-end asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6393–6397.

[23] Q. Li, Y. Zhang, D. Qiu, Y. He, L. Cao, and P. C. Woodland, "Improving confidence estimation on out-of-domain data for end-to-end speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6537–6541.

[24] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, P. C. Woodland, L. Cao, and T. Strohman, "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6388–6392.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.

[26] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus," in *Programme of the Workshop on Corpora for Research on Emotion and Affect*, 2008, p. 28.