Improving Speech Emotion Recognition through Hierarchical Classification and Text Integration for Enhanced Emotional Analysis and Contextual Understanding

Nawal Alqurashi
School of computer Science and Informatics
Cardiff University
Cardiff, UK
alqurashinm@cardiff.ac.uk

Yuhua Li
School of computer Science and Informatics
Cardiff University
Cardiff, UK
liy180@cardiff.ac.uk

Kirill Sidorov
School of computer Science and Informatics
Cardiff University
Cardiff, UK
sidorovk@cardiff.ac.uk

***Abstract*—Speech emotion recognition (SER) systems are designed to classify spoken audio samples into different emotion categories. However, misclassifying emotional samples and predicting them as neutral remains a challenging problem in these systems. One primary contributing factor to this issue is the limitation of speech features to recognize emotions from neutral spoken samples that convey emotional context, as these features do not account for contextual or meaning-based features. To address this issue and improve the recognition performance in SER, we propose a hierarchically structured classification model and integrate text features as a supportive modality to address the misclassification of emotional samples. Text-based features provide valuable contextual information that can aid in identifying emotional content in otherwise neutral speech. This work could be potentially applied in various fields, such as healthcare, education, and entertainment, where recognizing emotions from speech can be crucial for effective communication and decision-making.**

Keywords—speech emotion recognition, human-computer interaction, multimodal speech emotion recognition

I. INTRODUCTION

Speech Emotion Recognition (SER) systems are designed to categorize spoken audio samples into various emotion classes. Several researchers have employed various methods to enhance emotion classification performance across diverse datasets. Nevertheless, they commonly encounter a shared challenge: accurately discerning emotional samples and preventing their misclassification as neutral poses a significant obstacle [1] - [5]. A key factor contributing to this challenge is the inadequacy of speech features in capturing emotions from neutral spoken samples, which convey emotional context [1, 6].

As human beings, we judge the emotion from different modalities[7]. Individuals typically rely on their ability to recognize and interpret signals from various modalities. For example, if only the speech is available, determining the associated emotions can rely on the voice tone rather than the semantic content of the words. Alternatively, it could be based on comprehending the context irrespective of the voice tone, or the process often involves integrating information from multiple factors to form a more comprehensive understanding of the emotions being expressed. However, recognising emotions is subjective and may not be agreed upon by all people [8 , 9]. Additionally, each modality might reveal totally different emotions. The information obtained from each modality may not necessarily align, and different aspects of a person's behaviours or communication might convey disparate emotional experiences. For instance, when emotionally charged words are spoken with a neutral tone, the emotional context may not be immediately apparent from the spoken modality alone. The individual might express strong feelings or sentiments through their choice of words, but the neutral tone obscures the emotional intensity.

The discrepancy between the emotional content of the words and the neutral tone underscores the complexity of emotional communication and the need to interpret cues beyond the spoken language. In the literature, most of the previous work in SER neglects this fact and tries to train supervised learning models with speech features that cannot be used to extract underlying emotions from the context of words. Although some work in SER includes textual information alongside speech features to capture the emotions in context and enhance the recognition of speech [1, 10, 11], they don't account for the issue of conflicting, where different emotions can be revealed from the same speech instance through different modalities[5]. The rationale for emphasizing this case lies in the prevalence of neutrality in everyday conversations, as seen in real-life discussions, TV shows, and series. Despite the wealth of underlying emotions, neutrality often serves as the prevailing state. By developing SER systems that specifically address the commonly employed neutral voice tone, we may potentially streamline the research process, leading to improved recognition accuracy. This approach could also pave the way for further exploration into leveraging multiple modes of information. Mental health conditions, such as depression, often manifest in individuals' language use, particularly in the way they express emotions. Depressed individuals tend to employ neutral or toneless language while conveying highly emotional or charged meanings through their choice of words

and the context in which they are spoken. Consequently, this poses a challenge for conventional SER systems that heavily rely on the identification of emotional states based on acoustic features, such as pitch, intensity, and speech rate. This leads to potential misinterpretation or misclassification of emotions in depressed individuals. Furthermore, autistic individuals can exhibit a wide range of communication styles and patterns, and this can include speaking neutrally or without typical emotional inflection even when discussing emotionally charged topics. It's important to recognize and respect these differences in communication style and to focus on understanding the individual's perspective and emotions, rather than solely relying on traditional emotional cues such as tone of voice or facial expressions. Incorporating inclusion and diversity into human-computer interaction (HCI) using SER involves ensuring that the technology is designed to be sensitive to a wide range of users including autistic individuals with different emotional expressions. We seek to explore and provide solutions for the following research question: How can the problem of classifying samples with emotional content as neutral in SER systems be effectively mitigated through the integration of text-based information? In addressing this question, we aim to investigate the potential of leveraging textual cues and context to enhance the precision and reliability of speech emotion recognition. This paper extensively examines and focuses on the case of misclassifying samples of emotional classes as neutral class, despite their underlying emotional significance. We propose a hierarchically structured classification for SER considering the integrating of text features as a supportive modality. Extensive experiments are conducted to investigate the efficacy of the proposed model. In the proposed hierarchical classification system, we explore three methods including LSTM, wav2vec and wav2vec with automatic speech recognition (ASR) fine-tuning. We also introduce a consistency and confidence threshold approach to moderate the impact of text model on the proposed system.

The structure of this paper is as follows: the Section 2 introduces in detail the proposed multi-modal emotion recognition framework which mainly includes speech models and text model; The Section 3 explains the experimental results of the proposed system followed with results of the consistency and confidence threshold approach. The Section 4 summarizes the work content and outlines the direction of the future work.

II. Proposed Multimodal Framework

Fig. 1 illustrates the block diagram of the proposed multimodal emotion recognition framework. The subsequent sections offer a concise overview of the framework, including descriptions of the overall structure, as well as the speech and text-based models and systems integrated into this framework.

A. Framework overview

In this section we present the structure and hierarchy of the proposed hierarchical classification system and the process we applied in the training phase and the prediction phase.

1) Training phase

Training a single classifier for all emotions directly at once might be challenging due to the nuances and variations in emotional expressions. Therefore, we use a two-level hierarchy classification to organise classes and their relationships allowing for specialised training. In the first level we train the first classifier (Model 1) on all samples using speech features to classify between emotional samples and neutral samples, simplifying the problem by initially categorizing into broad emotional categories. This order aligns with the objectives of our study, particularly centred around the neutral class, as we intend to analyse it with text data later. Then in the second level, we applied two further classifiers. One classifier (Model 2) in the second level focuses on the more refined task of classifying specific emotions using the same speech features and is trained on only with emotional samples to classify between happy, sad and angry. The deliberate exclusion of the neutral class from Model 2 in the classification process is motivated by its inherent ambiguity and difficulty in identification using speech features[12, 6]. This intentional omission seeks to prevent the blending of neutral samples with other emotional classes, thereby contributing to the effective management of the overall complexity of the classification task. Another

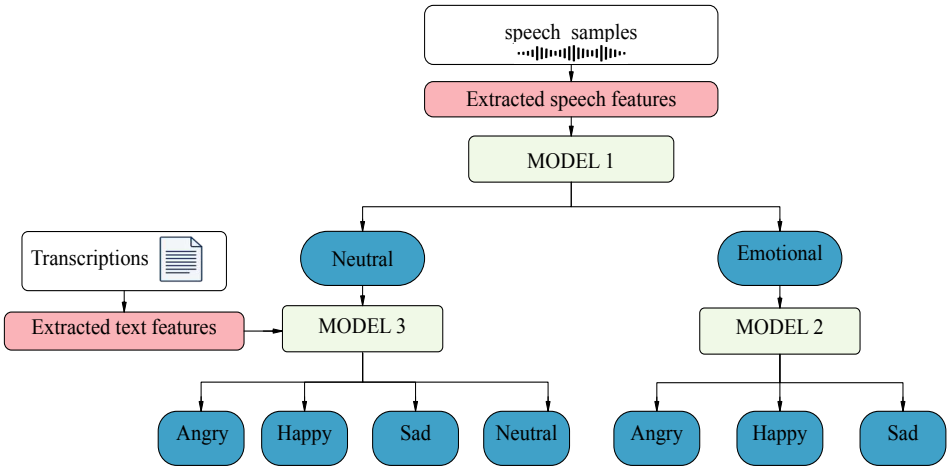

Fig. 1. Overall architecture of the proposed multimodal based hierarchical structure for speech emotion recognition.

classifier (Model 3) in the second level uses text features and is trained on all samples to classify between 4 classes (neutral, happy, sad and angry). The later text-based Model functions to reclassify the predicted classes as neutral from the previous level and assigns them into 4 classes based on text features. Beyond the possible conflicts that may arise when fusing audio and text in the first level, leading to interference between the combined modalities and ultimately compromising result accuracy[5], integrating text and audio does not align with our objective to analyse the influence of text features on predicting the class as neutral. Consequently, we opt not to merge text and speech in the first level, choosing instead to employ a specific text-based model in the second level. To do this hierarchical process, we adjust the dataset's annotation to align with the desired classifier outputs by keeping all neutral samples as they are and changing the class of all the three emotional classes to Emotional for the first level classification. In the second level classification, we keep the original annotation for the classes without changing as in this level we classify among emotional classes (angry, happy, and sad) in speech model and among four classes in text model (angry, happy, sad and neutral). It's important to highlight that our approach is adaptable and capable of being expanded to encompass a wider range of emotion categories, nevertheless, our selection prioritizes the emotions most prevalent in the SER field.

2) Prediction phase

In the prediction phase, the trained hierarchical classification model in the first level determines each sample from the test data whether they are emotional or neutral. Model 2 in the second level classification in the hierarchy takes all the samples in the first hierarchical level that are predicted as emotional and classifies them further into angry, happy, or sad classes. Model 3 in the second level classification in the hierarchy takes all the samples in the first hierarchical level that are predicted as neutral and reclassifies them further if they are still neutral or one of the emotional classes angry, happy, or sad. The classifiers in the second level are connected to the output of the first level including the misclassified instances of the first level for realistic/ applicable results.

B. Speech model

For speech models structures i.e. model 1 and model 2 in the hierarchical we employ three different variants of classifiers as follows.

1) LSTMs- based speech model

Both speech models in the hierarchical system applied LSTMs followed by two fully connected networks using the speech features set described below:

- Mel-frequency cepstral coefficients (MFCCs) are a representation of sound as heard by the human ear [13]. We use Librosa [14] toolkit to extract 39 dimensional (MFCCs) with 16000 Hz of sampling frequency and calculate the mean of the frames to produce 39-dimensional vector per utterance.
- The handcrafted speech features used in [15], including Pitch, Harmonics, Speech Energy, Pause and Central resulting eight speech feature vector per utterance We used the same toolkit Librosa for features extraction.
- The Geneva minimalistic acoustic parameter set [16] (GeMAPS) is a minimal feature set of eGeMAPS that proposed by Eyben et al [16]. GeMAPS set contains a compact set of 18 low-level descriptors (LLD) including frequency related parameters, amplitude related parameters and spectral parameters. We compute GeMAPS using the OpenSimle toolkit [17].

The batch size is set to 64 and we adopt the Adam method to optimize the parameters with cross entropy loss. To select the other hyperparameters we use Optuna optimization [18].

2) Pretrained wav2vec 2.0 without ASR fine-tuning

Transfer learning techniques have gained recent attention in SER[19, 20, 21]Consistent with these efforts, we investigate the application of the wav2vec 2.0 model for speech models within our proposed hierarchical system. Wav2vec 2.0 [22], a transformer-based model, is specifically trained to extract contextualized representations from raw audio signals. The architecture of the wav2vec2.0 model comprises three key sub-modules: the feature encoder, the transformer module, and the quantization module. The feature encoder operates as a multi-layer CNN, processing the input signal into fundamental low-level features. Utilizing this representation, the transformer module generates contextualized representations, while the quantization module discretizes these low-level features into a trainable codebook. During model training, a portion of the low-level features is masked from the transformer module. The primary objective is to accurately identify the quantized version of the masked features, utilising the contextual information available. For fine-tunning progress, we used the publicly released wav2vec2.0 model, which has been pretrained on a substantial dataset comprising 960 hours of Librispeech. The wav2vec2.0 model is comprised of two main components: a CNN-based feature encoder and a transformer-based contextualized encoder. To preserve the knowledge learned in the initial training, we freeze the CNN-based feature encoder, thereby keeping all parameters of these CNN blocks fixed. During fine-tuning, we exclusively update the parameters of the transformer blocks. The pretrained models are fine-tuned to serve as classifiers within our hierarchical speech models. Specifically, the classifier at the first level is tasked with distinguishing between neutral and emotional states, while the classifier at the second level is responsible for discerning between various emotional classes.

3) Pretrained wav2vec 2.0 with ASR fine-tuning

In this methodology, we employed wav2vec 2.0 with ASR fine-tuned model. It is the base model pretrained and fine-tuned on 960 hours of Librispeech. The learned representations are fine-tuned on labelled data, and a randomly initialized output layer is added atop the Transformer for character prediction, Aligning with the ASR objective. For our SER fine-tuning process, following a similar approach as previously outlined, the CNN-based feature encoder is frozen, keeping its parameters fixed. Throughout the fine-tuning, we update the parameters of the transformer blocks. Considering ASR fine-tuned models for wav2vec is deemed important due to the assumption that both ASR and emotion recognition rely on understanding acoustic features in speech. ASR fine-tuning enhances the model's sensitivity to these features, heightening its ability to discern subtle variations in speech patterns that could potentially convey emotional states.

C. Text model

For text model we use LSTMs followed by two fully connected networks in the second level as re-classifier using the text embeddings. For the text transcripts we use Embedding4BERT [23] for extracting word embeddings of pretrained language model (BERT) [24]. BERT allows the model to learn the context of a word based on all its surroundings. The result of the embedding process is a matrix with dimensions of t by 768, where t represents the length of the utterance and 768 is the embedding dimension of the BERT model for each word. The batch size is configured as 64. We employ the Adam method for parameter optimization, utilizing cross-entropy loss. Optuna optimization is employed to choose the remaining hyperparameters.

III. EXPERIMENTS

A. Dataset description

IEMOCAP [25] is an acted, multimodal, and multi-speaker database and contains five recorded sessions of conversations by 10 speakers. Each session contains utterances from two speakers (1 male and 1 female). The dataset consists of nearly 12 hours of audio-visual information along with provided transcriptions. The dataset's emotions are classified into 10 categories: neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited, and other. Aligning with similar studies, this study focuses on the four most frequently used emotions, namely anger, happiness, neutrality, and sadness, by merging the excitement dataset with the happiness dataset[26, 27, 28] In total, 5331 audio utterances and transcriptions are used in this research.

Two different split settings were employed for the experiments: Speaker-Dependent (SD) setting and Speaker-Independent (SI) setting. In SD setting, all the 5 sessions are merged and split by 80/10/10, the 80% data were used for training ,10% for validation set and 10% for test the model. In SI setting, we use 4 sessions for training the model (4 male, 4 female), while the last session was split into validation set containing only male samples and test set containing female samples to evaluate the model ensuring that there is no overlapping between the speakers in all sets.

B. Results and analysis

We set only speech-based baselines for each system to facilitate the comparisons and analysis with the proposed system, providing a reference point for evaluating the effectiveness of our approach. There is no text model added in the baselines and the results are based on only speech features. We designate the models including the baselines with the following abbreviations based on this organisation: (Classifier Name-Modalities used-Setting), with the classifier part we refer to the classifier approach in each model as discussed in the methodology section LSTM/w2v/w2vASR. The modality: S /ST baselines with S denotes speech only, while ST denotes to proposed model using speech and text modalities. The last part refers to the dataset setting either speaker dependent setting or speaker independent setting with D or I, respectively. For example, LSTM-S-D is the baseline model applied LSTM classifier using only speech features in speaker dependent setting, while w2vASR-ST-I refers to the model applied w2vec model with ASR fine-tuning in speaker independent setting using speech and text modalities. Fig. 2 represents the impact of text model on the misclassified emotional samples that were predicted as neutral by speech model for all three used methods and settings. For example, upon utilizing the text model in LSTM-ST-D model, 59.5% of the samples originally misidentified as neutral by the speech model were accurately reclassified. We can see also from LSTM-ST-I model that 44.6% of the incorrectly predicted samples by the speech model were reclassified and accurately predicted by the text model. A noteworthy observation is that many of these instances encompass emotional contexts. For example, expressions such as "Wow, are you so excited?" and "I am very excited" were identified as happy by the text model, whereas the speech model initially labelled them as neutral. Similarly, phrases like "I rather not remember somethings, I rather not hope for somethings" and "It must be really really hard to lose a child" were categorized as sad, by the text model, in contrast to the speech model's neutral labelling, despite the presence of explicit words associated with sad feelings and expressions. Table I shows additional examples of reclassified samples extracted from the used models, offering further insights into the correction process. As illustrated in Fig. 2 and detailed in Table I, the text model demonstrates efficacy in identifying expressions with emotional content, previously labelled as neutral by the speech model. This underscores the distinct knowledge captured through the text modality that is absent in the speech features. During the process of experimentation, we found that the experimental outcomes under the dependent speaker setting are better than in independent setting, particularly in terms of Unweighted Average (UA) accuracy. This can be attributed to several factors. In a dependent speaker setting, certain conditions or characteristics may enhance the performance of the system. One possible explanation is that the model is better adapted to handle specific speaker dependencies, leading to a more accurate recognition of unique speech patterns and emotional expressions with those speakers. Furthermore, we observed that implementing the wave2vec model significantly improved accuracy and achieved the best results in our system. However, the ASR fine-tuned model did not prove

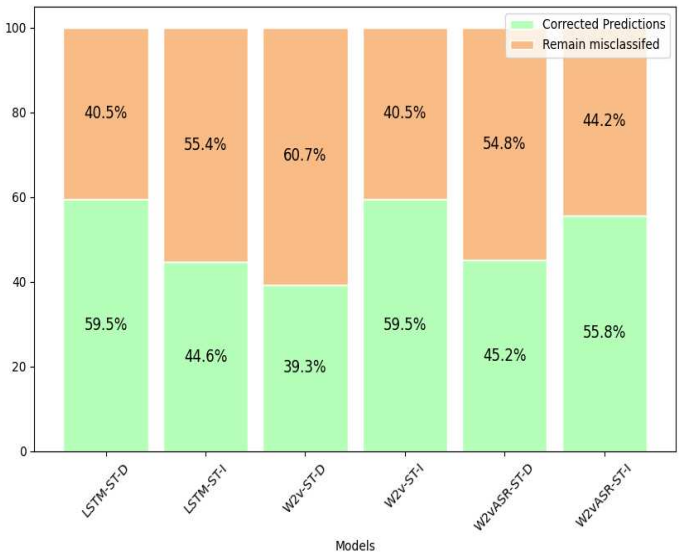

Fig. 2. Corrected and uncorrected misclassification samples by Text model in percentage.

TABLE I. OUTCOMES OF SOME SAMPLES, ILLUSTRATING PREDICTIONS FROM BOTH SM-SPEECH MODEL AND TM- TEXT MODEL

Sentence	SM's prediction	TM's prediction	Ture label
Well, if she does, then that's the end of it. But from her letters, I think she's forgotten him. I'll find out. And then we'll thrash it out with Dad, right? Mom don't avoid me"	neutral	sad	sad
Yeah, that's awesome	neutral	happy	happy
Wow, are you so excited?	neutral	happy	happy
I rather not remember somethings; I rather not hope for somethings	neutral	sad	sad
I am very excited	neutral	happy	happy
It must be really really hard to lose a child	neutral	sad	sad
Are you crazy?	neutral	angry	angry
We will have fun	neutral	happy	happy
Yeah. It's really cool	neutral	happy	happy
Oh, I am so glad	neutral	happy	happy
you are far too temperamental, try to control yourself.	neutral	angry	angry
Who-whoooo…,	neutral	happy	happy
really hard to imagine that like you are going to continue to live, and that person's just going to have stopped	neutral	sad	sad
It's a good idea	neutral	happy	happy
He cried hard	neutral	sad	sad
Congratulations…,	neutral	happy	happy
On the contrary, a child of two can get violently drunk on only one glass of brandy,	neutral	angry	angry

advantageous for SER task when compared to the performance of wave2vec without ASR fine-tuning. This observation is consistent with findings from other studies[21], suggesting that this lack of improvement is due to the loss of prosodic information during the ASR fine-tuning process. Table II provides the performance metrics for all models in our experiments and their baselines across the chosen classes (Angry, Happy, Sad, Neutral). The metrics include Precision, Recall, Unweighted Accuracy (UA), and Weighted accuracy (WA). Furthermore, Table III presents a comparison between our proposed approach and several existing studies in the literature focusing on SER. These studies specifically employ text modality and utilize the same dataset and classes as in our study.

C. Modality Consistency and Confidence threshold

In this study, precision is the measure of accurately predicted neutral instances relative to all instances predicted as neutral, and a high precision means a low rate of false predictions as neutral. Precision is particularly critical in our work as we are interested in minimizing the cases where a sample is mistakenly classified as neutral. High precision ensures a higher probability that content identified as neutral is genuinely neutral. As shown in Table II, the proposed system demonstrates enhancements in precision for the neutral class across all classifiers in comparison to baseline models. However, the overall accuracy of the proposed system did not exhibit an increase. A closer examination reveals that this lack of accuracy improvement is attributed to a decline in recall for neutral class, thereby influencing the overall accuracy. Although that our approach is more accurate and trustworthy in identifying genuinely neutral content, it's also important to consider the trade-off between precision and recall. In addressing considerations of robustness and achieving a harmonious balance between recall and precision within the neutral class, we introduce the Modality Consistency and Confidence algorithm based on probabilistic thresholds. Specifically, we employ consistency and confidence threshold to moderate the full impact of the text model on predicted samples categorised as neutral, originating from the speech model. This process aims to filter results, allowing the adoption of outcomes from the speech model only when the text model demonstrates sufficient consistency with speech model result and confidence in classifying an instance as neutral, devoid of emotional context. The algorithm determines the ultimate predicted class by evaluating whether the predicted class should be derived from the speech model (in our scenario, the neutral class) or from the text model, which includes emotional and neutral classes This decision is made by comparing the prediction probability associated with the neutral class in the text model for a given instance against a specified threshold. The threshold indicates the level of consistency required for the text model to align with the speech model in classifying the instance as neutral. If the text model is sufficiently consistent with the speech model's outcome, the predicted class is deemed neutral, affirming the speech model's prediction. Conversely, if the text model's results do not align adequately, the predicted class will be determined by the text model output. Algorithm 1 provides pseudo-code that implements this procedure. Different threshold values (0.25, 0.30, 0.35, 0.40) are systematically applied to ascertain an optimal threshold for each model. The precision-recall curve for the neutral class is visually depicted in Fig. 3 for each model, and Table II presents the corresponding results, emphasizing the superior accuracy achieved under specific threshold conditions.

Algorithm 1: Modality consistency and confidence threshold calculation

Input: Predicted probability from Text model *ModelT* for neutral class *of an instance* $P_{neutral}(I)$*, Threshold T.*

Output: The final predicted class based on threshold condition of instance I

Require: Predicted class from speech model *ModelS*, and Predicted class from text model *ModelT.*

set *Final prediction* ← predicted class from *ModelS*

if $P_{neutral}(I) > T$ **then**

    set Final prediction← predicted class from *ModelS*

else

    set Final prediction← predicted class from *ModelT*

end if

return *Final prediction*

end

TABLE II. SUMMARY OF THE EXPERIMENTS RESULTS IN TERM OF CLASS-WISE PRECISION AND RECALL, OVERALL ACCURACY OF THE BASELINES, PROPOSED SYSTEM, PROPOSED SYSTEM WITH DIFFERENT THRESHOLDS, UA-UNWEIGHTED ACCURACY, WA- WEIGHTED ACCURACY.

Model	**LSTM-S-D**		**LSTM-ST-D**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	66.0	66.0	65.0	70.0	65.0	70.0	65.0	70.0	66.0	69.0	67.0	69.0
Happy	50.0	56.0	49.0	65.0	49.0	65.0	50.0	65.0	51.0	65.0	52.0	65.0
Sad	57.0	79.0	58.0	84.0	58.0	84.0	58.0	84.0	58.0	84.0	58.0	84.0
Neutral	66.0	45.0	79.0	32.0	79.0	32.0	80.0	34.0	80.0	38.0	80.0	41.0
UA	58.84		59.20		59.92		59.74		59.74		59.56	
WA	61.35		62.80		63.36		63.22		63.22		63.08	
Model	**W2v-S-D**		**W2v-ST-D**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	79.0	78.0	75.0	82.0	75.0	82.0	75.0	82.0	75.0	82.0	75.0	82.0
Happy	71.0	67.0	66.0	75.0	69.0	75.0	68.0	75.0	68.0	75.0	66.0	75.0
Sad	70.0	71.0	55.0	81.0	58.0	80.0	56.0	80.0	55.0	80.0	55.0	81.0
Neutral	66.0	69.0	77.0	41.0	77.0	49.0	76.0	46.0	76.0	44.0	78.0	41.0
UA	70.93		66.78		68.95		68.05		67.50		66.96	
WA	71.41		69.77		71.35		70.64		70.22		69.91	
Model	**W2vASR-S-D**		**W2vASR-ST-D**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	76.0	80.0	73.0	84.0	73.0	84.0	73.0	84.0	73.0	84.0	73.0	84.0
Happy	67.0	58.0	64.0	68.0	66.0	67.0	66.0	67.0	66.0	67.0	65.0	68.0
Sad	70.0	71.0	51.0	75.0	55.0	75.0	53.0	75.0	52.0	75.0	51.0	75.0
Neutral	68.0	73.0	77.0	44.0	77.0	51.0	77.0	48.0	77.0	47.0	77.0	44.0
UA	70.21		64.98		67.32		66.24		65.88		65.16	
WA	70.69		67.66		69.47		68.62		68.34		67.80	
Model	**LSTM-S-I**		**LSTM-ST-I**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	56.0	58.0	52.0	60.0	54.0	60.0	53.0	60.0	52.0	60.0	52.0	60.0
Happy	40.0	52.0	42.0	60.0	42.0	58.0	42.0	58.0	42.0	58.0	42.0	60.0
Sad	57.0	65.0	51.0	72.0	53.0	69.0	53.0	70.0	52.0	71.0	52.0	71.0
Neutral	64.0	45.0	80.0	32.0	71.0	36.0	73.0	34.0	75.0	33.0	79.0	33.0
UA	52.88		52.20		52.37		52.37		52.03		52.37	
WA	54.80		56.06		55.70		55.89		55.74		56.06	
Model	**W2v-S-I**		**W2v-ST-I**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	64.0	63.0	40.0	85.0	64.0	63.0	55.0	76.0	48.0	78.0	40.0	83.0
Happy	71.0	57.0	60.0	74.0	64.0	72.0	64.0	73.0	62.0	74.0	60.0	74.0
Sad	73.0	58.0	69.0	72.0	75.0	70.0	75.0	70.0	70.0	71.0	69.0	72.0
Neutral	60.0	76.0	71.0	29.0	69.0	67.0	71.0	57.0	68.0	42.0	69.0	30.0
UA	65.08		58.30		68.30		66.77		62.03		58.30	
WA	63.41		65.05		68.28		66.26		61.81		64.84	
Model	**W2vASR-S-I**		**W2vASR-ST-I**		**Threshold 0.25**		**Threshold 0.30**		**Threshold 0.35**		**Threshold 0.40**	
Class	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***
Angry	69.0	54.0	42.0	83.0	69.0	54.0	58.0	71.0	69.0	54.0	69.0	54.0
Happy	68.0	43.0	60.0	67.0	65.0	65.0	64.0	65.0	69.0	58.0	69.0	58.0
Sad	61.0	66.0	60.0	74.0	63.0	73.0	62.0	73.0	62.0	67.0	62.0	67.0
Neutral	58.0	75.0	69.0	30.0	67.0	66.0	69.0	56.0	63.0	71.0	63.0	71.0
UA	61.52		56.94		65.59		64.06		64.57		64.57	
WA	59.45		63.68		64.35		66.03		62.70		62.70	

TABLE III. PERFORMANCE COMPARISON OF OUR PROPOSED APPROACH WITH DIFFERENT METHODS ON IEMOCAP

System	**UA**	**WA**
SVM tree ensembles [29]	67.4	67.4
E-vector + MCNN + LSTM [26]	65.9	64.9
(ENC1) +(ENC2)[27]	68.4	
Late Fusion-III [28]	59.3	61.2
ACO(+)Cepstrum(+) Cepstral-BoW+GSV-mean (+) Lex-BoW(+)Lex-eVector [30]	-	69.2
ASR-SER (Hierarchical co-attention) [31]	-	63.4
Ours (SD)W2v-ST-D	68.9	71.3
Ours (SI)W2v-ST-I	68.3	68.2

IV. CONCLUSION

The present study aims to investigate the feasibility of using text-based emotions classification in SER, with the ultimate goal of minimising the mistakenly emotional samples as neutral. The current study explores the potential means of reducing this type of error. In this regard, the IEMOCAP dataset, encompassing four distinct classes, is utilised to test the hypothesis and determine the effectiveness of our hierarchical classification. Additionally, three different classifiers, namely LSTM, wav2vec and wave2vec with ASR fine-tuning are trained and applied in the proposed system. To moderate the impact of text model, we propose the Modality Consistency and Confidence algorithm based on probabilistic thresholds. Experimental results reveal that our proposed method outperforms created speech-based baselines. The findings highlight that the incorporation of

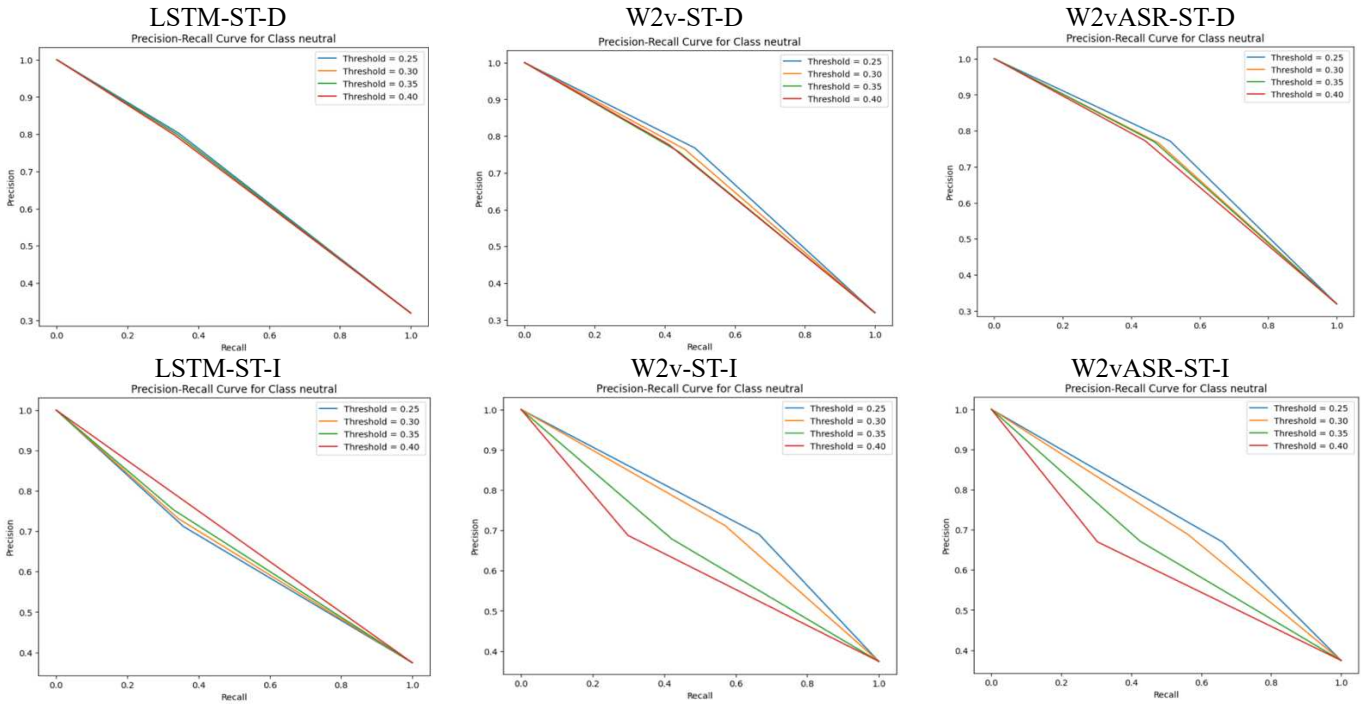

Fig. 3. Precision and recall curve of neutral class.

textual information enables a more comprehensive understanding of the emotional context underlying seemingly neutral spoken interactions. Future research could also extend the scope of individual sentences by considering the importance of incorporating contextual information from surrounding sentences to better infer emotions from neutrally spoken dialogues. By understanding the broader conversation or situational context, it becomes feasible to discern emotions that might not be explicitly evident in isolated sentences. Moreover, it would be valuable to investigate the feasibility and implications of learning hierarchical representations within a single unified model. Such an investigation holds promise in potentially streamlining the model complexity and reducing the need for maintaining multiple separate models.

REFERENCES

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

[2] Z. T. Liu, M. T. Han, B. H. Wu, and A. Rehman, "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning," *Applied Acoustics*, vol. 202, Jan. 2023.

[3] T. Han, Z. Zhang, M. Ren, C. Dong, X. Jiang, and Q. Zhuang, "Speech emotion recognition based on deep residual shrinkage network," *Electronics (Switzerland)*, vol. 12, no. 11, Jun. 2023.

[4] P. Bhattacharya, R. K. Gupta, and Y. Yang, "Exploring the contextual factors affecting multimodal emotion recognition in videos," *IEEE Trans Affect Comput*, pp. 1–12, 2021.

[5] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," *Inf Sci (N Y)*, vol. 619, pp. 679–694, Jan. 2023.

[6] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus training strategy for speech emotion recognition using self-supervised representations," *Applied Sciences (Switzerland)*, vol. 13, no. 16, Aug. 2023.

[7] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text Modalities: A systematic review of recent advancements and future prospects," *Expert Syst Appl*, p. 121692, Mar. 2023.

[8] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model," *Multimed Tools Appl*, vol. 82, no. 8, pp. 11897–11922, Mar. 2023.

[9] F. Huang, K. Wei, J. Weng, and Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 16, no. 3, pp. 1–19, Sep. 2020.

[10] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models," pp. 407–414, 2022.

[11] N. H. Ho, H. J. Yang, S. H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[12] N. Jia, C. Zheng, and W. Sun, "A multimodal emotion recognition model integrating speech, video and MoCAP," 2022.

[13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[14] B. McFee *et al.*, "librosa: audio and music signal analysis in python," *Proceedings of the 14th Python in Science Conference*, pp. 18–24, 2015.

[15] G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," 2019, *arxiv:1904.06022.*

[16] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans Affect Comput*, vol. 7, no. 2, pp. 190–202, 2016.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, no. May, pp. 835–838, 2013.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 2623–2631, Jul. 2019.

[19] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 1–5, Jun. 2023.

[20] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech* ,pp.3400-3404, 2021.

[21] Y. Wang, A. Boumadane, and A. Heba, “A fine-tuned wav2vec 2.0/Hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” *arXiv:2111.02735*, 2021.

[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” arXiv:2006.11477, Jun. 2020.

[23] Y. Chai, “embedding4bert: A python library for extracting word embeddings from pre-trained language models,” *GitHub repository*, 2020.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv:1810.04805*, 2018.

[25] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang Resour Eval*, vol. 42, no. 4, pp. 335–359, 2008.

[26] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, “Deep neural networks for emotion recognition combining audio and transcripts,” Proc. Interspeech 2018, vol. 2018-Septe, pp. 247–251, 2018.

[27] K. D. N and S. S. Reddy, “Multi-Modal Speech Emotion Recognition Using Speech Embeddings and Audio Features,” *Proc. The 15th International Conference on Auditory-Visual Speech Processing*, pp. 16-20, 2019.

[28] J. Sebastian, P. Pierucci, and T. L. Gmbh, “Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts,” *Interspeech*, pp. 51–55, 2019.

[29] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, “Ensemble of svm trees for multimodal emotion recognition,” *Proc. IEEE Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–4, 2012.

[30] Q. Jin, C. Li, S. Chen, and H. Wu, “Speech emotion recognition with acoustic and lexical features,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 4749–4753, 2015.

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