Emosical: An Emotion-Annotated Musical Theatre Dataset

Hayoon Kim[1] **Ahyeon Choi**[1] **Sungho Lee**[1] **Hyunjin Jung**[1] **Kyogu Lee**[1,2,3]
[1]Music and Audio Research Group, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]Artificial Intelligence Institute, Seoul National University
{hyway, chah0623, sh-lee, 3388jung, kglee}@snu.ac.kr

Abstract

This paper presents Emosical, a multi-modal open-source dataset of musical films. Emosical comprises video, vocal audio, text, and character identity paired samples with annotated emotion tags. Emosical provides rich emotion annotations for each sample by inferring the background story of the characters. To achieve this, we leverage the musical theatre script, which contains the characters' complete background stories and narrative contexts. The annotation pipeline includes feeding the speaking character, text, global persona, and context of the dialogue and song track into a large language model. To verify the effectiveness of our tagging scheme, we perform an ablation study by bypassing each step of the pipeline. The ablation results show the usefulness of each component in generating accurate emotion tags. A subjective test is conducted to compare the generated tags of each ablation result. We also perform a statistical analysis to find out the global characteristics of the collected emotion tags. Emosical would enable expressive synthesis and tagging of the speech and singing voice in the musical theatre domain in future research. Emosical is publicly available at https://github.com/gillosae/emosical.

1 Introduction

As a fundamental aspect of human experience, recognizing emotion requires a depth of understanding in the multi-domains. With the advances in deep learning techniques, various approaches delve into detecting emotion through expressed emotion from different modalities (Poria et al., 2017). Studies aim to detect emotion through the domain of text (Pang et al., 2008; Socher et al., 2013), speech (Satt et al., 2017; Akçay and Oğuz, 2020), and facial expressions (Zafeiriou et al., 2015; Li and Deng, 2020). Furthermore, approaches to integrate multiple modalities (Zadeh et al., 2017; Lian et al., 2023) have been introduced.

While those techniques have advanced the understanding of human emotions in different domains, emotion recognition still remains a challenge. First, it is hard to capture ground-truth emotion, which is expressed through both linguistic and non-linguistic elements. This complexity requires multi-modal data to capture the emotional context fully, but integrating and synchronizing such diverse inputs presents technical difficulties (Baltrušaitis et al., 2018). Although some models are trained using recorded multi-modal datasets (Busso et al., 2008), creating these datasets is resource-intensive and lacks the scalability and diversity needed to fully capture the complexity of human emotions.

Additionally, human emotions are complex and subtle, with the inherent ambiguity of expressed emotional cues and variations in cultural expressions (Russell, 2003). Thus, existing models struggle to recognize emotions from expressed features at a fine-grained level, mostly clustering emotions into coarse categories (Ververidis and Kotropoulos, 2006). We suspect that knowledge of the underlying personality or situation makes it easier to infer one's emotion more accurately and finely, which is a capability missing in most datasets.

We propose that *theatre* is a particularly effective medium for addressing these challenges. Theatre naturally encompasses multi-modal emotional expressions since actors and directors use various techniques such as dialogue, music, lighting, and stage design to communicate a wide range of emotions to the audience (Artaud, 1938). These emotions are typically expressed with an emphasis for greater impact (Stanislavski, 1948), making it easier to capture and analyze them effectively. Also, the theatre has the unique property of containing a 'script' for the play, which enables the inference of each character's storyline and personality. In this view, theatre contains a diverse range of expressed emotions conveyed through multiple modalities

Findings of the Association for Computational Linguistics: EACL 2024, pages 4169–4180
November 12-16, 2024 ©2024 Association for Computational Linguistics

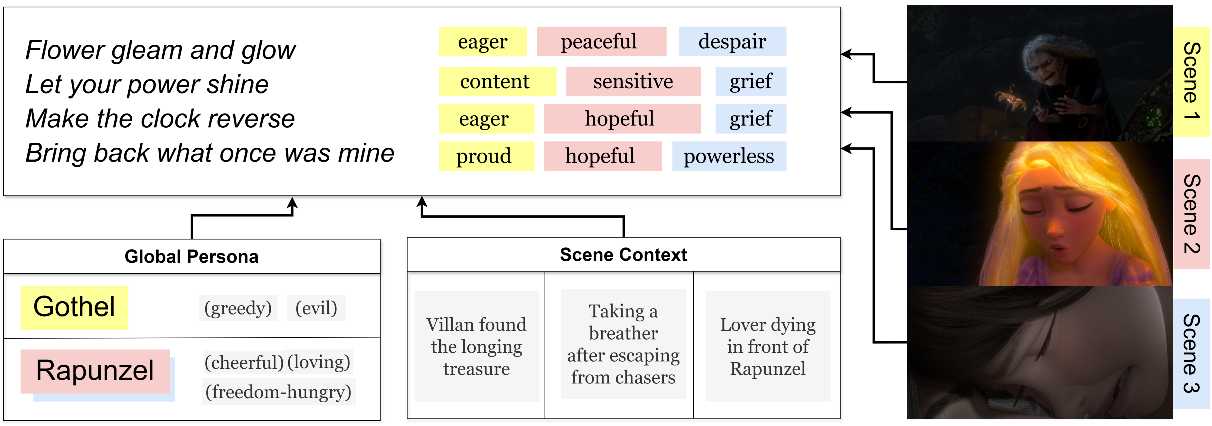

Figure 1: Emotion inferring process of 'Healing Incantation' in 'Tangled.'

that are deeply rooted in each character's narrative context, making it a suitable medium for emotion research.

In response, we build `Emosical`, an emotion-annotated theatre dataset. Specifically, we select *musical theatre* for our dataset in order to contain singing samples. There is no theatre dataset specifically curated for machine learning purposes. It might be due to theatre's inherent complexity, which combines spoken natural language, audio, and visual elements, making it hard to create complex recordings manually. Therefore, we design an annotation pipeline that infers emotion from the narrative by leveraging the script from each theatre.

Our annotation pipeline aims to analyze and annotate the narrative's emotions as automatically as possible. Thus, we crawl musical theatre films and apply them as pipeline input. We require theatre film video with the timestamp and speaker identity aligned `SRT` file as the inputs. In the pipeline, we feed the character, text, global persona, and context of the dialogue and song track into an LLM to obtain the corresponding emotion tag per sample.

Our annotation pipeline captures emotion effectively by capturing the narrative context through fed inputs. In musical theatre, some prominent songs tend to be reprised and presented multiple times throughout the act, conveying different emotional nuances. Figure 1 shows an example; the number 'Healing Incantation' of 'Tangled' is the case, which emerges once at an introductory moment, once at a highly-elated scene, and lastly at the ending part of the movie. In the figure, the three reprised songs have distinctly different emotions, clearly indicating diverse emotional expressions. 'Healing Incantation' is reprised triple times in the movie. Even though they all have the same lyrics our tagging pipeline tags corresponding singing emotions well by inferring emotions from the context and character's persona. Through our pipeline, even though the song consists of the same lyrics, the resulting tags are different due to different scene contexts fed to obtain the tags.

This pipeline has another advantage that it applies to building some types of emotion-paired datasets, which are hard to develop. For instance, there is no public singing data annotated with emotions aside from Livingstone and Russo (2018). Unlike speech, singing demands attention to nuances like pitch, tone, and emotional delivery (Sundberg, 1987). Even though singing is a powerful medium that conveys expressive emotional expression, it is hard to record a singing voice sample with specific emotions manually. However, our approach enables emotional inference on any kind of data, including singing - under the condition that the identity's background context is given.

Through text, video, and vocal-aligned samples with rich emotion annotations, we aim for `Emosical` to be primarily used to understand the relationship between theatre's multi-modal characteristics and emotions. Additionally, we hope the dataset could also be used for tasks such as emotion tagging and emotional synthesis for each modality; a baseline system for the former is developed and evaluated as an example. To this end, `Emosical` provides detailed and fine-grained emotion annotation for every short segment. Unlike emotion-annotated datasets (McKeown et al., 2012; Nojavanasghari et al., 2016), which tend to annotate emotions in broad groups over long segments, our dataset with dense emotion tags annotated to short data samples allows for more precise and temporal studies of emotional changes.

	Modality				Annotation		Dataset Size			
Dataset	Text	Speech	Singing	Video	Identity	Emotion	#Movies	#Samples	#Speakers	#Tags
ESD (Zhou et al., 2022)	✓	✓				✓	-	350	20	5
EmoDB (Burkhardt et al., 2005b)	✓	✓			✓	✓	-	535	10	7
RAVDESS (Livingstone and Russo, 2018)	✓	✓	✓	✓	✓	✓	-	2452	24	8
IEMOCAP (Busso et al., 2008)	✓	✓		✓		✓	-	10039	10	9
CMU-MOSEI (Zadeh et al., 2018)	✓	✓		✓	✓	✓	-	23453	1000	
MELD (Poria et al., 2019)	✓	✓		✓	✓	✓	-	13000	260	7
SEMAINE (McKeown et al., 2012)		✓		✓	✓	✓	✓	80	4	5
OMG-Emotion (Barros et al., 2018)	✓	✓		✓		✓	✓	7371		7
EmotiW (Dhall et al., 2023)		✓		✓		✓	✓	1809		7
VocalSet (Wilkins et al., 2018)	✓		✓				-	3560	20	-
OpenSinger (Huang et al., 2021)	✓		✓				-	43075	93	-
M4Singer (Zhang et al., 2022)	✓		✓				-	20942	20	-
MPII-MD (Rohrbach et al., 2015a)	✓			✓			94	68337	-	-
MovieQA (Tapaswi et al., 2016)	✓			✓			140	6771	-	-
V2C-Animation (Chen et al., 2022)	✓	✓		✓	✓	✓	26	10217	153	8
Cognimuse (Zlatintsi et al., 2017)	✓	✓		✓	✓	✓	13	11109		-
Emosical (ours)	✓	✓	✓	✓	✓	✓	20	27294	261	128

Table 1: Comparison of various open-source datasets with `Emosical`.

We summarize our contributions as follows:

- We present `Emosical`, the first open-source musical film dataset with emotion annotations.
- Our dataset contains singing voice samples with identity and emotion annotation, which most existing singing voice dataset lacks.
- We build an automatic emotion tagging pipeline that utilizes the musical film script to infer the background story of the singer.
- We provide a baseline tagging model trained on our dataset, which predicts emotion labels from the speech and singing voice signals.

2 Related Works

Table 1 summarizes several key characteristics of the related datasets and compares them with ours.

Multimodal Emotion Recognition Datasets. A multitude of datasets have been developed for multimodal emotion recognition by integrating various modalities such as video, audio, and text. The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset (Busso et al., 2008) includes audio-visual data from actors performing scripted and improvised scenarios designed to elicit specific emotions. Similarly, the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset (Bagher Zadeh et al., 2018) offers emotion-annotated video segments from YouTube videos. The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) consists of dialogue sequences from the TV series annotated with emotion labels, including synchronized video, audio, and textual data, appropriate for emotion recognition in conversational contexts.

The SEMAINE database (McKeown et al., 2012) contains audiovisual recordings of interactions between humans and an avatar designed to elicit emotional responses, including high-quality audio and video data with continuous annotations for emotion dimensions such as arousal and valence. The RÉCital Corpus for Multimodal Emotion Analysis (RECOLA) dataset (Ringeval et al., 2013) includes audio, video, and physiological data recorded from participants during team working tasks, annotated for continuous emotion dimensions, making it a comprehensive resource for studying dynamic emotional expressions. The OMG-Emotion dataset (Barros et al., 2018) contains video recordings of people reacting to predefined stimuli, with annotations for continuous emotion dimensions, providing continuous perspectives on emotional responses.

The Audio-Visual Emotion Challenge (AVEC) (Ringeval et al., 2019) provides datasets, including synchronized video and audio recordings annotated with emotional states. The Emotion Recognition in the Wild (EmotiW) challenge (Dhall et al., 2023) similarly features datasets capturing spontaneous expressions of emotions in real-world environments, including video, audio, and textual data, suitable for developing emotion recognition systems that work in naturalistic settings.

Speech Emotion Recognition Datasets. The Emotional Speech Database (EmoDB) (Burkhardt et al., 2005a) includes recordings of professional actors who simulated seven different emotions. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018) contains actors vocalizing two lexically matched statements. Each expression is labeled for one of eight emotional states, offering a rich dataset for both speech and song emotion recognition. The Speech Emotion Recognition (ESD) dataset (Zhou et al., 2022) is a multilingual dataset containing emotional speech data, which provides a diverse set of emotional speech samples for cross-lingual emotion recognition studies.

Film Datasets. Film-specific datasets offer extensive resources for analyzing the complex interplay of visual, auditory, and narrative elements in movies. The V2C-Animation dataset (Chen et al., 2021) focuses on animated videos and includes video clips with corresponding textual descriptions. The MPII Movie Description Dataset (Rohrbach et al., 2015b) is a large-scale collection of movie clips annotated with natural language descriptions. MovieQA (Tapaswi et al., 2016) is a dataset designed to test story comprehension through question-answering tasks based on movie plots, integrating visual, textual, and auditory information to evaluate narrative understanding. Cognimuse (Zlatintsi et al., 2017) is a comprehensive dataset that includes multimodal annotations (audio, visual, and textual) of Hollywood movies, with detailed annotations for scene boundaries, character interactions, and emotion.

3 Dataset

Overview. Table 2 shows the overall statistics of `Emosical`. It comprises 27294 samples, totaling 20 hours, from 20 distinct musical films, including animation musicals, musical movies, and theatre recordings. Each sample is a tuple of {vocal audio, video, text, character} accompanied by annotated emotion tags. Vocal audio samples include 21040 speech and 6254 singing samples. The overall structure of the dataset is shown in Figure 2. The dataset collection and the annotation process are shown in Figure 3.

Statistics	Count
Total # of films	20
Total # of video/audio segments	27294
The average length of segments	2.66s
Total # of distinct speakers	479
Total # of speech samples	21040
Total # of singing samples	6254
Total # of words in scripts	162277
Total # of unique words in scripts	15610
Total # of emotion tags	128

Table 2: Summary of `Emosical` dataset statistics.

```
.
├── data/
│   ├── raw/
│   │   ├── theatre/
│   │   │   ├── frozen.mov
│   │   │   └── ...
│   │   └── srt/
│   │       ├── frozen.srt
│   │       └── ...
│   ├── audio/
│   │   ├── frozen/
│   │   │   ├── 1.wav
│   │   │   └── ...
│   │   └── ...
│   ├── video/
│   └── text/
└── metadata/
    ├── number_info.csv
    ├── global_persona/
    │       ├── frozen.yaml
    │       └── ...
    └── scene_summarization/
            ├── frozen.yaml
            └── ...
```

Figure 2: Structure of `Emosical`.

Dataset Structure. Given that the movies are not freely available, we offer automated scripts to process the data and links for downloading each film. We provide subtitle files (in SRT extension; throughout the paper, we will denote them as SRT files) that contain characters and text aligned to the movie. These SRT files are different from the publicly available ones and have more precise timestamps. We also provide metadata that contains emotion and vocal type per sample, as well as noisy audio recordings, which will be cleaned with an enhancement model. To use the dataset, users will first place movie video files in the `data/raw/theatre` directory and place them along with corresponding subtitle files in the `data/raw/srt` directory. Then, users can run the provided code, which transforms the dataset into the compiled form with additional directories, including `data/audio/`, `data/video`, and `data/text`. We also provide metadata that shows which clip corresponds to which scene so users can access individual video/audio clips from specific scenes of each movie.

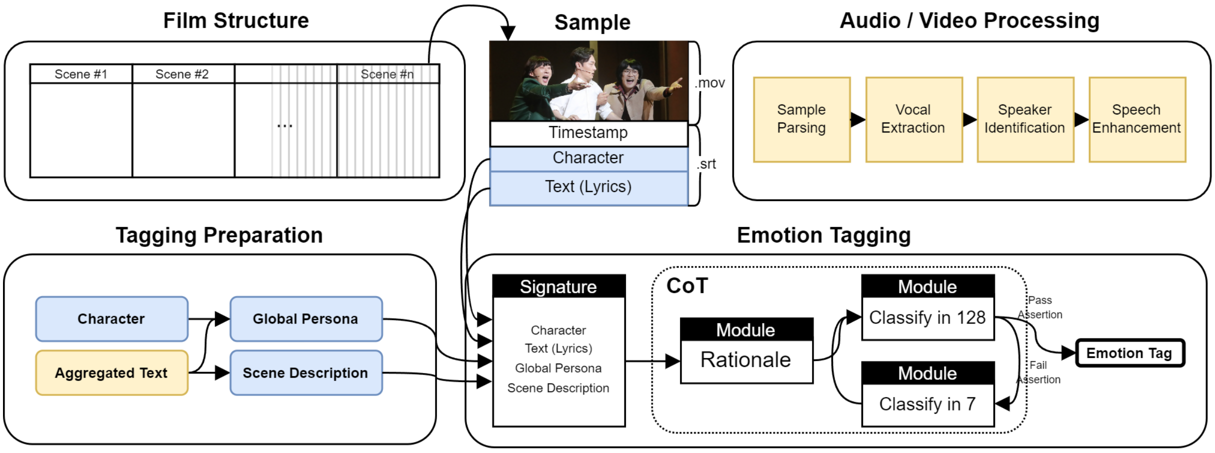

Figure 3: Dataset collection pipeline of Emosical.

3.1 Dataset Creation

We aim to develop a dataset suitable for multimodal emotion analysis of musical theatre. Additionally, we aim for our dataset to be applicable for multiple purposes, including voice synthesis and tagging tasks utilizing our audio dataset. To suit these purposes, we construct a data generation pipeline that is especially focused on audio processing. The pipeline can be automatically run when raw video files, prepared SRT files, and metadata are given.

Timestamp Correction. Our dataset leverages the publicly available original SRT files. SRT files contain the sequential number of current utterances, starting and ending points in the video timeline, and corresponding text. However, as they are for subtitles, their timestamps are not sufficiently precise for our purposes, e.g., splitting the video and audio with timestamps. Therefore, we need to precisely tune the timestamp and text of each SRT segment to contain the starting and ending point of each utterance properly. To achieve this, we first utilized a transcription alignment tool Gentle (Hawkins et al., 2024) to create the rough timestamps. Then, we manually post-processed those to ensure accuracy and to set each sample's length to be within 10 seconds. The provided SRT files are post-processed ones.

Video/Audio Segmentation. For each video, we utilize the MoviePy library (Zulko et al., 2024) to parse samples according to the starting and ending timestamps in its corresponding SRT file. We also extract the corresponding audio data and convert the stereo audio into mono by summing the left and right channels. These processings are done with the ffmpeg toolkit (Tomar, 2006).

Vocal Isolation and Enhancement. Considering the potential audio applications such as voice synthesis and tagging tasks, it is desirable to have clean voice signals without any background noise. Therefore, before segmenting the audio, we separate all the voice signals from the remaining using the open-source Demucs model (Rouard et al., 2023). After the segmentation, we perform the following two extra steps. First, we manually check for noisy audio files. For noisy audio, we further remove the background noise by employing the background noise reduction model, SGMSE (Richter et al., 2023). Second, we filter out audio clips that do not satisfy the following requirements: (i) no overlapping speech and singing voices, (ii) no residual noise, and (iii) negligible processing artifacts. We manually exclude these segments since we aim to curate an automatically processable dataset.

Speaker Identification. After collecting video clips, audio segments, and their corresponding text, each is annotated with the corresponding speaker's identity and matched against the SRT file. This is useful not only as an additional feature of the dataset but also for the emotion annotation pipeline; it helps large language models (LLMs) effectively distinguish each character and recognize the emotional nuances conveyed through the storyline; refer to Section 3.2 for the details. To identify speakers, we first select representative audio samples for each main character. Then, we use a pre-trained speaker diarization model and assign a speaker with the highest similarity to each speech/singing segment. Note that our audio samples include both speech and singing from the same speakers. This is called a *cross-genre* setup, which is known to be challenging, especially when singing voices are involved (Li et al., 2022). We used WeSpeaker (Wang et al., 2023), the only publicly available pretrained cross-genre model. However, it should be noted that the model was trained on a dataset in the Chinese language (Li et al., 2022), potentially leading to inaccurate estimations for the English films. Therefore, to ensure the accuracy of the speaker annotation, we manually checked each annotation and fixed it if necessary. Through this data collection pipeline, we finally gather a triplet of {vocal, text, speaker} for audio data. In the metadata, each voice segment is annotated with binary data that distinguishes singing voice segments from speech.

3.2 Emotion Annotation

As we collected the {video, audio (vocal), character, text} data through the aforementioned pipeline, we now aim to annotate the emotion for each sample. To achieve this, we focus on the storyline of the theatre to further infer the emotion of the character line by line. Note that this approach is similar to Bhattacharya et al. (2023), which generated story descriptions to handle downstream tasks. We leverage full text from the SRT file and feed it to a LLM. Specifically, the annotation process integrates four key components for each character: (i) global persona, (ii) scene summarization, (iii) visual description, and (iv) the text of each sample.

Global Persona. For each character, we define a global persona that encapsulates their overarching traits and narrative role. We obtain the global persona by feeding the whole script into the large language model and prompting it to summarize the character's overall storyline and personality. As the latter ablation study shows, this is crucial for understanding the emotional context of their actions and expressions throughout the movie.

Scene Separation and Summarization. We separate the entire text into multiple scenes to effectively summarize the context of each section of the film. Then, we feed the aggregated text of the scene into LLM to obtain a summarized story. Summarizing the scene helps infer the characters' emotional state when they commence certain utterances, thereby guiding the LLM in generating accurate emotion tags afterward. Overall, feeding global persona and context summarization helps LLM follow the storyline and understand the character's personality throughout the musical theatre.

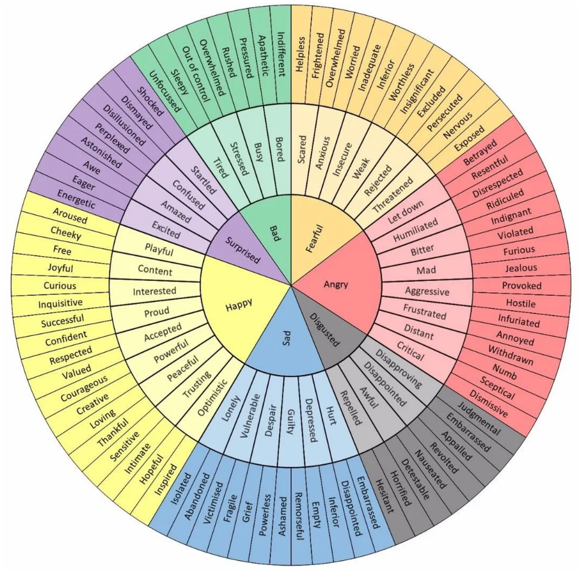

Figure 4: 128 emotion wheels with 7 primary, 40 secondary, and 81 tertiary emotions.

As a result, it aids LLM in successfully guessing the character's emotional state when saying specific text or singing specific lyrics.

The Emotion Wheel. Most emotion-annotated datasets categorize emotions into 4 to 8 groups (Zhou et al., 2022; Burkhardt et al., 2005a; Livingstone and Russo, 2018). However, it is desirable to use a more sophisticated taxonomy of emotion labels to capture the nuanced emotions conveyed in the musical film. This motivates us to leverage the emotion wheel. The Plutchik emotion wheel (Plutchik and Kellerman, 2013) is developed to categorize human emotions based on the idea that distinct emotions can be mixed to create other emotions. We use the expanded version of Plutchik's original emotion wheel. The "128 Emotion Wheel" is gradually structured with primary, secondary, and tertiary emotions to provide a more granular understanding of human emotional experiences (Roberts, 2024). These 128 emotions are sub-classes of the primary 7 emotions ('angry,' 'disgusted,' 'sad,' 'happy,' 'surprised,' 'bad,' and 'fearful'), making each label suitable for primary emotion clustering, enabling easy comparison with other datasets. Also, diverse tags can enrich the input language when training the model for prompting purposes.

LLM Prompting with DSPy. With the character's global persona, scene summarization, sample description, and text with the character ready at hand, we feed them with prompts into the LLM

(Chat-GPT 3.5 Turbo (Brown et al., 2020)) to generate emotion annotations for each line of the dataset.

We utilize DSPy (Khattab et al., 2023), a framework to optimize LLM prompts and weights with small training data. In the DSPy framework, we define the Signature, which outlines what information we provide to the LLM. Our Signature is defined as {character, text, visual description, scene context, global persona} to include all necessary details on a full picture of the scene and characters. We manually created 50 training samples consisting of Signature elements and the ground-truth emotion tag with its rationale.

Then, we proceed to fine-tune our LLM. The fine-tuning process involves the chain-of-thought (CoT) (Wei et al., 2022) method. Specifically, our CoT uses 3 DSPy modules. The first module asks LLM to generate a rationale about the situation that the speaker encountered based on the input Signature. This guides the LLM in following a concrete reasoning process. Next, the second module makes LLM classify the emotion based on the former output rationale, which leads to more accurate and contextually relevant tags.

However, our LLM sometimes predicts an emotion tag outside the 128 emotion wheel due to the unconstrained nature of the LLM. To mitigate this, we apply DSPy Assertion (Singhvi et al., 2024), which poses constraints to guide the LLM's output. If the emotion tag doesn't meet the assertions, DSPy.Suggest triggers the backtracking process. This LLM re-evaluates the input context to generate a new emotion tag. However, when the backtracking exceeds the pre-defined maximum number, the third module is introduced, prompting the LLM to classify the emotion into seven primary emotions. Lastly, we prompt the LLM to classify the emotion from the sub-emotions of the primary emotion. This fine-tuned LLM significantly exceeds the accuracy of the untrained baseline LLM.

3.3 Data Split

We provide additional metadata for the dataset split for development and test purposes. We split the 20 films into 16 and 4; the latter is used only for tests. The 16 movies are further split, where random 80% of each movie is used for the development while the remaining are used for the test. This split allows us to evaluate the models both in *seen* (in-distribution) and *unseen* (possibly out-of-distribution) film.

	AUC	**Precision**	**Recall**	**F-score**
Ours	0.690	0.739	0.690	0.710
Speech-based	0.173	0.355	0.173	0.184
Text-based	0.286	0.635	0.286	0.338

Table 3: Emotion-tagging pipeline vs. other models.

	Base		**Fine-tuned**	
	Train	**Test**	**Train**	**Test**
BootstrapFewShot	43.86	58.82	31.58	82.35
+ RandomSearch (ours)	43.86	58.82	45.61	85.29

Table 4: Baseline LLM vs. Fine-tuned LLM.

4 Evaluation and Analysis

4.1 Emotion Tagging Pipeline

We evaluate the effectiveness of our emotion-tagging pipeline versus other baseline models. For comparison with other emotion-tagging baselines, we change our pipeline to classify each sample into seven emotions (anger, disgust, fear, joy, neutral, sadness, surprise). The evaluation is performed on a human-annotated dataset we created from a short musical theatre film. We perform evaluation on the handcrafted dataset because our pipeline requires global persona and context input, which is only available from the full theatre script. We compare our pipeline with publicly available speech-based (Hartmann, 2022) and text-based (R-F, 2022) pre-trained models in Table 3.

4.2 LLM Finetuning

In the emotion-tagging pipeline, we fine-tune LLM (GPT-turbo 3.5), leveraging DSPy with human-annotated training data consisting of 50 training and 200 test samples. The final pipeline leverages chain-of-thought and BootstrapFewshotWithRandomSearch as a teleprompter. Fine-tuned LLM leveraging our pipeline significantly exceeds the baseline LLM, as shown in Table 4.

4.3 Ablation Study

To validate the usefulness of each component fed to an LLM in the pipeline, we conduct an ablation study by bypassing each component. Our proposed model feeds global persona, previous context, singer, and lyrics to LLM to bring out the final emotion tag. We bypass each step to compare the usefulness. Ablations are of four groups: Ablation1 (Text), Ablation2 (Text + Character), Ablation3 (Text + Character + Scene Summarization),

Lyrics	Ablation 1	Ablation 2	Ablation 3	Proposed
Anna: For the first time in forever	hopeful	hopeful	excited	hopeful
Anna: I could be noticed by someone	vulnerable	fearful	hopeful	hopeful
Anna: And I know it is totally crazy	excited	nervous	excited	playful
Anna: To dream I'd find romance	hopeful	optimistic	excited	excited
Anna: But for the first time in forever	fearful	fearful	optimistic	hopeful
Anna: At least I've got a chance	pressured	pressured	hopeful	optimistic
Elsa: Don't let them in, don't let them see	fearful	anxious	fearful	anxious
Elsa: Be the good girl you always have to be	overwhelmed	frustrated	pressured	pressured
Elsa: Conceal, don't feel, put on a show	numb	anxious	fearful	pressured
Elsa: Make one wrong move, and everyone will know	anxious	anxious	anxious	fearful

Ablation 1: Text only. **Ablation 2:** Text + Character. **Ablation 3:** Text + Character + Scene Summarization.
Proposed: Text + Character + Scene Summarization + Global Persona.

Table 5: Ablation results of musical film Frozen.

Ablation 1	Ablation 2	Ablation 3	Proposed
2.72 ± 0.07	3.01 ± 0.08	3.33 ± 0.08	$\mathbf{3.60 \pm 0.07}$

Table 6: Mean opinion scores (MOS) of tags from the tagging models with 95% confidence intervals.

and Proposed (Text + Character + Scene Summarization + Global Persona).

Table 5 shows the ablation results of the musical film 'Frozen.' From a qualitative analysis perspective, in Ablation 1, when only text is fed to the LLM, the model judges emotion solely based on lyrics. In Ablation 2, the speaker is also fed with text. Hence, LLM can distinguish two different singers, distinguishing the contrasted emotions of the two singers. In Ablation 3 and the proposed method, in which both previous contexts are fed, LLM understands the context of the singing, one character singing in joy while another faces the pressured situation.

We conduct subjective tests to evaluate the fitness of generated tags per each ablation and proposed tagging pipeline. We randomly selected samples from the dataset and tested 50 data samples with text, character, and generated emotion tags, 25 samples each for speech and singing. The test was conducted on 27 subjects. The results of the four groups are shown in Table 6. As shown in Table 6, the proposed tagging pipeline shows better tagging results than bypassed pipelines in ablations.

4.4 Dataset Analysis

Figure 5 shows the distributions of the frequency of the tags. In clustered tags of primary emotions, the top tag with the highest frequency is 'hopeful,' followed by 'curious,' which is a subset of the primary emotion 'happy.'. The least frequent tag is 'sleepy,' Figure 6 shows the word cloud of 128 emotion tags.

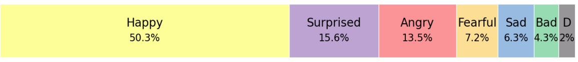

Figure 5: The distribution of primary emotion labels.

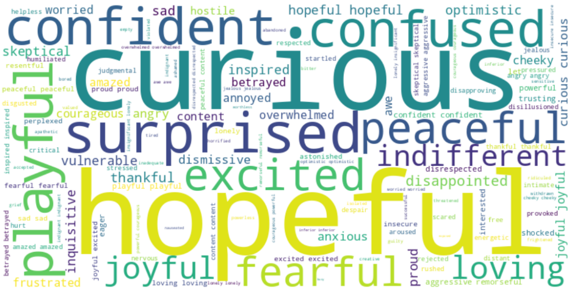

Figure 6: Word cloud of emotion tags in Emosical.

4.5 Tagging Model

We conduct vocal emotion tagging experiments using the Emosical dataset. We designed a simple baseline model for classifying both speech and singing voices into 7 primary emotions. The model is a convolutional neural network (CNN) architecture, starting with a convolutional layer with 32 filters, followed by batch normalization and ReLU activation. It includes three sequential residual blocks, each doubling the number of filters (64, 128, and 256) and incorporating batch normalization and shortcut connections. Adaptive average pooling reduces the feature map to a fixed size, followed by dropout for regularization. The fully connected layers reduce the features to 128 dimensions and finally to the 7 emotion classes, with the output using log softmax activation. The model is trained with the cross-entropy loss function and optimized using the AdamW optimizer (Loshchilov,

	AUC	F-score	Precision	Recall
Singing	0.598	0.219	0.146	0.178
Speech	0.573	0.153	0.225	0.221
Both	0.611	0.129	0.120	0.167

Table 7: Voice emotion tagging results with different dataset configurations.

2017) with a OneCycleLR learning rate scheduler (Smith and Topin, 2019). The performance of the baseline tagging model is reported in Table 7.

5 Conclusion

We presented a novel dataset, Emosical, the first open-source multimodal dataset specifically curated for musical films with comprehensive emotion annotations. By integrating video, audio, text, and character identity with emotion tags derived from a detailed narrative context, Emosical provides a rich resource for advancing research in emotion recognition, synthesis, and tagging in the musical theatre domain.

Our dataset leveraged a novel annotation pipeline, incorporating global persona, scene context, visual description, and dialogue or lyrics to generate nuanced emotion tags using a large language model (LLM). Through statistical analysis and a series of ablation studies, we demonstrated the effectiveness of our tagging scheme. Our subjective evaluations further validated the precision and reliability of our annotations.

Additionally, we proposed a baseline tagging model for emotion recognition in singing voices, setting a foundation for future research in this area. Emosical opens up new avenues for exploring the interplay between various modalities in conveying emotions and can serve as a valuable resource for developing more emotionally resonant systems.

Future work may include expanding the dataset to encompass more diverse genres and languages, refining the emotion tagging pipeline, and exploring its applications in various multimodal emotion recognition and synthesis tasks. We believe Emosical can contribute to further research in multimodal understanding of emotion expressions in musical theatre.

6 Limitations

Several limitations should be noted for future work and improvements in Emosical.

- *Diversity of Source Material.* The dataset is currently limited to 20 distinct musical films, which may not fully capture the wide range of emotional expressions and styles present across different musical theatre productions. So, we plan to expand the dataset to include more films, as well as musical recordings from live theatre performances to enhance the generalizability of models trained on this data.
- *Manual Intervention During Data Processing.* While we automated much of the data processing pipeline, certain steps, such as verifying SRT timestamp accuracy and checking speaker diarization results, still require human intervention. Further refinement and automation of these processes would improve the efficiency and scalability of dataset creation.
- *Emotion Tagging Granularity.* Although we employ an extensive set of 128 emotion tags based on the emotion wheel, this granularity can lead to challenges in ensuring consistent and accurate tagging across samples. In some cases, the subtleties between closely related emotions might be difficult to distinguish, leading to potential ambiguities.
- *Dependency to LLMs.* Our emotion tagging relies on LLMs' capabilities. While these models offer sophisticated natural language understanding, they are not infallible and can sometimes generate inaccurate or inconsistent tags, especially when faced with highly nuanced emotional expressions.
- *Bias and Representation.* The selected musical films may reflect certain cultural biases and predominantly represent Western musical theatre traditions. This limits the applicability of the dataset for studying emotions in a more global and culturally diverse context. Future efforts should include a more diverse range of films from various cultures and languages.
- *Temporal Context and Dynamics.* While the dataset includes scene summarization and global persona information, capturing the full temporal dynamics and evolution of emotions over longer periods within the films remains a challenge. Future work could focus on better integrating temporal context to understand how emotions develop and change over time.
- *Quality of Vocal Isolation.* We observed that the quality of isolated vocals varies, particu-

larly when background music or noise is complex. Improving vocal isolation methods or exploring alternative approaches could enhance the clarity and usability of the audio samples.

- *Evaluation Metrics and Human Subjectivity.* Emotions' subjective nature indicates that human evaluations can vary, impacting the consistency of our MOS tests and other evaluation metrics. Developing more objective and standardized evaluation methods would be beneficial for assessing the quality of annotations.

Addressing these limitations in future iterations of `Emosical` will help create a more robust and comprehensive dataset, ultimately contributing to the advancement of multimodal emotion recognition and synthesis research in the domain of musical theatre.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and partly by Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) [No. RS-2022-II220320, 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multimodal interactions].

References

Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Antonin Artaud. 1938. *The Theater and Its Double*. Grove Press, New York.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Pablo V. A. Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The omg-emotion behavior dataset. *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. *Preprint*, arXiv:2305.09758.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. 2005a. A database of german emotional speech. volume 5, pages 1517–1520.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005b. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Qi Chen, Yuanqing Li, Yuankai Qi, Jiaqiu Zhou, Mingkui Tan, and Qi Wu. 2021. V2c: Visual voice cloning. *Preprint*, arXiv:2111.12890.

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022. V2c: visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21242–21251.

Abhinav Dhall, Monisha Singh, Roland Goecke, Tom Gedeon, Donghuo Zeng, Yanan Wang, and Kazushi Ikeda. 2023. Emotiw 2023: Emotion recognition in the wild challenge. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 746–749.

Jochen Hartmann. 2022. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Hawkins et al. 2024. Gentle. Accessed: 2024-06-14.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-Singer: Fast multi-singer singing voice vocoder with a large-scale corpus. *Preprint*, arXiv:2112.10358.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang. 2022. Cn-celeb: multi-genre speaker recognition. *Speech Communication*, 137:77–91.

Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*.

Steven R. Livingstone and Frank A. Russo. 2018. 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*, 13(5):1–35.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, pages 137–144.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*, volume 1. Academic press.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Preprint*, arXiv:1810.02508.

R-F. 2022. Wav2vec english speech emotion recognition. `https://huggingface.co/r-f/wav2vec-english-speech-emotion-recognition`. Accessed: 2024-06-14.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. 2023. Speech enhancement and dereverberation with diffusion-based generative models. *Preprint*, arXiv:2208.05830.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic. 2019. Avec'19: Audio/visual emotion challenge and workshop. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2718–2719.

Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. pages 1–8.

Geoffrey Roberts. 2024. Feelings wheel. Accessed: 2024-06-14.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015a. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015b. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 23*.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Aharon Satt, Shai Rozenberg, Ron Hoory, et al. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093.

Arnav Singhvi, Manish Shetty, Shangyin Tan, Christopher Potts, Koushik Sen, Matei Zaharia, and Omar Khattab. 2024. Dspy assertions: Computational constraints for self-refining language model pipelines. *Preprint*, arXiv:2312.13382.

Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Constantin Stanislavski. 1948. *Building a Character*. Theatre Arts Books, New York.

J. Sundberg. 1987. *The Science of the Singing Voice*. Northern Illinois University Press.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. *Preprint*, arXiv:1512.02902.

Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Dimitrios Ververidis and Constantine Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. 2018. VocalSet: A singing voice dataset. In *International Society for Music Information Retrieval Conference*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

AmirAli Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. 2015. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. 2022. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Preprint*, arXiv:2105.14762.

Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. 2017. Cognimuse: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017.

Zulko et al. 2024. Moviepy. Accessed: 2024-06-14.