Emosical: An Emotion Annotated Musical Theatre Dataset

Anonymous ACL submission

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 derive the emotion tags, we leverage the musical theater script, which contains the characters' complete background stories and narrative contexts. The annotation pipeline includes feeding the singing character, text, global persona, and context of the dialogue and song track into a large language model (LLM). To verify the effectiveness of our tagging scheme, we perform an ablation study by bypassing each step of the pipeline. 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 singing voice in the musical theatre domain in future research.

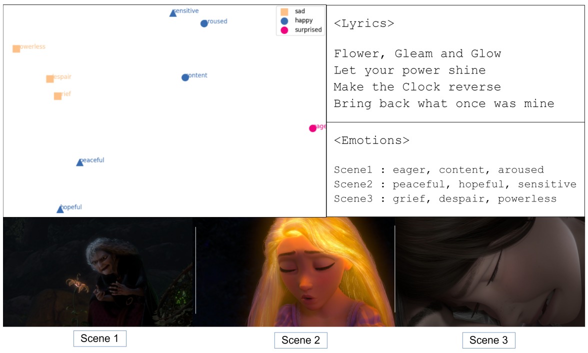

Figure 1: Emotion embedding visualization of 'Healing Incantation' in 'Tangled' using T-SNE. Different colors mean different primary emotions of detailed emotions drawn, and different markers indicate different songs in the film. 'Healing Incantation' is reprised triple times in the movie. Even though they all have the same lyrics our tagging pipeline tags corresponding singing emotions well by inferring emotions from the context and character's persona.

1 Introduction

Emotion is a fundamental aspect of the human experience, distinguishing us from machines. Many researchers are endeavoring to develop AI systems capable of inferring human emotions, which is being vigorously explored within the natural language processing (NLP) domain. Several studies employing various methodologies have focused on creating more emotionally engaging generative models using datasets labeled with emotion tags (Livingstone and Russo, 2018; Zaragozá et al., 2024). Additionally, numerous efforts have been made to understand emotions in multi-modal media (Barros et al., 2018; Zadeh et al., 2018), including YouTube-crawled videos annotated with emotions. Other studies have aimed to create comprehensive multimodal datasets with diverse sources and detailed emotion annotations (Busso et al., 2008a; Köprü and Erzin, 2020).

However, there are still difficulties in accurately extracting emotion tags or annotations. This is because it is challenging to identify 1) what the mediums for conveying emotions are and 2) how these emotions are conveyed through these mediums. These difficulties arise from the fact that emotions are fully realized through not only linguistic elements but also non-linguistic elements such as facial expressions, music, context, and gestures. We propose that *theatre* is a particularly effective medium for addressing these challenges, as it inherently integrates both linguistic and non-linguistic elements in conveying emotions.

As renowned actor Sanford Meisner once remarked, "The greatest piece of acting or music or sculpture or what-have-you always has its roots in the truth of human emotion." Theatre excels at conveying the emotions of the story to the audience. Actors and directors use various techniques such as dialogue, music, lighting, and stage design to communicate a wide range of emotions to the audience. In this view, as a complex art form, theatre is an unparalleled multimodal medium.

Despite this, there is no dataset specifically for theatre in emotion research. This absence is attributed to theatre's inherent complexity. As mentioned, theatre is a combination of text, audio, and visual elements. Unlike general emotional speech or recorded facial videos typically found in multimodal datasets, creating a comprehensive musical theatre dataset requires significant financial and time costs. For instance, capturing the full range of modalities involved in a theatrical performance requires sophisticated and often expensive recording equipment. Furthermore, theatrical performances are live events, making it to create consistent, high-quality recordings challenging.

In the case of musical theatre, the challenge is even greater. Since musical theatre incorporates singing, which itself is a powerful medium for emotional expression, the vocal characteristics in theatre should be categorized into spoken dialogue and sung lyrics, each requiring different recording and analysis techniques. Singing as unimodal data demands attention to nuances like pitch, tone, and emotional delivery, further complicating the data collection. That's why, currently, there is no public singing data annotated with emotions aside from (Livingstone and Russo, 2018). Therefore, creating a comprehensive dataset to study the relationship between theatre and emotion remains unfulfilled.

In response, we build a dataset of theatre, specifically for 'musical theatre,' which uniquely consists of elements 'music' and 'singing.' Given the complexity of collecting musical theatre data, we opted not to build the dataset from scratch but to crawl and analyze existing data. We also aim to design a pipeline that can analyze and annotate a narrative's emotions as automatically as possible.

We aim for this dataset to be primarily used to understand the relationship between theater's multimodal characteristics and emotions. Additionally, due to the multimodal dataset's nature and musical theatre's unique feature distribution, we hope it will also be used for tasks such as emotion tagging and emotional synthesis for each modality in musical theatre. Therefore, when constructing the dataset, we divided it into 5-second samples and provided detailed annotations. Most existing datasets annotate emotions in broad groups over long segments. Instead, we have applied dense emotion tags to short data samples, allowing for more precise and temporal studies of emotional changes.

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 singing voice tagging model that leverages this annotated dataset for emotion recognition in singing voices.

2 Related Works

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., 2008b) 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

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., 2008a)	✓	✓		✓		✓	-	10039	10	9
VocalSet (Wilkins et al., 2018)	✓		✓				-	3560	20	-
OpenSinger (Huang et al., 2021)	✓		✓				-	80 hours	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
Emosical (ours)	✓	✓	✓	✓	✓	✓	10	25354	261	128

Table 1: Open-Source Dataset Comparison

people reacting to predefined stimuli, with annotations for continuous emotion dimensions, providing continuous perspectives on emotional responses.

The Audio-Visual Emotion Challenge (AVEC) provides datasets including synchronized video and audio recordings annotated with emotional states. The Emotion Recognition in the Wild (EmotiW) challenge 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 in a neutral North American accent. 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 across multiple languages, which provides a diverse set of emotional speech samples for cross-linguistic 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

3.1 Overview

`Emosical` comprises n samples, totaling n hours, from 10 distinct musical films, including theater recordings and theater-like cinematics. Each sample is a tuple of {audio, video, text, character} accompanied by annotated emotion tags. Samples include n speech and n singing samples. Table 1 outlines several key characteristics of the dataset, including the types and numbers of annotated tags and the number of characters compared to relevant datasets.

3.2 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 raw subtitle files that contain characters and text aligned to the movie with metadata. The metadata contains emotion and vocal type per sample, as well as noisy samples to eliminate or run a speech enhancement model. In the raw dataset, users will place movie video files in the theatre directory, along with corresponding subtitle files in the `SRT` directory. After users place the movies in the speci-

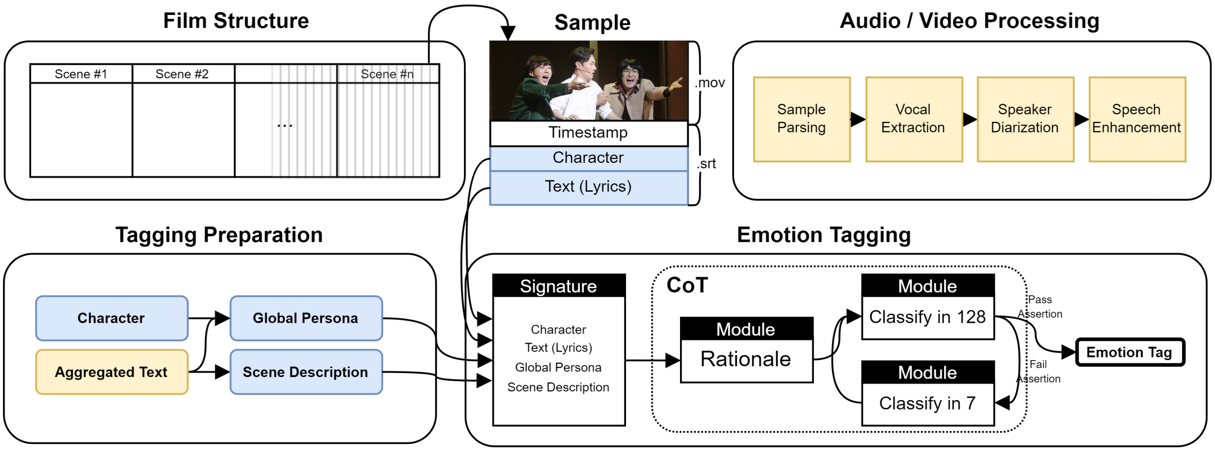

Figure 2: Dataset collection pipeline of `Emosical`. We use srt and raw video file to process the data. We parse audio and video samples according to the timestamp of srt file and process audio to get pure vocal. After text, character, audio, and video pair is readied we run the emotion tagging pipeline.

fied folder and compile the data using the provided code, the dataset structure transforms into the following compiled form, where audio and video files are organized by scenes within movie-specific directories. This structure allows users to access individual audio clips from specific scenes of each movie.

3.3 Dataset Collection

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.

Movie Gathering. First, we obtain the musical film video. We select 10 movies containing musical components, with their subtitle files (`SRT`) readily available. `SRT` files contain the sequential number of current utterances, starting and ending points in the video timeline, and corresponding text. Since we will split the video with `SRT` timestamp and align text with audio, we need to precisely tune the timestamp and text of each `SRT` segment to contain the starting and ending point of each utterance properly. 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 around 5 seconds.

Video Parsing. 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.

Audio Parsing and Vocal Isolation. In the case of audio sample processing, considering the multitude of purposes of the dataset, such as voice synthesis and tagging tasks, we processed our audio data to be pure voice without background audio. The initial phase of our audio processing pipeline involves decoupling of vocals from the video. We isolate the audio track from the movie file and extract the center channel, both utilizing the `ffmpeg` toolkit. We extract the center channel (sum of the left and right audio channels) to minimize the influence of background music, predominantly isolating the main characters' vocal utterances to the greatest extent possible. Then, we utilize the open-source `Demucs` algorithm (Rouard et al., 2023) to extract the singing vocal from the center channel. And with the assistance of `SRT` files, we segment the audio into discrete clips.

Speech Enhancement for Audio Samples. After chopping the audio into segments, we check for noisy audio files. For noisy audio, even after the vocal isolation, we note them additionally to purify the background noise further and employ the background noise reduction model (Kim and Hahn, 2019) to bring out the final audio. We then eliminate audio clips that don't match our requirements. These involve overlapping voices or singing voices with residual noise artifacts despite the preprocessing process. We manually exclude these segments since we aim to curate an automatically processable dataset.

Speaker Diarization for Audio Samples. After collecting audio data and its' corresponding text, each audio clip is annotated with the corresponding singer's identity and matched against the SRT file. This is for distinguishing unique singers and also enables large language models (LLMs) to effectively discern each character and categorize the emotional nuances conveyed through the storyline in the tagging process. The intricacies of employing singer-specific information for emotional tagging will be explained in detail in Section 3.4. To identify singing characters, we first use a pre-trained speaker diarization model (Wang et al., 2023) trained to identify speaker similarity in both singing and speech audio. We gather all vocal audio of talking characters in the movie and compute speaker similarity by all vocal segments. Then, we temporally assign a speaker with the highest similarity. However, due to the everchanging nature of musical theatre's speech and singing, the diarization result was not perfect, so we manually checked each line and modified the character annotations. Through this data collection pipeline, we finally gather a triplet of {vocal, text, singer} for audio data. In the metadata, singing audio is checked to distinguish it from speech audio.

3.4 Emotion Annotation

As we collected the {video, audio (vocal), character, text} data through the mentioned pipeline, now we aim to annotate the emotion for each sample. To this end, we focus on the storyline of the theatre to further infer the emotion of the character line by line, similar to the approach in (Bhattacharya et al., 2023), which generated story descriptions to handle downstream tasks. We leverage full text from the srt file, utilizing a LLM. The annotation process integrates four key components for each character: global persona, scene summarization, visual description, and the text of each sample.

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

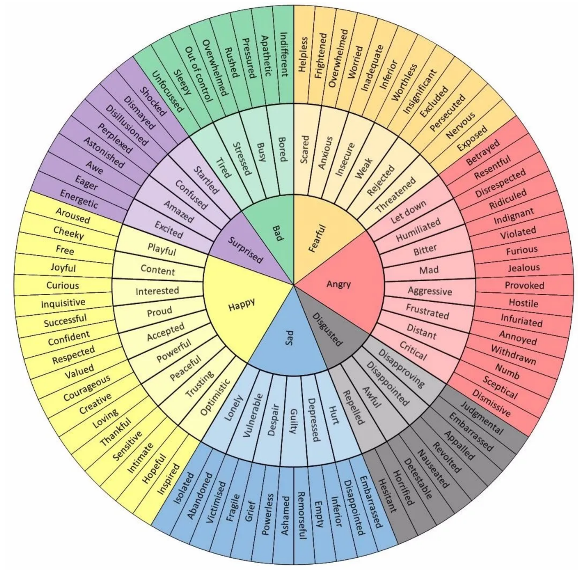

Figure 3: 128 emotion wheels with 7 primary, 40 secondary, and 81 tertiary emotions.

Scene Separation and Summarization. We separate scenes to effectively summarize the context of each chunk of film, which is done arbitrarily. Then, we feed the aggregated text of the scene into LLM to obtain a summarized story. Summarizing the scene helps infer 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 to LLM helps LLM follow the storyline and understand the personality of the character shown throughout the musical theater, aiding LLM to successfully guess the emotional state of the character when saying specific text or singing specific lyrics.

The Emotion Wheel. The majority of emotion-annotated datasets categorize emotions into 4 to 8 groups. However, to capture the meticulous, nuanced emotions conveyed in the musical film, we require emotion labels with sophisticated distinct emotions for annotation. So, we classify the emotion tags following the emotion wheel. The widely known Plutchik emotion wheel (Plutchik and Kellerman, 2013) is developed to categorize human emotions based on the idea that distinct emotions can be mixed and 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,' '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) to generate emotion annotations for each line of the dataset. We utilize the DSPy framework (Khattab et al., 2023) to facilitate optimizing language model prompts and weights. We define character, text, visual description, scene context, global persona with GT emotion tag as DSPy Signature and feed corresponding data for training the LLM. Then, we apply a chain of thoughts method to infer the emotion tag. The first model predicts rationale about input Signatures. The second model classifies the emotion using the rationale to classify the emotion of 128 tags. However, due to the unconstrained nature of LLM outputs, LLM tends to output tags out of emotion lists. So we added a dspy. Suggest constraints to the module (Singhvi et al., 2024), and when the module exceeds max backtracks, we use the second classification model, which acts as a teacher. The teacher module classifies emotion into 7 primary emotion tags and then passes the primary emotion as a hint to the 128 emotions classification module. We pre-train chain-of-thought modules with training sets from unseen musicals. The compiled module significantly exceeds the untrained baseline module. In summary, we annotate emotion tags with a pipeline containing - gathering global persona, scene summarization, visual description, and feeding singer and lyrics with LLM prompting. An ablation study of this annotation pipeline is presented in Section 4, detailing the impact and significance of each component in the emotion annotation accuracy. The dataset collection and the annotation process are further elaborated in Figure 1, providing a visual overview of the methodology.

Statistics	Count
Total # of films	10
Total # of video samples	12677
Total # of singing samples	2374
Total # of speech samples	10303
The average length of video samples	5.21s
Total # of distinct speakers	261
Total # of emotion tags	128
Total # of words in sentences	62792

Table 2: Summary of `Emosical` dataset statistics.

Figure 4: The tag frequency of the primary emotions is shown as a bar plot.

4 Dataset Analysis

4.1 Global Characteristics of Emotion Tags

Figure 4 shows the distributions of the frequency of the tags. In clustered tags of primary emotions, the top tag with the highest frequency is 'happy,' followed by 'surprised', and the least frequent tag is 'disgusted,' Figure 5 shows the word cloud of 128 emotion tags. Most tags are generated once, and the top tag with the highest frequency is 'playful' and 'excited', which is a subset of the primary emotion 'happy.'

4.2 Case Study of Tag Annotations

Among numerous movies and song tracks in our dataset, we choose 'Healing Incantation' from the movie 'Tangled' as an example of the tag annotations at the song level. In musical theater, some prominent songs tend to be reprised and emerge multiple times throughout the act, conveying different emotional nuances. The number 'Healing Incantation' is the case, which emerges two times throughout the movie, once at an introductory moment and once at a highly-elated scene of the movie. Figure 1 shows that even though the song consists of mostly the same lyrics, the resulting tags are different due to different scene contexts fed to obtain the tags. There is a noticeable difference of emotion tags between the three song tracks illustrated.

4.3 Ablation Study

To validate the usefulness of each step of the pipeline, we conduct an ablation study by bypassing each step of the pipeline. 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), and Proposed (Text + Character + Scene Summarization + Global Persona).

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

Table 3: Ablation results of musical film 'frozen'. Ablation 1: Text, Ablation 2: Text + Character, Ablation 3: Text + Character + Scene Summarization, Proposed: Text + Character + Scene Summarization + Global Persona.

Figure 5: Word cloud of emotion tags in Emosical.

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

Table 3 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, while in Ablation 2, when the speaker is fed with text, LLM recognizes two different singers, distinguishing the contrasted emotions of the two singers. While 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 one 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 samples of data with text, character, and generated emotion tags, 25 samples each for speech and singing. The test was conducted on 27 people. The results of the four groups are shown in Table 2. As shown in Table 2, the proposed tagging pipeline shows better tagging results than bypassed pipelines in ablations.

	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 5: Voice emotion tagging results with different dataset configurations.

5 Tagging Model

We performed 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 with a OneCycleLR learning rate scheduler. The performance of the baseline tagging model is elaborated in Table 4.

6 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.

7 Limitations

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

- *Diversity of Source Material.* The dataset is currently limited to 10 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, particularly 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.

References

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.

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.

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. 2008a. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008b. 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.

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.

Juntae Kim and Minsoo Hahn. 2019. Speech enhancement using a two-stage network for an efficient boosting strategy. *IEEE Signal Processing Letters*, 26(5):770–774.

Berkay Köprü and Engin Erzin. 2020. Multimodal continuous emotion recognition using deep multi-task learning with correlation loss. *arXiv preprint arXiv:2011.00876*.

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.

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.

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

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.

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*.

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.

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.

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.

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

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.

Lucía Gómez Zaragozá, Rocío del Amor, Elena Parra Vargas, Valery Naranjo, Mariano Alcañiz Raya, and Javier Marín-Morales. 2024. Emotional voice messages (EMOVOME) database: emotion recognition in spontaneous voice messages. *Preprint*, arXiv:2402.17496.

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.