Expanding the Singular Channel - MODALINK: A Generalized Automated Multimodal Dataset Generation Workflow for Emotion Recognition

Amany H. AbouEl-Naga[1], May Hussien[1], Wolfgang Minker[2], Mohammed A.-M. Salem[3] and Nada Sharaf[1]

[1]*Faculty of Informatics and Computer Science, German International University, New Capital, Egypt*

[2]*Institute of Communications Engineering, Ulm University, Ulm, Germany*

[3]*Faculty of Media Engineering and Technology, German University in Cairo, New Cairo, Egypt*

Keywords: Emotion Recognition, Multimodal Emotion Recognition, Dataset Generation, Artificial Intelligence, Affective Computing, Deep Learning, Machine Learning, Multimodality.

Abstract: Human communication relies deeply on the emotional states of the individuals involved. The process of identifying and processing emotions in the human brain is inherently multimodal. With recent advancements in artificial intelligence and deep learning, fields like affective computing and human-computer interaction have witnessed tremendous progress. This has shifted the focus from unimodal emotion recognition systems to multimodal systems that comprehend and analyze emotions across multiple channels, such as facial expressions, speech, text, and physiological signals, to enhance emotion classification accuracy. Despite these advancements, the availability of datasets combining two or more modalities remains limited. Furthermore, very few datasets have been introduced for the Arabic language, despite its widespread use (Safwat et al., 2023; Akila et al., 2015). In this paper, MODALINK, an automated workflow to generate the first novel Egyptian-Arabic dialect dataset integrating visual, audio, and text modalities is proposed. Preliminary testing phases of the proposed workflow demonstrate its ability to generate synchronized modalities efficiently and in a timely manner.

1 INTRODUCTION

Emotions are a collection of mental states triggered by different ideas and actions. People continually express their emotions when they communicate, and recognizing emotions is crucial for many facets of daily living and interpersonal communication. Emotion is vital in interpersonal interactions, insight, perception, and other areas of life (Radoi et al., 2021).

Although emotions are expressed differently across societies and situations, they can be seen as a universal language that transcends linguistic and cultural barriers. While cultural norms influence expressions of love or sorrow, fundamental similarities highlight universal human experiences (Kalateh et al., 2024b).

Identifying emotions is a key component of human communication, driving research into emotion recognition to develop systems capable of recognizing and understanding human emotions through the analysis of the modality that expresses the emotion itself (Ahmed et al., 2023). Emotion recognition has broad applications, from mental health monitoring to enhancing customer service, user experience, and natural human-robot interactions (Islam et al., 2024; Kalateh et al., 2024b; Ahmed et al., 2023).

For years, research has focused on unimodal emotion recognition methods, such as facial expression, speech, and text analysis. Despite achieving reasonable results, unimodal approaches are limited due to the inherently multifaceted nature of emotional expressions. A single modality provides restricted information, making it difficult to interpret emotions accurately in complex social contexts. Factors such as lighting, camera angles, stress, and physical activity can alter how emotions are perceived, while variations in speech signals due to language and cultural differences further challenge accuracy. Additionally, unimodal algorithms are vulnerable to input data fluctuation and noise (Kalateh et al., 2024b; Ahmed et al., 2023; Vijayaraghavan et al., 2024).

To address these shortcomings, researchers have

AbouEl-Naga, A. H., Hussien, M., Minker, W., Salem, M. A.-M., Sharaf and N.
Expanding the Singular Channel - MODALINK: A Generalized Automated Multimodal Dataset Generation Workflow for Emotion Recognition.
DOI: 10.5220/0013520100003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 415-422
ISBN: 978-989-758-758-0; ISSN: 2184-285X

turned to multimodal emotion recognition (MER), integrating modalities such as facial expressions, speech, and text. Recent studies demonstrate that MER outperforms unimodal approaches, as different modalities complement each other, providing a more comprehensive understanding of emotions (Vijayaraghavan et al., 2024; Kalateh et al., 2024b).

Many datasets have been introduced, primarily covering audio, visual, and text modalities, often integrating two or all three. These datasets facilitate the development and evaluation of MER models by providing diverse emotional states with careful curation and annotation (Ahmed et al., 2023; Vijayaraghavan et al., 2024).

Despite the availability of MER datasets, Arabic remains underrepresented. Most studies focus on English, Western European, and East Asian languages, leaving a gap in Arabic MER research. With an estimated 420 million Arabic speakers, including 310 million native speakers across eight major dialect groups, the lack of Arabic MER datasets is a significant challenge. Arabic is often considered a low-resource language in emotion recognition and NLP due to the limited availability of annotated datasets. This scarcity is particularly problematic given the linguistic complexity and dialectal diversity of Arabic, including the diglossia between Modern Standard Arabic and regional dialects (Arabiya, 2025; Al Roken and Barlas, 2023). The Egyptian dialect has unique phonological and lexical traits that may influence emotion expression and recognition. This work proposes an automated process for creating a multimodal dataset in Egyptian Arabic, integrating text, visual, and audio modalities for emotion recognition tasks. The MODALINK framework is designed to bridge the gap in Arabic MER research, specifically for the Egyptian dialect.

MODALINK ensures a wide range of emotions and participant diversity by automating the processing and annotation of Egyptian talk shows and series using advanced deep learning techniques. It incorporates natural language processing, facial expression analysis, and speech recognition to extract relevant features from each modality. The objective is to reduce the time and resources needed for dataset creation while maintaining high-quality, consistent classification and synchronization across modalities.

The rest of the paper is structured as follows: Section 2 reviews related work, Section 3 details the automated workflow and dataset structure, Section 4 presents the findings, and Section 5 concludes the study.

2 RELATED WORK

Emotion recognition has been studied for a long time, initially using textual data and gradually moving to facial expressions and, ultimately, multimodal techniques to boost accuracy(Kalateh et al., 2024a). This move was prompted by advances in AI and the need for improved emotion recognition. Researchers improved applications in affective computing, human-computer interaction, virtual assistants, mental health monitoring, and entertainment by combining text, facial expressions, and audio.

Multimodal emotion recognition overcomes the limitations of single-modality techniques and provides a more accurate understanding of emotions. High-quality, large-labeled datasets are critical for creating and testing AI emotion models (Kalateh et al., 2024a). Diverse datasets that account for demographic, cultural, and environmental variations enable advancement by facilitating new architectures and fusion methods for complicated emotion detection.

2.1 Popular Datasets in Multimodal Emotion Recognition

To facilitate experimental research and study, many datasets have been developed in multimodal emotion recognition (MER). Nevertheless, several areas still need more work and attention, even with the large diversity of datasets accessible in the literature. Every dataset comes with its own set of strengths and limitations. Here, we will provide an overview of the most famous MER datasets.

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) (Busso et al., 2008), this dataset was developed by the USC Signal Analysis and Interpretation Lab (SAIL) and is commonly utilized in research on emotion recognition. It contains video, speech, and text data, allowing for a comprehensive examination of emotional states in interactive environments. The dataset includes recordings from ten actors spanning approximately twelve hours, including scripted and spontaneous dialogue. It comprises 4,784 improvised and 5,255 written conversations across nine emotions: happy, sadness, rage, surprise, fear, disgust, frustration, excitement, and neutrality. IEMOCAP's positives are its good quality, synchronized multimodal data, and combination of acted and spontaneous emotional reactions. Its small size, however, restricts its potential for deep learning models, whereas the imbalance in emotion classes can affect model performance. Furthermore, as an English-only dataset, it lacks cultural diversity.

The CMU-MOSEI dataset is one of the largest avail-

able for multimodal sentiment analysis and emotion recognition (Bagher Zadeh et al., 2018). This dataset includes more than 12 hours of annotated segments from YouTube videos by more than 1,000 speakers on approximately 250 different topics. The emotions represented in the dataset are: anger, happiness, disgust, sadness, fear, and surprise. The data collection techniques employed ensure that every video involves only one speaker, further establishing the mono-logic nature of this dataset. Although the CMU-MOSEI dataset is well-known for its vast size and is still widely utilized for model development and validation, it has a few limitations. The dataset is limited to English, which may limit the ability of models trained on it to generalize to other languages. Furthermore, like with other emotion datasets, some emotions are underrepresented.

A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations(MELD) (Poria et al., 2018), by focusing on interactions between several participants, the MELD dataset offers a novel perspective on multimodal emotion recognition. Each statement is classified as anger, disgust, sadness, joy, neutrality, surprise, or fear. It includes 13,708 utterances and 1,433 conversations from the American television series Friends. This dataset addresses a research gap by assisting in the investigation of emotional dynamics in multi-party interactions. Its cultural significance and reality present difficulties, though. The conversation and emotions may be exaggerated because it is based on a written script, which lessens realism. Furthermore, models trained on MELD may not work well in other cultural contexts due to their English language and American cultural standards, underscoring the necessity for additional varied datasets in emotion recognition studies.

Sentiment and Emotion, Well-being, and Affective dynamics dataset (SEWA) (Kossaifi et al., 2019) is a valuable resource in the field. This dataset closes a significant gap in emotion recognition studies through incorporating linguistic and cultural diversity. It contains audio-visual data in six languages—English, German, French, Hungarian, Greek, and Chinese—that capture natural emotions and conversational dynamics from people aged 18 to 65. While it improves multilingual research, the small number of participants limits generalization, and many commonly utilized languages are underrepresented, emphasizing the need for additional expansion in the future.

2.2 Arabic Multimodal Emotion Recognition Dataset

Multimodal emotion recognition (MER) has received a lot of attention, thanks to increased research on cultural and language diversity. Despite its widespread use and numerous dialects, Arabic MER has received very little attention. The enormous variety among Arabic dialects makes it difficult to construct a universal emotion recognition model. This section focuses on the efforts accomplished to develop Arabic-specific MER datasets.

The first audio-visual Arabic emotional dataset is called (AVANEmo), One of the early efforts to create an Arabic MER dataset was introduced in (Shaqra et al., 2019). The dataset contains 3000 clips for video and audio data, covering one of the following emotions: Happy, Sad, Angry, Surprise, Disgust, Neutral. They focused on videos that contain personal stories or life experiments, we also considered the talk shows where speakers shared their perspectives on a specific issue. They ended up with 90 video sources and 82 speakers of different age groups. While this was a good start, the dataset only focused on audio and visual features. The dataset has a mix of Arabic which could be challenging during processing.

An Arabic Multi-modal emotion recognition dataset was introduced in (Al Roken and Barlas, 2023). This dataset has only visual and audio modalities. Clips were collected from more than 40 shows available on YouTube, featuring various guests from different backgrounds and using different Arabic dialects. The clips were annotated with one of the following emotions: Angry, Happy, Neutral, Sad, and Surprised. While this dataset seeks to fill a gap by providing an Arabic MER dataset, it faces several limitations. The dataset is relatively small, with a limited number of videos and speakers. Additionally, it lacks diversity, as it only includes clips featuring adults, and the number of male speakers exceeds that of female speakers. Finally, the model results presented in the cited work indicate that building dialect-specific datasets may be beneficial at this stage, as there is still more room for improvement.

Tables 1, 2 show that existing emotion recognition datasets have made significant contributions to the domain, yet they still do not represent Arabic dialects, Egyptian Arabic in particular, adequately. The majority of the datasets focus on Modern Standard Arabic (MSA) or a combination of dialects, which do not capture real speaking patterns and emotional expression of Egyptian Arabic. To fill this gap, we propose to create a multimodal emotion recognition dataset

Table 1: Overview of Emotion Recognition Datasets.

Dataset	Language	Emotions	Source	Modality
EMOCAP (Busso et al., 2008)	English	Happiness, sadness, anger, surprise, fear, disgust, frustration, enthusiasm, and neutral	Recorded	Audio, Video, Text
CMU-MOSEI (Bagher Zadeh et al., 2018)	English	Anger, happiness, disgust, sadness, fear, and surprise	Youtube	Audio, Video, Text
MELD (Poria et al., 2018)	English	Joy, sadness, anger, fear, disgust, surprise, and neutral	Friends TV Series	Audio, Video, Text
SEWA (Kossaifi et al., 2019)	English, German, French, Hungarian, Greek, and Chinese	Continuous valence and arousal	Recorded	Audio, Video
ANADemo (Shaqra et al., 2019)	Arabic	Happy, sad, angry, surprise, disgust, and neutral	Youtube	Audio, Video
Arabic MER dataset (Al Roken and Barlas, 2023)	Arabic	Angry, happy, neutral, sad, and surprised	Youtube	Audio, Video

Table 2: In-depth view of Arabic Emotion Recognition Datasets.

Dataset	Dialect	Size	Number of Speakers	Availability
ANADemo (Shaqra et al., 2019)	Multi-dialect and MSA	892 segments	82	Not publicly available
Arabic MER dataset (Al Roken and Barlas, 2023)	Multi-dialect and MSA	1336 segments	unknown	Not publicly available

for Egyptian Arabic spanning audio, visual, and text modalities. This dataset aims to naturally represent Egyptian cultural and linguistic features.

3 PROPOSED METHODOLOGY

MODALINK, the proposed automated workflow, generates a multimodal dataset from Egyptian Arabic video content for emotion recognition. The process includes video acquisition, audio extraction, speech and video segmentation, face detection, and speech-to-text transcription (Figure 1). MODALINK integrates audio, visual, and textual modalities to extract emotion-rich segments with transcriptions, ensuring high-quality data while addressing face visibility, speech detection, and transcription accuracy.

3.1 Data Collection

In this step, videos chosen from publicly available Egyptian TV-series, variety shows, and talk shows are downloaded from YouTube using the yt-dlp library. The video is downloaded in its highest quality available in MP4 format ensuring both the audio and video channels are captured into a single file. Only

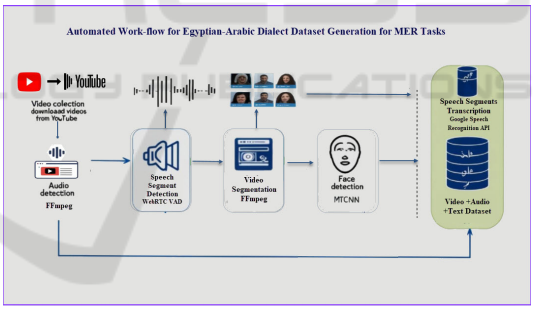

Figure 1: MODALINK: Proposed Automated Workflow for Egyptian-Arabic Dialect Dataset Generation.

videos with clear audio and visible faces are selected to maintain the quality of the dataset. Next, the video file is preprocessed to ensure consistent format.

3.2 Audio Extraction

After downloading the video, the audio track is extracted from the MP4 file using the FFmpeg library (Tomar, 2006). The audio is converted to a standardized PCM format with a sampling rate of 16kHz and a single mono channel to ensure the process of speech detection is optimal. A fallback mechanism is implemented to handle any corrupted or unusable audio files during the extraction.

3.3 Speech Segmentation

The extracted audio is segmented using a WebRTC Voice Activity Detection (VAD), where only the speech regions are retained and the non-speech regions are discarded. The audio signal is divided into frames of 30 ms then each frame is considered for the existence of speech.

Segments shorter than .5 seconds are discarded to ensure that there is sufficient duration and content in each segment to perform emotion recognition. Segments are split based on silence period threshold on .3 seconds. These parameters could be edited according to the need and the nature of the required data.

Continuous speech segments are identified based on the VAD output. MODALINK applies heuristics to merge closely spaced speech segments and discard very short segments, improving the overall quality of detected speech regions.

Detected speech segments information are stored as time intervals (start time and end time), these timestamps are vital to ensure synchronization between all modalities. This process helps to reduce the time and effort required for transcription as the transcription model will only process the detected speech segments.

3.3.1 Audio Feature Extraction and Speaker Change Detection

To enhance speech segmentation and identify speaker changes, MFCCs are extracted from overlapping windows within speech segments to capture key spectral characteristics used in speaker recognition. By analyzing the cosine similarity between consecutive MFCC features, significant changes are detected. Segments are then split at points where the similarity metric falls below a predefined threshold, indicating potential speaker changes. This threshold is adjustable to match the sensitivity required for the specific task.

3.4 Video Segmentation

To ensure face visibility per speech segment, the video segments are extracted from the original video using the timestamps obtained in the prior step. Video clips the match the time stamps are extracted using the FFmpeg library (Tomar, 2006).

The video is then used to extract individual frames at a specified frame rate, typically 1 frame per second. This step utilizes FFmpeg's capabilities to extract frames efficiently.

3.5 Face Detection and Speaker Matching

Each extracted frame undergoes pre-processing to optimize face detection. MODALINK down-scales the frame using a scale factor of 0.5 to enhance processing speed without significantly compromising detection accuracy. The down-scaled frame is then converted from BGR to RGB color space, as required by the face detection model.

The Multi-task Cascaded Convolutional Networks (MTCNN) face detection model is utilized to detect the presence of faces within the first 10 frames of each video segment (Zhang et al., 2016). The MTCNN algorithm uses a cascading series of neural networks to detect, align, and extract facial features from digital images with high accuracy and speed.

Segments without visible faces in the analyzed frames are discarded to ensure that visual data includes human expressions. The model was set to the aggressive mode to ensure that segments without clear human facial expressions are discarded.

A multiple face detection approach is employed, and each detected face is represented by its bounding box coordinates and confidence scores. A heuristic method is applied to match detected faces to the active speaker. The largest face in each segment is assumed to be the active speaker.

3.6 Speech-to-Text Transcription

Speech segments detected previously are fed to an automatic speech recognition model to generate text transcriptions. The Google Speech Recognition API is utilized as part of the MODALINK workflow with language configuration set to Egyptian-Arabic. Segments that fail transcription due to noise or API errors are flagged but not discarded entirely.

3.7 Metadata Generation

Each processed segment is annotated with synchronized multimodal data (audio, video, text) along with metadata.

Metadata Fields:

1. Timestamps: Start time and end time of each segment.
2. Video Clip: A short video clip corresponding to the speech segment is extracted and saved in MP4 format.
3. Audio Clip: The isolated audio for the speech segment is saved separately in WAV format.

4. Transcription: The text transcription of the speech segment is recorded.
5. Representative Frame: A single frame, typically from the midpoint of the segment, is extracted to visually represent the speaker.

MODALINK is implemented in Python, utilizing open-source tools like FFmpeg, MTCNN, WebRTC VAD, and the Google Speech Recognition API. Parallel processing with a thread pool executor accelerates the pipeline by handling multiple segments simultaneously. An error logging system ensures continuity without halting the process. By leveraging multi-core processors, MODALINK efficiently processes large videos, enabling scalable dataset generation while minimizing manual intervention.

4 RESULTS AND DISCUSSION

The proposed MODALINK workflow, explained in detail in the methodology section, combines several modules, including audio extraction, speech recognition, face recognition, transcription, and video downloading. MODALINK aims to automate the process of creating a well-structured Egyptian Arabic Multimodal Emotion recognition dataset.

4.1 Preliminary Results

Testing the proposed MODALINK workflow showed its ability to successfully :

- Download video content using yt-dlp and store it in MP4 format.
- Extract audio files from video files via FFmpeg
- Detect speech segments using WebRTC VAD
- Identify faces within video segments using the MTCNN model
- Transcribe audio clips using Google's Speech API customized for Egyptian Arabic.

Initial testing provided the following insights: Downloading the videos and extracting the corresponding audio streams was successful without any complications. In the next step, MODALINK successfully identifies several speech segments from the extracted audio files. Audio segments with loud background music, noises, and no speech were discarded. Temporal synchronization between audio and video was maintained through rigorous timestamp management. Next, MODALINK used MTCNN for reliable face detection in video frames. Any video segment without visible faces was filtered out. Finally, managing to transcribe the speech audio files and saving them along with the time stamps where they occurred. Figure 2 shows the various expressions detected in different video segments generated as output.

Figure 2: Sample of detected facial expressions in the generated video segments.

Figure 3 represents samples of the transcribed text obtained when we tested our workflow on an episode of a famous Egyptian TV series. Around 44 speech segments were extracted from the audio file of this video.

MODALINK provides extensive output, including segmented video clips with speech and visible faces, audio clips for each segment, detailed transcriptions with temporal alignment, and structured output files that maintain all component connections.

4.2 Optimization Strategies

To ensure efficient processing of long-form videos, the following optimizations were implemented:

1. **Parallel Processing:** Video and audio extraction, face detection, and transcription are parallelized using ThreadPoolExecutor. This approach maximizes CPU utilization and reduces overall execution time.
2. **Frame Rate Reduction:** Frames are extracted at 1 FPS instead of higher rates (e.g., 5 FPS). This reduces the number of frames processed without significantly impacting face detection accuracy.
3. **Downscaling:** Frames are down-scaled to 50% of their original resolution before face detection. This improves the efficiency of the MTCNN model while maintaining acceptable accuracy.
4. **Batch Processing:** Audio frames are processed in batches for VAD, and audio segments are transcribed in batches. This minimizes overhead and improves throughput.
5. **FFmpeg Optimization:** FFmpeg commands use

```
                                                    Segment: 9_0
                                                Time: 259.62 - 274.62
Video: /content/drive/MyDrive/Trial1stFeb1stTrial/video_clips/segment_9_0.mp4
Audio: /content/drive/MyDrive/Trial1stFeb1stTrial/audio_clips/segment_9_0.wav
 Frame: /content/drive/MyDrive/Trial1stFeb1stTrial/video_frames/frame_9_0.jpg
    Transcription: يا صباح الخير يالي معانا ومساء الخير عليكم يا اللي معانا
                                        Active Face: (547, 163, 145, 175)
```

(a) Sample 1 of transcribed text obtained from testing the proposed workflow.

```
                                                    Segment: 48_0
                                                Time: 681.66 - 686.25
 Video: /content/drive/MyDrive/Trial1stFeb1stTrial/video_clips/segment_48_0.mp4
 Audio: /content/drive/MyDrive/Trial1stFeb1stTrial/audio_clips/segment_48_0.wav
  Frame: /content/drive/MyDrive/Trial1stFeb1stTrial/video_frames/frame_48_0.jpg
      Transcription: فاكره فاكره يا علا كنا في الجامعه من قد ايه فاكره يا اختي
                                        Active Face: (749, 170, 161, 180)
```

(b) Sample 2 of transcribed text obtained from testing the proposed workflow.

Figure 3: Examples of transcribed text from the proposed workflow.

the -preset ultrafast option and -threads 4 to accelerate video and audio extraction. These settings optimize the trade-off between speed and quality.

4.3 Generalization of MODALINK

A key aspect of the MODALINK workflow is its potential for generalization across different languages and dialects. MODALINK's versatility is pivotal for broadening emotion recognition capabilities to underrepresented linguistic communities. MODALINK is designed to be modular, allowing for seamless modification of parameters to adjust various languages or dialects. Altering the transcription model's language is the main modification that must be made. Due to its adaptability, MODALINK can be used in a variety of linguistic contexts without requiring consequential reconfiguring. Furthermore, the transcription model itself can be easily swapped if a more accurate model becomes available for a specific language or dialect.
This guarantees that the process can take advantage of new developments in transcription models as they become available, in addition to increasing its adaptability. MODALINK provides a strong pipeline for building multimodal datasets that can facilitate emotion recognition research across a variety of languages and dialects by upholding an emphasis on modularity and adaptability.
This generalizability of MODALINK also unlocks a deeper understanding of how people from different cultures express emotions which is crucial for filling the resource gap. By enabling the creation of multimodal datasets in various languages, MODALINK helps in exploring the rich tapestry of human emotions across diverse linguistic and cultural landscapes.

4.4 Challenges and Observations

The segmentation is based on detected speech rather than fixed time intervals, resulting in segments of varying lengths. While this strategy assures that each segment includes relevant speech, it could complicate processing and analysis. Hence, a minimum segment length is exploited to ensure that we have meaningful text useful for emotion analysis.
Future iterations could implement more sophisticated speaker diarization techniques to handle multiple simultaneous speakers, however; these models are computationally intensive.
Also, some recognized speech chunks lack identifiable faces, leading to their elimination. This is to ensure both speech and facial emotions are present in each segment which is critical for further research.
One of the most challenging tasks was converting audio to text. Several established models and APIs were tested, but they all struggled with dialect-specific vocabulary, resulting in inaccurate transcriptions. Fine-tuned Wav2Vec models (Baevski et al., 2020) as well as Vosk models (Shmyrev and other contributors, 2020) pre-trained on Arabic language were tried. However, the transcription results were not satisfying. Google's speech recognition was an improvement over other approaches examined, and it has produced encouraging results so far; nonetheless, improving transcription accuracy remains an ongoing challenge.

5 CONCLUSIONS AND FUTURE WORK

The paper presents MODALINK to generate multimodal datasets leveraging visual, audio, and text modalities for emotion recognition. It addresses a significant gap in multimodal emotion recognition by tackling the scarcity of resources for low-resource languages, particularly Arabic, while capturing linguistic and cultural aspects. MODALINK utilizes advanced tools such as FFmpeg, MTCNN, WebRTC VAD, and Google Speech Recognition API to automate dataset generation efficiently, ensuring precise synchronization across modalities. Preliminary tests demonstrate its ability to process large-scale video data, extract emotion-rich segments, and produce synchronized outputs with minimal time, resources, and human intervention. Challenges remain in improving transcription accuracy for specific dialects and expanding diversity.The future goal is to develop a comprehensive, diverse Egyptian Arabic dataset incorporating all modalities for emotion recognition.

ACKNOWLEDGMENT

We acknowledge the use of AI tools to generate and enhance parts of the paper. The content was revised.

REFERENCES

Ahmed, N., Aghbari, Z. A., and Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. Appl.*, 17:200171.

Akila, G., El-Menisy, M., Khaled, O., Sharaf, N., Tarhony, N., and Abdennadher, S. (2015). Kalema: Digitizing arabic content for accessibility purposes using crowdsourcing. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 655–662. Springer.

Al Roken, N. and Barlas, G. (2023). Multimodal arabic emotion recognition using deep learning. *Speech Communication*, 155:103005.

Arabiya, S. (2025). 10 things you may not know about. Accessed: 16-Jan-2025.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Gurevych, I. and Miyao, Y., editors, *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.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Islam, M. M., Nooruddin, S., Karray, F., and Muhammad, G. (2024). Enhanced multimodal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach. *Biomed. Signal Process. Control.*, 94:106241.

Kalateh, S., Estrada-Jimenez, L. A., Hojjati, S. N., and Barata, J. (2024a). A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access*.

Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., and Barata, J. (2024b). A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. *IEEE Access*, 12:103976–104019.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Radoi, A., Birhala, A., Ristea, N., and Dutu, L. (2021). An end-to-end emotion recognition framework based on temporal aggregation of multimodal information. *IEEE Access*, 9:135559–135570.

Safwat, S., Salem, M. A.-M., and Sharaf, N. (2023). Building an egyptian-arabic speech corpus for emotion analysis using deep learning. In *Pacific Rim International Conference on Artificial Intelligence*, pages 320–332. Springer.

Shaqra, F. A., Duwairi, R., and Al-Ayyoub, M. (2019). The audio-visual arabic dataset for natural emotions. *2019 7th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 324–329.

Shmyrev, N. V. and other contributors (2020). Vosk speech recognition toolkit: Offline speech recognition api for android, ios, raspberry pi and servers with python, java, c#, and node. https://github.com/alphacep/vosk-api. Accessed: 2025-01-16.

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

Vijayaraghavan, G., T., M., D., P., and E., U. (2024). Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Inf. Fusion*, 105:102218.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503.