Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MODULAR EMOTION-AWARE MULTILINGUAL AI DUBBING FRAMEWORK WITH FER-GUIDED EXPRESSIVE SPEECH SYNTHESIS AND LIP SYNCHRONISATION

ANANDU M[1] SHERLY KK[2] JITHIN MATHEWS[3] WILLSON JOSEPH C[4]

[1,2,3,4]Rajagiri School of Engineering and Technology, Ernakulam, Kerala, India
[1]anandumaheedharan@gmail.com, [2]sherlykk@rajagiritech.edu.in, [3]jithinm@rajagiritech.edu.in, [4]willsonj@rajagiritech.edu.in

ABSTRACT

The development of emotion-aware multilingual dubbing solutions is a new promising research direction in media localization, accessibility, digital entertainment, and cross-lingual communication. Despite the advances in existing dubbing frameworks that concentrate mostly on enhancing transcriptions, translations, speech synthesis, and lip sync, there exists little work on capturing the emotional intent of the speaker during the entire process of dubbing. This may lead to the generation of linguistically correct but unnatural speech in terms of affective aspects from the original video. In order to address this challenge, it is necessary to introduce visual emotions into the speech generation process.In this study, we introduce a modular emotion-aware multilingual AI dubbing system based on Facial Emotion Recognition (FER), Whisper ASR, Neural Machine Translation (NMT), expressive speech synthesis with YourTTS, and lip-sync generation with Wav2Lip. In contrast to other existing methods, which make use of mostly text information, the introduced system directly uses facial emotion predictions to condition expressive speech synthesis, thus allowing us to keep the emotional context intact while keeping linguistic fidelity at a high level in multilingual settings. Another advantage of the proposed approach is its modular nature, which allows for individual optimization of every module independently of the others.The designed FER system uses the EfficientNet-B0 model and is trained with a mixed dataset that is created by combining the FER2013 five-class benchmark with carefully selected target-domain face images for the purpose of robustness against the video environment. Our experimental results on the FER2013 five-class test set give an emotion classification accuracy rate of 75% along with macro-F1 value of 0.68 and weighted-F1 of 0.75 in 20 stable epochs of training, thus providing reliable recognition despite problems such as class imbalance and inter-class similarity. Detected emotions are then used for the generation of emotional multilingual speech.The efficiency of the suggested approach is analyzed through objective and subjective measurements such as performance of FER classification, ASR precision, translation quality measured in BLEU, Mean Opinion Score (MOS), Perceptual Evaluation of Speech Quality (PESQ), and speaker similarity. The results obtained confirm that the explicit emotion-driven approach to speech synthesis could effectively improve expressive qualities of multilingual dubbing without losing flexibility and maintainability that are characteristic of modular system architecture. Thus, the suggested approach could serve as a basis for development of future emotion-driven dubbing systems supporting different languages and emotions.

Keywords: *Emotion-aware dubbing, Facial emotion recognition, EfficientNet-B0, Whisper ASR, Multilingual translation, YourTTS*

1. INTRODUCTION

This widespread practice of employing cross-language content in media platforms such as online video streaming sites, social networks, and educational media has led to the increasing necessity for efficient video localizations into different languages. The ultimate objective of automatic dubbing technology is to automate the post-production process via the integration of ASR, MT, TTS, and AV alignment. Although current neural models have significantly improved ASR performance in terms of recognition accuracy and naturalness of the synthesized speech, the representation of the speaker's emotionality is a challenging task.

It is evident from the literature review that current state-of-the-art dubbing systems focus more on semantic faithfulness, preservation of the speaker's identity, and lip-synchronization, whereas affective faithfulness remains an understudied research direction [1, 2, 3]. The

present work is thus justified, as high-quality dubbing requires not only the semantic faithfulness but also the appropriate emotions during voice synthesis.

In order to clarify the problem statement, we formulate three research questions. (RQ1) Can facial visual cues reliably guide emotional dubbing? (RQ2) Is the proposed modular ASR–MT–TTS pipeline linguistically reliable and emotionally controlled? (RQ3) Is the solution computationally efficient?

Nevertheless, conventional approaches to automatic dubbing tend to focus more on semantic fidelity in the sense that the generated speech must convey the same language message. The role of emotional expressivity cannot be underestimated, as it will have an effect on the effectiveness of the communication process. In reality, many automatic systems generate outputs with adequate linguistic meaning, yet in terms of prosodic appropriateness, they fail to do so. The gap between facial expression and voice translation may lead to a poor experience of using such technologies. In multilingual video dubbing research, there are several challenges, including prosodic transfer, synchronization, speaker identity preservation, and lip-syncing [1].

One of the critical limitations of existing solutions is the use of emotion modeling either with regard to text input or with regard to speech input, yet ignoring visual modality data about emotions. It is possible to use facial expressions of speakers as cues that indicate their emotional intentionality.

We suggest creating emotion-sensitive AI dubbing models using FER-based emotion recognition. In our approach, we follow the modular paradigm in which emotion predictions are used to modulate prosodic features of the generated speech. For emotion prediction, five realistic emotion types are considered, namely anger, happiness, sadness, neutral, and surprise, forming a compromise between expressiveness and stability of the generated speech.

EfficientNet-B0 architecture was chosen as the base for its efficiency measured by accuracy/number of parameters ratio. Emotion predictions per frame are estimated and then averaged over frames to obtain the segment-level prediction that is used further in the emotional TTS system built upon YourTTS framework.

Whisper is utilized as the foundation of multi-language ASR with the inclusion of NMT to generate the subtitles. The lip synchronization is achieved with the help of the Wav2Lip model, which makes sure the generated speech is synchronized with lip movements [4]. With the help of emotional consistency maintained during transcription, translation, synthesis of speech, and its synchronization with video data, the objective is to provide dubbed videos that are both linguistically and emotionally consistent.

The proposed system allows modularity, as opposed to multimodal systems that need a huge amount of jointly annotated datasets across all modalities. It becomes possible to fine-tune modules separately and incorporate emotion conditioning easily.

The main contributions of this work are as follows:

- We propose a modular, emotion-aware multilingual dubbing framework that contributes a practical FER-guided control mechanism for expressive speech synthesis within a standard ASR–MT–TTS pipeline.
- We construct a custom five-class FER dataset for the dubbing scenario by combining benchmark data with target-domain video samples, thereby reducing domain shift.
- We highlight the scientific contribution of the work by evaluating visual, speech, translation, and synthesis components separately, which makes the effect of emotion conditioning measurable and reproducible.
- We demonstrate that emotion conditioning improves affective consistency while preserving transcription accuracy, translation quality, and lip synchronization.

The remainder of the paper presents the proposed architecture, dataset construction methodology, experimental setup, quantitative evaluation, and the remaining limitations and future directions.

2. RELATED WORK

Multilingual dubbing has become a new field of interdisciplinary research that combines several technologies like ASR, NMT, expressive TTS, FER, and synchronization. In addition to

translating spoken content into another language, multilingual dubbing tries to maintain the speaker identity, emotions, and appearance. Survey papers classify multilingual dubbing systems into audio and video-based systems while mentioning several ongoing challenges including prosody transfer, speaker identity maintenance, emotion consistency, synchronization, and computational efficiency in unconstrained settings [1, 2, 3]. While audio-based systems concentrate on generation of natural and understandable speech, video-based systems pay attention to facial movements and lip movements using cross-modal alignment. Although substantial success has been achieved in both categories, maintaining emotion consistency during the entire process of dubbing is an open research problem.

One of the most important advancements in the realm of multilingual dubbing is the development of Wav2Lip, which treats lip sync as an audio-visual correspondence problem by leveraging lip-sync expert discriminators trained on a large number of unconstrained videos [4]. It considerably outperforms frame-based approaches for animating videos and has been widely adopted in many AI dubbing tools. However, other works have made it even more robust for difficult conditions with pose variations, lighting variations, occlusions, and head motions. However, multilingual dubbing imposes yet another challenge since the translation of speech into another language often leads to temporal discrepancies due to differences in rhythm and pronunciation.

Facial Emotion Recognition (FER) has emerged as an important building block for maintaining the emotionality of the original content when multilingual dubbing is done. Proper recognition of facial emotions ensures that the synthesized speech reflects the emotionality in line with the appearance of the speaker, leading to more immersive content generation. Nonetheless, FER still poses a difficult problem due to label ambiguities, inter-class variabilities, different facial poses, illumination changes, imbalanced training sets, and domain shift issues between lab-based datasets and unconstrained videos [5, 6]. The commonly used FER2013 dataset developed in the course of ICML 2013 Representation Learning Challenge consists of 48 by 48 sized grayscale facial images of seven classes and serves as a benchmark to assess FER techniques [7]. However, emotion recognition for multimedia applications usually employs simpler emotion classes.

The latest developments in FER have considered both the problem of discrete emotions classification and continuous affect prediction using valence and arousal space representation [8]. Deep convolutional neural networks have been the dominant solution in FER thanks to its capability of learning discriminative facial representations from image inputs. Specifically, EfficientNet has utilized compound scaling through the depth, width, and resolution of networks to achieve a good trade-off between efficiency and recognition accuracy when limited by computation resources [9]. The latter makes it a suitable choice for real-time multimedia applications combining FER with ASR, TTS, and lip sync modules.

Besides traditional CNN methods, there are recent researches that introduced new optimization and feature selection approaches to boost emotion recognition effectiveness. Joseph and Kathrine suggested a novel optimizer-based deep learning architecture for emotion detection and classification that improves the convergence and feature learning process using optimal training methodologies, resulting in improved accuracy of classification of emotions [10]. The work underlines the role of optimization approaches in boosting the robustness and generalization capacity of FER models. In another research, the same scientists introduced the Concurrent Black-winged Gated Network combined with Greylag Goose feature selection for emotion-based stress-level estimation in IoT-enabled health monitoring systems [11]. Even though the system was created for healthcare purposes, the presented framework shows the effectiveness of optimal feature selection and gated deep architectures in capturing the peculiarities of emotions and minimizing irrelevant features.

Automatic Speech Recognition has seen significant advancements with the use of large-scale transformer foundation models. Whisper is an example that shows multilingual generalization capability in weakly supervised pre-training using hundreds of thousands of hours of multilingual speech data, resulting in zero-shot transcription in various languages and acoustic scenarios [12]. Transformer encoder-decoder models have more robust performance than the traditional HMM-based models in terms of background noise, accent, and domain mismatch.

Whisper is therefore well suited for multilingual dubbing due to its unique properties.

The Neural Machine Translation framework has also been helped by the use of transformer architecture that manages to incorporate long-term dependency. The BLEU score has been the measure used in the evaluation of quality in translation. In order to enhance reproducibility and consistency in evaluation of results, there has been the adoption of SacreBLEU, which offers tokenization and consistency in reporting of results [13]. More recently, evaluation approaches have gone beyond just lexical similarity and now include semantic similarity and LLM-based methods that consider the translation quality of multimedia localization and dialogue translation [3]. This is especially important in multilingual dubbing where semantic similarity is more important than lexical similarity.

There have been fast advancements in neural speech synthesis using end-to-end systems like VITS and YourTTS, a multi-language variant of the former, that learn speaker and language embeddings and simultaneously model expressive prosody through conditioning [14]. In comparison to the earlier concatenative or statistical parametric TTS techniques, neural models yield much more natural speech and can perform flexible editing of parameters such as the speaker identity, speaking style, and emotion. Current studies also focus on disentanglement of the linguistic, speaker, and emotional information into separate representations, which leads to controllable generation of expressive speech. MOS is typically used for subjective evaluation based on the ITU-T P.800.1 recommendation, while PESQ is used for objective evaluation [15, 16].

Despite the significant advancements made in individual components like ASR, machine translation, FER, expressive TTS, and lip-sync, only a few papers in the literature have considered including visual emotional information into a full-fledged pipeline for multilingual dubbing in a modular and computationally efficient manner. Current multimodal dubbing approaches generally rely on multimodality in training, thus necessitating joint optimization among different modalities, and use huge multimodal corpora, making training challenging and computation-intensive [17]. On the other hand, our proposed framework leverages a modular approach based on FER where the predictions from different facial emotions are estimated in isolation and then used as conditioning inputs for generating expressive TTS. Such an architecture allows each component to be fine-tuned individually, is more scalable, upgradeable, and keeps the emotions consistent without re-training the whole multimodal model. Therefore, our proposed framework solves many limitations of current literature by providing a computationally efficient solution for emotion-aware multilingual AI dubbing.

3. PROPOSED SYSTEM

The proposed model for emotion-aware multilingual dubbing has five fundamental components that deal with individual steps involved in the complete process of dubbing. Each of these components performs its respective task sequentially but in a modular way so that all the components could be developed, optimized, and improved individually without having any effect on the rest of the components. The five main components include: (i) Pre-processing and synchronization of speech signal and visual content, (ii) Automatic Speech Recognition (ASR), (iii) Neural Machine Translation (NMT), (iv) Emotion-driven TTS, and (v) Lip-sync video creation.

The process pipeline starts with the extraction and synchronization of audio and video channels from the input multimodal content. Next, the speech signal is converted into text with the help of an ASR model, and the obtained text is translated into the desired language with the help of NMT model. In parallel, the facial frames are processed to recognize the emotions of the speaker that will act as an explicit condition for the expressive speech synthesis. The translated text and the detected emotion are fed to the multilingual TTS model to synthesize speech in the desired language. Finally, the generated speech is synchronized with the original facial video in order to obtain the dubbed video.

In this way, the framework proposed aims to preserve not only the correctness of the translation process, but also the emotions expressed by the speaker in the dubbing process. By combining visual emotion recognition, expressive speech synthesis, and audio-visual synchronization, the system strives to provide the generation of the natural multilingual dubbed videos with better emotional consistency and lip synchronization.

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Notation and Pipeline Overview

Let the input video be denoted as:

$$\mathcal{V} = \{(F_t, A_t)\}_{t=1}^{T} (1)$$

where F_t is the *t*-th video frame and A_t is the corresponding audio segment at timestamp *t*, and T is the total number of frames or time steps in the video.

The entire dubbing process is modelled as a composite transformation:

$$\mathcal{V} = \mathcal{L} \circ \mathcal{TTS} \circ \mathcal{E} \circ \mathcal{GT} \circ \mathcal{GSR}(\mathcal{V}) (2)$$

where:

- $ASR(\cdot)$ performs speech-to-text transcription from the audio stream A_t using Whisper.
- $MT(\cdot)$ translates the transcribed source language text into the target language.
- $E(\cdot)$ infers the predominant facial emotion from a sequence of frames $\{F_t\}$ using an EfficientNet-B0-based classifier.
- $TTS(\cdot)$ generates expressive speech conditioned on both translated text and predicted emotion using YourTTS.
- $L(\cdot)$ performs lip-synchronisation using Wav2Lip, aligning dubbed audio to facial motion.

The final output $\hat{V}$ is the rendered video containing translated, emotion-aware speech that is temporally aligned with facial articulation.

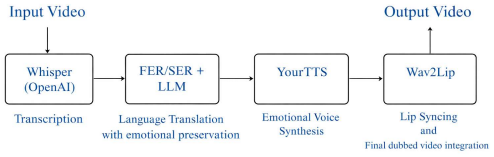


Figure 1: End-to-end pipeline of the proposed emotion-aware multilingual dubbing system.

3.2 Facial Emotion Recognition (FER)

To preserve expressive intent, emotion must be inferred from the speaker's face. Let $p_t \in \mathbb{R}^C$ denote the posterior probability vector at time *t*, where C = 5 represents the number of discrete emotion classes: *angry*, *happy*, *sad*, *neutral*, and *surprise*. The FER module outputs per-frame emotion distributions as:

$$p_t = EffNetB0(F_t) (3)$$

where EffNetB0(·) is the EfficientNet-B0 classifier pre-trained on FER2013 and fine-tuned on a custom hybrid dataset.

To obtain a stable segment-level emotion label, we compute the temporal average over T consecutive frame-level predictions:

$$\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p_t (4)$$

and select the dominant emotion label:

$$e = argmax_{c \in \{1,\dots,C\}} \bar{p}_c (5)$$

where $\bar{p}_c$ is the *c*-th component of $\bar{p}$. This averaging reduces noise and accounts for intra-segment expression variation.

3.3 Emotion-Conditioned Expressive TTS

The detected emotion *e* is used to guide prosody in speech synthesis. Let x_{text} denote the tokenised translated text input. We define an emotion embedding vector $\mathbf{e} \in \mathbb{R}^{de}$ corresponding to label *e*, where d_e is the dimension of the emotion embedding space.

The TTS encoder outputs a hidden text representation $h_{text} = f_{enc}(x_{text})$. Emotion conditioning is performed via additive bias:

$$h_{tts} = f_{enc}(x_{text}) + W_e\, e (6)$$

where:

- f_{enc}: $\mathbb{R}^L \rightarrow \mathbb{R}^{dh}$ is the TTS encoder (e.g., YourTTS encoder) mapping text to hidden states.
- $W_e \in \mathbb{R}^{dh \times de}$ is a learnable linear projection from emotion embedding space to TTS hidden state space.
- h_{tts} is the final TTS input encoding incorporating emotional bias.

This formulation allows flexible manipulation of speech prosody based on inferred affect, improving naturalness and emotional congruence in synthesis.

3.4 Lip Synchronisation via Wav2Lip

After synthesis, the generated audio must be temporally aligned with the speaker's lip movements. Wav2Lip takes as input the original silent face frames and the dubbed audio to generate synchronised talking-face frames.

Let $\phi_{lip}(\cdot)$ and $\phi_{audio}(\cdot)$ represent deep feature extractors for lips and audio, respectively. Wav2Lip minimises a synchronization loss:

$$\mathcal{L}_{sync} = \frac{1}{T}\sum_{t=1}^{T} \| \phi_{lip}(F_t) - \phi_{audio}(A_t) \|_2^2 (7)$$

where:

- F_t is the *t*-th video frame.
- $\hat{A}_t$ is the generated audio segment at time *t*.
- $\phi_{lip}(F_t) \in \mathbb{R}^{dl}$ and $\phi_{audio}(\hat{A}_t) \in \mathbb{R}^{dl}$ are extracted features from the lip region and audio at time *t*, respectively.
- L_{sync} ensures audio-visual coherence by enforcing that corresponding features are similar in latent space.

The output is a re-synthesised video $\hat{V}$ that preserves the original speaker's facial gestures while ensuring multilingual, emotionally expressive, synchronised speech.

4. DATASET AND EXPERIMENTAL SETUP

4.1 Facial Emotion Recognition Datasets

The FER experiments were carried out on two distinct datasets that include: (i) the remapping of a five-class version of the FER2013 dataset; and (ii) an experimental dataset developed for the target dubbing environment.

The FER2013 dataset is one of the benchmarking datasets for FER introduced in the ICML 2013 challenges. This dataset comprises 48×48 gray-scale images classified into seven emotion categories [7]. To conform to the requirements of the dubbing application, we use the remapped version of the FER2013 dataset, which comprises five discrete emotion types: *angry, happy, sad, neutral, surprise*.

4.1.1 Custom Dataset Construction

To make a data set which is more realistic compared to previous data sets, we build our own using 50 high-resolution videos. The videos have an average duration of between 1 and 3 minutes, meaning that each video has a total of around 8,000–12,000 frames. From the above videos:

- Faces are detected and cropped using a standard face detection pipeline.
- Frames are sampled at 3–5 fps to reduce redundancy.
- Emotion labels are manually annotated into the same five classes.

There were more than 20,000 faces captured by the model after preprocessing. The images are rescaled to size 224×224, and their normalization is done using the mean and variance of ImageNet. There is an ImageFolder-based partitioning scheme for dividing the dataset into three subsets: 70% for training, 15% for validation, and 15% for testing. The test support splits are:

FER2013 (5-class) test split = [145, 1088, 748, 730, 477] (N = 3188)

Custom dataset test split = [159, 1111, 761, 736, 482] (N = 3249)

4.2 FER Training Configuration

All FER models are trained under identical conditions to ensure fair comparison. The training configuration is summarised below:

- Backbone: EfficientNet-B0 (primary), ResNet-18, MobileNet-V2 (baselines)
- Loss: Cross-Entropy
- Optimizer: AdamW
- Learning rate: 1×10^{-4}
- Batch size: 16
- Epochs: 20
- num_workers: 4

Data augmentation includes RandomHorizontalFlip and RandomRotation (±10°). Evaluation metrics include accuracy, per-class precision, recall, F1-score, macro-F1, weighted-F1, and both raw and normalized confusion matrices.

4.3 ASR, Translation and Expressive TTS Setup

4.3.1 Automatic Speech Recognition (ASR)

Multilingual ASR is performed using Whisper [12]. Audio is resampled to 16 kHz mono before inference. Transcription quality is evaluated using:

- Word Error Rate (WER)
- Character Error Rate (CER)

WER and CER are computed using standard edit-distance formulations over reference transcripts.

4.3.2 Neural Machine Translation

The transcribed source-language text is translated into the target language using a neural machine translation module. Translation quality is evaluated using BLEU score:

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Tokenization and normalization follow standard SacreBLEU-style reporting for reproducibility.

4.3.3 Emotion-Conditioned Expressive TTS

Expressive synthesis is performed using YourTTS [14]. The predicted dominant emotion from FER is converted into a learned embedding and injected into the encoder hidden representation.

Evaluation metrics for TTS include:

- Mean Opinion Score (MOS) for perceptual quality
- Perceptual Evaluation of Speech Quality (PESQ)
- Speaker similarity score (cosine similarity in embedding space)

Subjective MOS evaluation is conducted with multiple listeners using a 1–5 rating scale.

4.4 Lip Synchronisation Module

Lip synchronization is performed using Wav2Lip [4]. The generated expressive speech waveform is aligned with the original face frames.

Synchronization quality is assessed through:

- Visual inspection of lip articulation
- Audio-visual feature consistency loss

The synchronization loss is defined as:

$$L_{sync} = \frac{1}{T}\sum_{t=1}^{T} \| \phi_{lip}(F_t) - \phi_{audio}(A_t) \|_2^2$$

where ϕ_{lip} and ϕ_{audio} are learned feature extractors.

4.5 End-to-End Evaluation

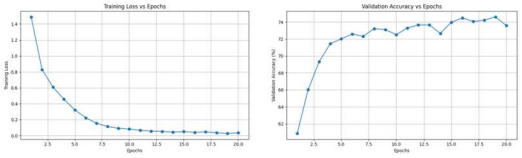
The full pipeline is evaluated across four dimensions:

1. Emotion classification accuracy (FER metrics)
2. Transcription accuracy (WER, CER)
3. Translation quality (BLEU)
4. Expressive synthesis quality (MOS, PESQ, speaker similarity)

This modular evaluation ensures that improvements in emotion conditioning can be quantitatively isolated from transcription or translation performance.

5. RESULTS AND ANALYSIS

Fig. 2 presents the FER training curves over 20 epochs. Training loss decreases steadily, while validation accuracy stabilises around 74–75%, indicating convergence without severe instability.



(a) Training loss curve (20 epochs). *(b) Validation accuracy curve (20 epochs).*

Figure 2: FER training convergence for EfficientNet-B0 (fixed training budget of 20 epochs).

On the FER2013 5-class test set (N = 3188), EfficientNet-B0 achieves 0.75 accuracy with macro-F1 0.68 and weighted-F1 0.75 (Table 1). The normalised confusion matrix (Fig. 3) shows strongest recognition for happy and surprise, while angry remains the most challenging class.

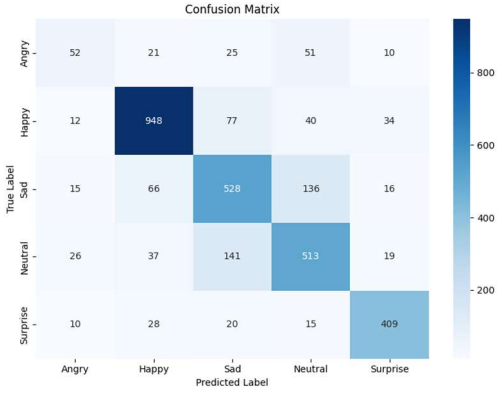


Figure 3: Confusion matrix for EfficientNet-B0 on FER2013 (5-class test split).

Baseline accuracy comparison on FER2013 data (Fig. 4) reveals the superior accuracy of EfficientNet-B0 compared to ResNet-18 and MobileNet-V2 (0.75 vs. 0.70 and 0.72 respectively). This confirms the efficacy of efficient model scaling. On our hybrid test data (N = 3249), we observe a comparable overall

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Statistical Comparison on FER2013 (5-Class).

Model	Params	Accuracy	Macro-F1	Weighted-F1
ResNet-18	11,689,512	0.70	–	–
MobileNet-V2	3,504,872	0.72	–	–
EfficientNet-B0	5,300,000	0.75	0.68	0.75

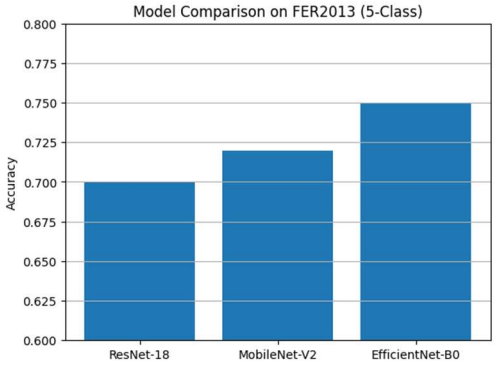
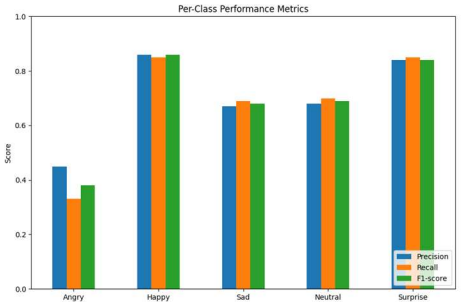
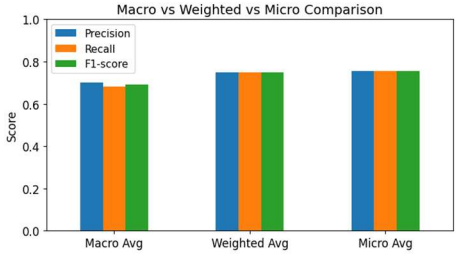


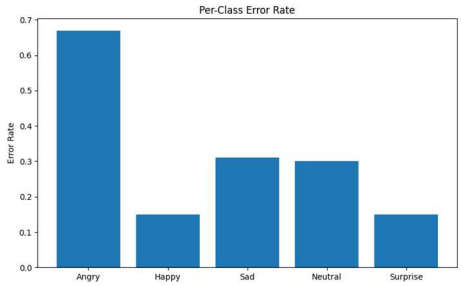
Figure 4: Accuracy comparison of ResNet-18, MobileNet-V2 and EfficientNet-B0 on FER2013 (5-class).



(a) Custom Dataset per-class metrics



(b) Custom dataset overall metrics



(c) Custom Dataset per-class error

Figure 5: Hybrid dataset evaluation artefacts for the 5-class FER task.

Error analysis: Anger displays low levels of recall and tends to confuse with neutral and sadness faces. Such pattern is consistent with the representation of the data set where anger is represented the lowest among the other emotion types, and the similarity of facial features, even when displayed at reduced image size. High rates of confusion for the sadness and neutral emotions suggest the possible need for aggregation prediction.

In terms of the speech pipeline, WER/CER is used to assess Whisper and BLEU for translation task in Fig. 6, while MOS, PESQ and speaker similarity metrics are reported for expressive TTS in Fig. 7. An example of the waveform is illustrated in Fig. 8.

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n) (8)$$

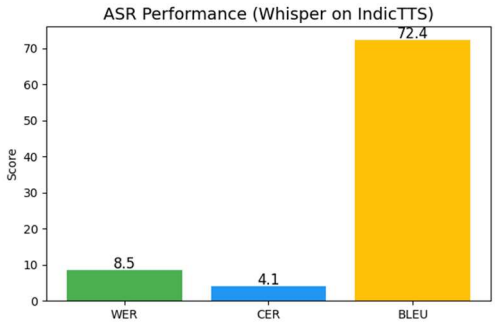


Figure 6: ASR and translation evaluation summary (WER, CER, BLEU) for the dubbing pipeline.

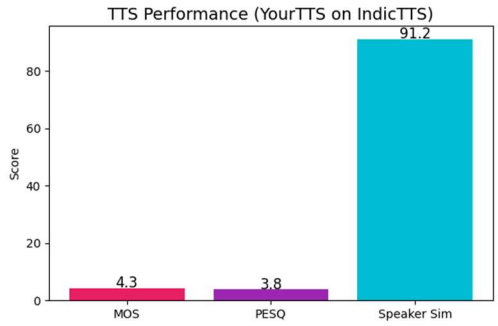


Figure 7: Expressive TTS evaluation summary (MOS, PESQ, speaker similarity).

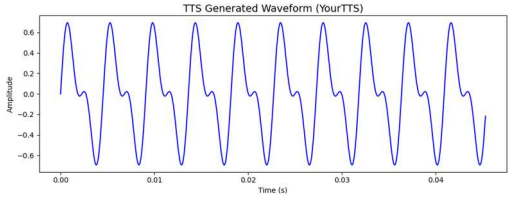


Figure 8: Example waveform plot of generated speech (illustrative).

6. DISCUSSION

It is clear from the above-mentioned results that adding explicit information about the emotional state of the face in the multilingual dubbing process substantially increases control over the speech expressiveness generation process without compromising the linguistic consistency of translation across different languages. The proposed pipeline is quite different from the traditional ones, since the latter use mostly the lexical or text-based emotional information as a conditioning signal, while our approach includes frame-based prediction of the emotional state of the speaker. In this way, it becomes possible to make synthesized speech more consistent with the visual emotional state.

As regards the task of facial emotion recognition, the EfficientNet-B0 model provides an optimal balance of computational load and accuracy due to the use of compound scaling techniques for network depth, width, and resolution of input images [9]. According to the experiment results, the model is able to achieve satisfactory accuracy on the five chosen classes of emotions while being lightweight enough to be integrated into multimedia systems. In our view, the main scientific novelty of the present study lies not only in using existing deep learning models but also in showing that visual emotion is a controllable signal through the entire multilingual dubbing process. Contrary to the fully-fledged multimodal end-to-end architectures which need expensive joint training of all components, the modular approach allows optimizing them independently.

While all of these results are encouraging, however, there are still some limitations to consider. Current five-class emotion categorization makes the process more simplistic than what actual human emotions can be in practice, with them being continuous, overlapping and dependent on context. Moreover, the relatively low recall rate for the *angry* emotion category illustrates the problems of class imbalance, similarity between classes, and the presence of fewer instances of minority emotion categories within face expression recognition data sets. While domain-specific sampling helps in achieving generalization, the domain gap between FER2013 and other benchmark data sets and unconstrained real-world videos still presents a significant problem [7]. Variations in facial poses, different illuminations, facial occlusion and blurring all serve to decrease the recognition accuracy during practical application. In addition, multilingual dubbing raises synchronization problems due to the differences in speaking rate and phoneme set between languages.

On the other hand, there are a number of reasons why this modular architecture might prove valuable from a practical standpoint. Given that the FER, ASR, NMT, expressive TTS, and lip synchronization modules are built separately, it is possible to replace any of the modules without the need for re-training of the whole system. Thus, this modularity leads to lower maintenance costs, easier integration of new models, as well as easier implementation of support for new languages and speakers. In our opinion, this is one of the most valuable practical contributions of the proposed architecture since it allows integrating new technologies into the system with little effort. As a result, the proposed framework might be useful in a number of practical applications such as multilingual media localization, dubbing for movies and TV series, educational content translation, translation for hearing impaired audience, virtual assistants, digital avatars, games, as well as production of short form content for online platforms.

Evaluating the methodology itself must be taken into account as well. While classification accuracy is useful in assessing the system performance in general, it cannot properly evaluate the performance of the systems under

imbalanced emotion distribution. Thus, additional evaluation metrics such as macro-F1 score, weighted-F1 score, precision, recall, confusion matrix, BLEU, SacreBLEU, and subjective evaluation like MOS must be considered in parallel to have a comprehensive evaluation of multilingual dubbing. The same applies to objective metrics for speech quality and synchronization, where their interaction can be quantified through evaluating them in conjunction.

Generally speaking, the results indicate that the inclusion of visual emotional cues in multilingual dubbing can greatly improve expressive speech production while retaining an efficient architecture design. Nonetheless, some research questions need to be addressed in the future. In particular, studies could consider continuous representation of emotions based on valence-arousal models, improved approaches for dealing with class imbalance using data augmentation and loss functions, better domain adaptation in actual video conditions, adaptive synchronization mechanisms that would take into account differences in timing between cross-linguistic settings, and broader perceptual studies including different languages and speakers.

7. CONCLUSION AND FUTURE WORK

The proposed framework in this paper is a modular approach of a multimodal emotion-aware AI dubbing system which makes use of modern technology components like automatic speech recognition, neural machine translation, facial emotion recognition, expressive text-to-speech synthesis and audio-visual synchronization. In particular, the suggested architecture uses Whisper ASR, transformer-based neural machine translation, EfficientNet-B0 for facial emotion recognition, YourTTS for multilingual expressive text-to-speech synthesis and Wav2Lip for lip synchronization [12, 14, 4]. As opposed to traditional systems of multilingual dubbing which operate mainly with textual information, the suggested approach uses visual emotional information explicitly as a condition for the generation of expressive speech. The suggested modular approach allows for the development, optimization and upgrades of each component individually while keeping compatibility with the overall dubbing process.

The emotion-aware speech generation was made possible by the fine-tuning of the facial emotion recognition model using a mixed dataset obtained from the FER2013 benchmark along with facial images from the target domain. The proposed EfficientNet-B0 classifier yielded an accuracy rate of 75% on the five-class FER2013 test set, showing robust recognition of the *happy* and *surprise* emotions while confusing the visually-similar *angry*, *sad*, and *neutral* categories [7]. The performance comparison of the model with ResNet-18 and MobileNet-V2 models shows its suitability for multimedia real-time applications [9, 18, 19].

The primary scientific contribution of this work is the use of visual facial emotion as an explicit control signal for multilingual expressive dubbing in an architecture that maintains its modularity. Unlike approaches based only on lexical information and costly end-to-end multimodal training processes, this approach introduces an interpretable relationship between the recognition of facial emotion and expressive speech synthesis. We believe that such a modular approach to emotion conditioning is an appropriate solution for future multilingual dubbing systems since it allows evolution of separate subsystems without compromising the performance of the entire system. Moreover, the proposed framework offers a flexible basis that will allow further development of the facial emotion recognition module, the speech synthesis module, or lip synchronization without any retraining of the whole pipeline.

Apart from the scientific significance of the method, the suggested approach is practical in the sense that it is easy to maintain, does not incur extra training costs, and allows quick addition of any new language, speaker, or more advanced deep neural network. Hence, it can easily be used in cases of multilingual media localization, educational content translation, accessibility software, virtual assistants, digital avatars, entertainment, games, and content creation. It is clear that there is considerable scope of practical application of emotion-aware multilingual dubbing in different multimedia applications.

However, some significant shortcomings persist. Firstly, the current five-emotion classification scheme is incapable of handling the dynamic nature of emotions, which includes various affective states at the same time. Moreover, class imbalance and overlap are still causing recognition issues, especially with the category of emotions such as *angry*. Secondly, the FER model suffers from sensitivity to illumination, pose change, occlusion, motion

blur, and domain discrepancy between benchmark datasets and unconstrained video clips. Lastly, dubbing in multiple languages creates additional synchronization problems, since translated speech may be longer, slower, or phonetically different from the original one. This proves that emotion-aware multilingual dubbing is an open problem.

Therefore, future research will consider several directions in particular. Firstly, continuous modeling of affects through valence-arousal representation will be analyzed in order to overcome discrete classes of emotions. Secondly, advanced temporal modeling techniques, attention models and learning from imbalanced data will be considered in order to improve emotions recognition in unconstrained environments. Thirdly, synchronization methods enabling dynamic compensation for cross-language duration discrepancy will be used in order to improve audiovisual synchrony in multilingual speech synthesis. Additionally, future research will perform experimental evaluation in terms of various languages, various speakers, adverse acoustical conditions, and will conduct detailed listening tests on humans. Finally, reproducibility will be improved through provision of complete setup of BLEU and SacreBLEU evaluations, release of evaluation scripts and adherence to benchmarking procedure according to the current practice in machine translation evaluation [20]. We hope that these research directions will allow us to develop more expressive, robust, scalable and practical multilingual AI dubbing systems.

Table 2: Training Hyperparameters for FER (All Models).

Hyperparameter	Setting
Epochs	20
Batch size	16
Optimiser	AdamW
Learning rate	1×10^{-4}
Input resolution	224×224
DataLoader workers	4
Augmentations	HFlip + Rotation (±10°)
Normalisation	ImageNet mean/std
Loss	Cross-Entropy

Table 3: Dataset Support (Test Split) for 5-Class FER.

Dataset	Angry	Happy	Sad	Neutral	Surprise	Total
FER2013 (5-class)	145	1088	748	730	477	3188
Hybrid (custom-curated)	159	1111	761	736	482	3249

REFERENCES

[1] D. Bigioi and P. M. Corcoran, "Multilingual video dubbing—a technology review and current challenges," Frontiers in Signal Processing, vol. 3, p. 1230755, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frsip.2023.1230755

[2] C. Hu, X. Tian, H. Lei, H. Lu et al., "Neural dubber: Dubbing for silent videos according to scripts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1283–1292.

[3] Z. Li, Y. Chen, Y. Zhang et al., "Pause-aware automatic dubbing using large language models and voice conversion," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, automatic dubbing with LLM and VC. [Online]. Available: https://arxiv.org/abs/2403.XXXXX

[4] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild," in Proceedings of the 28th ACM International Conference on Multimedia. ACM, 2020, pp. 484–492. [Online]. Available: https://arxiv.org/abs/2008.10010

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.

[6] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," Image and Vision Computing, vol. 65, pp. 23–36, 2017.

[7] I. J. Goodfellow, D. Erhan, P.-L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, Z. Xie, L. Romaszko, B. Xu, E. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," arXiv preprint, vol. arXiv:1307.0414, 2013. [Online]. Available: https://arxiv.org/abs/1307.0414

[8] H. Schneider et al., "Datasets for valence and arousal inference: A survey," arXiv preprint, vol. arXiv:2510.00738, 2025, survey of valence–arousal datasets for affect recognition.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114. [Online]. Available: https://arxiv.org/abs/1905.11946

[10] C. Willson Joseph and G. Jaspher Willsie Kathrine, "Improved optimizer with deep learning model for emotion detection and classification," Mathematical Biosciences and Engineering, vol. 21, no. 7, pp. 6631–6657, 2024.

[11] ——, "Concurrent black-winged gated network with greylag goose feature selection for emotion-based stress-level prediction on iot-based healthcare monitoring system," Biomedical Signal Processing and Control, vol. 113, p. 109066, 2026.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proceedings of the 40th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[14] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in Proceedings of the 39th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 2709–2720. [Online]. Available: https://proceedings.mlr.press/v162/casanova22a.html

[15] ITU-T, "Recommendation P.800.1: Mean opinion score (MOS) terminology," International Telecommunication Union, ITU-T Recommendation P.800.1, Jul 2016, series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks.

[16] "Recommendation p.862: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, International Telecommunication Union (ITU-T), Feb 2001.

[17] Y. Zhao, R. Liu, and G. Cong, "Towards expressive video dubbing with multiscale multimodal context interaction," arXiv preprint, vol. arXiv:2412.18748, 2024, m2CI-Dubber for prosody-expressive automatic video dubbing.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. [Online]. Available: https://arxiv.org/abs/1512.03385

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520. [Online]. Available: https://arxiv.org/abs/1801.04381

[20] M. Post, "A call for clarity in reporting BLEU scores," in Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[21] A. Mukherjee, S. Bansal, S. Satpal, and R. Mehta, "Text aware emotional text-to-speech with BERT," in Proceedings of Interspeech 2022, 2022, pp. 4601–4605.

[22] X. An, F. K. Soong, S. Yang, and L. Xie, "Effective and direct control of neural tts prosody by removing interactions between different attributes," Neural Networks, vol. 143, pp. 250–260, 2021.

[23] Z. Chen, N. Wang et al., "Lpips-attnwav2lip: Generic audio-driven lip synchronization for talking head generation in-the-wild," arXiv preprint, vol. arXiv:2501.01889, 2025, audio-driven lip synchronization.

[24] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in Proceedings of the Asian Conference on Computer Vision (ACCV), 2016, pp. 251–263.

[25] A. Vajpayee, P. Shrivastava et al., "A simple and efficient method for dubbed audio sync detection," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.