2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET) | 979-8-3503-5084-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/WISPNET61464.2024.10532876

A Survey on Voice Cloning and Automated Video Dubbing Systems

S. China Ramu
Professor, Department of CSE
Chaitanya Bharathi Institute of Technology
Telangana, India
chinaramu@gmail.com

Dhruv Saxena
Student, Department of CSE
Chaitanya Bharathi Institute of Technology
Telangana, India
dhruv162002@gmail.com

Vikram Mali
Student, Department of CSE
Chaitanya Bharathi Institute of Technology
Telangana, India
vikram595959@gmail.com

***Abstract*—In the context of today's interconnected world, where multilingual interactions are commonplace, the need for effective cross-language communication solutions is paramount. This survey explores innovative approaches to multilingual video dubbing, focusing on the integration of voice cloning and neural machine translation techniques for translating and dubbing videos across different languages while maintaining the authenticity of the original speaker's voice and ensuring seamless lip movement with the source video. Various advanced techniques, including voice cloning, speech emotion recognition, speech-to-lip synchronization, and neural machine translation, are employed in the examined literature. These techniques collectively aim to identify the source language, translate the content, and synthesize the speaker's voice in the target language that is synchronised with the source video. This survey also delves into the challenges associated with cross-language voice cloning and linguistic translation, exploring the collaborative potential of these technologies within the realm of video dubbing. The overarching goal is to contribute to the accessibility of multimedia content on a global scale, enabling viewers to enjoy videos in their preferred language without compromising the identity of the original speaker. Through comparative analyses of various algorithms, the effectiveness of these approaches in achieving high-quality, linguistically accurate, and emotionally resonant multilingual video dubbing is analyzed.**

***Index Terms*—Multilingual video dubbing, voice cloning, neural machine translation, speech emotion recognition, synchronizing speech with lips, cross-language voice synthesis.**

I. Introduction

In the ever-expanding global landscape of communication, transcending linguistic barriers has become a critical necessity. Video content, a potent medium for information distribution and cultural exchange, faces a considerable obstacle in the form of language diversity, hindering its widespread global distribution. The historical narrative in this field spans from the inception of video dubbing challenges to present-day efforts to bridge linguistic gaps. The need for such systems has grown exponentially with the global expansion of video content, prompting researchers to explore innovative solutions. From early attempts at lip syncing to recent breakthroughs in neural voice cloning, each development has contributed to the evolution of the field. Among these, multilingual video dubbing emerges as a pivotal solution, facilitating seamless translation while preserving the authenticity of the original speaker's voice[1][2].

This survey delves into the challenges and advancements in multilingual video dubbing, exploring a system designed to address these issues. Rooted in a comprehensive review of recent literature, our survey draws inspiration from groundbreaking work in pose-aware animated facial landmarks[1]. Novel data augmentation techniques, utilizing LSTM-based networks to decouple lip, jaw, and head motions, enhance the system's robustness[1]. Additionally, the integration of a unique neural fusion architecture, employing a unit concatenation method for improved speaker similarity and speech quality, forms the foundation of the audio synthesis system[3]. Insights from a CNN and ResNet based speaker extraction algorithm, incorporating visual cues and speech-lip synchronization, contribute to the adaptability in noisy environments with multiple speakers[2]. Exploration into cross-speaker emotion transfer and control for speech synthesis, disentangling prosody and timbre in emotional characteristics[4], informs our approach to maintaining emotional authenticity during translation.

The primary objective of this survey is to examine innovative approaches that seamlessly integrate advanced voice cloning and neural machine translation techniques. Drawing from methods such as facial landmark animation[1], lip syncing[5], and neural fusion for voice cloning[3], this survey aspires to analyze the landscape of real-time multilingual video dubbing with a focus on precision and emotional resonance[4]. The exploration opens new possibilities for real-time video dubbing across diverse domains, such as sports commentary, informative and entertainment videos, and news channels. Notably, the proposed solutions distinguish themselves by preserving the speaker's voice characteristics and offering an authentic viewing experience with support for Indian languages. The system processes input videos, pre-recorded or live, through concurrent tasks, including language identification[2], transcription and translation[6], lip synchronization[5][7], and real-time voice cloning[8].

The journey from early attempts at overcoming language barriers in video content to contemporary developments in real-time multilingual video dubbing illustrates the persistent efforts of researchers to enhance global communication and accessibility. This survey culminates with a clear research question, directing attention to the challenges and innova-

979-8-3503-5084-5/24/$31.00 ©2024 IEEE

tions in multilingual systems for an automated video dubbing pipeline.

II. Literature Survey

The literature survey in this paper is structured to provide a comprehensive overview of existing research in the domain of multilingual systems for automated video dubbing. To offer a clear understanding, we have organized the discussion into two main sections: video processing and audio processing. This division allows for a focused exploration of the intricate technologies and methodologies that contribute to the development of seamless multilingual video dubbing systems. In the realm of video processing, our survey delves into pioneering works that address challenges related to facial expressions, lip synchronization, and pose-aware dubbing. Each paper examined contributes valuable insights into the advancements that form the foundation of creating visually authentic multilingual video content. Transitioning to audio synthesis, our survey explores key papers that focus on voice cloning, speech synthesis, and neural machine translation techniques. Understanding the intricacies of these audio-centric aspects is crucial for a holistic comprehension of the challenges and innovations in achieving linguistically accurate and emotionally resonant multilingual videos.

A. Lip Syncing and Video Processing Systems

The paper [5] addresses the challenge of accurately generating lip-synced videos in unconstrained, real-world scenarios. The key problem it tackles is the inaccuracy of current approaches to lip-syncing unconstrained talking face videos, which are often limited by the range of identities, voices, or vocabulary they can effectively synchronize with. The paper proposes a novel approach to address this issue by introducing a pre-trained, precise lip-sync model that ensures the creation of natural lip motion, enhancing the quality of the lip-sync videos. The proposed model, Wav2Lip, leverages the expertise of a pre-trained lip-sync model to generate lip movements that are synchronized with the target speech.

It also incorporates a visual quality discriminator to further enhance the visual fidelity of the generated lip movements. The model's architecture and training process enable it to effectively synchronize lip movements with arbitrary speech segments in unconstrained talking face videos without being limited by specific identities, voices, or vocabulary. The evaluation metric uses a loss function that combines a reconstruction loss, an adversarial loss, and a visual quality loss to train the generator network. Additionally, the paper proposes two new metrics, to evaluate lip-sync accuracy in unconstrained videos. LSE-D measures the average error between the lip and audio representations in terms of their distance, while LSE-C measures the average confidence score of the lip-sync accuracy.

To address the problem of dealing with unseen videos with varied lip movements, an attention-based mechanism was introduced in [6]. The proposed AttnWav2Lip model incorporates spatial and channel attention modules, to focus on lip region reconstruction rather than unimportant regions of the face image.

The Spatial Attention Module determines where to emphasise or suppress in the feature maps across the spatial axes by inferring a spatial attention map. Pooling operations are carried out to compute the spatial attention map. This is then followed by concatenation, a convolution layer, and a sigmoid layer. The Channel Attention Module learns to suppress less helpful channels and emphasise informative ones in order to concentrate on ”what” is more important. The model architecture consists of three main components: a visual feature extractor, an audio feature extractor, and a lip-syncing network. The proposed model was evaluated on LRW, LRS2, and LRS3 datasets and demonstrated superior performance compared to baseline metrics measured by LSE-D and LSE-C as proposed in [5] as shown in Table 1.

TABLE I
Comparison of Performance of Lip Syncing Models

Method	LSE-D	LSE-C
Wav2Lip	7.521	6.406
AttnWav2Lip	**7.339**	**6.530**

A major limitation of this paper is that it only focuses on lip synchronization and does not consider other aspects of talking face generation, such as facial expressions and head movements, which may not always generate accurate videos.

The proposed system in [7] aims to improve the accuracy and visual quality of speech to lip generation by addressing four challenges, namely the dataset, audio feature extraction, face feature extraction, and lip-sync discriminator, which were prevalent in [5][6]. To overcome these challenges, the system uses attention mechanisms, including channel attention, spatial attention, and coordinate attention, as well as a new design of a visual quality discriminator. The system has been tested on three audio-visual datasets and has shown superior performance over prior works in terms of accuracy and visual quality, as seen in Table 2.

TABLE II
Comparison of performance of Lip Syncing models over the LRS2 dataset

Method	LSE-D	LSE-C
Wav2Lip	7.206	6.787
AttnWav2Lip	7.339	6.530
CA-Wav2Lip	**7.039**	**6.925**

Metrics such as SSIM and MS-SSIM, which are improvements over traditional metrics, are utilised because they take into account the structural information and the multi-scale nature of the human visual system.

[1] presents a pipeline to create a realistic talking videos from a single image and target audio, in contrast to earlier studies that dealt with a video or multiple image frames to achieve video dubbing. The pipeline, predicts the 3D facial landmarks by combining a RNN and a CNN. While the

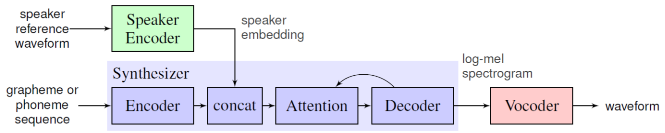

Fig. 1. Approach for Zero shot learning

RNN creates the matching 3D facial landmarks, the CNN extracts features from the input audio signal. The pipeline also incorporates a data augmentation technique that allows pose-aware landmark generation to be generated at inference by segmenting head, jaw, and lip motion.

This paper uses CNNs over GANs because they are well-suited for this task as they can extract features from the input audio signal and generate the corresponding 3D facial landmarks, while a GAN is more suited for image-to-image translation tasks, as shown in [5][6]. When compared to other models, this paper uses a structural-based method, which produces better-quality videos but may not be as fast or easily deployable as end-to-end systems.

B. Voice Cloning and Audio Synthesis Systems

A suitable audio stream that can be synced with the source video is a prerequisite for all of the aforementioned solutions. Two more limitations can be placed on such systems in the case of multilingual automated dubbing. First and foremost, the audio that needs to be synced needs to have similar speech characteristics and the same emotions as the original speech. The audio synthesis system also needs to be able to recognise a speaker's voice from a small number of samples. In the sections that follow, the survey describes the different text-to-speech and voice cloning systems in order to meet the aforementioned requirements. [3] introduces the idea of voice cloning to effectively mimic a speaker's voice in order to generate speech from text while maintaining the same qualities as the speaker's voice. For text-to-speech applications, the paper suggests a neural fusion architecture that includes a unit concatenation technique to enhance speaker similarity and speech quality.

The model is designed to address the challenges of traditional TTS models, such as slow inference speed and complex training procedures, by utilizing a non-autoregressive transformer architecture, which allows for parallel generation of mel-spectrograms from input text without the need for autoregressive decoding. Using a prosody extractor, the phoneme-level unit embeddings are taken out of the original speech. The suggested neural fusion architecture does not rely on a hybrid unit concatenation system, in contrast to previous works, and is still an end-to-end neural network model. However, building a pure hybrid unit selection system based on the scant training data is challenging because voice cloning frequently uses limited audio data.

In addition to the work proposed in [9] and models such as [10], several studies, such as [8], have also tried to incorporate zero-shot learning for an unseen dataset of multiple speakers. Applications like speech-to-speech translation, producing natural-sounding speech from text in low-resource environments, and helping users who have lost their voice to converse again will find this especially helpful.

Using frame of reference speech from a target speaker, the speaker encoder network creates a fixed-dimensional embedding vector that captures the features of the speaker. By training the speaker encoder on a sizable and varied speaker set, the model makes use of transfer learning from speaker verification to capture the knowledge of speaker variability. This enables the model to synthesise natural speech in the voices of unseen speakers and generalise well to them. The study shows how the model can represent gender, pitch, and formant ranges while projecting speaker characteristics onto unseen speakers. The paper does, however, also acknowledge its shortcomings. It notes that mismatched accents can hinder performance on speaker similarity and that certain subtleties associated with distinctive prosody may be lost. Apart from guaranteeing that the cloned voice maintains its emotions, there's a chance that the original audio contains background noise or multiple speakers, commonly referred to as adverse acoustic environments. As suggested in [2], a straightforward method to get around this problem is to use visual cues—specifically, between lip movements and speech—to enhance speaker extraction. The proposed approach involves acquiring speech-lip synchronisation embeddings by employing a pre-training approach and subsequently implementing them in the reentry model, which emulates human visual attention. For this, it comprises three primary components: the attractor and speaker encoder, and a speaker extractor to provide top-down attention to the speaker extractor while characterising the identified speaker. While the paper presents a novel approach to speaker extraction, potential limitations may include the need for extensive labelled data for training and the generalizability of the approach to diverse acoustic environments.

The paper [11] introduces a multilingual TTS model based on Tacotron 2 [10], which utilizes an attention-based sequence-to-sequence architecture to generate log-mel spectrogram frames. The architecture of Tacotron2 is based on a modified version of the sequence-to-sequence framework. Figure 2 describes a Tacotron model consisting of three main components: an attention-based decoder, an encoder and a post-processing network.

The encoder is a convolutional neural network (CNN) that processes the input character sequence and produces a sequence of hidden states. The attention-based decoder is an RNN that takes the encoder's hidden states as input and generates the output spectrogram one frame at a time, using a modified attention mechanism called location-sensitive attention. The post-processing network is a CNN that smooths the output spectrogram and converts it into a format that can be used by the WaveNet vocoder[12]. Notably, even when the training data contains multiple speakers per language, the model uses an explicit language embedding to enable moderate control of speech accent independent of speaker identity.

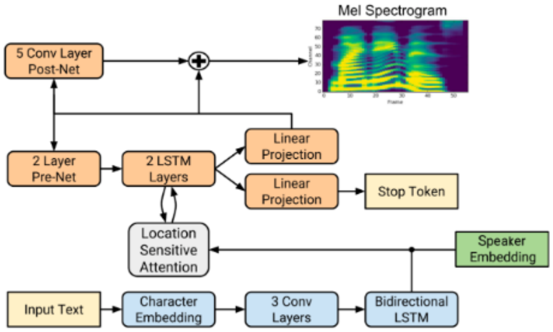

Fig. 2. Tacotron2 Architecture

Among all the voice cloning models mentioned earlier, the Mean Opinion Score (MOS) is the most common numeric metric and is used to evaluate speech synthesis quality. It ranges from 0 to 5, where a higher score indicates a more human-like speech. It is denoted by:

$$\text{MOS} = \frac{1}{N}\sum_{i=1}^{N}\text{Score}_i$$

Here, N denotes the number of samples and $Score_i$ represents the individual score for the i-th assessment given by an observer. Table 3 describes the comparison of the MOS scores obtained for a single speaker by various proposed models.

TABLE III
COMPARISON OF PERFORMANCE OF LIP SYNCING MODELS

Model	**Mean Opinion Score**
Cross-language voice cloning [13]	4.40 ± 0.07
iEmoTTS [4]	4.09 ± 0.04
Neural voice cloning [9]	4.66 ± 0.06
Transfer learning for TTS synthesis [8]	4.67 ± 0.04

III. DISCUSSION

The literature survey in this paper deeply examines the complex fields of audio and video processing, examining the many facets of multilingual systems for automated video dubbing. In the domain of video processing, the survey unveils challenges in facial expressions, lip synchronization, and pose-aware dubbing. Noteworthy among these is the CA Wav2Lip model, as presented in [7], which adeptly tackles the challenge of generating accurate lip-synced videos in real-world scenarios. By introducing a pre-trained model and incorporating attention-based mechanisms, CA Wav2Lip surpasses its counterparts, demonstrating superior performance in speech-to-lip generation. While it exhibits significant strides, considerations must be given to potential limitations, such as its reliance on pre-training and susceptibility to biases in the training data.

Transitioning to the realm of audio synthesis, the survey delves into the distinctions between voice cloning, speech synthesis, and neural machine translation techniques essential for achieving linguistically accurate and emotionally resonant multilingual videos. Notable in this domain is the neural fusion architecture proposed in [3], which introduces a unit concatenation technique to enhance speaker similarity and speech quality. This technique stands out for its ability to effectively mimic a speaker's voice and maintain the same qualities, addressing challenges posed by traditional text-to-speech models. The paper [4] proposes a system that disentangles prosody and timbre to preserve the emotions associated with a speaker's speech. Additionally, the task of cross-language speech transfer, exemplified by the work in [11], enables voice transfer across languages without the need for bilingual or parallel examples, showcasing its versatility in addressing language imbalances in training data.

The surveyed literature demonstrates a rich history of advancements in developing multilingual systems for automated video dubbing. The CA Wav2Lip [7] model shines in the video processing domain, surpassing its counterparts in addressing lip synchronization challenges, while the cross-language voice cloning and iEmoTTS architectures [11][4] showcase prowess in voice cloning, emphasizing speaker similarity, speech quality, and emotion retention. Collectively, these findings indicate the progress towards intelligent and emotionally sensed automated video dubbing systems, with each study offering unique perspectives on the developing field.

IV. CHALLENGES

In the dynamic landscape of multilingual systems for automated video dubbing, persistent challenges demand meticulous examination to fortify the efficiency of existing solutions. While the preservation of emotional authenticity remains a focal point, exemplified by systems like IEmoTTS [4] adept at disentangling prosody and timbre, there exists a need for further refinement to accommodate diverse linguistic nuances and cultural variations. Cross-language voice cloning presents another formidable challenge, evident in the commendable efforts of Cross Language TTS [11]. However, the scarcity of language-specific training data and the intricate correlation between speaker identity and language pose enduring hurdles. The challenges of lip-syncing models, typified by Wav2Lip [5] and AttnWav2Lip [6], persist in accurately synchronizing lip movements amid real-world intricacies, including background noise and varied lip motions in unseen videos. The dependence on pre-training lip-sync experts introduces computational and data dependency concerns [5]. These challenges collectively underscore the need for advancements to unleash the full potential of multilingual video dubbing, ensuring linguistic precision, emotional resonance, and adaptability across diverse linguistic and cultural landscapes.

In addition to the aforementioned challenges, fine issues intricately contribute to the complexities faced by multilingual video dubbing systems. Adverse acoustic environments present an inherent challenge, and although visual cues enhance speaker extraction, as suggested in Selective Listening [2], the approach's efficacy in diverse real-world scenarios remains uncertain. The strategy, employed to learn speech-lip synchronization embeddings, may require extensive labeled data for training, raising concerns about its generalizability. The challenge of cross-speaker emotion transfer and control, addressed in IEmoTTS [4], introduces specific constraints. While successful in disentangling prosody and timbre, the system's primary evaluation on Chinese speech leaves uncertainties about its adaptability to other languages. The reliance on crowd-sourced Mean Opinion Score (MOS) for evaluating emotional characteristics may limit generalizability, prompting a need for objective metrics in multilingual contexts.

The challenge of maintaining linguistic accuracy during speech synthesis across various languages remains intricate. Despite the innovative multilingual TTS model proposed in Cross Language TTS [11], challenges persist regarding accent preservation and the impact of limited training data per language. The model introduces an additional loss for cross-lingual voice transfer, yet mitigating potential losses in linguistic nuances requires continued research. The generalization of voice cloning models to handle unseen speakers and diverse linguistic characteristics is a persistent concern. While models like Transfer Learning For MultiSpeaker [8] showcase promising zero-shot learning capabilities, challenges remain in capturing finer nuances of speaker characteristics. Inherent limitations, such as potential loss of characteristic prosody and constraints in handling accent variations, necessitate ongoing exploration and refinement.

These challenges encompass a spectrum of issues, from handling adverse acoustic environments and preserving emotional authenticity to ensuring cross-lingual voice cloning accuracy and addressing limitations in emotion transfer across different languages. Each challenge presents a unique set of complexities requiring further targeted research and innovation to propel the field of multilingual video dubbing forward.

V. Conclusion

In conclusion, this survey paper has explored the complex field of multilingual systems for automated video dubbing and provided a thorough analysis of the developments made as well as the ongoing obstacles that continue to influence this ever-evolving field. Our investigation has shed light on the practicalities of lip synchronisation in practical contexts in addition to emphasising the advancements made in maintaining emotional authenticity and accomplishing cross-language voice cloning. Prominent systems such as IEmoTTS and Cross Language TTS have made remarkable strides, but there are still significant obstacles to overcome, such as delicately managing linguistic details, managing sparse training data, and guaranteeing lip synchronisation accuracy.

Diving deeper into the intricacies, our survey has unearthed challenges encompassing adverse acoustic environments, the intricate dynamics of cross-speaker emotion transfer, and the delicate balance required to maintain linguistic accuracy. Addressing these challenges necessitates innovative solutions, according to the papers surveyed, such as [4], [7], and [11], to address context-aware dubbing, adaptive lip-syncing through coordinate attention, and efficient handling the prevalence of code-switching sentences to recognize complex sentences with a shared linguistic origin can further elevate the precision and cultural sensitivity of multilingual video dubbing.

As we reflect on these findings, it becomes evident that the path forward involves not only tackling technical challenges relevant to training and gathering data but also delving into the details of emotional expression and cultural relevance. This holistic approach not only refines the existing systems but also sets the stage for future research endeavors. Collaborative efforts and innovative solutions will be essential in unlocking the full potential of multilingual video dubbing, offering a seamless and resonant experience across the diverse linguistic and cultural landscapes that define the global audience. The journey ahead promises a convergence of technical precision and emotional authenticity, shaping a more sophisticated and culturally aware era of multilingual video dubbing.

References

[1] D. Bigioi, H. Jordan, R. Jain, R. McDonnell, and P. Corcoran, “Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing,” *IEEE Access*, vol. 10, pp. 133357–133369, 2022, doi: 10.1109/ACCESS.2022.3231137.

[2] B. Chen, C. Du, and K. Yu, “Neural Fusion for Voice Cloning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1993–2001, 2022, doi: 10.1109/TASLP.2022.3171971.

[3] Z. Pan, R. Tao, C. Xu, and H. Li, “Selective Listening by Synchronizing Speech With Lips,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1650–1664, 2022, doi: 10.1109/TASLP.2022.3153258.

[4] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, “iEmoTTS: Toward Robust Cross-Speaker Emotion Transfer and Control for Speech Synthesis Based on Disentanglement Between Prosody and Timbre,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1693–1705, 2023, doi: 10.1109/TASLP.2023.3268571.

[5] Y. Wang et al., “Tacotron: Towards End-to-End Speech Synthesis,” in *Proc. Interspeech 2017*, pp. 4006–4010, 2017, doi: 10.21437/Interspeech.2017-1452.

[6] KR 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*, pp. 484–492, 2020.

[7] G. Wang, P. Zhang, L. Xie, W. Huang, and Y. Zha, “Attention-based lip audio-visual synthesis for talking face generation in the wild,” *arXiv preprint arXiv:2203.03984*, 2022.

[8] Y. Jia et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in neural information processing systems*, vol. 31, 2018.

[9] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural voice cloning with a few samples,” in *Advances in neural information processing systems*, vol. 31, 2018.

[10] K.-C. Wang et al., “CA-Wav2Lip: Coordinate Attention-based Speech To Lip Synthesis In The Wild,” in *2023 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 1–8, 2023.

[11] J. Shen et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783, 2018.

[12] S. Ronanki, S. Reddy, B. Bollepalli, and S. King, “DNN-based Speech Synthesis for Indian Languages from ASCII text,” *arXiv preprint arXiv:1608.05374*, 2016.

[13] Y. Zhang et al., “Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning,” *arXiv preprint arXiv:1907.04448*, 2019.