

Towards Automatic Face-to-Face Translation

Prajwal K R*
prajwal.k@research.iiit.ac.in
IIIT Hyderabad

Rudrabha Mukhopadhyay*
radrabha.m@research.iiit.ac.in
IIIT Hyderabad

Jerin Philip
jerin.philip@research.iiit.ac.in
IIIT Hyderabad

Abhishek Jha
abhishek.jha@research.iiit.ac.in
IIIT Hyderabad

Vinay Namboodiri
vinaypn@iitk.ac.in
IIT Kanpur

C. V. Jawahar
jawahar@iiit.ac.in
IIIT Hyderabad

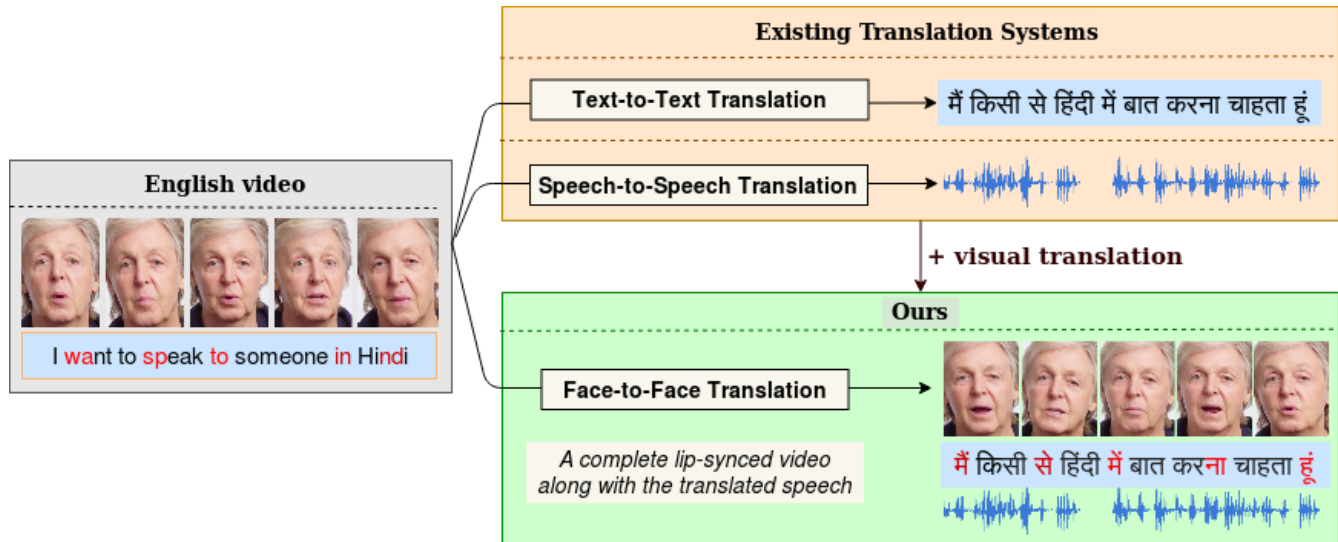


Figure 1: In light of the increasing amount of audio-visual content in our digital communication, we examine the extent to which current translation systems handle the different modalities in such media. We extend the existing systems that can only provide textual transcripts or translated speech for talking face videos to also translate the visual modality i.e. lip and mouth movements. Consequently, our proposed pipeline produces fully translated talking face videos with corresponding lip synchronization.

ABSTRACT

In light of the recent breakthroughs in automatic machine translation systems, we propose a novel approach that we term as "Face-to-Face Translation". As today's digital communication becomes increasingly visual, we argue that there is a need for systems that can automatically translate a video of a person speaking in language A into a target language B with realistic lip synchronization. In this work, we create an automatic pipeline for this problem and demonstrate its impact in multiple real-world applications. First, we

build a working speech-to-speech translation system by bringing together multiple existing modules from speech and language. We then move towards "Face-to-Face Translation" by incorporating a novel visual module, LipGAN for generating realistic talking faces from the translated audio. Quantitative evaluation of LipGAN on the standard LRW test set shows that it significantly outperforms existing approaches across all standard metrics. We also subject our Face-to-Face Translation pipeline, to multiple human evaluations and show that it can significantly improve the overall user experience for consuming and interacting with multimodal content across languages. Code, models and demo video are made publicly available.

*Both authors contributed equally to this research.

*Equal Contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '19, October 21–25, 2019, Nice, France

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6889-6/19/10...\$15.00

<https://doi.org/10.1145/3343031.3351066>

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Machine translation*; *Learning from critiques*.

KEYWORDS

Lip Synthesis; Translation systems; Cross-language talking face generation; Neural Machine Translation; Speech to Speech Translation; Voice Transfer

ACM Reference Format:

Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C. V. Jawahar. 2019. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3343031.3351066>

1 INTRODUCTION

Communicating effectively across language barriers has always been a major aspiration for humans all over the world. In recent years, there has been tremendous progress by the research community towards this goal. Neural Machine Translation (NMT) systems have become increasingly competent[5, 31, 33] in automatically translating foreign languages without the need for a human in the loop. The success of the recent NMT systems not only impacts plain text-to-text translation but also plays a pivotal role in speech-to-speech translation systems. The latter problem is also of great interest because a large part of our communication with others is oral. By cascading speech recognition, neural machine translation and speech synthesis modules, current systems can generate a translated speech output for a given source speech input[10, 19]. In this work, we argue that it is possible to extend this line of research further with a visual module that can greatly broaden the scope and enhance the user experience of existing speech translation systems.

The motivation to incorporate a visual module into a translation system arises from the fact that the majority of the information stream today, is increasingly becoming audio-visual. YouTube, the world's largest online video sharing platform generates 300 hours of video content every minute¹. The meteoric rise of video conferencing [22] also exemplifies the preference for rich audio-visual communication. Existing systems can only translate such audio-visual content at a speech-to-speech level and hence possess some major limitations. Firstly, the translated voice sounds very different from the original speaker's voice. But, more importantly, the generated speech when directly overlaid on the original video produces unsynchronized lip movements with respect to the speech, leading to poor user experience. Thus, we build upon the speech-to-speech translation systems and propose a pipeline that can take a video of a person speaking in a source language and output a video of the same speaker speaking in a target language such that the voice style and lip movements justifies the target language. By doing so, the translation system becomes holistic, and as shown by our human evaluations in this paper, significantly improves the user experience in creating and consuming translated audio-visual content.

Our pipeline is made up of five modules. In the scope of this paper, we work with two widely spoken languages: English and Hindi. For speech to speech translation, we use an automatic speech recognizer[3] to transcribe text from the original speech in language L_A . We adapt state-of-the-art neural machine translation and text-to-speech models[25, 31] to work for Indian languages and generate translated speech in language L_B . We also personalize the voice[14]

generated by the TTS model to closely match the voice of the target speaker.

Finally, to generate talking faces conditioned on the translated speech, we design a novel generative adversarial network, *LipGAN* in which we employ an adversary that measures the extent of lip synchronization in the frames generated by the generator. Furthermore, our system is capable of handling faces in random poses without the need for realignment to a template pose. Our intuitive approach yields realistic talking face videos from any audio with no dependence on language. We achieve state-of-the-art scores on the LRW test set across all quantitative metrics. Using our complete pipeline, we show a proof-of-concept on multiple applications and also propose future directions in this novel research problem. Different resources for this work along with demo videos are available publicly². In summary, our contributions are as follows:

- (1) For the first time, we design and train an automatic pipeline for the novel problem of face-to-face translation. Our system can automatically translate a talking face of a person into a given target language, with realistic lip synchronization.
- (2) We propose a novel model, LipGAN, for generating realistic talking faces conditioned on audio in any language. Our model outperforms existing works in both quantitative and human-based evaluation.
- (3) In the process of creating a face-to-face translation pipeline, we also achieve state of the art neural machine translation results in the Hindi-English language pair by incorporating recent advancements in the area.

The rest of the paper is organized as follows: In section 2, we survey the recent developments in speech, vision and language research which enable our work. Following this, the adaptation of the existing methods in speech and language to our problem setting is described in section 3. Section 4 explains in detail our novel contributions to bring improvements in lip synchronization. We produce a few applications deriving from our work in Section 5 and conclude our findings in Section 6.

2 BACKGROUND

Given a video of a speaker speaking in language L_A , our aim is to generate a lip-synchronized video of the speaker speaking in language L_B . Our system brings together multiple modules from speech, vision, and language to achieve face to face translation for the first time.

2.1 Automatic Speech Recognition

We make use of recent works on Automatic Speech Recognition (ASR)[3] to convert the speech of the source language L_A into the corresponding text. Speech recognition for English has been extensively investigated, owing to the existence of large open-source speech recognition datasets[23, 26] and trained models[3]. We employ the DeepSpeech 2 model to perform English speech recognition in this work.

¹<https://merchdope.com/youtube-stats/>

²<http://cvit.iiit.ac.in/research/projects/cvit-projects/facetoface-translation>

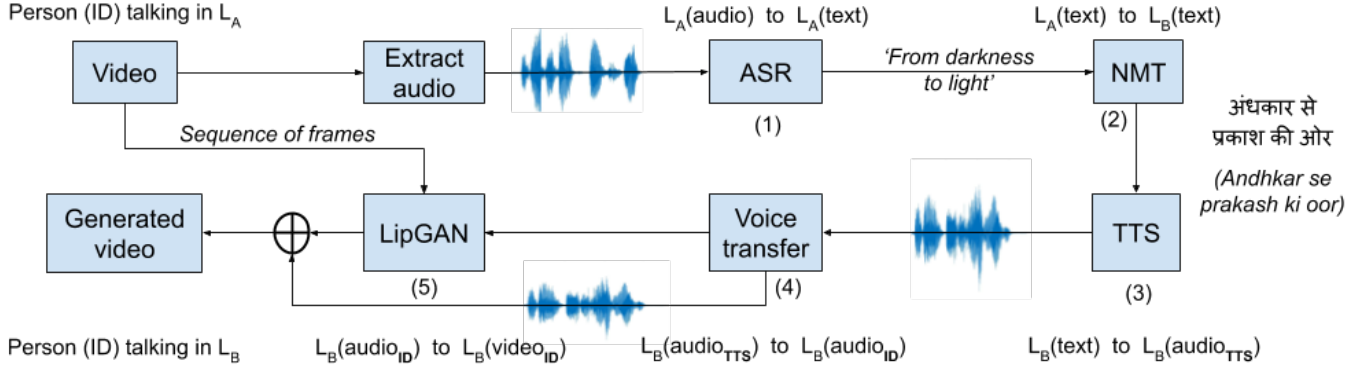


Figure 2: Block diagram of the overall pipeline of our network. In our case, L_A is English and L_B is Hindi. We decompose our problem into: (1) recognize speech in the source language L_A , (2) translate the recognized text in L_A to a target language L_B , (3) synthesize speech from the translated text (5) generate realistic talking faces in language L_B from the synthesized speech. Additionally, to obtain personalized speech for a speaker, we employ a Voice transfer module (4).

2.2 Neural Machine Translation

NMT is often modelled as a sequence to sequence problem which was first introduced with neural networks in Sutskever et al. [28]. Further improvements were brought about with attention mechanisms by Bahdanau et al. [5] and Luong et al. [20]. More recently, Vaswani et al. [31] introduced transformer network which relies only on an attention mechanism to draw global dependencies between input and output. The transformer network outperforms its predecessors by a healthy margin and hence we decided to adopt this into our pipeline. It has also been observed in works like Johnson et al. [13] that training multilingual translation systems also improve the performance especially for low resource languages. Thus, in this work, we also follow a similar path where we use state of the art architectures extended to multilingual learning setups.

2.3 Text to Speech

There has been a lot of work in the area of text-to-speech (TTS) synthesis, starting with the most commonly used HMM-based models[34]. These models can be trained with lesser data to produce fairly intelligible speech, but fail to capture aspects like prosody that is evident in natural speech. Recently, researchers have achieved natural TTS by training neural network-based architectures [25, 27] to map character sequences to mel-spectrograms. We adopt this approach, and train Deep Voice 3[25] based models to achieve high-quality text-to-speech synthesis in our target language L_B . Our implementation of DeepVoice 3 also makes use of a recent work on guided-attention [30] allowing it to achieve high-quality alignment and faster convergence.

2.4 Voice Transfer in Audio

Multiple recent works [4, 25] make use of multi-speaker TTS models to generate voice conditioned on speaker embeddings. While these systems offer the advantage of being able to generate novel TTS voice samples given a few seconds of reference audio, the quality of TTS is inferior [25] compared to single-speaker TTS models. In our system, we employ another recent work[14] that uses a CycleGAN architecture to achieve good voice transfer between two human

speakers with no loss in linguistic features. We train this model to perform a cross-language transfer of a synthetic TTS voice to a natural target speaker voice. We evaluate our models and show that by using just about ten minutes of a target speaker’s audio samples, we can emulate the speaker’s voice and significantly improve the experience of a listener.

2.5 Talking Face Synthesis from Audio

Lip synthesis from a given audio track is a fairly long-standing problem, first introduced in the seminal work of Bregler et al. [6]. However, realistic lip synthesis in unconstrained real-life environments was only made possible by a few recent works [17, 29]. Typically, these networks predicted the lip landmarks conditioned on the audio spectrogram in a time window. However, it is important to highlight that these networks fail to generalize to unseen target speakers and unseen audio. A recent work by Chung et al. [8] treated this problem as learning a phoneme-to-viseme mapping and achieved generic lip synthesis. This leads them to use a simple fully convolutional encoder-decoder model. Even more recently, a different solution to the problem was proposed by Zhou et al. [35], in which they use audio-visual speech recognition as a probe task for associating audio-visual representations, and then employ adversarial learning to disentangle the subject-related and speech-related information inside them. However, we observed two major limitations in their work. Firstly, to train using audio-visual speech recognition, they use 500 English word-level labels for the corresponding spoken audio. We observed that this makes their approach language-dependent. It also becomes hard to reproduce this model for other languages as collecting large video datasets with careful word-level annotated transcripts in various languages is infeasible. Our approach is a fully self-supervised approach that learns a phoneme-viseme mapping, making it language independent. Secondly, we observe that their adversarial networks are not conditioned on the corresponding input audio. As a result, their adversarial training setup does not directly optimize for improved lip-sync conditioned on audio. In contrast, our LipGAN directly optimizes for improved lip-sync by employing an adversarial network

that measures the extent of lip-sync between the frames generated by the generator and the corresponding audio sample. Additionally, both Zhou et al. [35] and Chung et al. [8] normalize the pose of the input faces to a canonical pose, thus making it difficult to blend the generated faces in the original input video. Proposed LipGAN tackles this problem by providing additional information about the pose of the target face as an input to the model thus making the final blending of the generated face in the target video fairly straightforward.

3 SPEECH-TO-SPEECH TRANSLATION

In the previous section, we surveyed the possibility of using state of the art models in speech and language to suit our problem setting. There are not many existing systems reported for speech recognition, machine translation and speech synthesis available for Indian languages. In this section, we describe the current state of the art architectures we use for text and speech, and how we adapt them to our data.

3.1 Recognizing speech in source language L_A

We use publicly available state-of-the-art ASR systems for generating text in language L_A . A publicly available pre-trained model using Deep Speech 2 is used for speech recognition in English. This model was trained on LibriSpeech dataset and achieves WER% of 5.22% on the LibriSpeech test set. Once we have text, recognized in a source language, we translate it into a target language using an NMT model, which we discuss next.

3.2 Translating to target language L_B

We use the re-implementation of Transformer-Base [31] available in fairseq-py³. The language pairs we attempt our problem on contains a low resource language, Hindi. To create a NMT system which works well for Hindi as well as English, we resort to training a multiway model to maximize learning[2, 21]. We closely follow Johnson et al. [13] in training a multi-way model whose parameters are shared across all seven languages - Hindi, English, Telugu, Malayalam, Tamil, Telugu, Urdu. Details of the translation system has been reported in [24]. In Table 1, we report evaluation metrics

Direction	our-BLEU	Online-G
Hindi to English	22.62	19.58
English to Hindi	20.17	17.87

Table 1: NMT Evaluation Scores.

for language directions which are within the scope of this paper. We indicate the size of training data used and the evaluated scores using the widely used Bilingual Evaluation Under Study (BLEU) obtained on the test split of IIT-Bombay Hindi-English Parallel Corpus [18]. We compare against Google Translate⁴ in this test set, which is indicated in Table 1 as Online-G. We achieve an increase of 3 BLEU points on the test set compared to Google Translate.

Next, we describe our methods of generating speech from the target text in L_B , obtained after translating source text in language L_A .

³<https://github.com/pytorch/fairseq>

⁴compared in the first week of April, 2019

3.3 Generating Speech in language L_B

For our Hindi text-to-speech model, we adapt a re-implementation of the DeepVoice 3 model proposed by Ping et al. [25]. Due to the lack of publicly available large scale dataset for Hindi, we curate a dataset similar to LJSpeech by recording Hindi sentences from crawled news articles.

We adopt the NYANKO-BUILD⁵ implementation of DeepVoice 3 to train our Hindi TTS model. We trained on about 10,000 audio-text pairs and evaluated on 100 unseen test sentences. Griffin-Lim algorithm [11] was used to generate waveforms from the spectrograms produced by our model. We evaluate this model by conducting a user study with 25 participants using our unseen test set. The average Mean Opinion Scores (MOS) scores with 95% confidence intervals are reported in Table 2. In the next section, we describe how we can modify the voice of the TTS model to a given target speaker.

Sample Type	MOS
DeepVoice 3 Hindi	3.56
Ground truth Hindi	4.78

Table 2: The MOS for our Hindi TTS is comparable to the same architecture trained on the LJSpeech English TTS dataset.

3.4 Personalizing speaker voice

Voice of a speaker is one of the key elements of her acoustic identity. As our TTS model only generates audio samples in a single voice, we personalize this voice to match the voice of different target speakers. As collecting parallel training data for the same speaker across languages is infeasible, we adopt the CycleGAN architecture [14] to work around this problem.

For a given speaker we collect about 10 minutes of audio clips, which can be easily obtained as we need only a non-parallel dataset. Using our trained TTS model, we generate 5000 samples amounting to about 3 hours worth of synthetic TTS speech. For each speaker, we train a CycleGAN for about 50K iterations with a batch size of 16. The other hyperparameters are the same as used in Kaneko and Kameoka [14]. During inference, given a TTS generated audio sample, the model preserves the linguistic features and generates speech in the voice of the speaker it was trained on.

Speaker	Quality	Similarity	MOS	No Transfer
Modi	4.21	3.56	3.89	1.85
Andrew Ng	3.45	4.1	3.78	1.91
Obama	3.64	2.9	3.27	1.61

Table 3: MOS scores for Voice Transfer of Hindi TTS on various target speakers. Using the CycleGAN approach, we are able to consistently achieve reasonable cross-language voice transfer from the TTS generated voice to a given speaker.

We evaluate our Voice Transfer models in a similar fashion to Kaneko and Kameoka [14], with the help of 30 participants. We use 20 generated TTS samples each of which are voice transferred across 5 famous personalities. Table 3 reports the result of this study.

⁵https://github.com/r9y9/deepvoice3_pytorch

In the next section, we describe how we generate realistic talking face videos.

4 TALKING FACE GENERATION

Given a face image I containing a subject identity and a speech A divided into a sequence of speech segments $\{A_1, A_2, \dots, A_k\}$, we would like to design a model G , that generates a sequence of frames $\{S_1, S_2, \dots, S_k\}$ that contains the face speaking the audio A with proper lip synchronization. Additionally, the model must work for unseen languages and faces during inference. As collecting annotated data for various languages is tedious, the model must also be able to learn in a self-supervised fashion. Table 4 compares our model against recent state-of-the-art approaches for talking face generation.

Method	Works for any face?	Cross-language	No manual labeled data	Smooth blending into target video
Suwajanakorn et al. [29]	×	×	✓	✓
Kumar et al. [17]	×	×	×	✓
Zhou et al. [35]	✓	×	×	×
Chung et al. [8]	✓	✓	✓	×
LipGAN (Ours)	✓	✓	✓	✓

Table 4: Comparison of recent works on talking face synthesis against our LipGAN model. Ours is the first model that generates *realistic* in-the-wild talking face videos across languages without the need for any manually labeled data.

4.1 Model Formulation

We formulate our talking face synthesis problem as “learning to synthesize by testing for synchronization”. Concretely, our setup contains two networks, a generator G that generates faces by conditioning on audio inputs and a discriminator D that tests whether the generated face and the input audio are in sync. By training these networks together in an adversarial fashion, the generator G learns to create photo-realistic faces that are accurately in sync with the given input audio. The setup is illustrated in Figure 3.

4.2 Generator network

The generator network is a modification of Chung et al. [8] and contains three branches: (i) Face encoder, (ii) Audio encoder and a (iii) Face Decoder.

4.2.1 The Face Encoder. We design our face encoder a bit differently from Chung et al. [8]. We observe that during the training process of the generator in [8], a face image of random pose and its corresponding audio segment is given as input and the generator is expected to morph the lip shape. However, the ground-truth face image used to compute the reconstruction loss is of a completely different pose, and as a result, the generator is expected to change the pose of the input image without any prior information. To mitigate this, along with the random identity face image I , we also provide the desired pose information of the ground-truth as input to the face encoder. We mask the lower half of the ground truth face image and concatenate it channel-wise with I . The masked ground truth image provides the network with information about the target pose while ensuring that the network never gets any

information about the ground truth lip shape. Thus our final input to the face encoder is a $H \times H \times 6$ image. The encoder consists of a series of residual blocks with intermediate down-sampling layers and it embeds the given input image into a face embedding of size h .

4.2.2 Audio Encoder. The audio encoder is a standard CNN that takes a Mel-frequency cepstral coefficient (MFCC) heatmap of size $M \times T \times 1$ and creates an audio embedding of size h . The audio embedding is concatenated with the face embedding to produce a joint audio-visual embedding of size $2 \times h$.

4.2.3 Face Decoder. This branch produces a lip-synchronized face from the joint audio-visual embedding by inpainting the masked region of the input image with an appropriate mouth shape. It contains a series of residual blocks with a few intermediate deconvolutional layers that upsample the feature maps. The output layer of the Face decoder is a sigmoid activated 1×1 convolutional layer with 3 filters, resulting in a face image of $H \times H \times 3$. While Chung et al. [8] employ only 2 skip connections between the face encoder and the decoder, we employ 6 skip connections, one after every upsampling operation to ensure that the fine-grained input facial features are preserved by the decoder while generating the face. As we feed the desired pose as input during training, the model generates a morphed mouth shape that matches the given pose. Indeed, in our results, it can be seen that we preserve the face pose and expression better than Chung et al. [8] and only change the mouth shape. This allows us to seamlessly paste the generated face crop into the given video without any artefacts, which was not possible with Chung et al. [8] due to the random uncontrollable pose variations.

4.3 Discriminator network

While using only an L2 reconstruction loss for the generator can generate satisfactory talking faces [8], employing strong additional supervision can help the generator learn robust, accurate phoneme-viseme mappings and also make the facial movements more natural. Zhou et al. [35] employed audio-visual speech recognition as a probe task to associate the acoustic and visual information. However, this makes the setup language-specific and offers only indirect supervision. We argue that directly testing whether the generated face synchronizes with the audio provides a stronger supervisory signal to the generator network. Accordingly, we create a network that encodes an input face and audio into fixed representations and computes the L2 distance d between them. The face encoder and audio encoder are the same as used in the generator network.

4.4 Joint training of the GAN framework

Our training process is as follows. We randomly select a T millisecond window from an input video sample and extract its corresponding audio segment A , resampled at a frequency F Hz. We choose the middle video frame S in this window as the desired ground-truth. We mask the mouth region (assumed to be the lower-half of the image) of a person in the ground truth frame to get S_m . We also sample a negative frame S' , i.e., a frame outside this window which is expected to not be in sync with the chosen audio segment A .

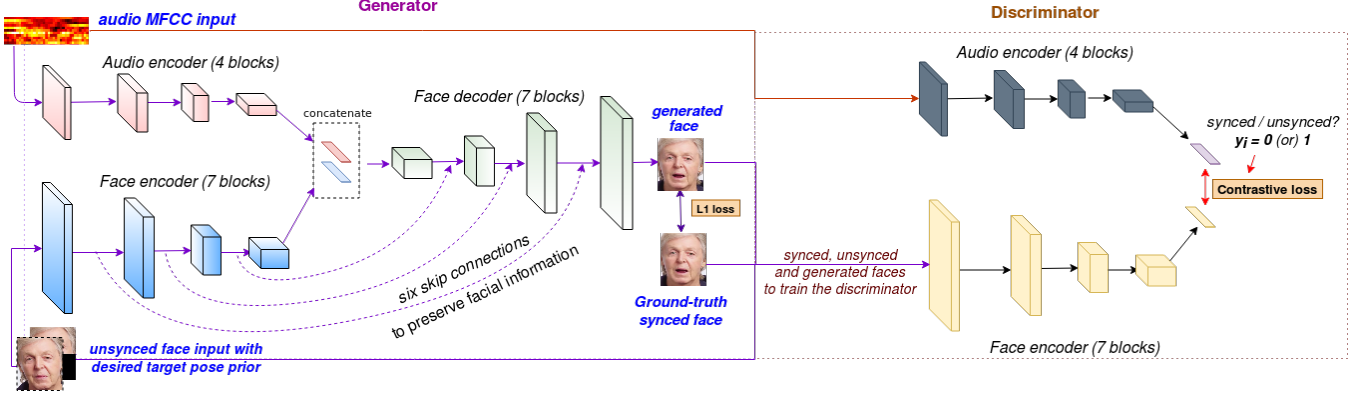


Figure 3: We train our LipGAN network in an intuitive GAN setup. The generator generates face images conditioned on the audio input. The discriminator checks whether the generated frame and the input audio are in sync. Note that while training the discriminator, we also feed extra ground-truth synced / unsynced samples to ensure that the discriminator learns to specifically check for superior lip-sync and not just the image quality.

At each training batch to the generator, the unsynced face S' concatenated channel wise with the masked ground truth face S_m and the target audio segment A is provided as the input. The generator is expected to generate the synced face $G([S'; S_m], A) \approx S$. Each training batch to the discriminator contains three types of samples: (i) Synthetic samples from the generator ($G(S', A), A$); $y_i = 1$, (ii) Actual frames synced with audio (S, A); $y_i = 0$ and (iii) Actual frames out of sync with audio (S', A); $y_i = 1$. The third sample type is particularly important to force the discriminator to take into account the lip synchronization factor while classifying a given input pair as real / synthetic. Without the third type of sample, the discriminator would simply be able to ignore the audio input and make its decision solely on the quality of the image. The discriminator learns to detect synchronization by minimizing the following contrastive loss:

$$L_c(d_i, y_i) = \frac{1}{2N} \sum_{i=1}^N (y_i \cdot d_i^2 + (1 - y_i) \cdot \max(0, m - d_i)^2) \quad (1)$$

where m is the margin, which we set to 2. The generator learns to reconstruct the face image by minimizing the L1 reconstruction loss:

$$L_{Re}(G) = \frac{1}{N} \sum_{i=1}^N \|S - G(S', A)\|_1 \quad (2)$$

We train the generator G and discriminator D using the following GAN objective function:

$$L_{\text{real}} = \mathbb{E}_{z, A} [L_c(D(z, A), y)] \quad (3)$$

$$L_{\text{fake}} = \mathbb{E}_{S', A} [L_c(D(G([S'; S_m], A), A), y = 1)] \quad (4)$$

$$L_a(G, D) = L_{\text{real}} + L_{\text{fake}} \quad (5)$$

where $z \in \{S, S'\}$. Here, G tries to minimize L_a and L_{Re} and D tries to maximize L_a . Thus, the final objective function is:

$$G^* = \arg \min_G \max_D L_a(G, D) + L_{Re} \quad (6)$$

4.5 Implementation Details

We use the LRS 2 dataset [1] which contains over 29 hours of talking faces in the provided train split in the dataset. We train on four NVIDIA TITAN X GPUs with a batch size of 512. We extract 13 MFCC features from each audio segment ($T = 350, F = 100$) and discard the first feature similar to Chung et al. [8]. We detect faces in our input frame using dlib [15] and resize the face crops to $96 \times 96 \times 3$. We use the Adam [16] optimizer with an initial learning rate of $1e-3$ and train for about 20 epochs.

4.6 Results and Evaluation

We evaluate our novel LipGAN architecture quantitatively and also with subjective human evaluation. During inference, the model generates the talking face video of the target speaker frame-by-frame. The visual input is the current frame concatenated with the same current frame with the lower-half masked. That is, during inference, we expect the model to morph the input shape and preserve other aspects like pose and expression. Along with each of the visual inputs, we feed a $T = 350ms$ audio segment. In Figure 4, we compare the talking faces generated by 3 models on audio segments actually spoken by Narendra Modi and Elon Musk.

4.6.1 Quantitative evaluation. To evaluate our lip synthesis quantitatively, we use the LRW test set [9]. We follow the same inference method mentioned above, but with one change. Instead of feeding the current frame as input as mentioned above, we feed a random input frame of the speaker, concatenated with the masked current frame for the pose prior. This is done to ensure we do not leak any lip information to the model while computing the quantitative metrics. In Table 5, we report the scores obtained using standard metrics: PSNR, SSIM [32] and Landmark distance [7]. As can be seen in Table 5, our model significantly outperforms existing works across all quantitative metrics. These results highlight the superior quality of our generated faces (judged by PSNR) and also a highly accurate lip synthesis (LMD, SSIM). The noted increase in

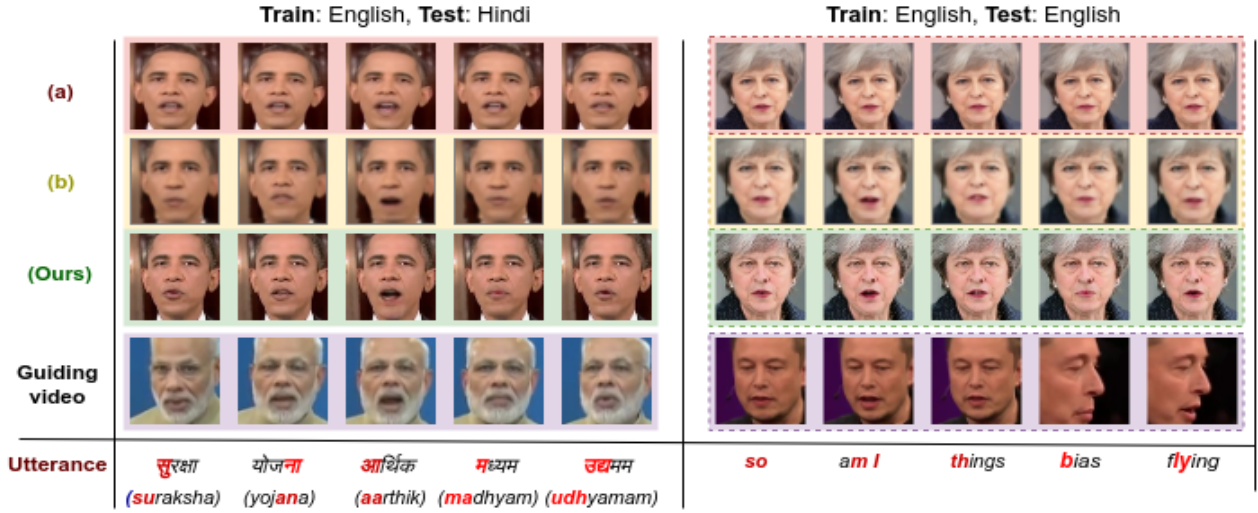


Figure 4: Visual comparison of faces generated by different models when they try to speak specific segments of the words shown in the last row. The audio segments corresponding to these word segments are extracted from the guiding video and fed into each of the models compared above. From top to bottom row: (a) Zhou et al. [35] (b) Chung et al. [8] and Our LipGAN model. While (a) achieves poor lipsync across languages, and (b) generates unnatural lip movements, our LipGAN model produces consistent accurate, natural talking faces across languages.

SSIM and the decrease in LMD can be attributed to the direct lip-synchronization supervision provided by the discriminator, which is absent in prior works.

Algorithm	PSNR	SSIM	LMD
Chung et al. [8]	28.06	0.460	2.22
Zhou et al. [35]	26.80	0.884	-
LipGAN (Ours)	33.4	0.960	0.60

Table 5: Our proposed LipGAN model achieves significant improvements over existing competitive approaches across all standard quantitative metrics.

4.6.2 Importance of the lip sync discriminator. To illustrate the effect of employing a discriminator in the LipGAN network that tests whether the generator faces are in sync, we conduct the following experiment. We train the talking face generator network separately on the same train split of LRS 2 without changing any of the other hyperparameters. We feed the unseen test images shown in Figure 5 along with unseen audio segments as input to our LipGAN network and the plain generator network that was trained without the discriminator. We plot the activations of the penultimate layer of the generator in both these cases. From the heatmaps in Figure 5, it is evident that our LipGAN network learns to attend strongly on the lip and mouth regions compared to the one that is not trained with a lip-sync discriminator. These findings also concur with the significant increase in the quantitative metrics as well as the natural movement of the lip regions in the generated faces.

4.6.3 Human evaluation. Talking face generation is primarily done for direct human consumption. Hence, alongside the quantitative metrics, we also subject it to human evaluation. We choose 10 audio samples, with an equal number of English and Hindi speech

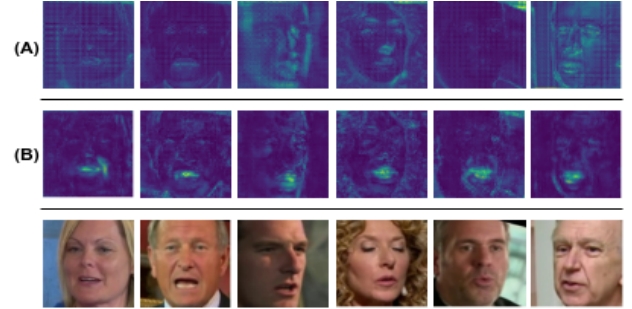


Figure 5: Activation heatmaps from the penultimate layer of two generator networks, one trained without a lip-sync discriminator (A) and the LipGAN network (ours) with a discriminator (B). Our network with the discriminator is highly attentive towards lip and mouth regions.

Approach	Lip-sync rate	Realistic rate
Zhou et al. [35]	2.41	2.42
Chung et al. [8]	2.95	3.10
Ours	3.68	3.73

Table 6: LipGAN achieves significantly higher scores for both realistic rate and the extent of lip synchronization

videos. For each audio sample, we generate talking faces using three different models for 5 popular identities to yield a total of 150 samples. We compare the faces generated by three different models: (i) Chung et al. [8], (ii) Zhou et al. [35] and (iii) Our LipGAN model. We conduct a user study with the help of 20 participants who are asked to rate each of the videos on a scale of 1 to 5 based on the extent of lip synchronization and realistic nature. As shown in

Table 6, our model obtains significantly higher scores compared to existing works.

4.7 Evaluating the complete pipeline

Finally, our Face-to-Face translation pipeline with all the components put together is evaluated based on its impact on the end-user experience. We choose 5 famous identities and generate talking face videos in Hindi of Andrew Ng, Obama, Modi, Elon Musk and Chris Anderson using our complete pipeline. We do this by choosing short videos of each of the above speakers speaking in English. We use our ASR and NMT modules to recognize the speech in English and translate it to Hindi. We use our Hindi TTS model to obtain speech in Hindi. We convert this speech to the voices of each of the above speakers using our CycleGAN models. Using these final voices, we generate talking face videos using our LipGAN network. We compare these videos against videos with (i) English speech and automatically translated subtitles (ii) automatic dubbing to Hindi (iii) automatic dubbing with voice transfer and (iv) automatic dubbing with voice transfer + lip synchronization. Additionally, we also benchmark human performance for speech-to-speech translation: (v) Manual dubbing and (vi) Manual Dubbing + automatic lip synchronization. We ask the users to rate the videos on a scale of 1 – 5 for two attributes. First one is "Semantic consistency" to check whether the automatic pipelines preserve the meaning of the original speech and the second attribute is the "Overall user experience", where the user considers factors such as the realistic nature of the talking face and his/her comfort level. The results of this study are reported in Table 7.

Method	Semantic Consistency	Overall Experience
Automatic Translated Subtitles	3.45	2.10
+ Automatic Dubbing	3.22	2.21
+ Automatic Voice Transfer	3.16	2.54
+ lip-sync	3.16	2.96
Manual dubbing	4.79	4.18
+ lip-sync	4.80	4.55

Table 7: User ratings for different ways to consume cross-language multimedia content.

The results present three major takeaways. Firstly, we observe that there are significant scopes for improvement in each of the modules of automatic speech-to-speech translation systems. Future improvements in each of the speech and text translation systems will improve the user study scores. Secondly, the increase in user scores by using lip synchronization after manual dubbing again validates the effectiveness of the LipGAN model. Finally, note that adding each of our automatic modules increases the user experience score, emphasizing the need for each of them. Our complete proposed system improves the overall user experience over traditional text-based and speech-based translation systems by a significant margin.

5 APPLICATIONS

Our face-to-face translation framework can be used in a lot of applications. The demo video available here⁶ demonstrates a proof-of-concept for each of these applications.

5.1 Movie dubbing

Movies are generally dubbed by dubbing artists manually. The dubbed audio is then overlaid to the original video. This causes the actors' lips to be out of sync with the audio, thus affecting the viewer experience. Our pipeline can be used to automate this process at different levels with different trade-offs. We demonstrate that we can synthesize and synchronize lips in manually dubbed videos, thus automatically correcting any dubbing errors.

We also show a proof-of-concept for performing automatic dubbing using our translation pipeline. That is, given a movie scene in a particular language our system can potentially be used to dub it to a different language. However, as shown by the user study scores in Table 7, significant improvements to the speech-to-speech pipeline is necessary to achieve realistic dubbing of complex speech present in movies.

5.2 Educational videos

As reported before in [12], a large amount of online educational content is present in English in the form of video lectures. They are often aided with subtitles of foreign languages. But this increases the cognitive load of the viewer. Dubbing these videos with just speech-to-speech systems creates a visual discrepancy between the lip motion and the dubbed audio. However, with the help of our face to face translation system, it is possible to automatically translate such educational video content and also ensure lip synchronization.

5.3 Television news and interviews

Automatic Face-to-Face Translation systems can potentially allow viewers to access and consume important information from across the globe irrespective of the underlying language. For example, a Hindi or German viewer can watch an English interview of Obama in the language of his/her choice with lip synchronization.

6 CONCLUSION

We extend the problem of automatic machine translation to face to face translation with a focus on audio-visual content, i.e., where input and output are talking face videos. Beyond demonstrating the feasibility of a Face-to-Face translation pipeline, we also introduce a novel approach for talking face generation. We also contribute towards several language processing tasks (such as textual machine translation) for resource-constrained languages. Finally, we manifest our work on practical applications such as automatically dubbing educational videos, movie clips and interviews. Our attempt of "Face-to-Face Translation" also opens up a number of research directions in computer vision, multimedia processing, and machine learning. For instance, the duration of the speech gets naturally modified upon translation. This demands the transformation of the corresponding gestures, expressions, and background content. In addition to improving the existing individual modules, we believe that the above directions are all open for exploration.

⁶<http://cvit.iiit.ac.in/research/projects/cvit-projects/facetoface-translation>

REFERENCES

- [1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).
- [2] Roei Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. *arXiv preprint arXiv:1903.00089* (2019).
- [3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. 173–182.
- [4] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*. 10040–10050.
- [5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [6] Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: driving visual speech with audio.. In *Siggraph*, Vol. 97. 353–360.
- [7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 520–535.
- [8] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that? *arXiv preprint arXiv:1705.02966* (2017).
- [9] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.
- [10] Christian Federmann and William D Lewis. 2016. Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *International Workshop on Spoken Language Translation*.
- [11] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243.
- [12] Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Cross-Language Speech Dependent Lip-Synchronization. (2019). To appear in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- [13] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5 (2017), 339–351.
- [14] Takuhiro Kaneko and Hirokazu Kameoka. 2017. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293* (2017).
- [15] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.
- [16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [17] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. 2017. Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442* (2017).
- [18] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- [19] Will Lewis. 2015. Skype Translator: Breaking Down Language and Hearing Barriers. In *Proceedings of Translating and the Computer (TC37)*. <https://www.microsoft.com/en-us/research/publication/skype-translator-breaking-down-language-and-hearing-barriers/>
- [20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
- [21] Graham Neubig and Junjie Hu. 2018. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 875–880.
- [22] NPD. 2016. 52 Percent of Millennial Smartphone Owners Use their Device for Video Calling, According to The NPD Group. <https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use-their-device-for-video-calling-according-to-the-npd-group/>
- [23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5206–5210.
- [24] Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2019. A Baseline Neural Machine Translation System for Indian Languages. *arXiv preprint arXiv:1907.12437* (2019).
- [25] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654* (2017).
- [26] Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus.. In *LREC*. 125–129.
- [27] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4779–4783.
- [28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
- [29] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.
- [30] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4784–4788.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.
- [32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [33] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
- [34] Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. *speech communication* 51, 11 (2009), 1039–1064.
- [35] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2018. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. *arXiv preprint arXiv:1807.07860* (2018).