LIPVOICER: Generating Speech From Silent Videos Guided by Lip Reading ICASSP'23

Visual Speech Recognition for Multiple Languages in the Wild ICLR'24

Yewon Min (25/05/28)

Agenda

- LIPVOICER
- VSR

LIPVOICER
(ICASSP’23)

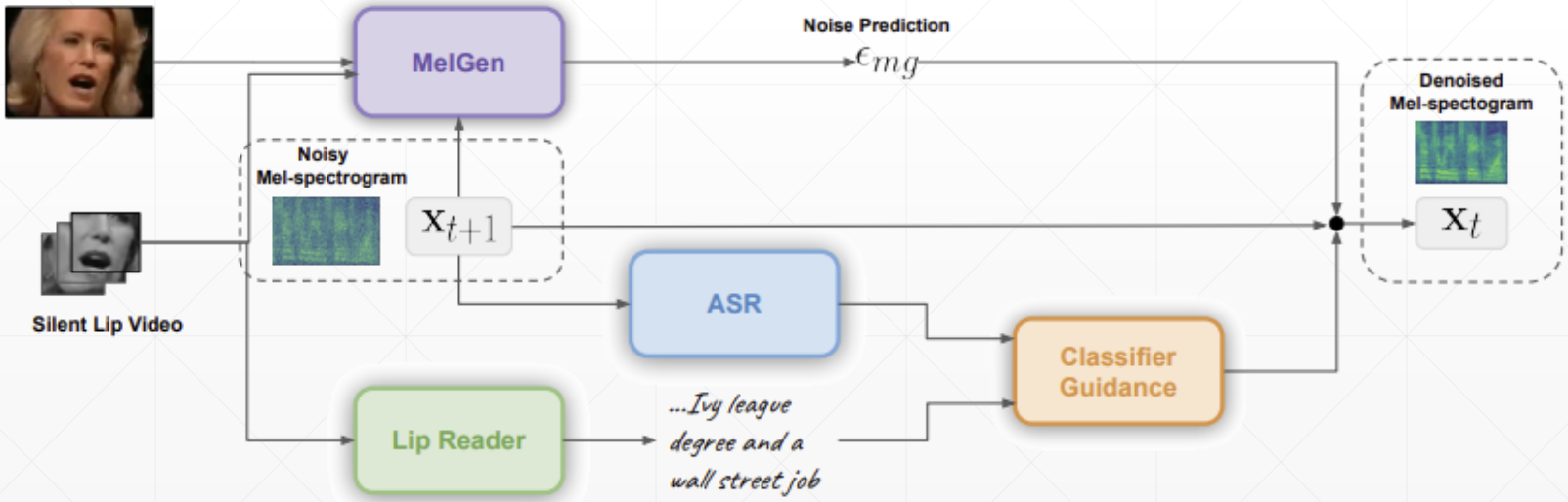
Motivation

- Lip-to-speech
  - **Generating a natural-sounding speech** synchronized with **a soundless video of a person talking**
    - intelligibility, synchronization with lip motion, naturalness, and alignment with the speaker's characteristics such as age, gender, accent, and more
  - Significant progress in recent years
    - partly due to advancements made in deep generative models
- Despite recent advances
  - Produce satisfying results only when applied to a limited number of speakers, and constrained vocabularies
    - GRID (Cooke et al., 2006) and TCD-TIMIT (Harte & Gillen, 2015)
  - Struggle to reliably generate natural speech with high levels of intelligibility
    - on more challenging datasets such as LRS3

Introduction

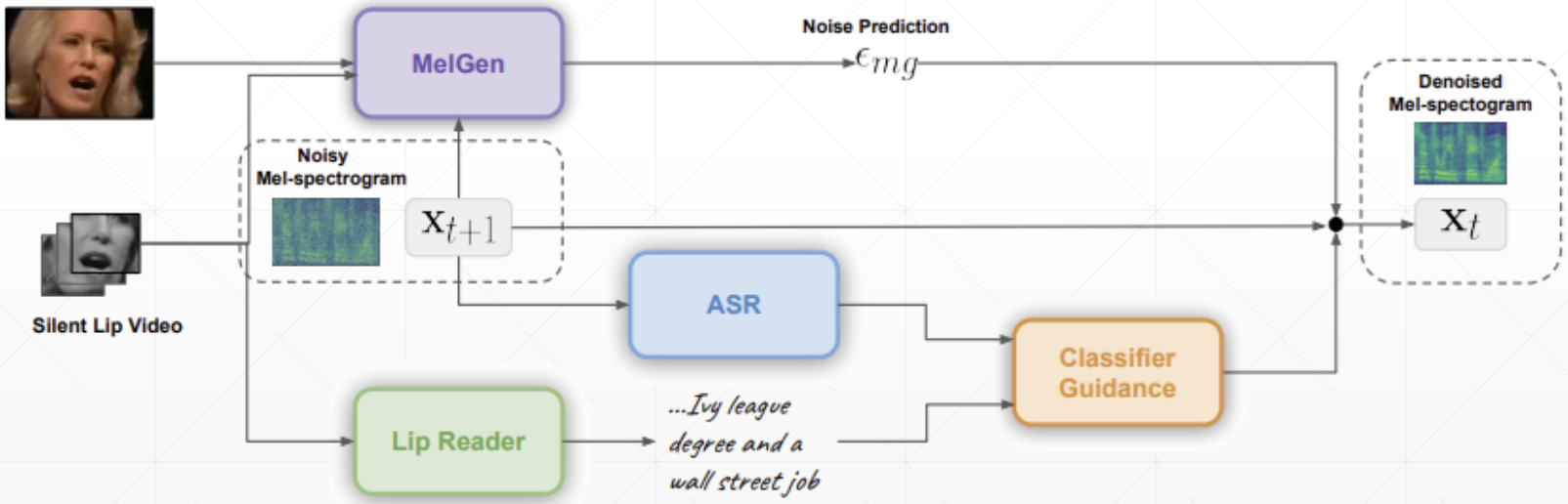
LipVoicer

- A novel method for **producing high-quality speech for silent videos**
  - even for in-the-wild and rich datasets
  - achieves state-of-the-art results for highly challenging in-the-wild datasets
- The first method **to use text inferred by lip-reading** to enhance lip-to-speech synthesis



Method (1/3)

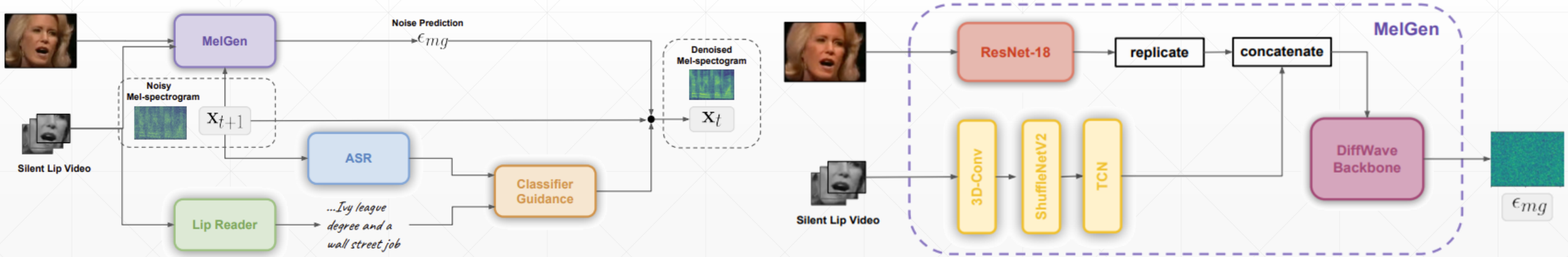
- LipVoicer
  - Input: silent talking-face video *V*
  - Output: a mel-spectrogram that corresponds to a high likelihood underlying speech signal



Method (2/3)

MelGen

- A **mel**-spectrogram **gen**erator that learns to create a mel-spectrogram image from *V*
- A conditional denoising diffusion probabilistic models (ddpm) MODEL
  - Trained to generate a mel-spectrogram waveform x conditioned on the video *V* without the text modality
- DiffWave
  - Residual backbone for MelGen



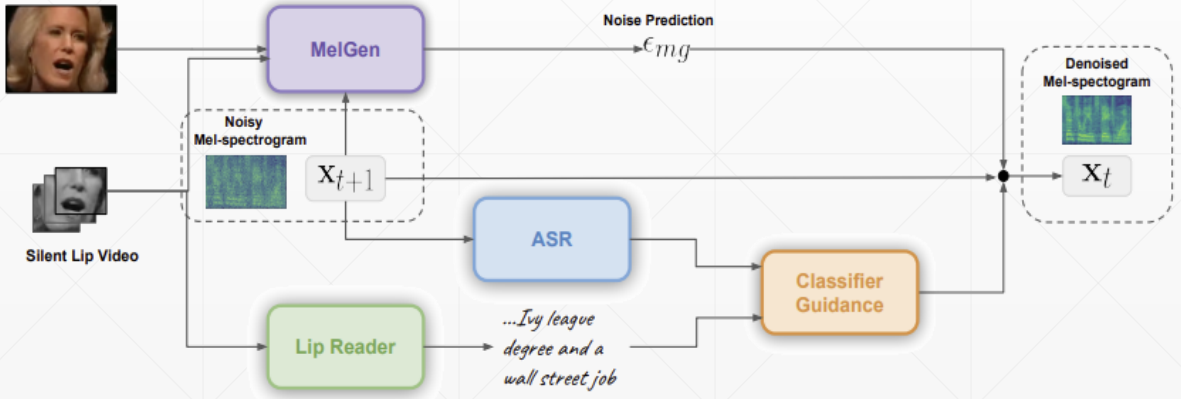
Method (3/3)

Lip Reader

- A pre-trained lip-reading network that infers the most likely text from the silent video
- Text modality as an additional source of guidance at inference time
  - helps the diffusion model to focus on creating naturally synced speech

ASR + Classifier Guidance

- A pre-trained **A**utomatic **S**peech **R**ecognition system that anchors the mel-spectrogram recovered by MelGen to the text predicted by the lip-reader
- Guiding MelGen to match the text as determined by the ASR



Experiments (1/2)

- Datasets
  - LRS2
    - 142,000 Oxford BBC Speech Videos of British English
  - LRS3
    - 151,000 TED Speech Videos
    - English but with different accents, including non-native ones
  - Selected for their diverse range of real-world scenarios
    - with variations in lighting conditions, speaker characteristics, speaking styles, and speaker-camera alignment

Experiments (2/2)

Baselines

- (1) SVTS (de Mira et al., 2022)
  - A transformer-based video to mel-spectrogram generator with a pre-trained neural vocoder
- (2) VCA-GAN (Kim et al., 2021)
  - A GAN model with a visual context attention module that encodes global representations from local visual features
- (3) Lip2Speech (Kim et al., 2023)
  - A multi-task learning framework that incorporates ground truth text to form an additional loss to enforce the correspondence between the text predicted from the generated speech and the target text at train time

Results (1/2)

Human Evaluation Results

- Given 50 random samples, the listeners were asked to rate the videos
- Measured by the mean opinion score (MOS)
  - On a scale of 1-5, for ***Intelligibility***, ***Naturalness***, ***Quality***, and ***Synchronization***
- LipVoicer outperforms all three baseline methods
  - It is also remarkably close to the ground truth scores

Figure 4: MOS evaluations. Participants rank the samples generated by VCA-GAN (Kim et al., 2021), SVTS (de Mira et al., 2022), LIP2SPEECH (Kim et al., 2023), and LIPVOICER (OURS), and the ground-truth. The order of appearance is randomized.

	Intelligibility	Naturalness	Quality	Synchronization
GT	4.38 ± 0.03	4.45 ± 0.03	4.42 ± 0.03	4.36 ± 0.03
LIP2SPEECH (Kim et al., 2023)	2.21 ± 0.08	2.20 ± 0.09	2.01 ± 0.07	2.69 ± 0.08
SVTS (de Mira et al., 2022)	2.17 ± 0.08	2.15 ± 0.09	1.99 ± 0.07	2.71 ± 0.09
VCA-GAN (Kim et al., 2021)	2.19 ± 0.08	2.20 ± 0.09	2.08 ± 0.08	2.71 ± 0.08
LIPVOICER (OURS)	$\mathbf{3.44 \pm 0.07}$	$\mathbf{3.52 \pm 0.07}$	$\mathbf{3.42 \pm 0.08}$	$\mathbf{3.56 \pm 0.07}$

Table 2: LRS3 Human evaluation (MOS).

	Intelligibility	Naturalness	Quality	Synchronization
GT	4.33 ± 0.04	4.43 ± 0.04	4.34 ± 0.04	4.39 ± 0.04
LIP2SPEECH (Kim et al., 2023)	2.07 ± 0.08	1.98 ± 0.08	1.93 ± 0.08	2.66 ± 0.10
VCA-GAN (Kim et al., 2021)	1.77 ± 0.08	1.85 ± 0.09	1.77 ± 0.08	2.34 ± 0.09
LIPVOICER (OURS)	$\mathbf{3.53 \pm 0.07}$	$\mathbf{3.54 \pm 0.08}$	$\mathbf{3.69 \pm 0.08}$	$\mathbf{3.82 \pm 0.07}$

Table 1: LRS2 Human evaluation (MOS).

Results (2/2)

Objective Evaluation Results

- WER (Word Error Rate)
  - Proposed method significantly improves over competing baselines
- STOI-Net (Short-Time Objective Intelligibility) & DNSMOS (Deep Noise Suppression Mean Opinion Score)
  - LipVoicer generates much more intelligible and higher quality speech compared to the competitors
- High-quality content
  - LipVoicer demonstrates commendable synchronization scores, ensuring that the generated natural speech aligns seamlessly with the accompanying video

	WER ↓	STOI-Net ↑	DNSMOS ↑	LSE-C ↑	LSE-D ↓
GT	1.5%	0.91	3.14	6.840	7.194
LIP2SPEECH	51.4%	0.70	2.37	**6.815**	**7.370**
VCA-GAN	100.7%	0.51	2.26	3.369	10.703
LIPVOICER (OURS)	**17.8%**	**0.91**	**2.89**	6.600	7.840

Table 3: Performance comparison between LipVoicer and the baselines on LRS2.

	WER ↓	STOI-Net ↑	DNSMOS ↑	LSE-C ↑	LSE-D ↓
GT	1.0%	0.93	3.30	6.880	7.638
LIP2SPEECH	57.4%	0.67	2.36	5.231	8.832
SVTS	82.4%	0.65	2.42	6.018	8.290
VCA-GAN	90.6%	0.63	2.27	5.255	8.913
LIPVOICER (OURS)	**21.4%**	**0.92**	**3.11**	**6.239**	**8.266**

Table 4: Performance comparison between LipVoicer and the baselines on LRS3.

Limitations and Social Impacts

- LipVoicer
  - A powerful lip-to-speech method that has the potential to bring about some social impacts
  - Positive
    - It can help restore missing or corrupt speech in important videos
  - Negative
    - The generated speech may introduce the risk of misrepresentation or manipulation

VSR
(ICLR'24)

Intro

Intro (1/3)

- Visual speech recognition (VSR)
  - The task of **automatically recognizing speech from video** based only on lip movements
    - also known as lipreading
- Limitations of VSR
  - (1) The lack of large transcribed audio-visual datasets resulted in models
    - only recognize a limited vocabulary and work only in a laboratory environment
  - (2) The use of handcrafted visual features prevented the development of high-accuracy models

Intro (2/3)

- Recently Advancements
  - Large audio-visual transcribed datasets have become available
    - like LRS2 [3] and LRS3 [4]
    - allowed the development of a large vocabulary and robust models
  - Advances in deep learning have made possible the use of end-to-end models
- However, these advances are usually due to the larger training sets rather than the model design
  - demonstrate that designing better models is equally as important as using larger training sets

Intro (3/3)

Proposed Approach

- (1) Addition of prediction-based auxiliary tasks to a VSR model
- (2) Appropriate data augmentations
- (3) Hyperparameter optimization of an existing architecture

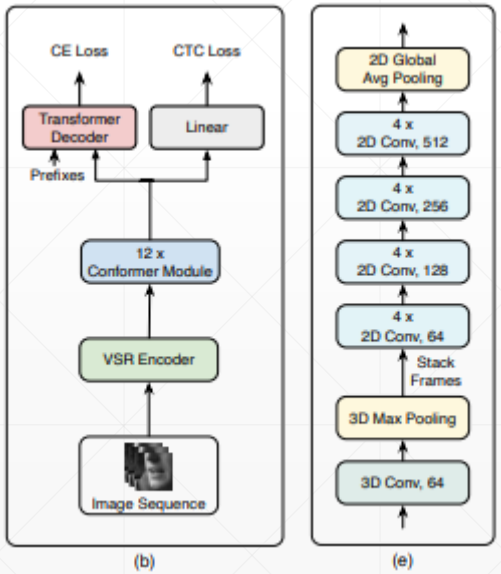
Contribution

- To propose **a novel method for VSR** that outperforms state-of-the-art methods trained on publicly available data by a large margin
- To do so with **a VSR model with auxiliary tasks** that jointly performs VSR and prediction of audio and visual representations
- To demonstrate that the proposed VSR model performs well, not only in **English**, but also in other languages, such as **Spanish, Mandarin, Italian, French** and **Portuguese**

Experiment

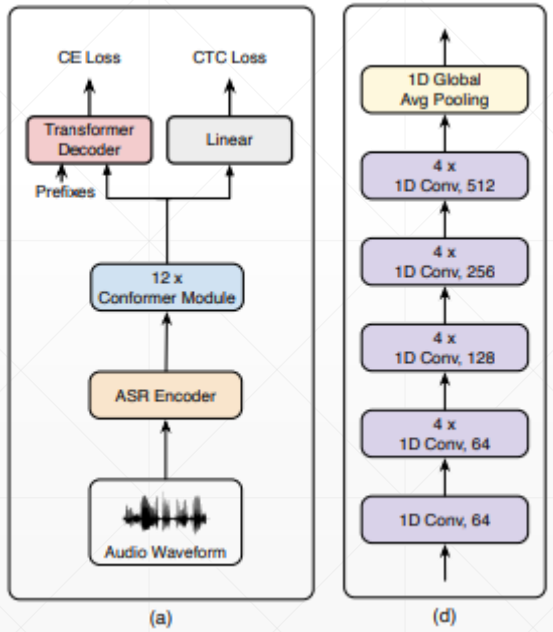
Baseline VSR Model

- *End-to-end audiovisual speech recognition with conformers* [ICASSP'21]
  - A three-dimensional (3D) convolutional layer with a receptive field of five frames
    - a 12-layer Conformer model [11] and a transformer decoder
  - This model achieves state-of-the-art VSR performance on the LRS2 and LRS3 datasets



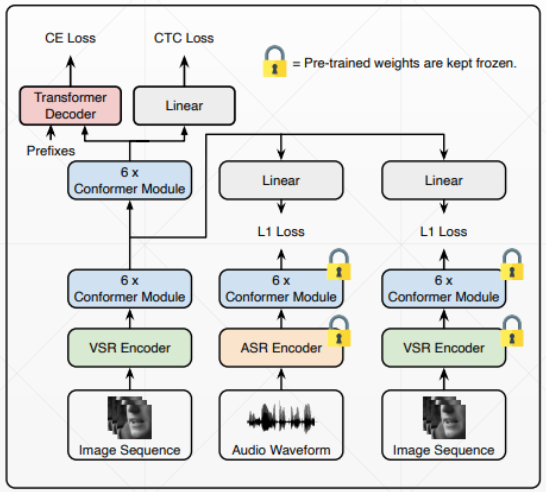
Baseline ASR Model

- *End-to-end audiovisual speech recognition with conformers* [ICASSP'21]
  - An 1D ResNet-18, a 12-layer Conformer model and a transformer decoder
  - This is the state-of-the-art ASR model on the LRS2 and LRS3 datasets



Proposed Approach

- Focus on improving the performance by carefully designing a model without relying on additional data
  - 1) **Optimize hyperparameters and improve the language model (LM)**
  - 2) Introduce **time-masking**, which is a temporal augmentation method that is commonly used in ASR models
    - improves the VSR performance by forcing the model to rely more on contextual information
  - 3) Use **a VSR model with auxiliary tasks** where the model jointly performs VSR and prediction of audio and visual representations extracted from pre-trained VSR and ASR models



(c)

Visual Speech Recognition

Datasets

- LRS2
  - a large-scale audio-visual English dataset collected from BBC programmes
  - 144,482 video clips with a total duration of 224.5 h
- LRS3
  - the largest publicly audio-visual English dataset collected from TED talks
  - 438.9 h with 151,819 utterances
- CMLR
  - a large-scale audio-visual Mandarin dataset collected from a Chinese national news programme
  - 102,072 clips with transcriptions
- CMU-MOSEAS
  - large-scale dataset that contains multiple languages and was collected from YouTube videos
  - 40,000 transcribed sentences and includes Spanish, Portuguese, German and French
- Multilingual TEDx
  - a multilingual corpus collected from TEDx talks
  - covers eight languages with manual transcriptions and has a total duration of 765 h
- AVSpeech
  - a large-scale audio-visual dataset consisting of 4,700 h of video in multiple languages

Performance Metrics

WER

- measures how close the predicted word sequence is to the target word sequence
  - *S*: the number of substitutions
  - *D*: the number of deletions
  - *I*: the number of insertions needed to get from the predicted to the target sequence
  - *N*: the number of words in the target sequence

$$WER = \frac{S + D + I}{N}$$

CER

- Measures how close the predicted and target character sequences are
  - *S*, *D*, and *I*: computed at the character level
  - *N*: the total number of characters

Pre-processing

- Detect 68 facial landmarks
  - RetinaFace face detector
  - Face Alignment Network (FAN)
- Normalization
  - Remove translation and scaling variations
  - The faces were then registered to a neutral reference frame using a similarity transformation
- Crop the mouth region of interest
  - A bounding box of 96 × 96, centred on the mouth centre

Hyperparameter optimization

- Aims to improve the performance of a model by fine-tuning the values of the parameters
  - used to control the training process or the model architecture
- Each hyperparameter was optimized independently based on the WER
- Significant impact on performance
  - *Batch size*
    - increasing the batch size from 8 to 16 led to reduced WER on the validation set of the LRS2 dataset

Improving LMs

- The Role of Language Models (LMs)
  - Determines the probability of a given sequence of characters
  - Used during decoding and favours sequences that are more likely to occur
- Increasing the capacity of the LM (for reduced WER)
  - Increase Multiple text corpora for training
  - Increase the number of sequences considered during decoding
    - Beam size is set to 40

Time Masking

Purpose

- To prevent overfitting by enlarging training datasets
- Most existing VSR augmentation uses image-based methods
  - e.g., random cropping, flipping
  - these spatial augmentations ignore the temporal nature of visual speech
- **Temporal context** is crucial to disambiguate visually similar phonemes

Method

- Works by randomly **masking n consecutive frames** by replacing them with the mean sequence frame
  - more effectively use contextual information and can better disambiguate similar lip movements
  - robust to short missing segments
- 1 mask per second of video during training
  - each mask randomly mask up to 40% of consecutive frames

Prediction-based Auxiliary Tasks

- Limitation of Generalization
  - The standard approach to VSR relies on end-to-end training, which allows the entire model to be optimized towards the desired target
    - challenges the quality of the intermediate representation
- The use of **auxiliary tasks** in the form of additional losses applied to intermediate layers of the model
  - Acts as regularization, which helps the model learn better representations and leads to better generalization on test data
  - The prediction from intermediate layers of audio and visual representations
    - learned by pre-trained ASR and VSR models
  - Layer 6 was found to be the optimal level based on the performance on the validation set

$$L_{AUX} = \beta_a \| h_a(f^l(x_v)) - g_a(x_a) \|_1 + \beta_v \| h_v(f^l(x_v)) - g_v(x_v) \|_1$$

Implementation

- Adam optimizer
  - $\beta_1 = 0.9,\ \beta_2 = 0.98\ \text{and}\ \epsilon\ = 10^{-9}$
- Learning rate of 0.0004
  - 25,000 steps
- 50 epochs with a batch size of 16
- Model averaged over the last ten checkpoints for evaluation

Result

Results (1/4)

- Training Set
  - LRS2: An English audio-visual dataset
- Results
  - Outperforms all existing works by a large margin, even when it is trained on smaller amounts of training data

Method	Pre-training Set	Training Set	Training Sets Total Size (hours)	Mean±Std.	Best
Using Publicly Available Datasets					
MV-WAS [3]	-	LRS2	223	-	70.4
CTC/Att. [12]	LRW	LRS2	380	-	63.5
KD + CTC [13]	VoxCeleb2^{clean}+LRS3	LRS2	995	-	51.3
KD-seq2seq [14]	LRW+LRS3	LRS2	818	-	49.2
TDNN [15]	-	LRS2	223	-	48.9
CM-seq2seq [10]	LRW	LRS2	380	-	37.9
Ours	-	LRS2	223	**33.6±0.5**	**32.9**
Ours	LRW	LRS2	380	**29.5±0.4**	**28.7**
Ours	LRW+LRS3	LRS2	818	**27.6±0.2**	**27.3**
Ours	LRW+LRS3+AVSpeech	LRS2	1 459	**25.8±0.4**	**25.5**
Using Non-Publicly Available Datasets					
TM-seq2seq [4]	MVLRS+LRS3	LRS2	1 391	-	48.3

Results (2/4)

- Training Set
  - LRS3: Another an English audio-visual dataset
- Results
  - Proposed approach substantially outperforms all existing works that are trained using publicly available datasets
  - Performs worse than [6], which presents a model trained on 90 000 hours, which is 62 times more training data

Method	Pre-training Set	Training Set	Training Sets Total Size (hours)	Mean±Std.	Best
Using Publicly Available Datasets					
KD+CTC [13]	VoxCeleb2^{clean}	LRS3	772	-	59.8
KD-seq2seq [14]	LRW+LRS2	LRS3	818	-	59.0
CM-seq2seq [10]	LRW	LRS3	595	-	43.3
Ours	-	LRS3	438	**38.6±0.4**	**37.9**
Ours	LRW	LRS3	595	**35.8±0.5**	**35.1**
Ours	LRW+LRS2	LRS3	818	**34.9±0.2**	**34.7**
Ours	LRW+LRS2+AVSpeech	LRS3	1 459	**32.1±0.3**	**31.5**
Using Non-Publicly Available Datasets					
TM-seq2seq [4]	MVLRS+LRS2	LRS3	1 391	-	58.9
V2P [5]	-	LSVSR	3 886	-	55.1
RNN-T [16]	-	YT-31k	31 000	-	33.6
ViT3D-TM [6]	-	YT-90k	90 000	-	25.9
ViT3D-CM [17]	-	YT-90k	90 000	-	17.0

Results (3/4)

- Training Set
  - CMLR: A Mandarin audio-visual dataset
- Results
  - Achieve an absolute improvement of 12.9 % over the state of the art [9]

Table 3: Results on the CMLR dataset. 'Mean±Std.' refers to the mean character error rate over ten runs and the corresponding standard deviation, while "Best" denotes the best (lowest) CER.

Method	Pre-training Set	Training Set	Training Sets Total Size (hours)	Mean±Std.	Best
LipCH-Net [7]	-	CMLR	61		34.0
CSSMCM [8]	-	CMLR	61	-	32.5
LIBS [18]	-	CMLR	61	-	31.3
CTCH [9]	-	CMLR	61	-	22.0
Ours	-	CMLR	61	**9.1±0.05**	**9.1**
Ours	LRW+LRS2+LRS3	CMLR	879	**8.2±0.06**	**8.1**
Ours	LRW+LRS2+LRS3+AVSpeech	CMLR	1 520	**8.1±0.05**	**8.0**

Results (4/4)

- Training Set
  - CMU-MOSEAS-Spanish: An audio-visual Spanish dataset
  - Pre-trained the model on English datasets and then fine-tuned it using the Spanish videos only
- Results
  - Observe that proposed approach results in a 7.7 % absolute reduction in the WER

Table 4: Results on the CMU-MOSEAS-Spanish (CM_{es}) dataset. 'Mean±Std.' refers to the mean word error rate over ten runs and the corresponding standard deviation, while "Best" denotes the best (lowest) WER.

Method	Pre-training Set	Training Set	Training Sets Total Size (hours)	Mean±Std.	Best
CM-seq2seq [10]	LRW	CM_{es}+MT_{es}	244	58.9±0.8	58.1
Ours	LRW	CM_{es}+MT_{es}	244	**51.5±0.8**	**50.4**
Ours	LRW+LRS2+LRS3	CM_{es}+MT_{es}	905	**47.4±0.2**	**47.2**
Ours	LRW+LRS2+LRS3+AVSpeech	CM_{es}+MT_{es}	1 546	**44.6±0.6**	**43.9**

Conclusion

- Approach for VSR
  - Demonstrated that state-of-the-art performance can be achieved
  - Not only by using larger datasets but also by carefully designing a model
- Importance
  - Hyperparameter optimization
  - Time-masking
  - New architecture based on auxiliary tasks
    - The VSR model also predicts audio-visual representations learned by pre-trained ASR and VSR models
- This approach outperforms all existing VSR works
  - Trained on publicly available datasets in English, Spanish and Mandarin, by a large margin

Q&A