ECE 590SIP

An overview of speech-visual multimodal learning

Xinsheng Wang

November 2, 2020

School of Software Engineering, Xi'an Jiaotong University, China

Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands

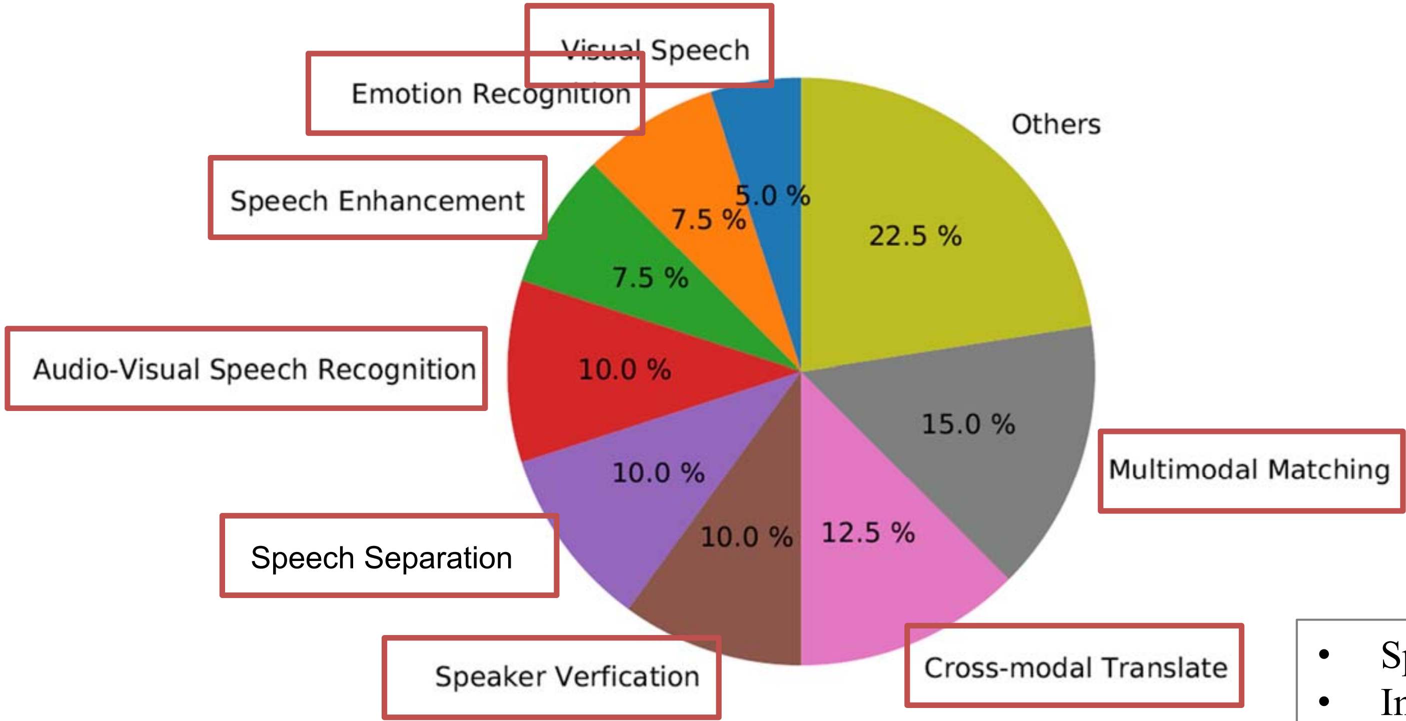
Importance of acoustic and visual information

Learn from watching and listening

Looking at the lips helps understand what people are saying

Overview of visual-speech related papers in Interspeech 2020

- A total of 40 papers published in Interspeech 2020



- Retrieval
- Segment

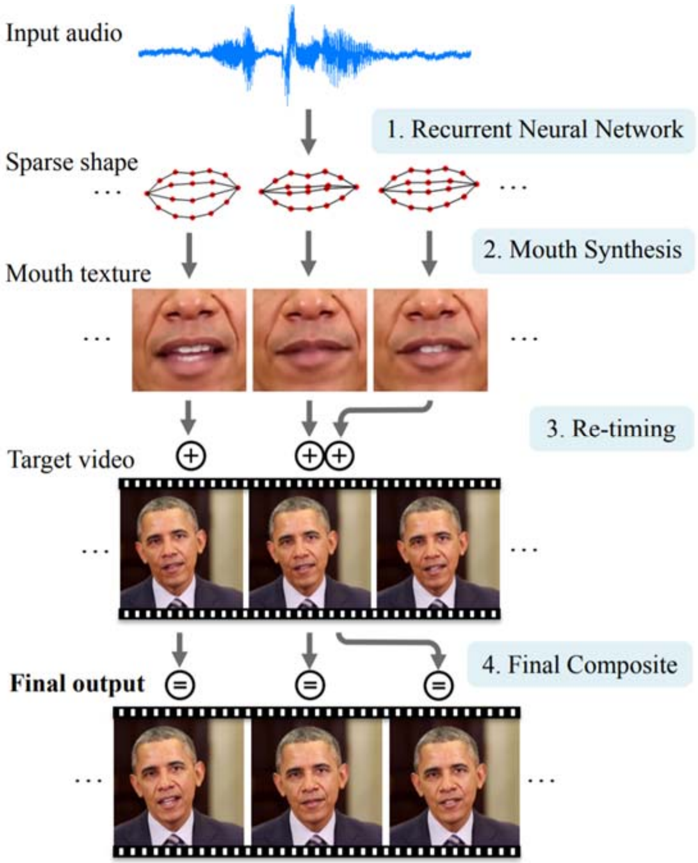
- Speech2Image
- Image2speech

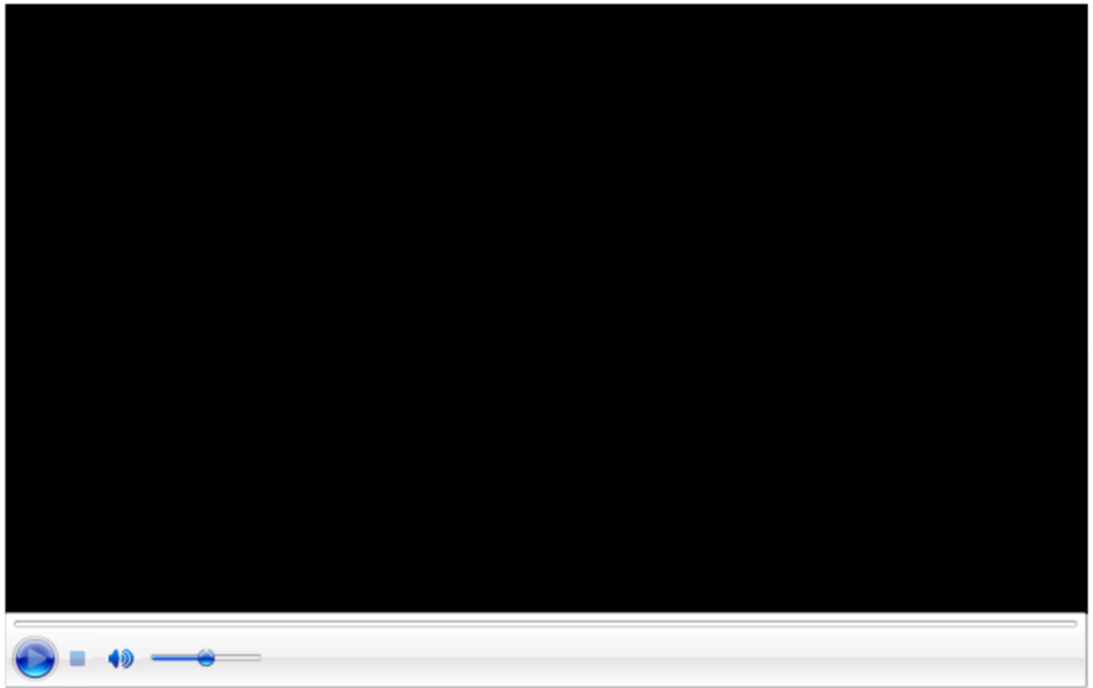
Cross-modal translate

- Cross-modal translate
  - Speech2image
    - Speech2face
      - Face reenactment
      - Infer the face behind the voice
    - Content-based
      - Generate images based on spoken descriptions
  - Image2speech
    - Image caption
      - Phoneme-based
      - End-to-end

Speech-to-Image: Speech2face

- Face reenactment

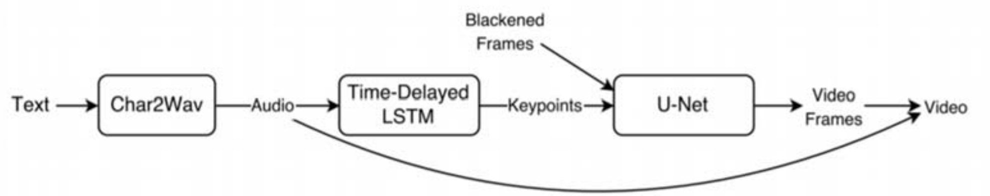


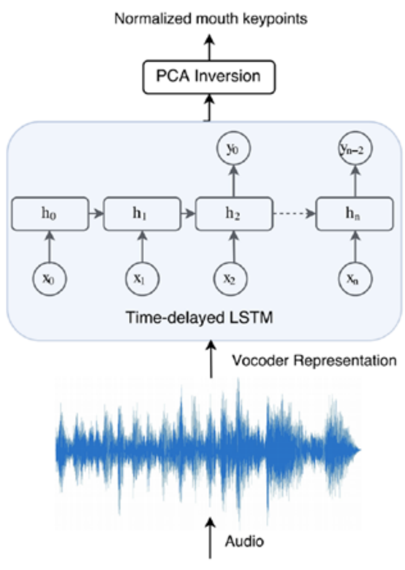


S. Suwajanakorn, et al., "Synthesizing Obama: Learning Lip Sync from Audio," SIGGRAPH, 2017.

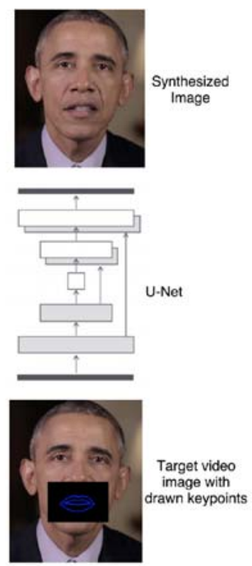
Speech-to-Image: Speech2face

- Face reenactment





a) Keypoint Generation Network

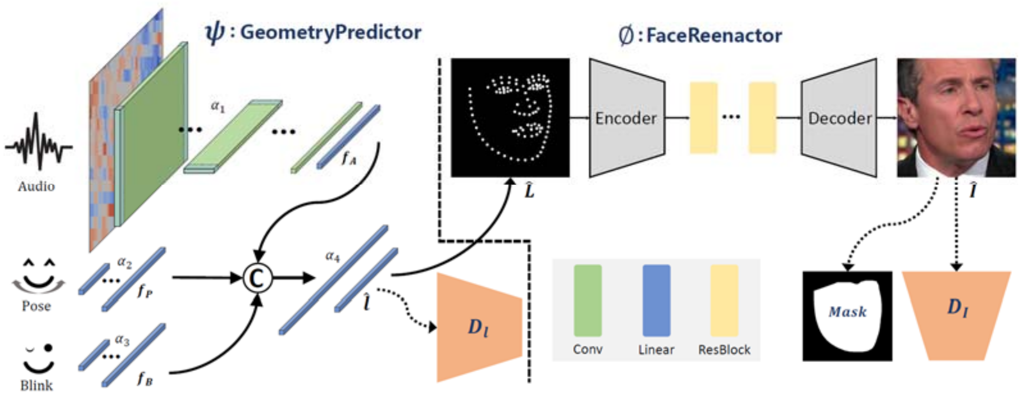


b.) Video Generation Network

R. Kumar, et al., ObamaNet: Photo-realistic lip-sync from text. NeurIPS workshop 2017

Speech-to-Image: Speech2face

➢ Face reenactment



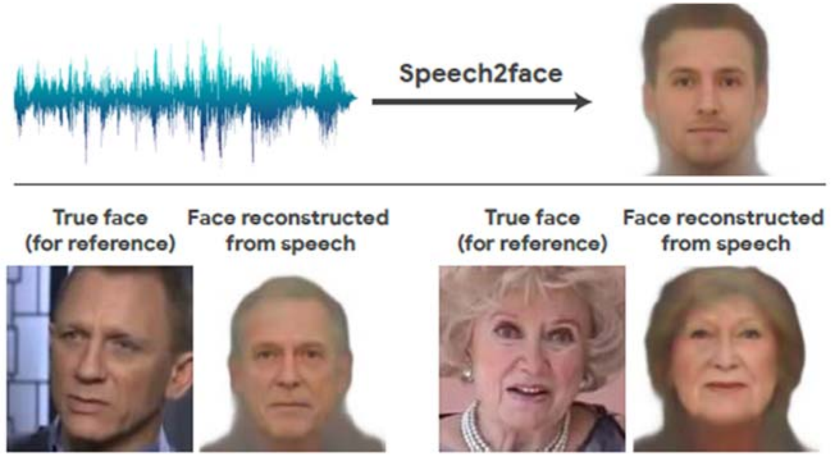
New Database: AnnVI

Face reenactment results

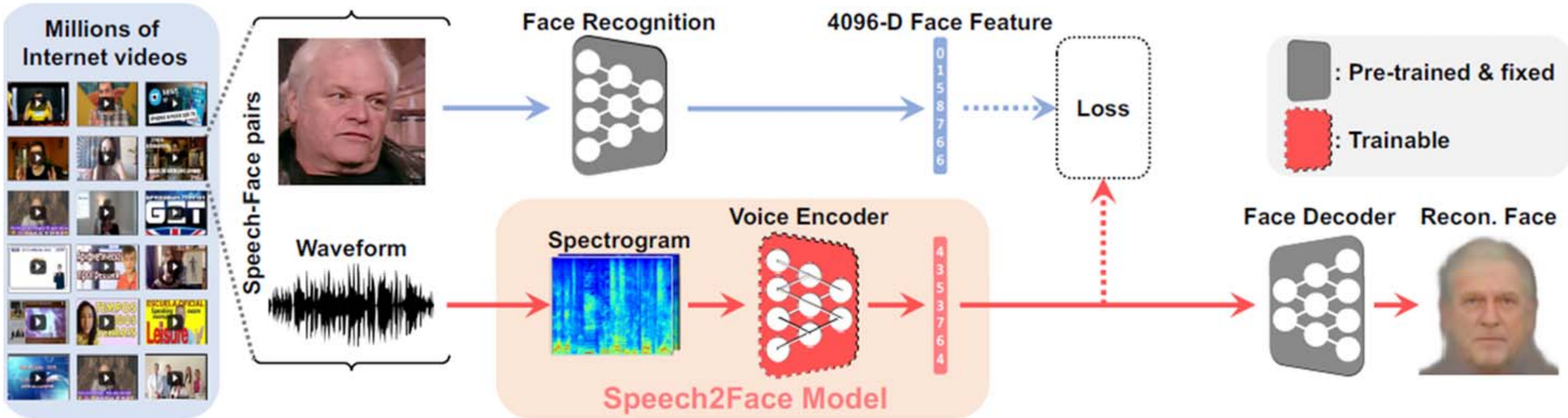
J. Zhang, et al., APB2FACE: audio-guided face reenactment with auxiliary pose and blink signals. ICASSP 2020

Speech-to-Image: Speech2face

- Infer the face



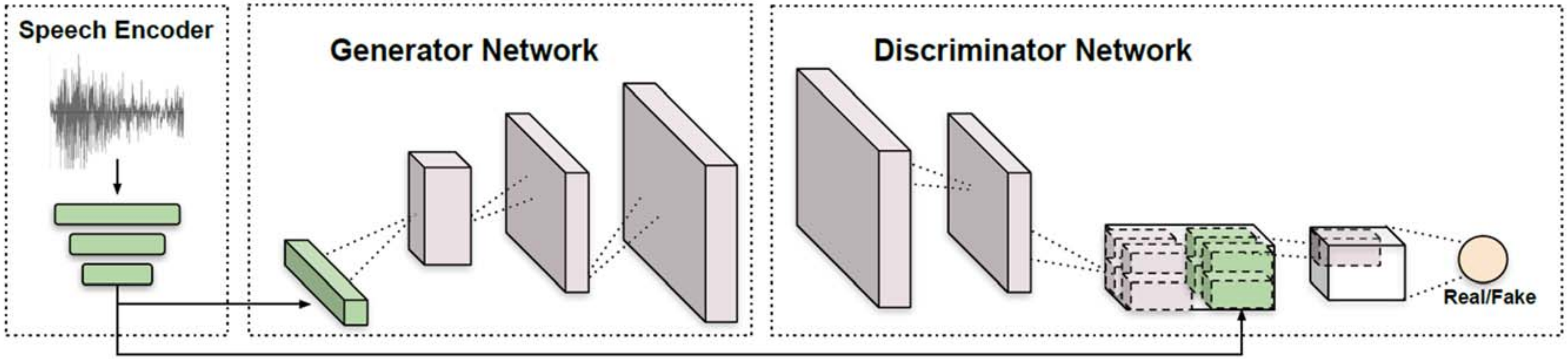
T. Oh, et al., Speech2Face: Learning the Face Behind a Voice. CVPR 2019



Speech-to-Image: Speech2face

- Infer the face

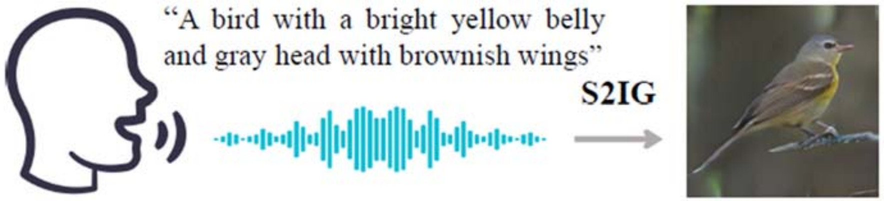
A. Duarte, et al., WAV2PIX: speech-conditioned face generation using generative adversarial networks. ICASSP 2019



Speech-to-Image

- Content-based speech2image

X. Wang, et al., S2IGAN: Speech-to-Image Generation via Adversarial Learning. Interspeech 2020



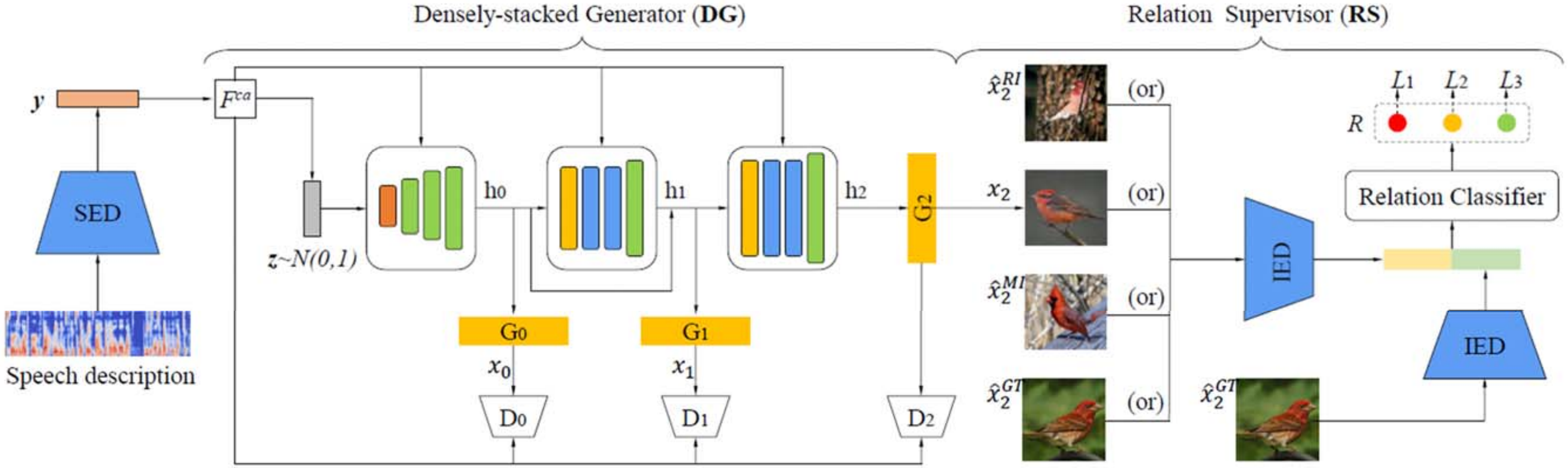
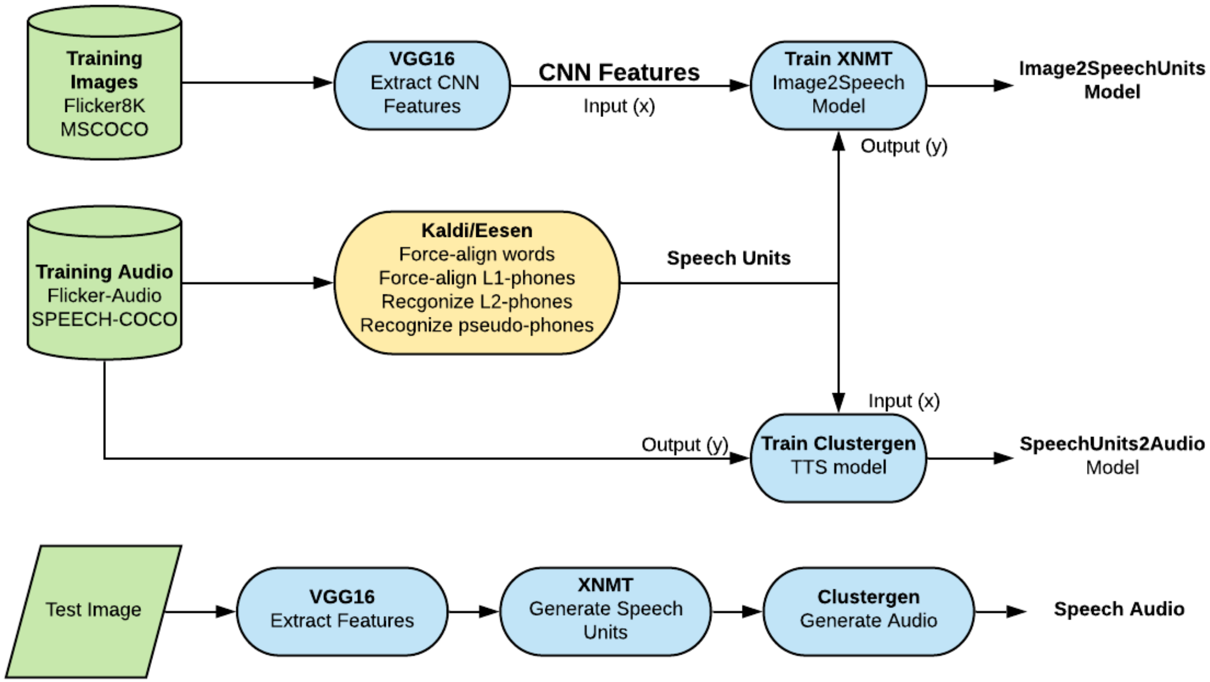


Image-to-speech synthesis: image captioning

➢ Phoneme-based method



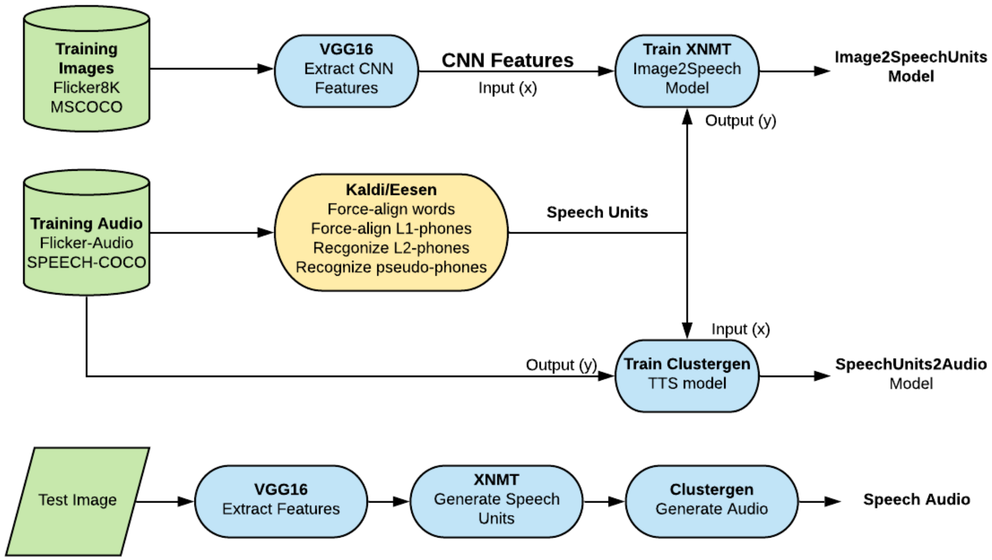
Steps of image-to-speech units-to-speech:

- Extract image features (VGG16)
- Translate image features to speech units (XNMT machine translation toolkit)
- Generate audio from each speech unit sequence (ClusterGEN)

M. Hasegawa-Johnson, et al., "Image2speech: Automatically generating audio descriptions of images," ICNLSSP, 2017.

Image-to-speech synthesis

➢ Phoneme-based method



M. Hasegawa-Johnson, et al., "Image2speech: Automatically generating audio descriptions of images," ICNLSSP, 2017.

Speech Units:

- **L1-phones**: generated using a same language ASR (can not be applicable to unwritten languages)
- **L2-phones**: generated by an ASR that has been trained in some other languages.
- **Pseudo-phones**: generated by an unsupervised acoustic unit discovery system.

	Validation		Test	
Dataset, Targets	BLEU	UER	BLEU	UER
Flickr8k, Words	4.7%	91.3%	3.7%	130%
Flickr8k, L1-Phones	13.7	87.9	13.7	84.9**
Flickr8k, L2-Phones	5.4	115	6.1	101
MSCOCO, Words	4.8		5.5	88.5
MSCOCO, L1-Phones	15.1		16.3	78.8**
MSCOCO, Pseudo-Ph.	2.2		1.4	123

UER: unit error rates

Image-to-speech synthesis

- Phoneme-based method

Speech Units: L1-phones

Figure 3: Examples of a very good and a bad caption. *Left image (rated 6.4) captioned:* "EY G R UW P AX F S K IY R Z AXR S K IY IX NG D AW N EY S N OW IY HH IH L" *("A group of skiers are skiing down a snowy hill.").* *Right image (rated 2.0) captioned:* "EY M AE N IH N EY Y EH L OW SH ER T IH Z S T AE N D IX NG AA N AX S T R IY T" *("A man in a yellow shirt is standing on a street.")*

J. van der Hout, et al., "Evaluating Automatically Generated Phoneme Captions for Images," INTERSPEECH, 2020.

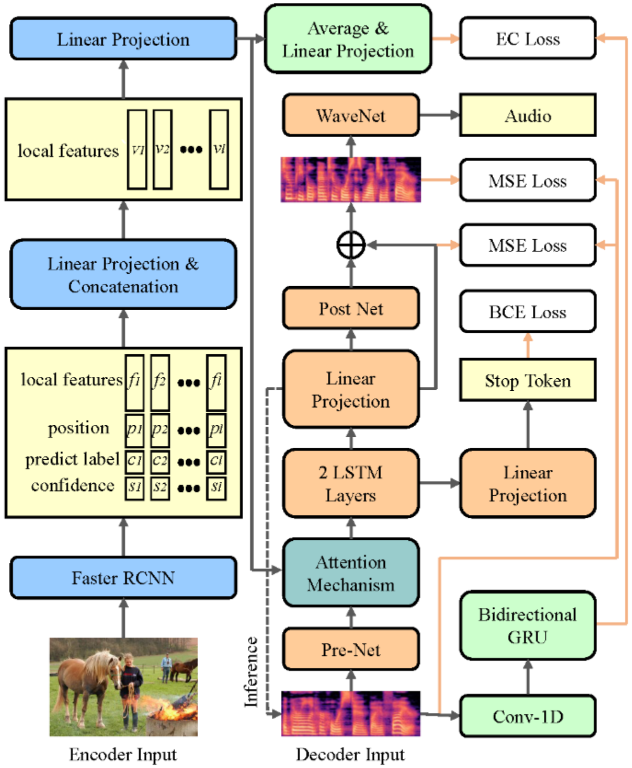
Correlation between different evaluation metrics and human ratings

Metric	Score	r	$r_{actions}$	$r_{objects}$
MTurk	3.40		0.569	0.627
BLEU1	82.6	0.155	0.214	0.195
BLEU2	61.3	0.355	0.388	0.411
BLEU3	46.4	0.425	0.446	0.486
BLEU4	36.1	0.435	0.449	0.494
BLEU5	24.6	0.429	0.435	0.484
BLEU6	18.2	0.410	0.406	0.451
BLEU7	13.7	0.378	0.373	0.423
BLEU8	9.3	0.340	0.319	0.376
METEOR	29.4	0.258	0.265	0.322
ROUGE-L	49.3	0.425	0.416	0.485
CIDEr	42.4	0.272	0.305	0.315
PER	71.4	−0.361	−0.363	−0.381

Image-to-speech synthesis

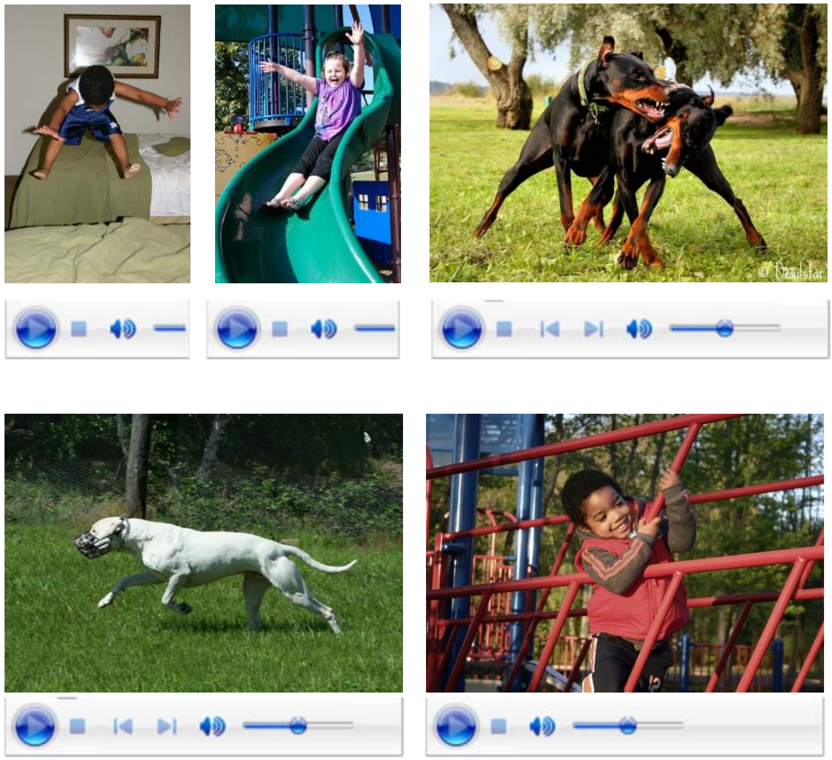
More examples, database, and source code can be found from: https://xinshengwang.github.io/projects/SAS/

- End-to-end method



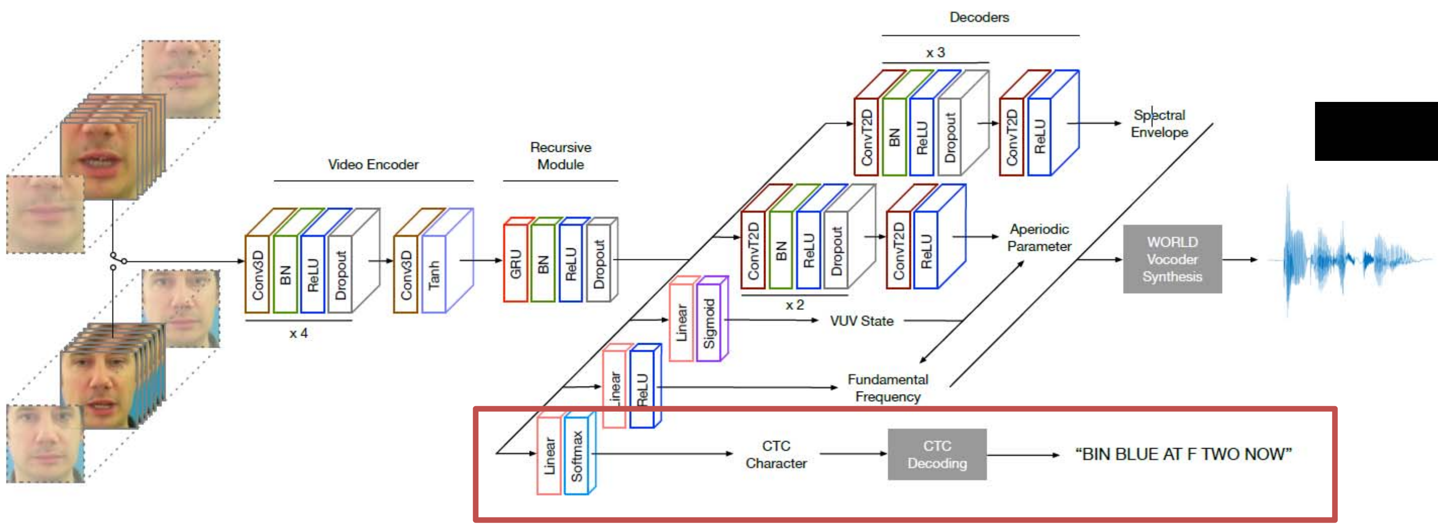
X. Wang, et al., "Show and speak: directly synthesize spoken description of images," submitted to ICASSP 2021

Subjective Results



Video-to-speech synthesis

- Lip reading
  - Synthesize speech from the silent video of a talker



Visual speech recognition (VSR)

D. Michelsanti, et al., "Vocoder-Based Speech Synthesis from Silent Videos," Interspeech 2020.

Cross-modal matching

Multimodal matching

Cross-modal retrieval

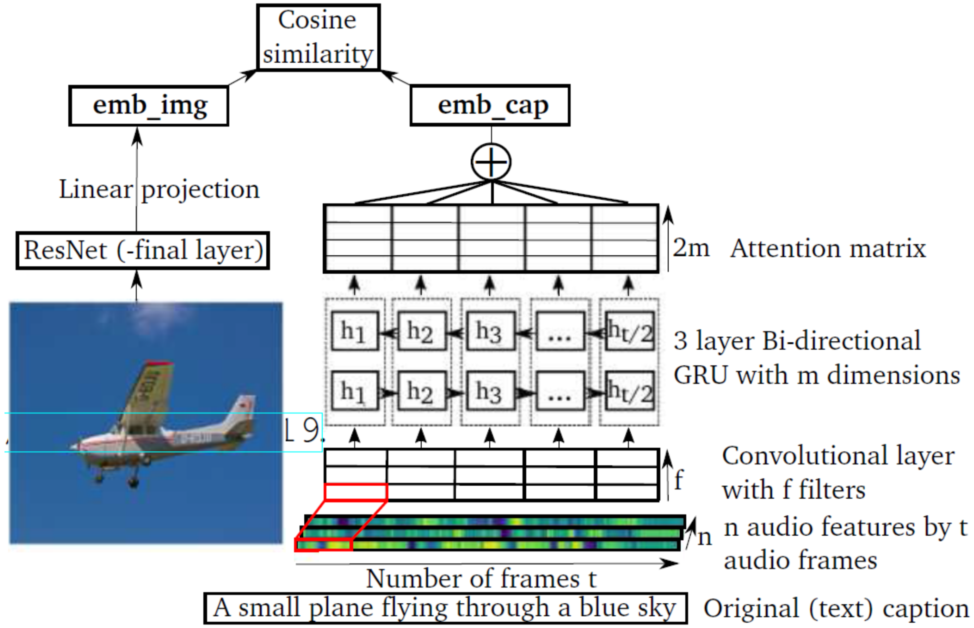
- V. Krishnamohan, et al., "Audiovisual Correspondence Learning in Humans And Machines," Interspeech, 2020.
- M. Zhang, et al., "Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition", Interspeech 2020
- M. Mortazavi., "Speech-Image Semantic Alignment Does Not Depend on Any Prior Classification Tasks", Interspeech 2020
- L. Nortje & H. Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images", Interspeech 2020
- Y. Ohishi, et al., "Pair Expansion for Learning Multilingual Semantic Embeddings using Disjoint Visually-grounded Speech Audio Datasets", Interspeech 2020

Cross-modal segment

- L. Wang, et al., "A DNN-HMM-DNN Hybrid Model for Discovering Word-like Units from Spoken Captions and Image Regions," Interspeech, 2020.

Multimodal matching

- Cross modal retrieval



D. Merkx, et al., "Language learning using Speech to Image retrieval," Interspeech 2019.

Masked Margin Softmax (MMS) Loss

G. Ilharco, et al., "Large-scale representation learning from visually grounded untranscribed speech," ICASSP 2019.

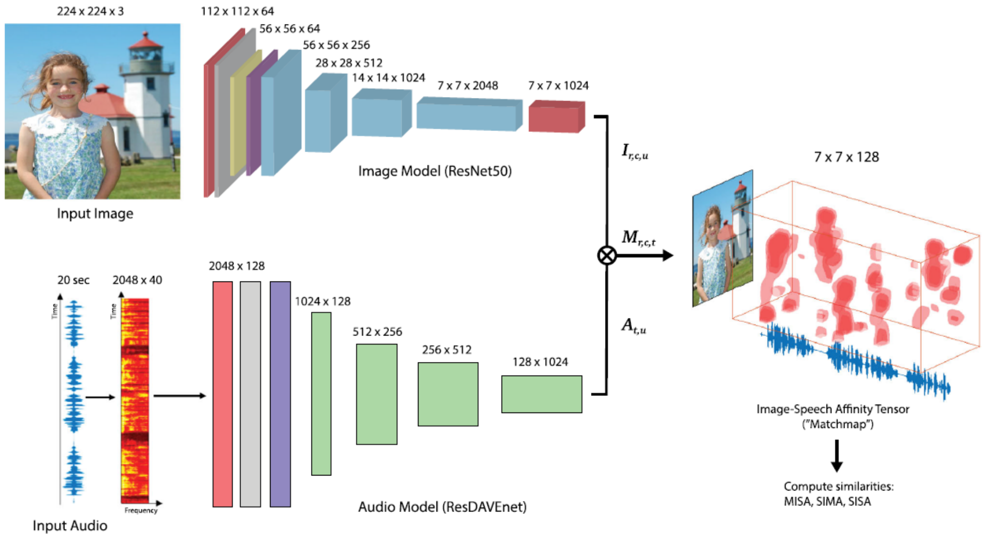
$$\mathcal{L}_{\text{MMS}} = \mathcal{L}_{xy} + \mathcal{L}_{yx}$$

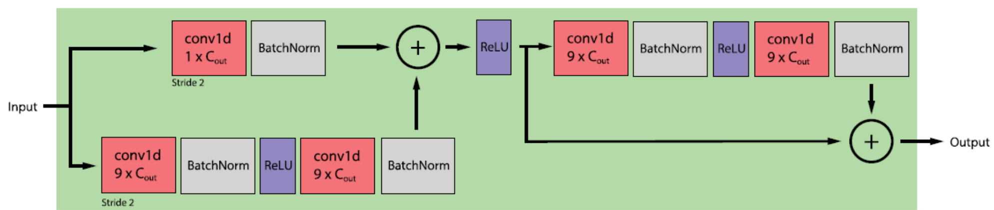
$$\mathcal{L}_{xy} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{\mathbf{Z}_{ii}-\delta}}{e^{\mathbf{Z}_{ii}-\delta} + \sum_{j=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}}$$

$$\mathcal{L}_{yx} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{e^{\mathbf{Z}_{jj}-\delta}}{e^{\mathbf{Z}_{jj}-\delta} + \sum_{i=1}^{B} \mathbf{M}_{ij} e^{\mathbf{Z}_{ij}}}$$

Multimodal matching

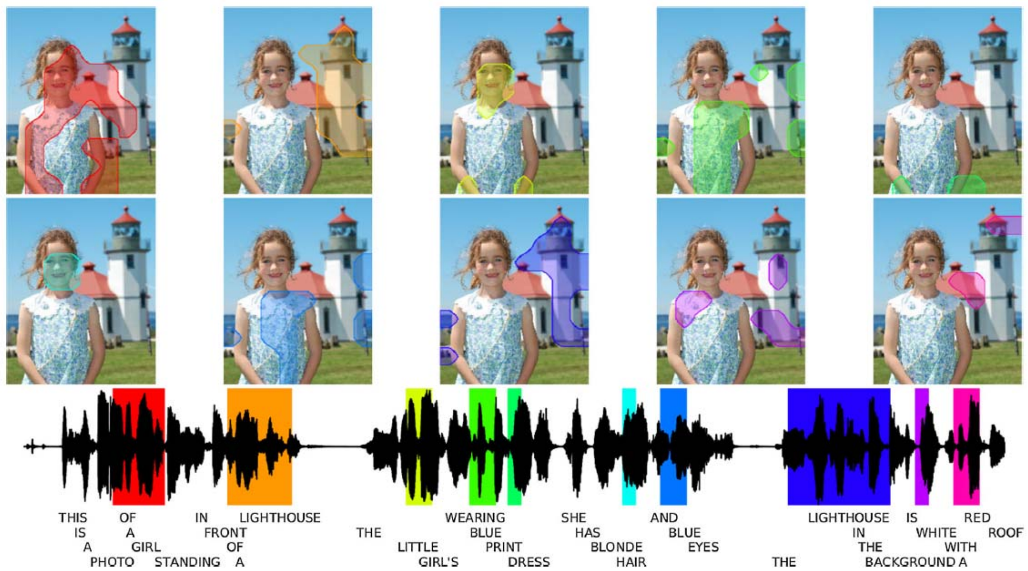
- ➢ Cross modal retrieval & segment





DAVEnet

D. Harwath1, et al., "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input," IJCV 2020.



Multimodal matching

- Cross modal segment (Unit discovery)

$$\mathbf{z}^*, \phi^*, \mathbf{i}^* = \arg\max_{\mathbf{z},\phi,\mathbf{i}} p(\mathbf{z}, \phi, \mathbf{i} | \mathbf{x}, \mathbf{y}).$$

z: image concepts

Φ: phone clusters

i: hidden alignments between each image region and subsets of phone segments that describe that image region

x: acoustic features

y: image region features

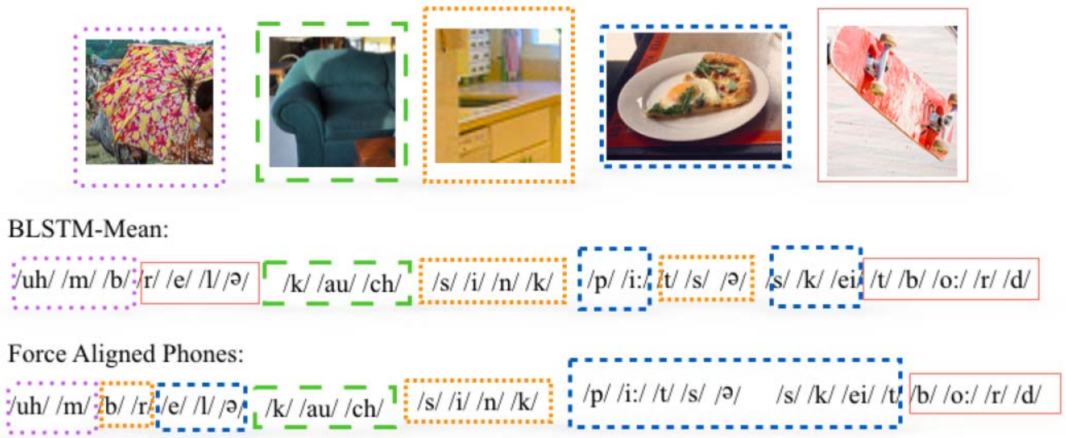
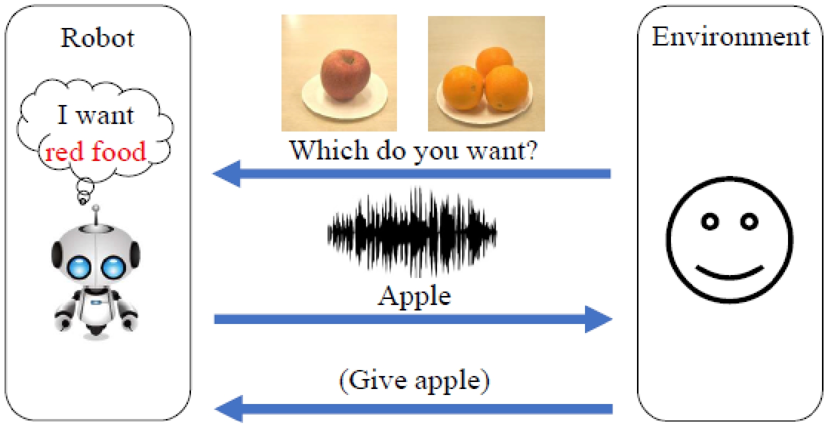


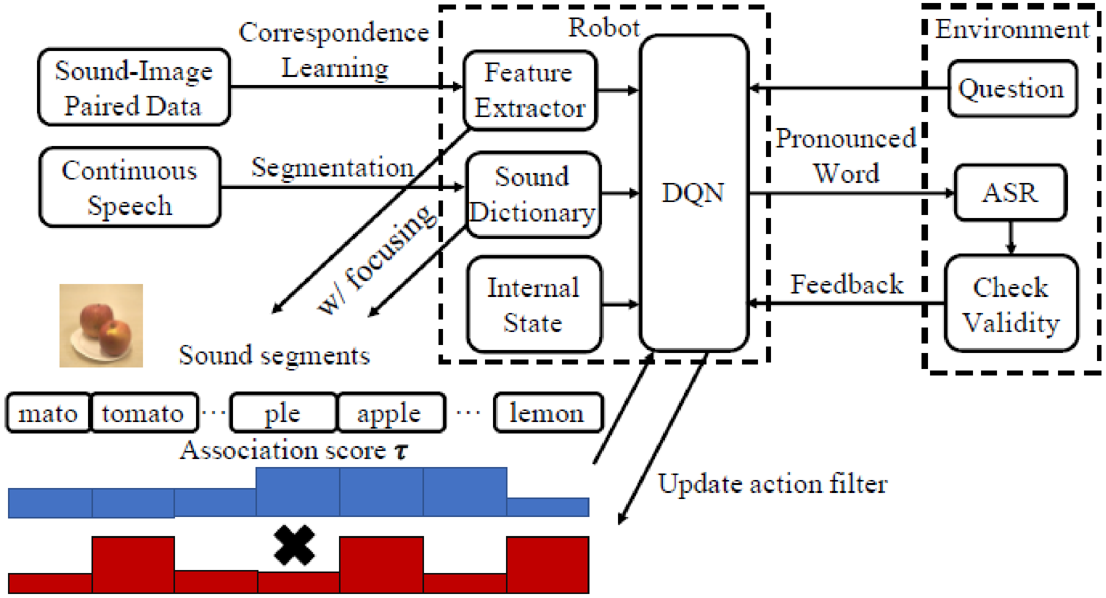
Figure 1: *An example of the image-to-audio word discovery result. The inputs of the algorithm are acoustic phone segments and image regions. The ground truth phone labels are not available during training and only shown for clarity. The phone segment and image region with matching color frames are aligned by the models.*

L. Wang & M. Hasegawa-Johnson, "A DNN-HMM-DNN Hybrid Model for DiscoveringWord-like Units from Spoken Captions and Image Regions," Interspeech 2020.

Multimodal matching

➢ Spoken Language Acquisition Task





M. Zhang, et al., "Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition," Interspeech 2020.

Speaker-related recognition

Speaker-related recognition

- Speaker Verification
- Emotion Recognition

Speaker Verification

- R. Tao, et al., “Audio-visual Speaker Recognition with a Cross-modal Discriminative Network,” Interspeech, 2020.
- G. Antipov, et al., “Automatic Quality Assessment for Audio-Visual Verification Systems. The LOVe submission to NIST SRE Challenge 2019”, Interspeech 2020
- S. Shon & J. Glass, “Multimodal Association for Speaker Verification”, Interspeech 2020
- Z. Chen, “Multi-modality Matters: A Performance Leap on VoxCeleb”, Interspeech 2020

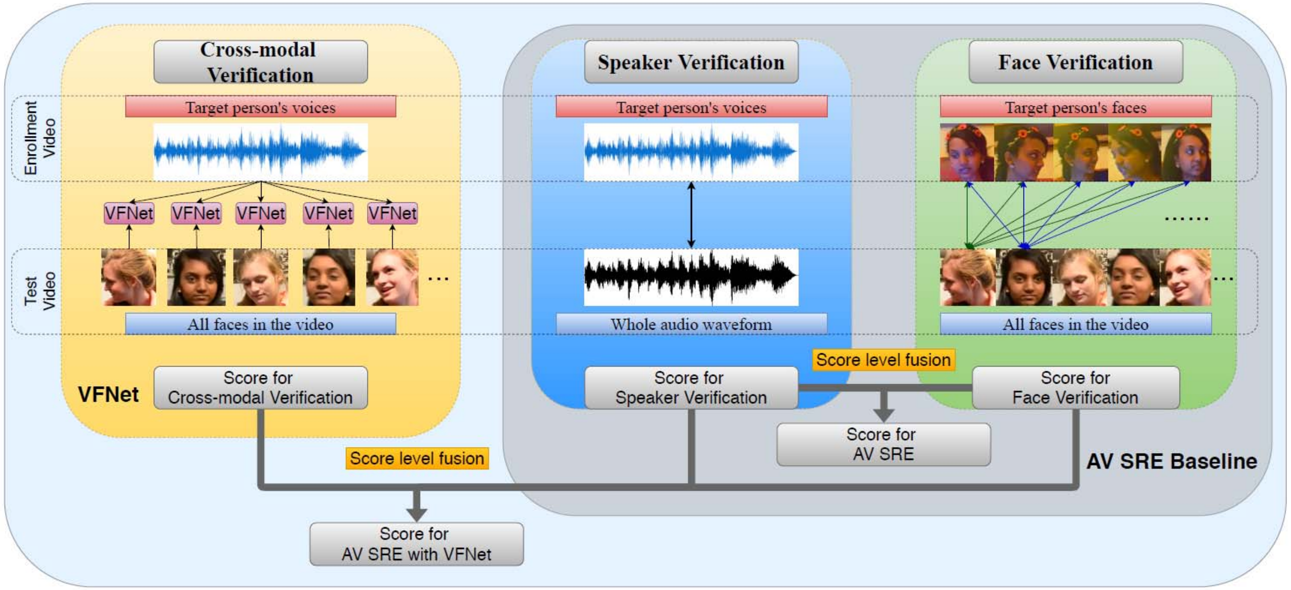
Emotion Recognition

- A. Khare, et al., “Multi-modal embeddings using multi-task learning for emotion recognition”, Interspeech, 2020.
- Z. Pan, et al., “Multi-modal Attention for Speech Emotion Recognition”, Interspeech, 2020.
- J. Zhang, et al., “Multimodal Deception Detection using Automatically Extracted Acoustic, Visual, and Lexical Features”, Interspeech, 2020

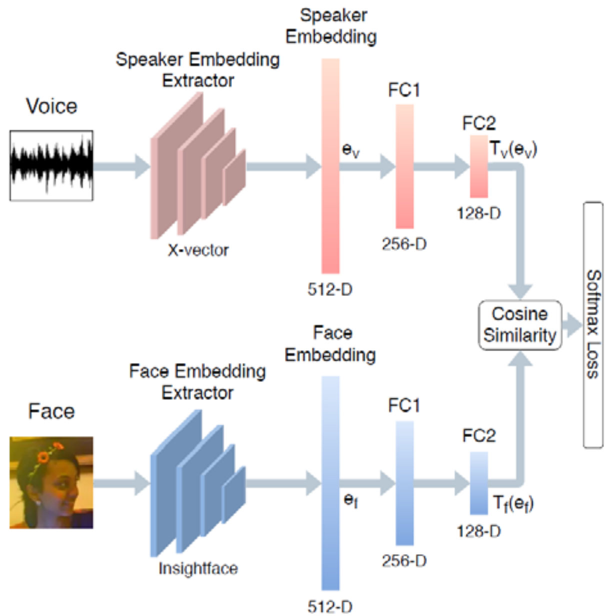
Speaker recognition evaluation (SRE): given an enrollment video, SRE is to determine whether the target person is presented in a given test video.

Speaker-related recognition

- Speaker recognition evaluation (SRE).
  - Take the cross-modal correlations into consideration



R. Tao, et al., “Audio-visual Speaker Recognition with a Cross-modal Discriminative Network,” Interspeech 2020

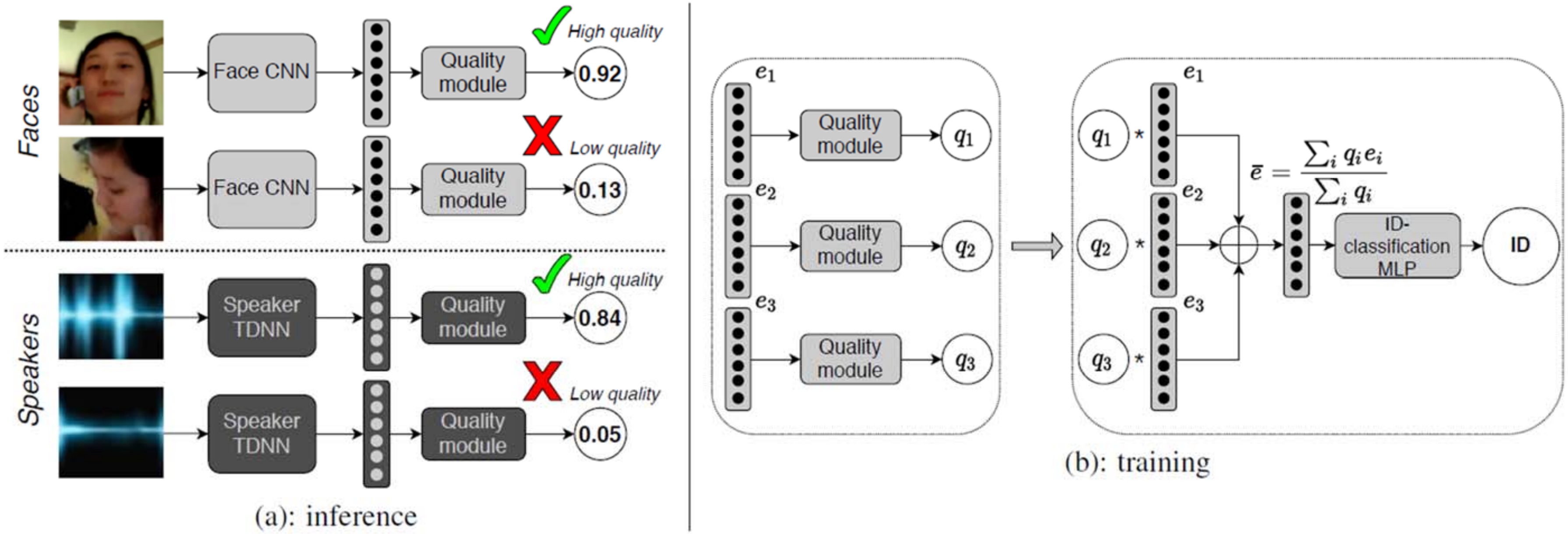


VFNet

Speaker-related recognition

➤ Speaker recognition evaluation (SRE).

- Weighted fuse



G. Antipov, et al., "Automatic Quality Assessment for Audio-Visual Verification Systems. The LOVe submission to NIST SRE Challenge 2019," Interspeech 2020

Speaker-related recognition

➢ Emotion Recognition

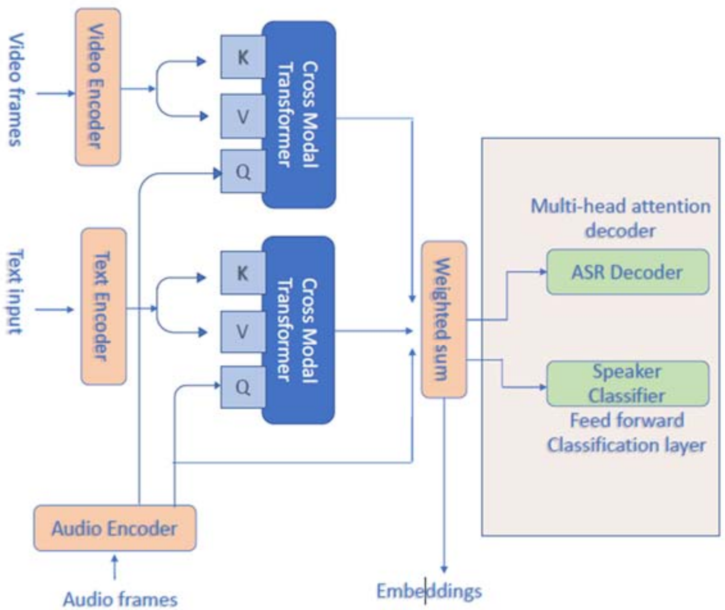


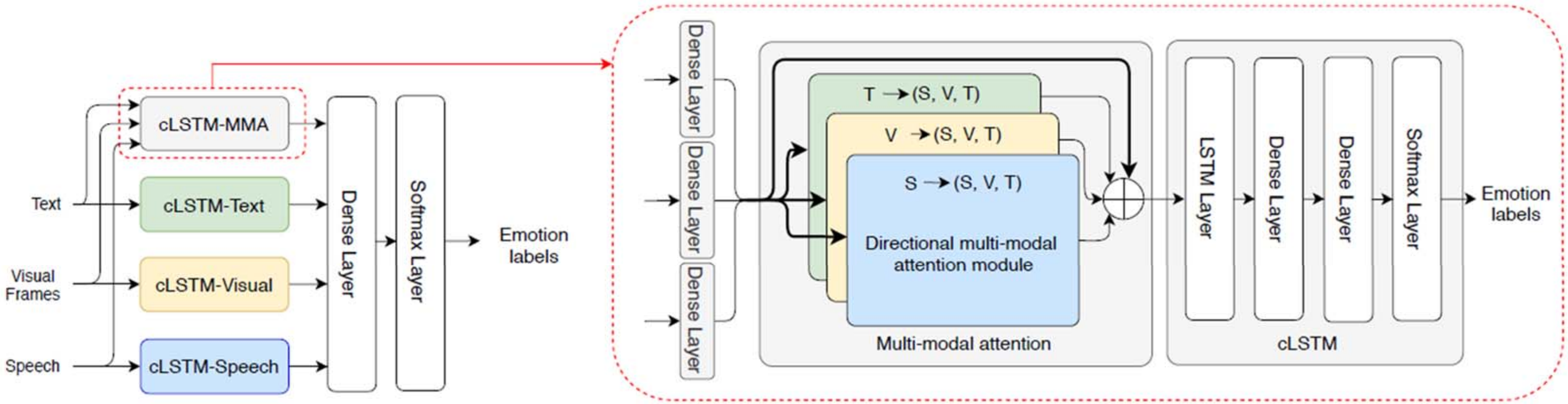
Figure 1: *Transformer based multi-task architecture*

- Step1: Learning fused features in a mutli-task model
- Step2: Emotion recognition based on achieved features

A. Khare, et al., "Multi-modal embeddings using multi-task learning for emotion recognition", Interspeech 2020

Speaker-related recognition

- Emotion Recognition



Z. Pan, et al., "Multi-modal Attention for Speech Emotion Recognition," Interspeech 2020

Speech enhancement and separation

Enhancement / separation

- Speech Enhancement
- Speech Separation

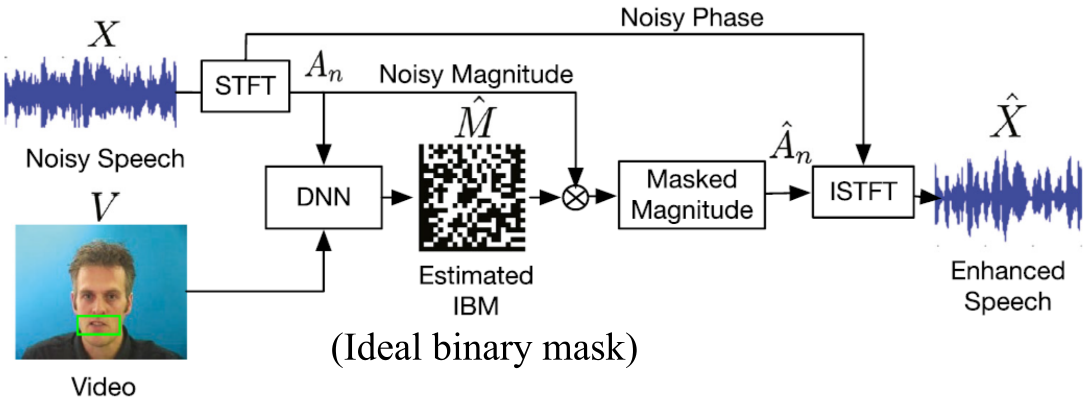
Speech Enhancement

- Z. Fu & J. Chen et al., “Congruent Audiovisual Speech Enhances Cortical Envelope Tracking during Auditory Selective Attention”, Interspeech, 2020.
- S. Chuang, et al., “Lite Audio-Visual Speech Enhancement”, Interspeech, 2020.
- M. Gogate, et al., “Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System”, Interspeech, 2020

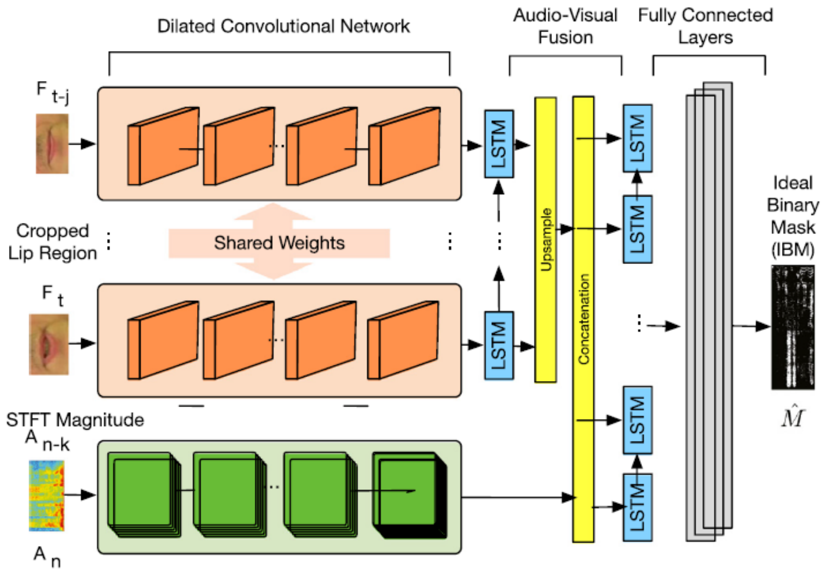
Speech Separation

- J. Yu, et al., “Audio-visual Multi-channel Recognition of Overlapped Speech”, Interspeech, 2020.
- S. Chung, et al., “FaceFilter: Audio-visual speech separation using still images”, Interspeech, 2020.
- C. Li & Y. Qian, “Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation”, Interspeech, 2020
- L. Qu, “Multimodal Target Speech Separation with Voice and Face References”, Interspeech, 2020

Speech enhancement



(Ideal binary mask)



M. Gogate, et al., "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement", Information Fusion, 2020

Speech enhancement

- New database

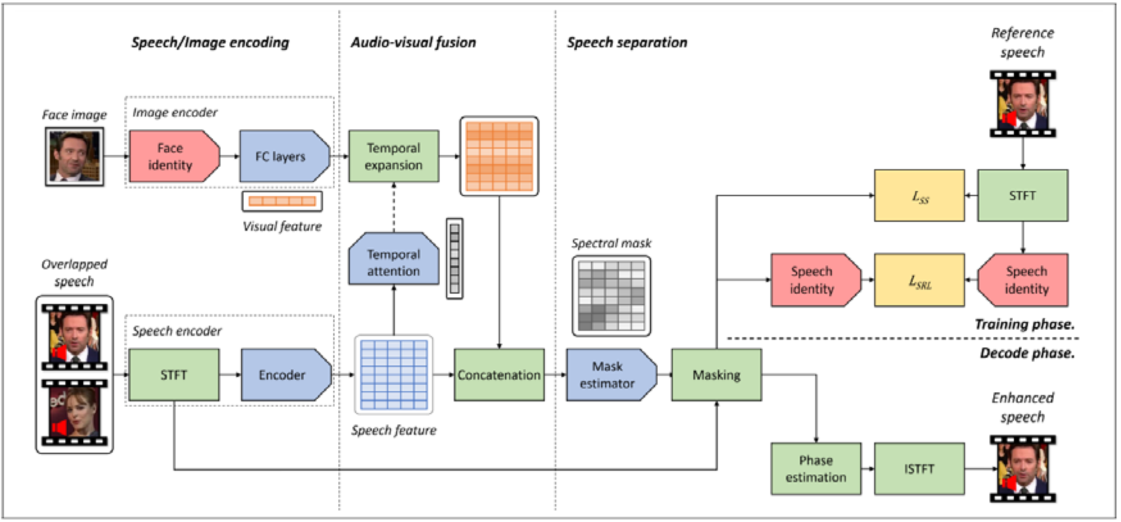
Figure 1: *Sample Frames from the VISION Corpus*

Dataset	Modality	Speakers	Real Noisy Environment	Noise types
COSINE [17]	A-only	133	Yes	Cafeteria, Streets
VOICES [18]	A-only	300	No	Television, Speech
GRID [12]	AV	34	-	No noise
Mandarin Sentences [5]	AV	1	-	No noise
AVSPEECH [13]	AV	-	-	No noise
BANCA [14]	AV	208	Yes	Speech noise only
AVICAR [15]	AV	100	Yes	Car noise only
ASPIRE [10]	AV	3	Yes	Cafeteria, Restaurant, Speech
VISION	**AV**	**209**	**Yes**	**Social gathering, Street, Cafeteria, Speech**

Table 1: *Comparison of VISION with state-of-the-art A-only and AV Corpora*

M. Gogate, et al., "Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System," Interspeech 2020

Speech separation



S. Chung, et al., "FaceFilter: Audio-visual speech separation using still images," Interspeech 2020

Speech recognition

Speech Recognition

- AVSR
- VSR

AVSR

- S. Liu, et al., “Exploiting Cross-Domain Visual Feature Generation for Disordered Speech Recognition”, Interspeech 2020
- M. Wand & J. Schmidhuber., “Fusion Architectures for Word-based Audiovisual Speech Recognition”, Interspeech 2020
- H. Liu, et al., “Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion”, Interspeech 2020
- G. Sterpu, et al., “Should we hard-code the recurrence concept or learn it instead ? Exploring the Transformer architecture for Audio-Visual Speech Recognition”. Interspeech 2020

Audio-visual speech recognition (AVSR) is to exploit complementary visual information to improve the accuracy of ASR systems.

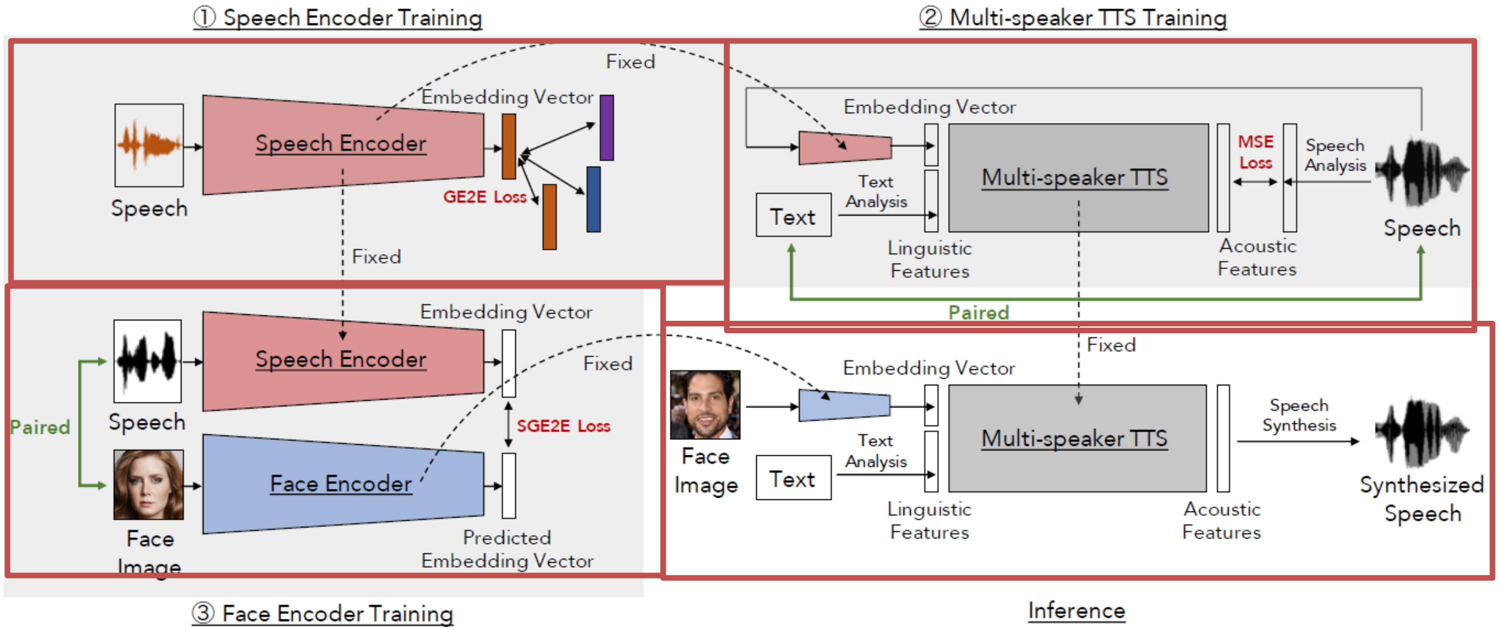
VSR

- A. Koumparoulis, et al., “Resource-adaptive Deep Learning for Visual Speech Recognition”, Interspeech, 2020.
- T. Afouras, et al., “Now you’re speaking my language: Visual language identification” , Interspeech, 2020.

Visual speech recognition (VSR): lipreading

Others

- Face2Speech
  - Multi-speaker TTS: using face image to represent speaker identity



[Samples]

S. Goto, et al., "Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image," Interspeech 2020

Related papers in Interspeech 2020

❑ Cross-modal Translate

- J. van der Hout, et al., "Evaluating Automatically Generated Phoneme Captions for Images," Interspeech, 2020.
- K. Papadimitriou & G. Potamianos, "Multimodal Sign Language Recognition via Temporal Deformable Convolutional Sequence Learning", Interspeech 2020
- X. Wang, et al., "S2IGAN: Speech-to-Image Generation via Adversarial Learning", Interspeech 2020
- W. Li, et al., "TMT: A Transformer-based Modal Translator for Improving Multimodal Sequence Representations in Audio Visual Scene-aware Dialog", Interspeech 2020
- D. Michelsanti, et al., "Vocoder-Based Speech Synthesis from Silent Videos", Interspeech 2020

❑ Multimodal Matching

- V. Krishnamohan, et al., "Audiovisual Correspondence Learning in Humans And Machines," Interspeech, 2020.
- M. Zhang, et al., "Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition", Interspeech 2020
- M. Mortazavi., "Speech-Image Semantic Alignment Does Not Depend on Any Prior Classification Tasks", Interspeech 2020
- L. Nortje & H. Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images", Interspeech 2020
- Y. Ohishi, et al., "Pair Expansion for Learning Multilingual Semantic Embeddings using Disjoint Visually-grounded Speech Audio Datasets", Interspeech 2020
- L. Wang, et al., "A DNN-HMM-DNN Hybrid Model for Discovering Word-like Units from Spoken Captions and Image Regions," Interspeech, 2020.

Speaker Verification

- R. Tao, et al., “Audio-visual Speaker Recognition with a Cross-modal Discriminative Network,” Interspeech, 2020.
- G. Antipov, et al., “Automatic Quality Assessment for Audio-Visual Verification Systems. The LOVe submission to NIST SRE Challenge 2019”, Interspeech 2020
- S. Shon & J. Glass, “Multimodal Association for Speaker Verification”, Interspeech 2020
- Z. Chen, “Multi-modality Matters: A Performance Leap on VoxCeleb”, Interspeech 2020

Emotion Recognition

- A. Khare, et al., “Multi-modal embeddings using multi-task learning for emotion recognition”, Interspeech, 2020.
- Z. Pan, et al., “Multi-modal Attention for Speech Emotion Recognition”, Interspeech, 2020.
- J. Zhang, et al., “Multimodal Deception Detection using Automatically Extracted Acoustic, Visual, and Lexical Features”, Interspeech, 2020

Speech Separation

- J. Yu, et al., “Audio-visual Multi-channel Recognition of Overlapped Speech”, Interspeech, 2020.
- S. Chung, et al., “FaceFilter: Audio-visual speech separation using still images”, Interspeech, 2020.
- C. Li & Y. Qian, “Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation”, Interspeech, 2020
- L. Qu, “Multimodal Target Speech Separation with Voice and Face References”, Interspeech, 2020

Speech Enhancement

- Z. Fu & J. Chen et al., “Congruent Audiovisual Speech Enhances Cortical Envelope Tracking during Auditory Selective Attention”, Interspeech, 2020.
- S. Chuang, et al., “Lite Audio-Visual Speech Enhancement”, Interspeech, 2020.
- M. Gogate, et al., “Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-based Baseline System”, Interspeech, 2020

AVSR

- S. Liu, et al., "Exploiting Cross-Domain Visual Feature Generation for Disordered Speech Recognition", Interspeech 2020
- M. Wand & J. Schmidhuber., "Fusion Architectures for Word-based Audiovisual Speech Recognition", Interspeech 2020
- H. Liu, et al., "Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion", Interspeech 2020
- G. Sterpu, et al., "Should we hard-code the recurrence concept or learn it instead ? Exploring the Transformer architecture for Audio-Visual Speech Recognition". Interspeech 2020

VSR

- A. Koumparoulis, et al., "Resource-adaptive Deep Learning for Visual Speech Recognition", Interspeech, 2020.
- T. Afouras, et al., "Now you're speaking my language: Visual language identification" , Interspeech, 2020.

Others

- T. Purohit, et al., "An investigation of the virtual lip trajectories during the production of bilabial stops and nasal at different speaking rates", Interspeech, 2020.
- S. Lin & Xinyuan Qian, "Audio-Visual Multi-Speaker Tracking Based On the GLMB Framework" , Interspeech, 2020.
- J. Effendi, et al., "Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework" , Interspeech, 2020.
- V. Konda, et al., "Caption Alignment for Low Resource Audio-Visual Data", Interspeech 2020
- S. Goto, et al., "Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image", Interspeech 2020
- W. Chiu, et al., "Investigating the Visual Lombard Effect with Gabor Based Features", Interspeech 2020
- K. Teplansky., "Tongue and Lip Motion Patterns in Alaryngeal Speech", Interspeech 2020
- I. Dourous., "Using Silence MR Image to Synthesise Dynamic MRI Vocal Tract Data of CV", Interspeech 2020
- J. Deadman & Jon. Barker., "Simulating realistically-spatialised simultaneous speech using video-driven speaker detection and the CHiME-5 dataset", Interspeech 2020

Thank you!