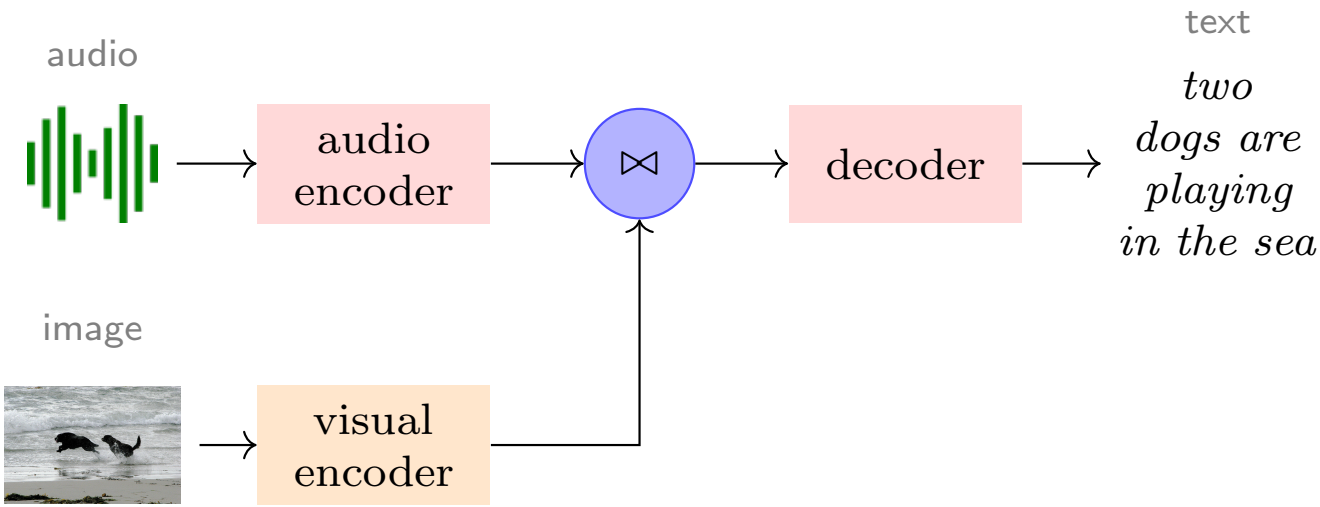
Improving multimodal speech recognition by data augmentation and speech representations

Dan Oneață Horia Cucu

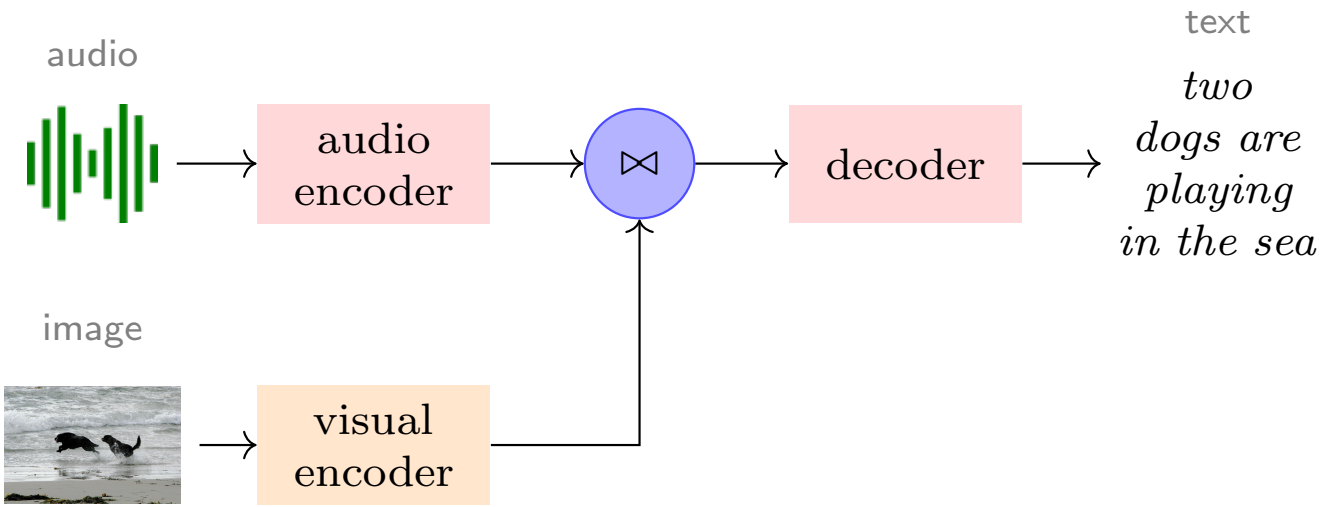
Politehnica University of Bucharest, Romania

Overview



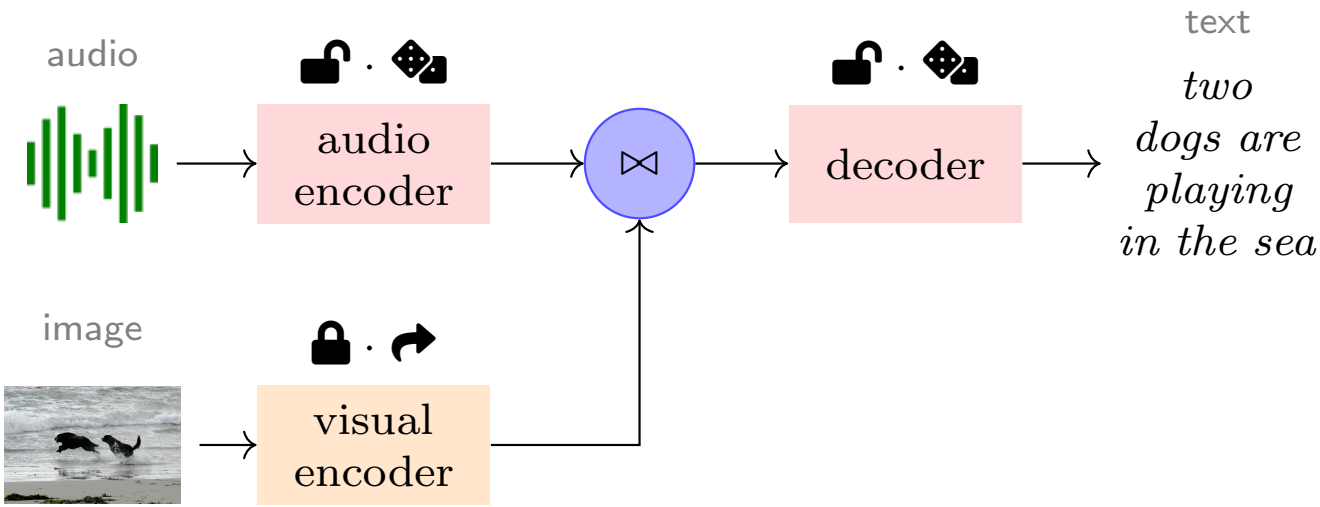
- Goal: Multimodal speech recognition
  - Transcribe spoken utterances
  - Leverage contextual visual information

Overview



- Goal: Multimodal speech recognition
  - Transcribe spoken utterances
  - Leverage contextual visual information
- Main contribution: Improve the speech component
  - How much does it account to the final performance?
  - Does the visual component improve over the speech-only system?

Our approach: Transferring audio representations

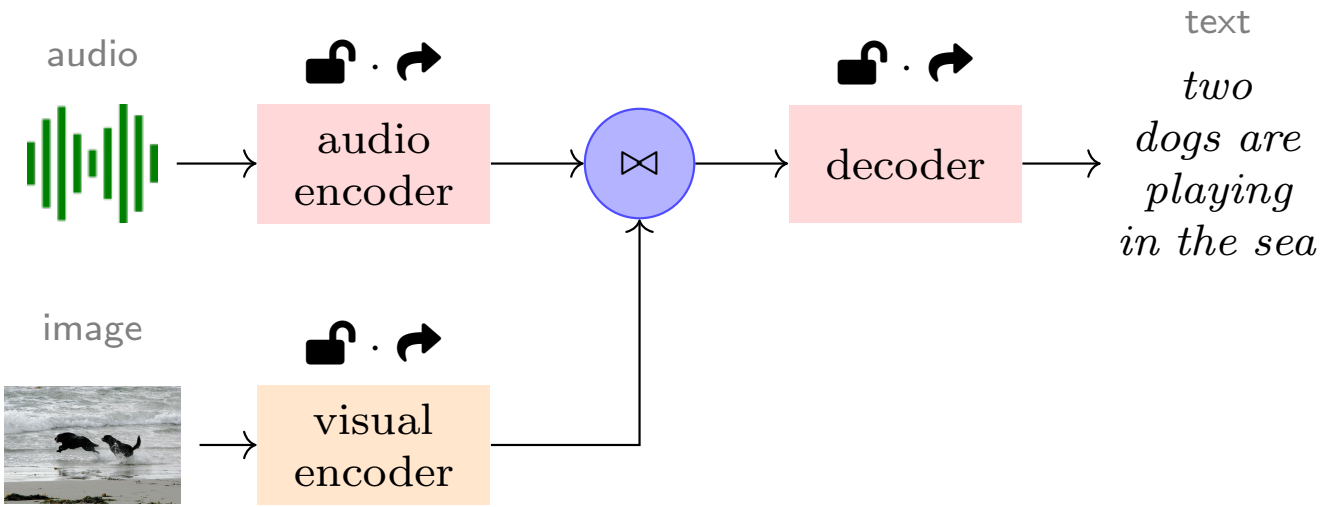


training: 🔒 – frozen; 🔓 – finetune

weights: ↱ – transferred; 🏷 – random

- Transfer visual representations; trained on
  - objects (Palaskar et al., 2018; Srinivasan et al., 2020b)
  - scenes (Miao and Metze, 2016; Gupta et al., 2017; Srinivasan et al., 2020a)
  - actions (Caglayan et al., 2019; Paraskevopoulos et al., 2020)
  - faces (Miao and Metze, 2016; Moriya and Jones, 2019)

Our approach: Transferring audio representations

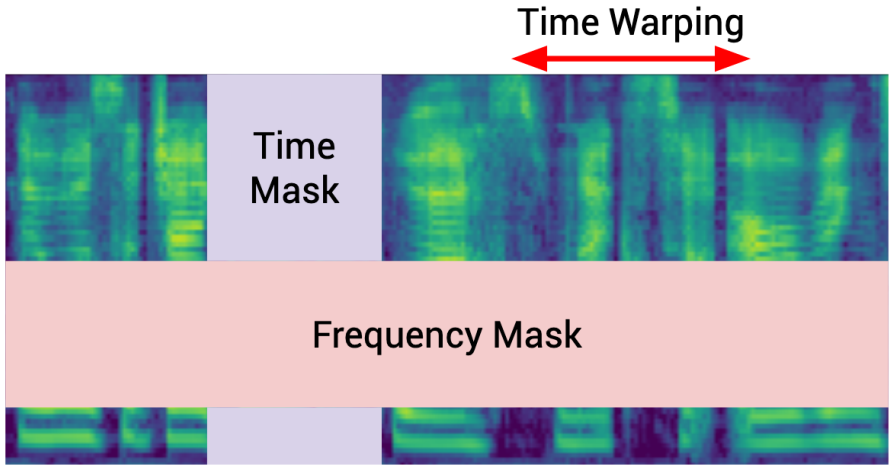


training: 🔒 – frozen; 🔓 – finetune

weights: ↱ – transferred; 🎲 – random

- Transfer visual representations; trained on
  - objects (Palaskar et al., 2018; Srinivasan et al., 2020b)
  - scenes (Miao and Metze, 2016; Gupta et al., 2017; Srinivasan et al., 2020a)
  - actions (Caglayan et al., 2019; Paraskevopoulos et al., 2020)
  - faces (Miao and Metze, 2016; Moriya and Jones, 2019)
- Transfer state-of-the-art audio representations
  - Transformer speech recognition system (Karita et al., 2019)
  - Pretrained on LibriSpeech (Panayotov et al., 2015)

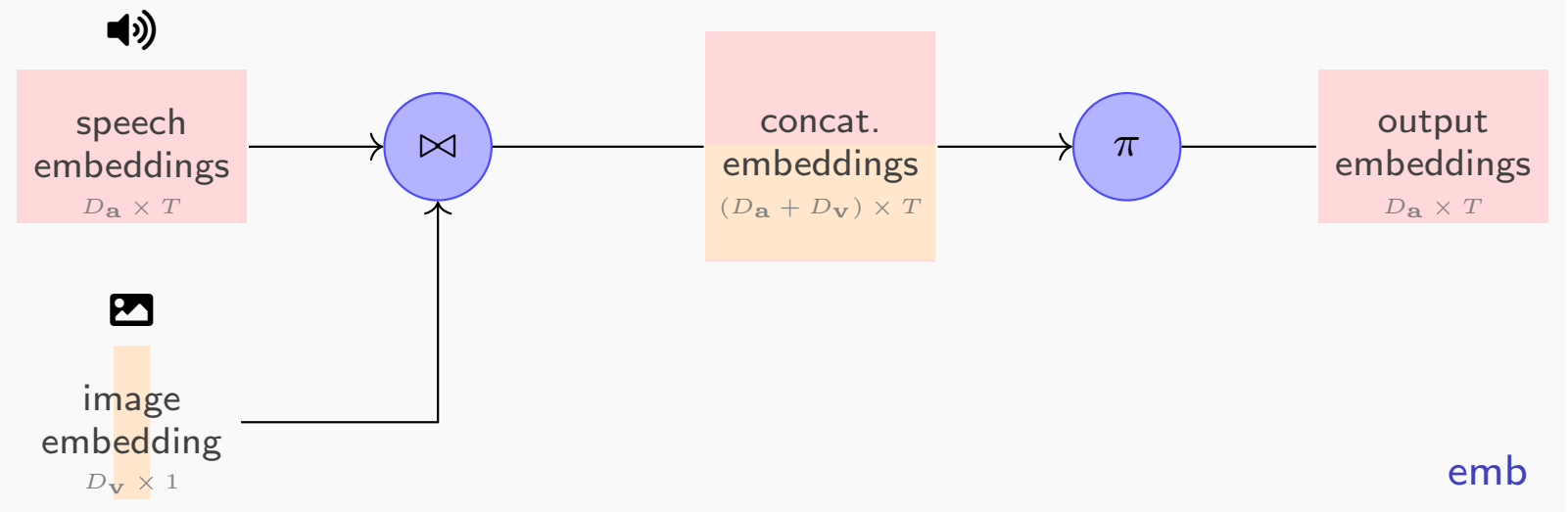
Our approach: Data augmentation



- SpecAugment: Perturbations of the spectrograms (Park et al., 2019)
  - time warping
  - time masking
  - frequency masking
- Encourage the multimodal system to latch onto the visual stream

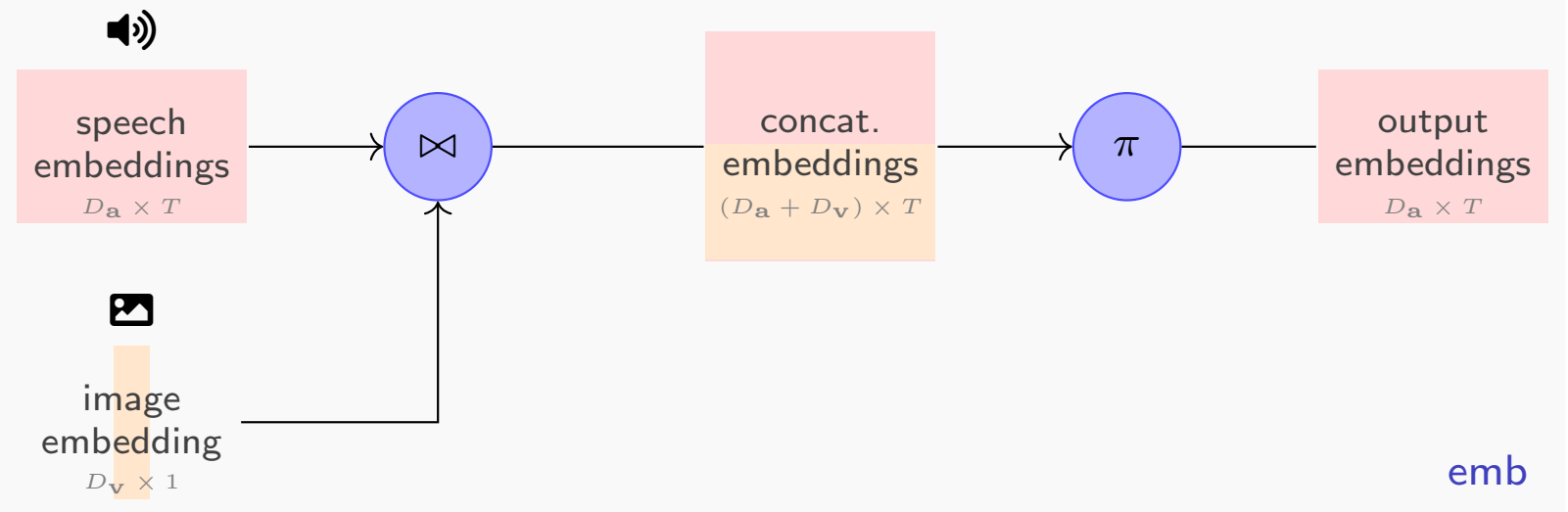
Our approach: Fusion mechanisms

- emb: concatenate along embedding dimension

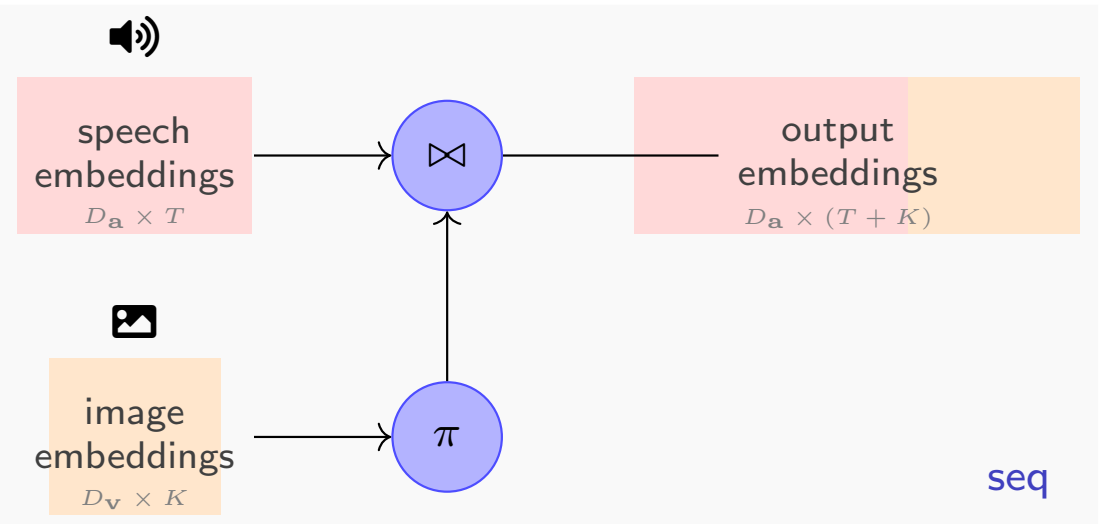


Our approach: Fusion mechanisms

- emb: concatenate along embedding dimension



- seq: concatenate along sequence dimension



Main results

- Report results on three multimodal datasets
- Metric: Word error rate ↓

method	visual	fuse	aug.	Flickr8K	How2	Loc. Nar.
(Srinivasan et al., 2020b)				13.6	—	—
	✔			14.1	—	—
(Ghorbani et al., 2021)				—	17.7	—
	✔			—	17.2	—
pretrain				11.1	26.9	49.3
finetune				**3.8**	**11.8**	**4.3**

- Strong speech-only baseline

Main results

- Report results on three multimodal datasets
- Metric: Word error rate ↓

method	visual	fuse	aug.	Flickr8K	How2	Loc. Nar.
(Srinivasan et al., 2020b)				13.6	—	—
	✔			14.1	—	—
(Ghorbani et al., 2021)				—	17.7	—
	✔			—	17.2	—
pretrain				11.1	26.9	49.3
finetune				**3.8**	11.8	4.3
finetune	✔	emb	✔	4.3	11.1	**3.9**
finetune	✔	seq	✔	4.7	**10.8**	4.0

- Strong speech-only baseline
- Adding visual signal helps for two of the datasets

Main results

- Report results on three multimodal datasets
- Metric: Word error rate ↓

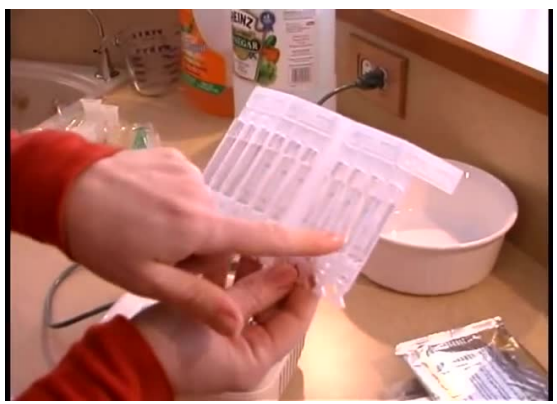
method	visual	fuse	aug.	Flickr8K	How2	Loc. Nar.
(Srinivasan et al., 2020b)				13.6	—	—
	✔			14.1	—	—
(Ghorbani et al., 2021)				—	17.7	—
	✔			—	17.2	—
pretrain				11.1	26.9	49.3
finetune				**3.8**	11.8	4.3
finetune	✔	emb		4.8	11.8	4.1
finetune	✔	emb	✔	4.3	11.1	**3.9**
finetune	✔	seq		4.0	11.8	4.2
finetune	✔	seq	✔	4.7	**10.8**	4.0

- Strong speech-only baseline
- Adding visual signal helps for two of the datasets
- Speech augmentation is key for multimodal learning

Qualitative samples: Successful cases

r: mix it up really good because that egg white is thick it's really thick
u: mix it up really good because that *eight* white is thick it's really thick
m: mix it up really good because that egg white is thick it's really thick

---



r: so each vial here is actually one use
u: so each *vile* here is actually one use
m: so each vial here is actually one use

r – reference; u – unimodal model; m – multimodal model

Qualitative samples: Failure cases

r	here's an example of a nicely bound script
u	here's an example of a nicely bound script
m	here's an example of a nicely *balance* script

r	five more keep breathing deep expanding
u	five more keep breathing deep expanding
m	five more *key breathe* than deep expanding

r – reference; u – unimodal model; m – multimodal model

Discussion and conclusions

- Speech component has a great impact on the final performance
- The visual component yields improvements

- Performance limited by the type of errors the ASR makes
- How does the multimodal system use the visual channel?
  (Srinivasan et al., 2019; Wu et al., 2021)

References I

Caglayan, O., Sanabria, R., Palaskar, S., Barraul, L., and Metze, F. (2019). Multimodal grounding for sequence-to-sequence speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8648–8652.

Ghorbani, S., Gaur, Y., Shi, Y., and Li, J. (2021). Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations. In *IEEE Spoken Language Technology Workshop*, pages 621–628.

Gupta, A., Miao, Y., Neves, L., and Metze, F. (2017). Visual features for context-aware speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5020–5024.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., and Zhang, W. (2019). A comparative study on transformer vs RNN in speech applications. In *Workshop on Automatic Speech Recognition and Understanding*, pages 449–456.

Miao, Y. and Metze, F. (2016). Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, pages 3414–3418.

Moriya, Y. and Jones, G. J. (2019). Multimodal speaker adaptation of acoustic model and language model for ASR using speaker face embedding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8643–8647.

References II

Palaskar, S., Sanabria, R., and Metze, F. (2018). End-to-end multimodal speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5774–5778.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Paraskevopoulos, G., Parthasarathy, S., Khare, A., and Sundaram, S. (2020). Multiresolution and multimodal speech recognition with transformers. In *Association for Computational Linguistics*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617.

Srinivasan, T., Sanabria, R., and Metze, F. (2019). Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In *The How2 Challenge: New Tasks for Vision & Language, ICML*.

Srinivasan, T., Sanabria, R., and Metze, F. (2020a). Looking enhances listening: Recovering missing speech using images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6304–6308.

Srinivasan, T., Sanabria, R., Metze, F., and Elliott, D. (2020b). Fine-grained grounding for multimodal speech recognition. In *Empirical Methods in Natural Language Processing*.

References III

Wu, Z., Kong, L., Bi, W., Li, X., and Kao, B. (2021). Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Association for Computational Linguistics*.