Multimodal Speech Recognition with Unstructured Audio Masking

Tejas Srinivasan
Language Technologies Institute
Carnegie Mellon University
tsriniva@andrew.cmu.edu

Ramon Sanabria
CSTR, ILCC
University of Edinburgh
r.sanabria@ed.ac.uk

Florian Metze
Language Technologies Institute
Carnegie Mellon University
fmetze@andrew.cmu.edu

Desmond Elliott
Department of Computer Science
University of Copenhagen
de@di.ku.dk

Abstract

Visual context has been shown to be useful for automatic speech recognition (ASR) systems when the speech signal is noisy or corrupted. Previous work, however, has only demonstrated the utility of visual context in an unrealistic setting, where a fixed set of words are systematically masked in the audio. In this paper, we simulate a more realistic masking scenario during model training, called RandWordMask, where the masking can occur for any word segment. Our experiments on the Flickr 8K Audio Captions Corpus show that multimodal ASR can generalize to recover different types of masked words in this unstructured masking setting. Moreover, our analysis shows that our models are capable of attending to the visual signal when the audio signal is corrupted. These results show that multimodal ASR systems can leverage the visual signal in more generalized noisy scenarios.

1 Introduction

Jointly modelling linguistic and visual signals is beneficial for several language processing tasks, such as machine translation (Sulubacak et al., 2019), visual question-answering (VQA) (Antol et al., 2015), summarization (Palaskar et al., 2019) and automatic speech recognition (ASR) (Palaskar et al., 2018; Sanabria et al., 2018). However, it is unclear exactly how the visual signals are useful for these tasks. For example, in VQA, it has been observed that models can ignore the visual context and instead rely on linguistic biases in the dataset (Ramakrishnan et al., 2018; Grand and Belinkov, 2019); in machine translation, it has been shown that some models are not affected by incorrect visual signals (Elliott, 2018); and in multimodal ASR, the visual signals were shown to act as a regularizer instead of useful disambiguating context (Caglayan et al., 2019). Given these uncertainties, there is a need to clarify the circumstances in which visual signals are useful.

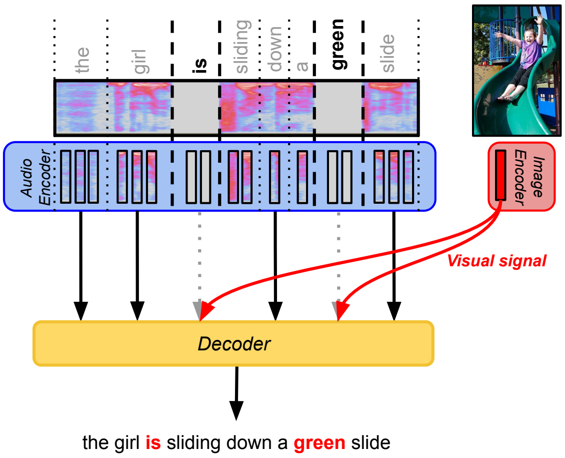


Figure 1: We propose to train multimodal speech recognition models while randomly masking different types of words in the speech signal. The model learns to use the visual signal to correctly predict the masked words.

Previous work in multimodal machine translation (Caglayan et al., 2019) and ASR (Srinivasan et al., 2020) shows that the visual signal is useful when the linguistic signal is degraded by dropping the input. In this setting, multimodal models leverage the visual signals to recover the missing language information. The results in (Srinivasan et al., 2020) are a promising start towards *verifiably useful* multimodality for robust speech recognition. However, the experiments were conducted with structured noise that focused on a predetermined set of groundable entities (*i.e.*, nouns and places). In real world scenarios, however, noise occurs in a more unstructured manner. Therefore, it is important that multimodal models can use the visual signal in a wider variety of situations.

In this work, we study multimodal ASR in more realistic noisy scenarios. We follow the methodology from (Srinivasan et al., 2020) but we randomly

Proceedings of the First International Workshop on Natural Language Processing Beyond Text, pages 11–18
Online, November 20, 2020. ©2020 Association for Computational Linguistics
http://www.aclweb.org/anthology/W23-20%2d

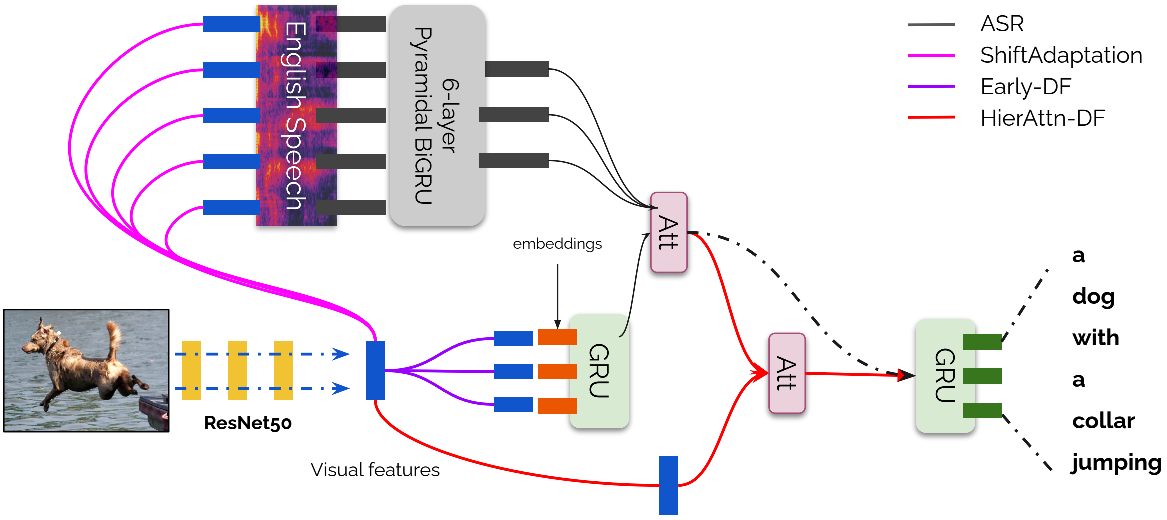


Figure 2: Our unimodal ASR model, along with several of our fusion methods for integrating a visual context vector (in blue) into the ASR model. The two fusion methods not displayed above, Weighted-DF and Middle-DF, were constructed similar to Early-DF and HierAttn-DF respectively

mask words in an unstructured manner in the audio signal (we refer to this as RandWordMask). This is in contrast to the structured masking in (Srinivasan et al., 2020), where the masked audio corresponds to only entities (which we refer to as EntityMask). The example in Figure 1 shows that RandWordMask can mask any words in the audio signal, whereas EntityMask would only mask entities like “girl” and “slide”. We apply masking both during training and testing.

The main contributions of this work are:

- We simulate a more realistic masking scenario, called RandWordMask[1], during training and testing of our ASR models (Section 2).
- We propose several multimodal models (Section 2.2), and show that training with RandWordMask improves their ability to recover masked words (Section 4).
- We show that our multimodal ASR models are right for the right reasons through several quantitative analyses (Section 4.1, 4.2, 4.4).

The results show that visual signals improve speech recognition in this more difficult, unstructured setting where random words are masked. Our models are not only able to recover masked entities, but they also recover words from other syntactic categories, *e.g.*, adjectives, cardinals, and verbs. Furthermore, our analysis shows that our models when trained using RandWordMask attend to the visual signal when the audio signal is unavailable. This confirms that the visual context can be leveraged when the primary audio signal is masked.

2 Methodology

In this section, we describe the different ASR models and our technique for simulating unstructured audio masking.

2.1 Unimodal ASR Model

Our unimodal ASR model is a word-level (Palaskar and Metze, 2018) sequence-to-sequence model with attention (Bahdanau et al., 2016; Chan et al., 2016), identical to the model used in (Srinivasan et al., 2020). The encoder (**E**) consists of 6 bidirectional LSTM layers (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) with temporal sub-sampling (Chan et al., 2016) in the middle two layers. The decoder is a two-layer conditional gated-recurrent-unit (Cho et al., 2014) which computes attention over the encoder states **E**.

$$\mathbf{h}_\mathbf{t}^\mathbf{dec1} = \mathrm{GRU}_1(\mathbf{y}_{\mathbf{t-1}}, \mathbf{h}_{\mathbf{t-1}}^\mathbf{dec1}) \tag{1}$$

$$\mathbf{z}_\mathbf{t} = \mathrm{Attention}(\mathbf{E}, \mathbf{h}_\mathbf{t}^\mathbf{dec1}) \tag{2}$$

$$\mathbf{h}_\mathbf{t}^\mathbf{dec2} = \mathrm{GRU}_2(\mathbf{z}_\mathbf{t}, \mathbf{h}_{\mathbf{t-1}}^\mathbf{dec2}) \tag{3}$$

[1] We note that RandWordMask is different from robust ASR (Barker et al., 2018) scenarios, where the whole signal is corrupted with stationary noise.

2.2 Multimodal ASR Models

We explore several fusion methods to integrate a visual feature vector $\mathbf{v}$ into the unimodal ASR model.

Encoder Feature Fusion: We use a visual adaptation method similar to (Caglayan et al., 2019), which we call **Shift Adaptation**. The visual feature vector $\mathbf{v}$ is projected down to the speech feature dimension; the resulting "shift vector" $\mathbf{s}$ is then added to the input speech features at all timesteps.

$$\mathbf{s} = \mathbf{W_v f} + \mathbf{b} \tag{4}$$

$$\mathbf{x_t} = \mathbf{x_t} + \mathbf{s} \qquad \forall \mathbf{t} \in \{1, ..., T\} \tag{5}$$

Decoder Feature Fusion: Instead of integrating the visual features into the encoder, we can integrate them in the decoder. We hypothesize that this will bias the ASR's language modelling capacity. Anastasopoulos *et al.*(Anastasopoulos et al., 2019) explore several strategies for incorporating visual features into an LSTM language model. We employ similar fusion methods in our decoder.

1. **Early Decoder Fusion (Early-DF):** At each timestep, we concatenate $\mathbf{v}$ to the input embedding $\mathbf{y_t}$, which is then projected down to the embedding dimension.

$$\mathbf{y_t} = \mathbf{W_{proj}}[\mathbf{y_t}; \mathbf{v}] \tag{6}$$

2. **Weighted Early Decoder Fusion (Weighted-DF):** We calculate a timestep-dependent weighted scalar between the input embedding $\mathbf{y_t}$ and the embedded visual features $\mathbf{v}$ (Eqn. 7), which scales the contribution of the visual features in the concatenated input (Eqn. 8):

$$\lambda = \sigma(\mathbf{y_t} \cdot \mathbf{v}) \tag{7}$$

$$\mathbf{y_t} = \mathbf{W_{proj}}[\mathbf{y_t}; \lambda\mathbf{v}] \tag{8}$$

3. **Middle Decoder Fusion (Middle-DF):** In this approach, fusion occurs between the GRU layers at $\mathbf{z_t}$ (Eqn. 2), which is the input to the 2nd decoder layer:

$$\mathbf{z_t} = \mathbf{W_{proj}}[\mathbf{z_t}; \mathbf{v}] \tag{9}$$

4. **Hierarchical Attention over Features (HierAttn-DF):** In this approach, we add a hierarchical attention layer (Libovický and Helcl, 2017) that attends between the encoder context vector $\mathbf{z_t}$ (Eqn. 2) and the visual feature vector $\mathbf{v}$. The hierarchical context vector $\mathbf{z_t^{hier}}$ is the input to the second decoder layer (Eqn. 3):

$$\mathbf{z_t^{hier}} = \text{Attention}(\{\mathbf{z_t}, \mathbf{v}\}, \mathbf{h_t^{dec1}}) \tag{10}$$

By conditioning the hierarchical attention on the output of the first decoder layer, the attention layer learns to decide which of the audio and visual modalities is more important for decoding at a given timestep.

2.3 Unstructured Masked Audio: RandWordMask

We simulate a degradation of the audio signal by randomly masking words in the audio with silence. This approach differs from (Srinivasan et al., 2020), where they masked a fixed set of words corresponding to entities, i.e., nouns and places. Figure 1 shows an example of an audio spectrogram with **RandWordMask**. The intuition behind random word masking, as opposed to entity-based word masking, is that noise in the audio signals is unlikely to systematically occur when someone is speaking about an entity. Our multimodal ASR models need to be responsive to audio that drops outside systematically expected regions.

In real-world settings, the rate at which the speech is masked (unavailable) is highly variable. Therefore, we train the models with an augmented version of the dataset: for each audio utterance, we create four masked audio samples, where words are masked with 0%, 20%, 40% and 60% probability. Note that the text transcript ($\mathbf{y_{1...N}}$) and image modality ($\mathbf{v}$) remain intact. This approach to augmenting the dataset will result in models that can adapt to different amounts of corruption in the audio signal during evaluation.

3 Experimental Setup

3.1 Dataset

We perform experiments on the Flickr 8K Audio Caption Corpus (Harwath and Glass, 2015), which contains 40,000 spoken captions (total 65 hours of speech) corresponding to 8,000 natural images from the Flickr8K dataset (Hodosh et al., 2015). The augmented dataset that we use for training and testing (as described in Section 2.3) consists of 160,000 spoken captions.

In addition, we use the SpeechCOCO dataset (Havard et al., 2017) for pretraining. SpeechCOCO contains over 600 hours of *synthesised* speech paired with images.

3.2 Implementation Details

3.2.1 Audio Features

We extract 43-dimensional filter bank features in an identical manner to (Srinivasan et al., 2020). In order to mask the audio, we first extract word-audio alignments from a pre-trained GMM-HMM model and expand the start and end timing marks by 25% of the segment duration to account for misalignments. We mask words in the audio by replacing word segments with 0.5 seconds silence.

3.2.2 Visual Features

We extract visual features from a ResNet-50 CNN (He et al., 2016) pre-trained on ImageNet. Specifically, we extract features from the 2048-dim average pooling layer, and project these to 256-dim through a learned linear layer: $\mathbf{v} = \mathbf{W} \cdot \text{CNN}(\mathbf{img})$

3.2.3 Model Implementation

We use the same model hyperparameters as in (Srinivasan et al., 2020). Models are trained using the *nmtpytorch* framework (Caglayan et al., 2017). We first pre-train our models for 25,000 minibatches on the SpeechCOCO dataset. This pre-training step, inspired by (Ilharco et al., 2019), was crucial to ensure stable training of our models on the Flickr 8K dataset.

3.3 Evaluation Metrics

Our model evaluation (Table 1a) has been conducted on the development set of Flickr8k-Audio, while the rest of our analysis is conducted on the test set. We report **WER** for all our models. For datasets where words have been masked in the audio signal, we compute **Recovery Rate** (Srinivasan et al., 2020), which measures the percentage of masked words which have been correctly recovered in the transcription.

In addition, we can determine the contribution of the visual signal when decoding each word in the HierAttn-DF model. We do this by inspecting the weights of the audio and visual modalities in the hierarchical attention mechanism. We introduce a new metric to quantify this: **Grounding Rate (G.R.)**.

$$\text{G.R.} = \frac{\#\text{recovered words where visual attn} > 0.5}{\#\text{correctly recovered masked words}}$$

We choose 0.5 as the threshold since above this value, more attention was given to the visual modality than the audio. G.R. thus represents the percentage of recovered words where the model was focusing more on the visual context while decoding.

4 Results and Analysis

In Table 1a, we summarize the performance of our unimodal ASR and proposed multimodal ASR models. Our development set is constructed similarly to our training set described in Section 2.3, consisting of samples with 0%, 20%, 40% and 60% of words masked. We examine performance on this Augmented dataset, as well as datasets at each individual masking level.

We see that the Decoder-Fusion (DF) multimodal models outperform unimodal ASR on both WER and RR. However, the best-performing models on both metrics differ: Weighted-DF achieves the lowest WER, with an improvement of 1.40% on the augmented dataset. HierAttn-DF has the best Recovery Rate, with an absolute improvement of 4% over the Unimodal model. These trends hold across all masking levels. Moreover, we observe that as the amount of masking in the audio signal increases, the WER and RR gains of our models increase. The ShiftAdapt model, which integrates the visual features with the speech encoder input, does not show any improvements over unimodal ASR. We observe that ShiftAdapt shows improvements when trained and tested on clean data, which aligns with the regularization signal previously observed in (Caglayan et al., 2019).

The results in Table 1a show that multimodality can recover words which were masked in an unstructured manner. We now turn our attention to analysing which types of words are recovered better. We conduct this analysis across seven categories: five syntactic (nouns, verbs, adjectives, adverbs and cardinals) and two semantic (places and colors).[2] For each category, we create a new test set where we mask all word occurrences. We note that these categories are varying degrees of "groundable", which we define as how easily identifiable they are in the visual modality - the more groundable a category, the easier it is to identify words belonging to that category in the visual context. Nouns and places are the most groundable categories, while adjectives and colors are also frequently easy to identify in the image. Verbs and adverbs, however, are less groundable categories.

In Table 1b, we compare the Recovery Rate of

[2] Words for the syntactic categories were found by POS tagging the dataset and keeping the top 100 frequent words.

	↑ Recovery Rate (%)				↓ Word Error Rate (%)				
Masking Perc.	Aug.	20%	40	60%	Aug.	0% (Clean)	20%	40%	60%
Unimodal	29.3	36.5	30.9	24.7	34.0	13.7	26.3	40.7	57.1
ShiftAdapt	29.3	36.5	31.3	25.1	34.0	13.5	25.9	40.5	57.1
Early-DF	32.0	38.2	33.2	28.7	33.3	13.7	25.9	39.7	55.3
Weighted-DF	33.0	38.8	34.5	29.6	**32.6**	**13.4**	**25.5**	**38.9**	**53.9**
Middle-DF	32.4	37.9	34.1	29.7	34.1	14.6	26.9	40.3	55.3
HierAttn-DF	**33.5**	**40.3**	**35.2**	**30.1**	33.2	13.9	25.9	39.3	54.7

(a) Recovery Rate (RR) and Word Error Rate (WER) of the ASR models on the FACC development set.

Metric	Model	Nouns	Places	Adj.	Colors	Verbs	Adverbs	Cardinals
RR (%)	Unimodal	37.2	28.0	26.0	26.6	26.0	30.4	56.7
	HierAttn-DF	47.9	40.0	29.7	30.4	27.9	29.2	58.1
Rel. Δ RR (%)	-	28.8	42.8	14.2	14.3	7.3	-3.9	2.4
G.R. (%)	HierAttn-DF	92.7	92.5	76.8	75.5	67.6	33.5	82.3

(b) Comparison of Recovery Rates of unimodal and HierAttn-DF ASR on various syntactic and semantic word categories.

Table 1: Recovery Rate, Word Error Rate, and Grounding Rates for the proposed models on the FACC dataset.

the unimodal ASR and HierAttn-DF (the best multimodal model in terms of RR) on the different word types. We observe that on the groundable entities *i.e.*, nouns and places, there is a relative improvement of at least 25% compared to the Unimodal model. Adjectives and colors, which are also groundable in the visual modality, are recovered around 14% better than the Unimodal model. The relative RR improvement for verbs is around 7%, whereas adverbs recovery is 4% worse. These results show that visual context can recover words from a variety of categories, even though it is better at recovering entities, and struggles with words that are less groundable in the image.

4.1 Hierarchical Attention Analysis

In Table 1b, we also summarize the Grounding Rate of HierAttn-DF when recovering different types of words. We find that the most groundable words (nouns and places), have a Grounding Rate $> 90\%$. This means that 90% of the time the nouns/places were correctly recovered, the visual modality was being attended to. Adjectives and verbs, which are also groundable, have a grounding rate of $\approx 76\%$. These trends confirm that the model's improvements in masked word recovery are coming from using the visual signal.

In addition to calculating the Grounding Rate, we also check whether the model learns to "look" at the visual modality when it encounters a masked word. In Figure 3, we plot the average visual attention weight at the masked word timestep, as well as the two preceding and proceeding timesteps. We see that the more groundable the word category, the more attention it learns to pay to the visual modality when the word is masked.

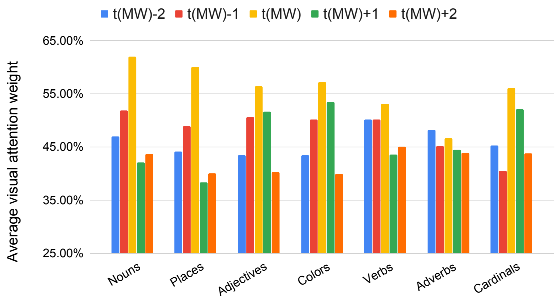


Figure 3: Average visual attention weight preceding and proceeding the onset of the masked word at timestep t(MW).

In Table 2, we present some qualitative examples where we visualize how the attention to each modality evolves with time. We observe that the timesteps corresponding to masked words in the signal have significantly higher visual attention. We see that in the first example, all masked words are correctly recovered. In the second example,

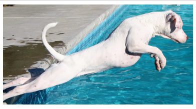
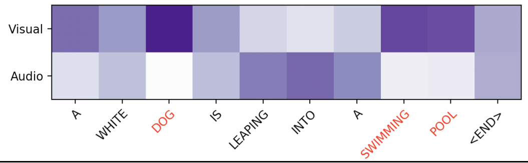
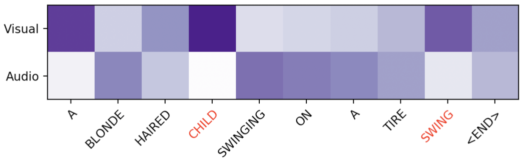
Image	Reference Caption	Hierarchical Attention Map w/ Hypothesis Decoding
	A white dog is leaping into a swimming pool	
	A blonde haired toddler swinging on a tire swing	

Table 2: Examples of the HierAttn-DF model attending to the visual modality to recover masked words

however, the model replaces the word *toddler* with *child*, which are semantically similar and visually identical, showing that the model knows what to recover in the image but does not always recover it in the correct form.

4.2 Utility of RandWordMask Training

We compare our RandWordMask training scheme with the EntityMask training mechanism from (Srinivasan et al., 2020). In EntityMasking, only entities (nouns) are masked during training, and we hypothesize that this makes the model better at recovering entities but unable to generalize to other word types. Since RandWordMask training involves masking words at random, we expect the model should be able to generalize better to other words types. In Table 3, we compare the performance of the HierAttn-DF model when trained with three different training mechanisms: (i) None: no words are masked during training, (ii) EntityMask: top 100 frequent nouns are masked, (iii) RandWordMask. As expected, when trained without masking in the training set (None), the model recovers almost none of the masked words. While EntityMasking shows strong performance on recovering nouns and places (which are closely related), it doesn't generalize to the other syntactic/semantic word types. RandWordMask results in slightly worse performance on noun recovery, but it generalizes much better to other word categories.

Masked Word	None	EntityMask	RandWordMask
Nouns	4.3	59.1	47.9
Places	2.4	43.1	40.0
Adjectives	0.7	4.7	29.7
Colors	1.3	3.4	30.3
Verbs	0.7	11.9	27.9
Adverbs	1.1	4.6	29.2
Cardinals	3.5	4.3	58.1

Table 3: RR (%) of different training schemes

Silence Masking		
Masking %	Unimodal	HierAttn-DF
20%	36.6	40.6
40%	31.1	36.0
60%	25.7	31.3
Whitenoise Masking		
Masking %	Unimodal	HierAttn-DF
20%	33.1	37.4
40%	26.7	32.1
60%	21.5	28.1

Table 4: RR (%) of unimodal and HierAttn-DF ASR models when trained and tested on silence and white noise masked audio, at different masking levels

4.3 Silence vs Whitenoise Masking

Our results in Tables 1a and 1b are performed in the experimental setting where words are masked with silence. However, another masking strategy explored in (Srinivasan et al., 2020) is white noise masking, where the masked word is replaced with white noise in the audio signal. (Srinivasan et al., 2020) had reported results in both masking scenarios, and noted that the improvements of the multimodal ASR model were similar in both scenarios. We further verify this by training unimodal and HierAttn-DF ASR models using RandWordMask, but with white noise masking instead of silence.

In Table 4, we report the Recovery Rates of both ASR models in both silence and white noise

Masking %	Congruent	Incongruent
20%	40.6	29.3
40%	36.0	24.7
60%	31.3	20.2

Table 5: Recovery Rates (%) for the HierAttn-DF model when provided with correct (congruent) and misaligned (incongruent) image

masking scenarios. We observe that while recovery is generally harder with white noise masking (evidenced by lower RR of both unimodal and multimodal ASR models), the HierAttn-DF model shows approximately the same absolute improvements in RR over the unimodal ASR. This indicates that the multimodal model can be applied to the more difficult white noise masking as well.

4.4 Congruency Analysis

We perform a sanity check of our model by misaligning audio utterances and images while decoding the trained model (Elliott, 2018). This evaluation quantifies the sensitivity of the model towards the visual modality. A model that is sensitive to the visual context would perform significantly worse when presented with an unrelated (*incongruent*) image during evaluation. Since the model has been trained to actively use the image, it is likely to extract incorrect information. In Table 5, we see that the HierAttn-DF model is substantially affected by the unrelated images (the recovery rate drops on average by 7%). This verifies that our multimodal models are sensitive to the image modality.

5 Conclusions

We show that visual signals improve multimodal speech recognition when the audio signal is subject to unstructured masking. RandWordMask simulates a wider range of noisy scenarios by masking different types of words in the audio signal during training and evaluation, as opposed to previous work that only masked groundable entities (Srinivasan et al., 2020). Future work involves developing new models that attend over visual features extracted from object proposals, which provide better visual signals.

Acknowledgments

This work used the computational resources of the PSC Bridges cluster at Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014)

References

Antonios Anastasopoulos, Shankar Kumar, and Hank Liao. 2019. Neural language modeling with visual features. *arXiv preprint arXiv:1903.02930*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *Interspeech*.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *Prague Bulletion of Math. Linguistics*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context multimodal machine translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loïc Barrault, and Florian Metze. 2019. Multimodal Grounding for Sequence-to-Sequence Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *Automatic Speech Recognition and Understanding (ASRU)*.

William Havard, Laurent Besacier, and Olivier Rosec. 2017. Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set. In *International Workshop on Grounding Language Understanding (GLU)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2015. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). In *International Joint Conference on Artificial Intelligence IJCAI*.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Computational Natural Language Learning (CoNLL)*.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Association for Computational Linguistics (ACL)*.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Association for Computational Linguistics (ACL)*.

Shruti Palaskar and Florian Metze. 2018. Acoustic-to-word recognition with sequence-to-sequence models. In *Spoken Language Technology Workshop (SLT)*.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. 2018. End-to-end multimodal speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Workshop on Visually Grounded Interaction and Language (ViGIL), NeurIPS*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2020. Looking enhances listening: Recovering missing speech using images. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2019. Multimodal machine translation through visuals and speech. *arXiv preprint arXiv:1911.12798*.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science and Engineering*, 16(05):62–74.