Visual Keyword Spotting with Attention

K R Prajwal*
prajwal@robots.ox.ac.uk
Liliane Momeni*
liliane@robots.ox.ac.uk
Triantafyllos Afouras
afourast@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

In this paper, we consider the task of spotting spoken keywords in silent video sequences – also known as visual keyword spotting. To this end, we investigate Transformer-based models that ingest two streams, a visual encoding of the video and a phonetic encoding of the keyword, and output the temporal location of the keyword if present. Our contributions are as follows: (1) We propose a novel architecture, the *Transpotter*, that uses full cross-modal attention between the visual and phonetic streams; (2) We show through extensive evaluations that our model outperforms the prior state-of-the-art visual keyword spotting and lip reading methods on the challenging LRW, LRS2, LRS3 datasets by a large margin; (3) We demonstrate the ability of our model to spot words under the extreme conditions of isolated mouthings in sign language videos.

1 Introduction

In recent years, there has been significant progress in automatic visual speech recognition (VSR) due to the availability of large-scale annotated datasets and the development of powerful neural network-based learners [2, 7, 17]. These methods are continually improving and becoming more sophisticated, by incorporating better visual models, stronger language modelling and training on larger datasets. Indeed the best industrial grade lip reading models today are far superior to humans, and achieve error rates approaching Automatic Speech Recognition (ASR) performance [33, 39].

However, for many applications it is not necessary to transcribe every word that is spoken in a silent video (the task of VSR), rather only specific utterances or keywords need to be recognised. This is for example the case in "wake word" recognition, where only particular keywords need to be spotted over long input sequences. A further drawback of VSR methods is that they are heavily reliant on language modelling; in general, their performance decreases significantly when context is limited (e.g. short utterances) or parts of the input are occluded, e.g. from the speaker's hands or a microphone. In this work, we focus instead on the task of

* Equal Contribution.

Visual Keyword Spotting (KWS), where the the goal is to detect and localise a *given* keyword in (silent) spoken videos.

Automatic visual KWS enables a diverse range of practical applications: indexing archival silent videos by keyword to enable content-based search; helping virtual assistants (e.g. Alexa and Siri) and smart home technologies respond to wake words and phrases; assisting people with speech impairment (e.g. amyotrophic lateral sclerosis patients) or aphonia in communication [54]; and detecting mouthings in sign language videos [5].

KWS differs in complexity from VSR primarily because in KWS we are armed with the keyword we need to recognise, whereas VSR has the harder task of recognising every word from scratch. The core hypothesis motivating this work is that *this additional knowledge renders visual KWS an easier task* than VSR; and it is therefore expected that KWS should achieve a higher performance than VSR, and generally be more robust to challenging and adversarial situations. Nevertheless, visual KWS remains a *very difficult* task and shares similar challenges to VSR methods: first, some words sound different but involve identical lip movements ('man', 'pan', 'ban'), these *homopheme* words cannot be distinguished using only visual information. Second, speech variations such as accents, speed, and mumbling can alter lip movements significantly for the same word. Third, co-articulation of the lips between preceding and subsequent words in continuous speech also affects lip appearance and motion.

In this paper, we make the following three contributions: (i) We propose a novel Transformer-based architecture, the *Transpotter* (a portmanteau of *Trans*former and *Spotter*), that is tailored to the visual KWS task. The model takes as input two streams, one encoding visual information from a video and the other providing a phonetic encoding of the keyword; the heterogeneous inputs are then fused using full cross-modal attention. (ii) Through extensive evaluations, we show that our Transpotter model outperforms the prior state-of-the-art visual KWS and VSR methods on the challenging LRW, LRS2 and LRS3 lip reading datasets by a large margin. (iii) We test our best model under extreme conditions: finding words in mouthings of people communicating using sign language. Signers sometimes mouth words as they sign as an additional non-manual signal to disambiguate and help understanding [59]. This new task is extremely challenging as there is a significant domain shift between full spoken sentences (in our training and test sets) and mouthings, where the context is sporadic and phonemes of the keyword may be missing – as sometimes only parts of words are mouthed [11]. Our approach outperforms previous KWS models in this challenging, practical use-case. Video examples are available at the project's webpage: www.robots.ox.ac.uk/~vgg/research/transpotter.

2 Related work

Our work relates to prior work on KWS, lip reading, visual grounding, and applications of Transformers for text and video. We present a brief discussion of these topics below.

KWS. KWS in audio (speech) is a well studied problem with a long history, spanning several decades. Prior to the establishment of deep learning models, KWS methods were based on Hidden Markov Models [48, 63], dynamic time warping [31, 51, 73] or indexing of ASR lattices [13]. A number of works have since used deep architectures suitable for sequence modelling (e.g. RNNs, CNNs, or graph convolutional networks) [6, 15, 22, 30, 35, 36, 43, 50, 58, 62, 74], including encoder-decoder approaches [8, 49, 71, 76]. Berg *et al.* [9] recently proposed using a Transformer model for the same task. Different from ours, this work uses a single input stream (audio) and only learns to spot a fixed vocabulary of keywords. In contrast, we use Transformers to temporally process, then fuse the multi-modal inputs, building a model that can eventually perform open-set KWS. Visual KWS has also received attention recently. The proposed methods include query-by-example [32] approaches, sliding window classification [67], or looking up phonetic queries in lip reading feature sequences [41, 57], while audio-visual methods [19, 41, 65] that fuse the two modalities to improve robustness to noise have also been proposed. Our method builds upon these approaches: we address various weaknesses and propose superior video-text modelling as well as explicit keyword localization, resulting in significantly improved performance.

Lip reading. Early works in lip reading usually relied on hand-crafted pipelines and features [25, 44, 46, 75]. The availability of large scale lip reading datasets [1, 17] and the development of deep neural network models resulted in major performance improvements, initially in word-level lip reading [16, 56] and constrained sentences [7]. Sentence level models were subsequently developed, using sequence-to-sequence architectures based on RNNs [17], CTC-based [54] approaches, or a hybrid of the two [45]. Replacing RNNs with Transformers resulted in better performing architectures [2, 26, 72]. Joint audio-visual training and cross-modal distillation [3, 37, 68] have also been investigated. The current state-of-the-art model uses Transformers in the visual front-end and achieves remarkable results with word error rates reaching as low as 30.7% [53].

Visual grounding. Our work is also related to tasks such as natural language grounding in videos [14, 23, 24, 28, 38, 66, 69, 70] and subtitle alignment in sign language clips [12].

Transformers. Since their introduction for machine translation, Transformers [61] have become ubiquitous and are used today in a wide range of applications from natural language processing [18, 47] and speech recognition [20, 34, 40] to visual representation learning [10, 21, 53, 64]. In this work, we rely on Transformers as our building blocks for their strong sequence modelling capability and inherent potential for localisation through attention.

3 Visual KWS with Attention

In this section, we describe our proposed method shown in Figure 1. We outline the architecture of our model (Section 3.1), our training procedure (Section 3.2) and differences to prior work (Section 3.3). We refer the reader to the arXiv version of the paper for further details.

3.1 The Transpotter Architecture

Our model ingests two input streams: (i) a textual keyword $q = (q_1, q_2 \cdots, q_{n_p})$, and (ii) a silent video clip $v \in \mathbb{R}^{T\times H\times W\times 3}$ in which we need to spot the keyword. For each of the inputs, we have separate encoders that learn initial modality-specific representations. This is followed by a joint multi-modal Transformer that learns cross-modal relationships between the video and text features. The joint transformer predicts two outputs: (i) a sequence-level probability of the keyword occurring in the video and (ii) frame-level probabilities indicating the location of the keyword in the video if present. We describe each of the modules next.

Text Representations. Our textual input is a phonetic representation of the keyword, obtained using a pronunciation dictionary. The input phoneme sequence of length n_p is mapped to a sequence of learnable embedding vectors $Q \in \mathbb{R}^{n_p\times d}$. Sinusoidal positional encodings are added to the input phoneme feature vectors, and the result is passed through a Transformer Encoder [61] with N_t layers to capture temporal information across the phoneme sequence:

$$Q_{enc} = encoder_q(Q + PE_{1:n_p}) \in \mathbb{R}^{n_p\times d}.$$

Video Representations. We use a pre-trained visual front-end (either a CNN [2] or VTP [53]) to extract a feature vector for each input video frame, $V \in \mathbb{R}^{T\times d}$. Similar to the text encoder

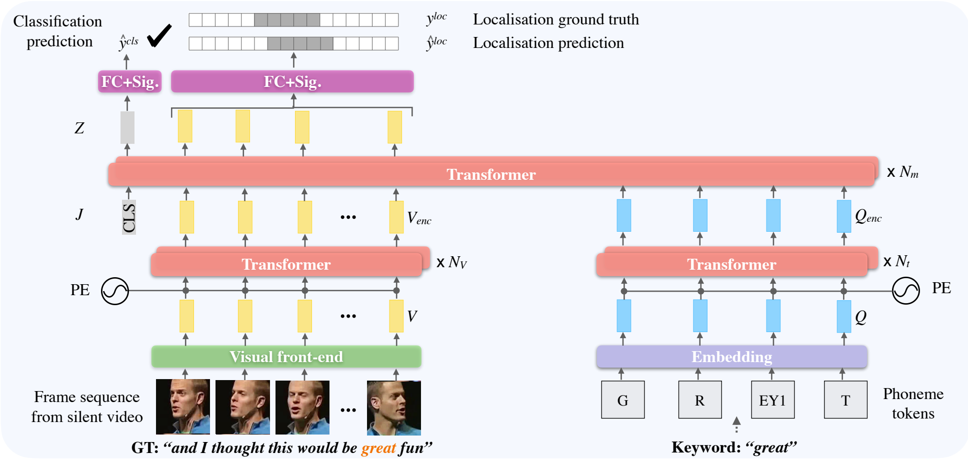

Figure 1: **The Transpotter architecture:** Video frames are inputted to a visual front-end (CNN [2] or VTP [33]) to extract low-level visual features, which are then passed to N_v Transformer layers to encode temporal information. The keyword in the form of a phoneme sequence is encoded using N_t Transformer layers. The text and visual features are finally concatenated in time and processed using a joint multi-modal Transformer which predicts: (i) the probability the keyword occurs in the video, (ii) frame-level probabilities indicating the location of the word. PE corresponds to positional encoding.

Q_{enc}, we pass the visual features through a Transformer Encoder with N_v layers to capture temporal information, after adding positional encodings:

$$V_{enc} = encoder_v(V + PE_{1:T}) \in \mathbb{R}^{T \times d}.$$

Joint Video-Text Representations. The uni-modal representations V and Q are concatenated along the time dimension to produce a single sequence of feature vectors. A learnable $[CLS]$ token embedding (such as in BERT [18] and ViT [21]) is then prepended to the result:

$$J = ([CLS]; V_{enc}; Q_{enc}) \in \mathbb{R}^{(1+T+n_p) \times d}.$$

We use a Transformer encoder with N_m layers to jointly learn the relationships across video and phoneme vectors:

$$Z = encoder_{vq}(J + PE_{1:(1+T+n_p)}) \in \mathbb{R}^{(1+T) \times d}.[1]$$

Prediction heads. The $[CLS]$ output feature vector Z_1 serves as a joint aggregate representation for the video-text pair. An MLP head for binary classification, f_c is attached to Z_1 to predict the probability of the keyword being present in the video:

$$\hat{y}^{cls} = \sigma(f_c(Z_1)) \in \mathbb{R}^1,$$

where σ denotes a sigmoid activation. To localise the keyword, we attach a second MLP head f_l that is shared across all the video output states from the multi-modal joint Transformer:

$$\hat{y}^{loc} = \sigma(f_l(Z_{2:(T+1)})) \in \mathbb{R}^T.$$

The output y_t^{loc} at each video frame time-step $t \in T$ indicates the probability of the frame t being a part of the keyword utterance.

[1] the n_p outputs corresponding to the phonetic embeddings are dropped.

3.2 Training

Optimisation objectives. Given a training dataset $\mathcal{D}$ consisting of tuples (v, q, y^{cls}, y^{loc}) of silent video clips, text queries, class labels and location labels (indicating the position of the keyword within the clip), we define the following objectives:

$$\mathcal{L}^{cls} = -\mathbb{E}_{(v,q,y^{cls})\in\mathcal{D}}\, BCE(y^{cls}, \hat{y}^{cls}) \tag{1}$$

$$\mathcal{L}^{loc} = -\mathbb{E}_{(v,q,y^{cls},y^{loc})\in\mathcal{D}}\, y^{cls}\left[\frac{1}{T}\sum_{t=1}^{T} BCE(y_t^{loc}, \hat{y}_t^{loc})\right] \tag{2}$$

$$BCE(y, \hat{y}) = y \log \hat{y} + (1-y)\log(1-\hat{y}), \tag{3}$$

where BCE stands for the binary cross-entropy loss. The labels y^{cls} are set to 1 when the given keyword occurs in the video and 0 otherwise; the frame labels y^{loc} are set to 1 for the frames where the keyword is uttered and 0 otherwise. We train the model the optimise the total loss $\mathcal{L} = \lambda\mathcal{L}^{cls} + (1-\lambda)\mathcal{L}^{loc}$, where λ is a balancing hyper-parameter.

3.3 Discussion

Compared to prior approaches, the design of our model offers several important advantages.
Stronger Visual Representations. Previous works [41, 57] model temporal relationships between video frames using RNNs. In contrast, we employ Transformers [61], which are far more effective in modeling temporal relationships [4, 29].
Joint Video-text Modeling. Prior works such as KWS-Net [41] follow a late-fusion strategy. In our model each frame-wise video feature can attend to any keyword token (phoneme) and vice-versa. The information exchange across the modalities occurs at every layer, without restrictions on the receptive field for either modality.
Stronger keyword localisation. Fine-grained localisation of the keyword in the video can be important for applications such as sign spotting [5]. Existing methods [41, 57] “weakly” localise the keyword by taking the sequence-level prediction to be the maximum probability over all the video time-steps. We instead provide stronger frame-level supervision, by enforcing the model to predict the exact temporal extent of the keyword in the video.

3.4 Implementation details

Pre-training the visual front-end. We explore two different visual front-end architectures for the Transpotter: (1) a CNN, highly similar in architecture to TM-seq2seq [2] and (2) VTP [33], the current state-of-the-art for lip reading (trained only on public data). Both models are trained end-to-end on two-word video clips of LRS2 [17] and LRS3 [1] for lip reading. We refer the reader to the arXiv version of the paper for the exact CNN architecture and training hyper-parameters. We refer the reader to [33] for architectural hyper-parameters and training protocols for VTP. We pre-compute the visual features for each backbone for both datasets and then train directly on them for faster training. All our models and ablations use the pre-trained CNN features, unless otherwise stated.
Sampling. We form the training dataset $\mathcal{D}$ by randomly sampling with 50% probability a positive or negative video clip v for each query q. Each video v contains word boundary annotations, which allows (i) performing data augmentation by randomly cropping video segments during training, and (ii) creating frame labels y^{loc}, as described in 3.2.
Misc. The keyword q is mapped to a phoneme sequence using the CMU dictionary [55]; words not present in the dictionary are discarded from training $\mathcal{D}$. We set $\lambda = 0.5$.

4 Experiments

This section is structured as follows: We first present the datasets used as well the evaluation protocols that we follow in our experiments (Section 4.1). Next, we compare the performance of our proposed Transpotter model against strong baselines (Section 4.2) and then present a comprehensive study ablating our design choices (Section 4.3). Finally, we perform further performance analysis and provide qualitative results (Section 4.4).

4.1 Datasets and Evaluation Protocol

Datasets. All models and baselines are trained and evaluated on LRS2 [17] and LRS3 [1] lip reading datasets. LRS2 contains BBC broadcast footage from British television and LRS3 is based on TED/TEDx videos downloaded from YouTube (refer to the arXiv version of the paper for detailed statistics). The video clips for both datasets are tightly cropped face-tracks of active speakers only. For each clip, a full transcription of the utterance as well as word boundary alignments are provided. The number of videos, number of keyword instances and keyword vocabulary for each of the test sets is shown in Table 1.

Evaluation Protocol. Evaluation is performed for every test dataset as follows: First, the vocabulary of test keywords is determined, by considering all the words occurring in the test set transcriptions with above a certain phoneme length n_p. If not specified, we use $n_p \geq 3$. Every word in the query vocabulary is then searched for in all the test set videos.

Metrics. Given ground truth video-keyword samples, we assess the performance of our model in two ways. First, we assess classification performance, *i.e.* whether the model can accurately predict whether the keyword occurs in the video or not. We compute accuracy (Acc$^{Cls}_{@k}$) and mean average precision (mAP^{Cls}) metrics, where Acc$^{Cls}_{@k}$ measures how often a given keyword occurs in any of the top-k retrievals, and mAP^{Cls} is obtained with the above criterion (where every word in the test keyword vocabulary is considered as a separate class).

Second, we assess the model's localisation capability, *i.e.* whether the model can accurately localise the keyword in the video clip. We follow common practice from the detection literature: we consider a keyword accurately detected when the intersection-over-union (IOU) between the prediction $\hat{y}^{loc}$ and ground truth label y^{loc} is above a certain threshold, and calculate the mean average precision mAP^{Loc}. To calculate the IOU, we binarise the model's predictions using a threshold $\tau = 0.5$.

4.2 Comparison to baselines

We compare our model's performance against a state-of-the-art VSR model and KWS-Net [41], the previous state-of-the-art visual KWS model.

VSR baseline. We use an improved version of the TM-seq2seq [2] VSR model, with the same pre-trained CNN backbone (Section 3.4) that we use for the KWS models. The model is trained with the curriculum training strategy of [2] (details in the arXiv version of the paper). The VSR model achieves state-of-the-art Word Error Rate (WER) performance of 36.9% and 48.0% on the LRS2, LRS3 test sets respectively. Since the VSR model only produces text transcriptions of a given video, but no localisation prediction, we can only evaluate its classification performance (Acc$^{Cls}_{@k}$,mAP^{Cls}). We follow the method detailed in [27] to estimate the posterior probability that the keyword occurs in a video clip.

KWS-Net. As a KWS baseline we use the state-of-the-art model of Momeni *et al.* [41]. For fair comparison, here too we use the same CNN backbone that is also used for our model.

State-of-the-art KWS. We report our model's performance and compare it with strong baselines in Table 1. It is clear that our model outperforms both baselines. On the last row,

	LRS2				LRS3			
	1.2K vids. / 4.3K inst. / 1.6K vocab.				1.3K vids. / 6.1K inst. / 1.9K vocab.			
Model	$\mathrm{Acc}^{Cls}_{@1}$	$\mathrm{Acc}^{Cls}_{@5}$	mAP^{Cls}	mAP^{Loc}	$\mathrm{Acc}^{Cls}_{@1}$	$\mathrm{Acc}^{Cls}_{@5}$	mAP^{Cls}	mAP^{Loc}
KWS-Net [41]	36.1	61.2	41.0	36.2	29.8	54.6	34.3	29.2
VSR	63.7	76.3	64.3	-	52.3	66.0	50.3	-
Transpotter	65.0	87.1	69.2	68.3	52.0	77.1	55.4	53.6
Transpotter (VTP)	**68.7**	**90.7**	**72.5**	**71.6**	**55.7**	**78.5**	**58.2**	**56.1**

Table 1: **Comparison to baselines:** We outperform the current state-of-the-art KWS and VSR methods by a large margin. Our Transpotter model is particularly effective in localising the keyword in the video. Moreover, by using the recently proposed VTP [33] architecture as the Transpotter's visual backbone instead of a CNN, we achieve even better performance.

we show the boost in performance by replacing the CNN with the recently proposed VTP backbone [33], resulting in state-of-the art performance on both the LRS2 and LRS3 datasets.
Evaluation on LRW. We also compare the performance of KWS-Net [41] with our proposed Transpotter model on the LRW [16] test set following the same evaluation protocol. The test set contains $25K$ single-word video clips spanning a vocabulary of 500 words (50 instances per word). Note that KWS-Net has been pretrained on the LRW training split, but the Transpotter has only been trained on LRS2 and LRS3. As we can see in Table 2, the Transpotter outperforms the previous state-of-the-art baseline KWS-Net by a large margin. We refer the reader to the arXiv version of the paper for a qualitative error analysis in this setting.

Model	$\mathrm{Acc}^{Cls}_{@1}$	$\mathrm{Acc}^{Cls}_{@5}$	mAP^{Cls}
KWS-Net [41]	66.6	89.0	33.0
Transpotter	**85.8**	**99.6**	**64.1**

Table 2: **Comparison on LRW [16]:** The Transpotter outperforms the previous state-of-the-art KWS model on the LRW test set, despite not having been trained on LRW data. The localization metric mAP^{Loc} is not reported as the input videos are single-word clips.

4.3 Architecture ablations

To assess our design choices for the Transformer skeleton, we perform a number of ablations considering variations of the model architecture. We briefly explain the alternative approaches below; more details can be found in the arXiv version of the paper.

In particular we consider two alternative encoder-decoder architectures, with the video input at the encoder side and the text query at the decoder (Enc_{vid}-Dec_{text}) and vice versa (Enc_{text}-Dec_{vid}). Since the latter model outputs at the temporal resolution of the video input, it can explicitly localise the keyword (in the same way as the Transpotter), while the former can only perform classification. We also consider a variant of the Transpotter, where the model does not output localisation predictions (hence no $\mathcal{L}^{loc}$ is used for its training). We show the results in Table 3. The selected Transpotter architecture outperforms all variants. In particular, by comparing rows 2 and 4, we observe that training with a localisation head and loss $\mathcal{L}^{loc}$ also improves classification (e.g. 64.0 vs 69.2 mAP^{Cls}).

4.4 Transpotter performance analysis

In this section, we analyse the performance of our proposed method when varying the keyword length and the size of the surrounding visual context.
Keyword length. In Figure 2a, we plot the model's performance on the LRS2 test set against the minimum keyword length in phonemes n_p. As expected, longer keywords are easier

	LRS2				LRS3			
Model	$\text{Acc}^{Cls}_{@1}$	$\text{Acc}^{Cls}_{@5}$	mAP^{Cls}	mAP^{Loc}	$\text{Acc}^{Cls}_{@1}$	$\text{Acc}^{Cls}_{@5}$	mAP^{Cls}	mAP^{Loc}
Enc_{vid}-Dec_{text}	52.5	80.0	57.9	-	40.3	66.9	43.2	-
Transpotter w/o loc.	59.4	84.1	64.0	-	46.5	72.1	49.8	-
Enc_{text}-Dec_{vid}	63.8	86.8	68.4	67.8	**52.1**	76.6	54.9	53.1
Transpotter	**65.0**	**87.1**	**69.2**	**68.3**	52.0	**77.1**	**55.4**	**53.6**

Table 3: **Model ablations:** Our approach of jointly modeling text and video sequences with a localisation head for stronger supervision outperforms other architectural designs.

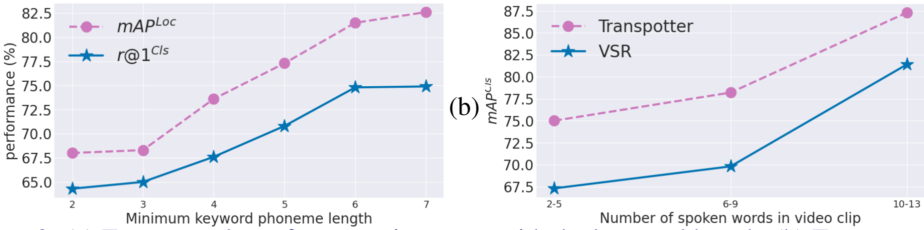

Figure 2: (a) Transpotter's performance increases with the keyword length; (b) Transpotter performs far better than VSR with limited context. Both methods improve with more context.

to spot and therefore result in better retrieval performance. Indeed for long 7-phoneme keywords, mAP^{Loc} reaches as high as 82.5. We note however that even for very challenging short keywords with only 2 phonemes (such as "my", "to", "at"), mAP^{Loc} stays high at 67.5.

Context. The visual appearance of spoken words can be highly ambiguous [2], therefore recognising isolated words from visual input alone may be very challenging. Current lip reading models utilise the surrounding visual context to resolve this ambiguity. In Figure 2b, we illustrate how the performances of our Transpotter KWS model and our VSR baseline vary based on the amount of contextual information available. We plot the mAP^{Cls} against the number of words in the video clip. We observe that both models benefit from larger surrounding context, with the Transpotter outperforming the VSR baseline consistently.

Qualitative analysis. In Figure 3, we show qualitative examples from the LRS2 and LRS3 test sets. It is clear that the model produces smooth predictions that precisely indicate the full location of the word. In the bottom right corner we observe a failure case where the model's confidence is low – the keyword "that's" in this case is short.

Model response to homophemes. We further probe our Transpotter model for failure cases. In visual-only keyword spotting, a common failure case is due to homophemes, *i.e.* words with identical lip movements. To investigate the response of our model to such cases, we construct a list of keywords from the LRS2 test set sentences that are known to have homopheme counterparts (*e.g. mark*, which has two matching homophemes, *bark* and *park*) and then for each test set clip that contains one of the keywords, we query that keyword along with its corresponding homophemes and plot the model's outputs. We illustrate several examples in Figure 4. We observe that in such cases, the model spots the keyword as well as its homophemes at the same (ground truth) location.

5 Mouthing Spotting in Sign Language videos

In this section, we investigate the application of our method for spotting mouthed words in sign language videos. This is an important application of visual KWS, as it has enabled an entire line of work on sign language recognition [6, 42, 60].

Figure 3: **Qualitative results on LRS2 and LRS3:** The Transpotter accurately localises the keyword in most examples. In the bottom right example, the model's confidence is low, most likely because it is a short word. The IOU is zero since we threshold at $\tau = 0.5$.

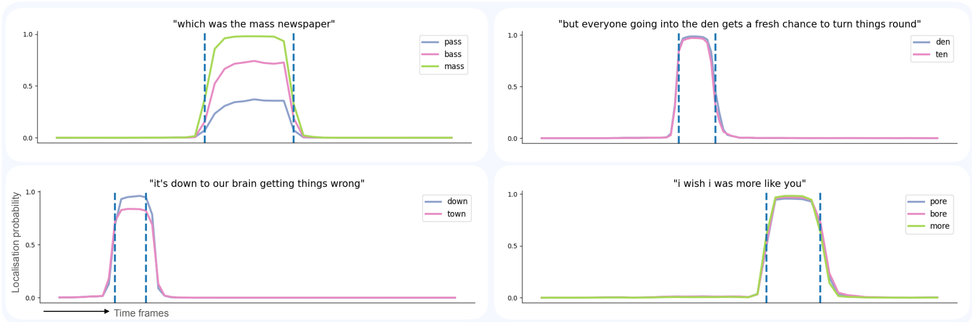

Figure 4: **Model's response to homophemes:** We query words and their corresponding homophemes for LRS2 test set clips. We observe that the model spots the words and their homophemes at the same (ground truth) location.

Data description & evaluation protocol. Here, we use a subset of BSL Corpus [52, 53] as a test set. BSL Corpus is a large public dataset containing videos of conversations conducted in sign language by deaf signers, from various regions across the UK. We extend the dataset's annotations by adding a *Mouthing* tier and asking a deaf annotator to identify and localise mouthing occurrences that correspond to visible signs. We obtain 383 mouthing instances, from 29 different signers, over a keyword vocabulary size of 187. We use a pre-processing pipeline similar to [17] to obtain face-cropped tracks around the faces of the signers. To evaluate KWS performance, we take 8-second video clips centered around the annotated mouthings and follow the same evaluation protocol described in Section 4.

Results. We summarise the evaluation results in Table 4. The Transpotter model is far superior to the prior state-of-the-art KWS baseline, achieving a great improvement in performance (e.g. 29.6 vs 15.6 mAP^{Cls} score). To complete this analysis, we also show qualitative examples of the spotted mouthings in Figure 5.

Discussion. We note that sign language mouthings are often very different from equivalent spoken words. Words may be partially mouthed and can be occluded by the signing hands. There is therefore a significant domain gap between the BSL-Corpus signing videos and our lip reading training videos. However, we note that our proposed model greatly outperforms the KWS-Net baseline – a variant of which has been successfully deployed for detecting mouthings in order to bootstrap learning of sign spotting methods [5, 42, 60]. This indicates the potential of our proposed method to greatly improve these pipelines.

Model	$\mathrm{Acc}^{Cls}_{@1}$	$\mathrm{Acc}^{Cls}_{@5}$	mAP^{Cls}
KWS-Net [41]	12.4	29.6	15.6
Transpotter	**22.5**	**47.1**	**29.6**

Table 4: **Spotting mouthings in BSL-Corpus:** The Transpotter is far more accurate than the current state-of-the-art in spotting keywords in videos.

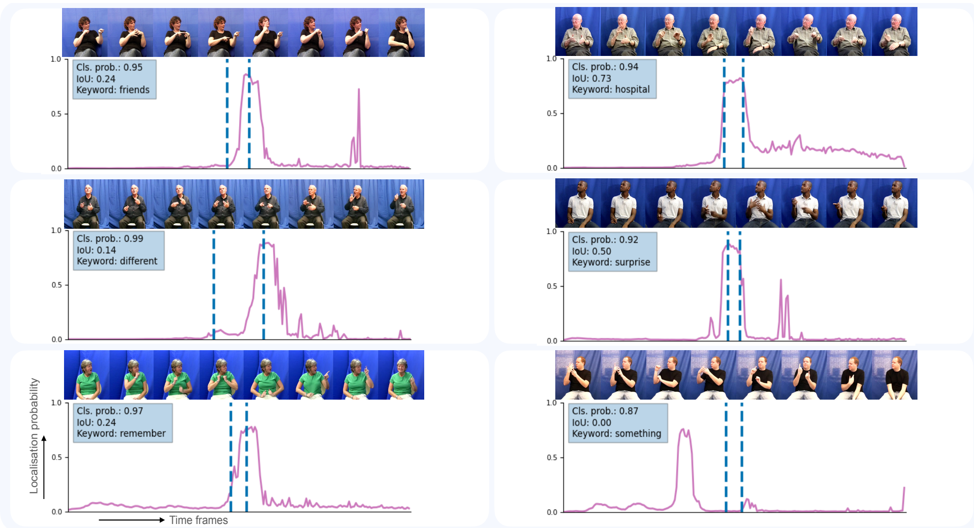

Figure 5: **Qualitative results on BSL-Corpus:** Despite the large domain shift from our training examples and additional challenges such as partially mouthed words and hand occlusions, the Transpotter succeeds in correctly spotting mouthings in these challenging conditions. We observe a failure case (bottom right) where the localisation is incorrect. We note that contrary to LRS2 and LRS3, where word boundaries are obtained through robust audio-based forced alignment, the annotations for BSL-Corpus are noisier as they are performed manually.

6 Conclusion

We have presented the *Transpotter*, a cross-modal attention based architecture for visual keyword spotting. Our method surpasses the performance of the previous best visual keyword spotting approach by a large margin, as well as that of a state-of-the-art lip reading baseline. We demonstrate the ability of our model to generalise to sign language videos where it can be used to spot mouthings, enabling automatic annotation of sign instances. In future work, we plan to further improve our method's performance by incorporating keyword semantics and context of the surrounding words.

Acknowledgements. Funding for this research is provided by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems, the Oxford-Google DeepMind Graduate Scholarship, the EPSRC Programme Grant VisualAI (EP/T028572/1) and the Royal Society Research Professorships 2019 RP/R1/191132. We thank Samuel Albanie for his invaluable help in applying our method to signer mouthings.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE PAMI*, 2019.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[4] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166, 2019.

[5] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020.

[6] Sercan Arik, Markus Kliegl, Rewon Child, Joel Hestness, Andrew Gibiansky, Chris Fougner, Ryan Prenger, and Adam Coates. Convolutional recurrent neural networks for small-footprint keyword spotting. In *INTERSPEECH*, 2017.

[7] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *arXiv:1611.01599*, 2016.

[8] Kartik Audhkhasi, Andrew Rosenberg, Abhinav Sethy, Bhuvana Ramabhadran, and Brian Kingsbury. End to-end asr-free keyword search from speech. *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[9] Axel Berg, Mark O'Connor, and Miguel Tairum Cruz. Keyword transformer: A self-attention model for keyword spotting. *arXiv preprint arXiv:2104.00769*, 2021.

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. ICML*, 2021.

[11] P Boyes Braem and RL Sutton-Spence. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press, 2001. ISBN 3927731838.

[12] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *arxiv*, 2021.

[13] Dogan Can and M. Saraçlar. Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:2338–2347, 2011.

[14] Jingyuan Chen, Xinpeng Chen, L. Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018.

[15] Xi Chen, Shouyi Yin, Dandan Song, Peng Ouyang, Leibo Liu, and Shaojun Wei. Small-footprint keyword spotting with graph convolutional network. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[16] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Proc. ACCV*, 2016.

[17] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proc. CVPR*, 2017.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[19] Runwei Ding, Cheng Pang, and Hong Liu. Audio-visual keyword spotting based on multidimensional convolutional neural network. In *IEEE International Conference on Image Processing*, 2018.

[20] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[22] Santiago Fernandez, Alex Graves, and Jurgen Schmidhuber. An application of recurrent neural networks to discriminative keyword spotting. In *International Conference on Artificial Neural Networks*, 2007.

[23] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

[24] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *NAACL-HLT*, 2019.

[25] John N Gowdy, Amarnag Subramanya, Chris Bartels, and Jeff Bilmes. Dbn based multi-stream models for audio-visual speech recognition. In *2004 IEEE International conference on acoustics, speech, and signal processing*, volume 1, pages I–993. IEEE, 2004.

[26] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, 2020.

[27] Yanzhang He, Rohit Prabhavalkar, Kanishka Rao, Wei Li, Anton Bakhtin, and Ian McGraw. Streaming small-footprint keyword spotting using sequence-to-sequence models. In *Proceedings of IEEE ASRU*, 2017.

[28] Lisa Anne Hendricks, O. Wang, E. Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[29] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[30] Kyuyeon Hwang, Minjae Lee, and Wonyong Sung. Online keyword spotting with a character-level recurrent neural network. *arXiv:1512.08903*, 2015.

[31] Fumitada Itakura. *Minimum Prediction Residual Principle Applied to Speech Recognition*, page 154–158. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1558601244.

[32] Abhishek Jha, Vinay P. Namboodiri, and C. V. Jawahar. Word spotting in silent lip videos. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[33] Prajwal K R, Triantafyllos Afouras, Andrew Zisserman, et al. Sub-word level lip reading with visual attention. *arXiv preprint arXiv:2110.07603*, 2021.

[34] Shigeki Karita, Nelson Yalta, Shinji Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *INTERSPEECH*, 2019.

[35] Taejun Kim and Juhan Nam. Temporal feedback convolutional recurrent neural networks for keyword spotting. *arXiv:1911.01803*, 2019.

[36] Chris Lengerich and Awni Hannun. An end-to-end architecture for keyword spotting and voice activity detection. In *NIPS 2016 End-to-End Learning for Speech and Audio Processing Workshop*, 2016.

[37] Wei Li, Sicheng Wang, Ming Lei, Sabato Marco Siniscalchi, and Chin-Hui Lee. Improving audio-visual speech recognition performance with cross-modal student-teacher training. In *Proc. ICASSP*, pages 6560–6564. IEEE, 2019.

[38] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[39] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2019.

[40] Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

[41] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. Seeing wake words: Audio-visual keyword spotting. In *BMVC*, 2020.

[42] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *Proc. ACCV*, 2020.

[43] Dimitri Palaz, Gabriel Synnaeve, and Ronan Collobert. Jointly learning to locate and classify words using convolutional networks. In *INTERSPEECH*, 2016.

[44] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(3):423–435, 2009.

[45] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. In *IEEE Spoken Language Technology Workshop*, pages 513–520, 2018.

[46] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 2003.

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[48] Richard C Rose and Douglas B Paul. A hidden markov model based keyword recognition system. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 129–132. IEEE, 1990.

[49] Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, and Michael Picheny. End-to-end speech recognition and keyword search on low-resource languages. In *ICASSP*, 2017.

[50] Tara N. Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *INTERSPEECH*, 2015.

[51] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, 26:43–49, 1978.

[52] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the British Sign Language Corpus. *Language Documentation & Conservation*, 7:136–154, 2013.

[53] Adam Schembri, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition), 2017. URL http://www.bslcorpusproject.org.

[54] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas. Large-Scale Visual Speech Recognition. In *INTERSPEECH*, 2019.

[55] Speech Group at Carnegie Mellon University. CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2014.

[56] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. In *INTERSPEECH*, 2017.

[57] Themos Stafylakis and Georgios Tzimiropoulos. Zero-shot keyword spotting for visual speech recognition in-the-wild. In *ECCV*, 2018.

[58] Ming Sun, Anirudh Raju, George Tucker, Sankaran Panchapagesan, Gengshen Fu, Arindam Mandal, Spyridon Matsoukas, Nikko Strom, and Shiv Vitaladevuni. Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting. *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[59] Rachel Sutton-Spence. Mouthings and simultaneity in british sign language. *Amsterdam studies in the theory and history of linguistic science, series 4*, 281:147, 2007.

[60] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[62] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard Lyon, and Rif Saurous. Trainable frontend for robust and far-field keyword spotting. In *ICASSP*, 2017.

[63] Jay Wilpon, Chin-Hui Lee, and Lawrence Rabiner. Application of hidden markov models for recognition of a limited set of words in unconstrained speech. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 254 – 257 vol.1, 06 1989. doi: 10.1109/ICASSP.1989.266413.

[64] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020.

[65] Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, and Xuewu Zhang. A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. *IEEE Transactions on Multimedia*, 2016.

[66] Huijuan Xu, Kun He, Bryan A. Plummer, L. Sigal, S. Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019.

[67] Yue Yao, Tianyu Wang, Heming Du, Liang Zheng, and Tom Gedeon. Spotting visual keywords from temporal sliding windows. In *Mandarin Audio-Visual Speech Recognition Challenge*, 2019.

[68] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP*, pages 6984–6988, 05 2020. doi: 10.1109/ICASSP40776.2020.9054127.

[69] Yitian Yuan, T. Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019.

[70] Runhao Zeng, H. Xu, W. Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020.

[71] Haitong Zhang, Junbo Zhang, and Yujun Wang. Sequence-to-sequence models for small-footprint keyword spotting. *arXiv:1811.00348*, 2018.

[72] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proc. ICCV*, pages 713–722, 2019.

[73] Yaodong Zhang and James Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 398 – 403, 01 2010. doi: 10.1109/ASRU.2009.5372931.

[74] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *CoRR*, 2017.

[75] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605, 2014.

[76] Yimeng Zhuang, Xuankai Chang, Yanmin Qian, and Kai Yu. Unrestricted Vocabulary Keyword Spotting Using LSTM-CTC. In *INTERSPEECH*, 2016.