2025 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3315-7486-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/SpeD67700.2025.11252119

Improved visually prompted keyword localisation in real low-resource settings

Leanne Nortje
Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
nortjeleanne@gmail.com

Dan Oneata
National University of Science and Technology
POLITEHNICA *Bucharest*
Bucharest, Romania
dan.oneata@gmail.com

Gabriel Pîrlogeanu
National University of Science and Technology
POLITEHNICA *Bucharest*
Bucharest, Romania
gabriel.pirlogeanu@upb.ro

Herman Kamper
Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
kamperh@sun.ac.za

***Abstract*—Given an image query, the goal in visually prompted keyword localisation (VPKL) is to find occurrences of the depicted word in a speech collection. This can be useful when transcriptions are not available for a low-resource language (e.g. if it is unwritten). Previous work showed that VPKL can be performed with a visually grounded speech model trained on paired images and unlabelled speech. But all experiments were done on English. Moreover, transcriptions were used to get positive and negative pairs for the contrastive loss. This paper introduces a few-shot learning scheme to mine pairs automatically without transcriptions. On English, this results in only a small drop in performance. We also – for the first time – consider VPKL on a real low-resource language, Yorùbá. While scores are reasonable, here we see a bigger drop in performance compared to using ground truth pairs because the mining is less accurate in Yorùbá.**

***Index Terms*—visually grounded speech models, multimodal learning, keyword localisation, speech–image retrieval**

I. Introduction

Developing applications that can search through speech data is challenging in low-resource languages where transcriptions are difficult or impossible to collect. One line of research has been looking at visually grounded speech models to address this [1]. These models learn from paired images and unlabelled spoken captions and can therefore be trained without transcriptions [2]–[7]. One way to perform speech search with such a model is to provide an image query depicting a word of interest. Formally, the task of visually prompted keyword localisation (VPKL) involves detecting whether an image query – which depicts a keyword – occurs in a spoken utterance, and if so, where it occurs [8]. An English example is shown on the left in Fig. 1.

Previous work [8] formalised the VPKL task and showed that it is possible at a reasonable level with a visually grounded speech model. However, there were two major shortcomings. First, all experiments were carried out on English datasets, treating it as an artificial low-resource language. Second – and more importantly – English transcriptions were used to obtain positive and negative pairs for the contrastive loss used in the visually grounded model. This reliance on transcriptions severely limits the applicability of the approach to a real low-resource setting. In this paper, we address these shortcomings by performing experiments on Yorùbá, a real low-resource language spoken by 44M people in Nigeria. A Yorùbá example is given in the right part of Fig. 1. We also adapt the original approach to work without using transcriptions, making it usable in the low-resource case.

Concretely, we turn to few-shot learning to mine training pairs [9]. We use a support set that contains a small number of isolated spoken examples of the keywords that we want to learn. Based on this set, we use a spoken query-by-example method to predict which keywords occur in the spoken captions of the speech–image training data. These predictions are used

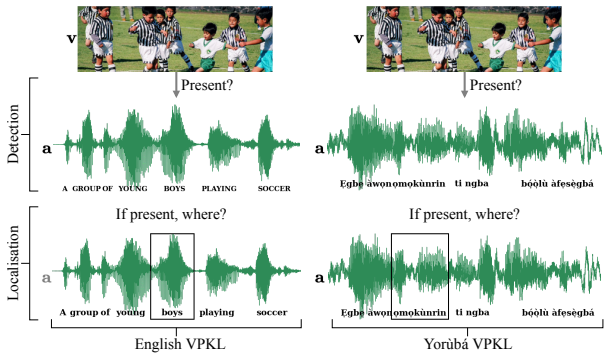

Fig. 1. The goal in visually prompted keyword localisation is to detect and locate a given query keyword (given as an image) within spoken utterances. On the right, the Yorùbá word for "boys" is "ọmọkùnrin".

This work was supported in part by a grant of the Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project number PN-IV-P2-2.1-TE-2023-1632, within PNCDI IV.

to automatically construct positive and negative examples for the contrastive loss of the visually grounded speech model. E.g. the English speech–image training data might have a caption "the boys playing soccer in the park" paired with only a single image. Using few-shot mining, we can also now pair this utterance with the utterance "a dad throws a ball at his boys" as a positive example in the loss. This encourages the model to not only focus on utterances as a whole but to learn within-utterance differences between keywords. Images that co-occur with mined utterances are similarly used to construct contrasting pairs.

We compare this few-shot mining method to an approach where a visual tagger is used to automatically annotate training images with text labels of words likely occurring in an image [8]. These generated tags can then be used to sample positive and negative image–caption pairs that contain the same or different keywords, which can again be used in the contrastive loss. On English data, we show that the few-shot mining approach consistently outperforms this visual tagger scheme in terms of VPKL localisation and detection performance. We also quantify the drop in performance compared to when transcriptions are used to construct perfect pairs: starting from 50–53% in the idealised case, detection and localisation F1 drops by roughly 11%.

We then turn to the actual low-resource language, Yorùbá, where we present VPKL results for the first time. Here we see a larger drop in performance when using few-shot mining compared to using transcriptions. This is because the query-by-example matching approach used for mining relies on a self-supervised speech model that is less tailored to Yorùbá than English. We also show that it is essential to pretrain the audio branch of the visually grounded speech model on unlabelled Yorùbá data – without this, the approach fails, even with perfect pairs. Qualitative analyses show that, while some scores like precision are modest, the proposed approach provides reasonable outputs on a real low-resource language. Code will be released upon acceptance.

II. Visually prompted keyword localisation

The task of visually prompted keyword localisation (VPKL) involves two steps: (i) detecting and (ii) localising a given keyword (specified through an image) in speech utterances. In the detection step (Fig. 1-middle), the model is shown an image query $\mathbf{v}$ depicting a keyword and predicts if the keyword occurs anywhere in a spoken utterance $\mathbf{a}$. In the localisation step (Fig. 1-bottom), the model predicts the time when the query occurs within the utterance $\mathbf{a}$.

To perform VPKL, we assume we have a dataset of speech and image pairs. This enables the training of a visually grounded speech model (Sec. II-A), which learns a similarity between images and spoken utterances. But this is not enough to enable precise detection of specific keywords. So we further assume access to a small support set of spoken keyword examples. Based on this set, we automatically mine more training pairs (Sec. II-B), which are used for learning to detect the desired keywords. For localisation, we don't have explicit training data, but we perform it in a weakly-supervised manner by extracting the time frame of the audio that is most similar to the query image.

A. Visually grounded speech model and loss

The model that we use consists of a vision and an audio branch, connected with an attention mechanism, as shown in Fig. 2. The vision branch is the AlexNet network [10] and it encodes an image $\mathbf{v}$ as a sequence of embeddings $\mathbf{y}_{\mathrm{v}}$. The acoustic branch uses an acoustic network, pretrained on unlabelled speech with contrastive predictive coding (CPC) [11], and it is followed by two BiLSTM layers; these networks encode a spoken input $\mathbf{a}$ as a sequence of frame embeddings $\mathbf{y}_{\mathrm{a}}$. The vision and audio branches are connected by a matchmap attention mechanism [7] that computes the dot product between each audio embedding in $\mathbf{y}_{\mathrm{a}}$ and each vision embedding in $\mathbf{y}_{\mathrm{v}}$, yielding a similarity matrix $\mathcal{M}$. To predict at which frames an image query occurs, we take the maximum over the image axis of $\mathcal{M}$ and obtain a similarity score for each frame. To get the overall similarity score S for VPKL detection, we take the maximum over the entire $\mathcal{M}$. We refer to this model as LocAttNet. Our model is similar to that of [8], but the latter employed a much more intricate approach to obtain the similarity score by using context vectors on top of the matchmap, while here we just get the detection score directly.

LocAttNet is trained as follows. Paired images and spoken captions in our dataset are used as anchor pairs $(\mathbf{a}, \mathbf{v})$. For each anchor, we sample positive utterances $\mathbf{a}_i^+$ and images $\mathbf{v}_i^+$, and negative utterances $\mathbf{a}_i^-$ and images $\mathbf{v}_i^-$. Positives and negatives are sampled based on a particular keyword. E.g. if the keyword is "boys", then the anchor image $\mathbf{v}$ and each positive image $\mathbf{v}_i^+$ contain visual depictions of BOYS somewhere in each image; similarly, on the audio side, the anchor utterance $\mathbf{a}$ and each positive utterance $\mathbf{a}_i^+$ contain "boys" somewhere within each utterance. The visual or spoken representations of "boys" do not occur in the negative images $\mathbf{v}_i^-$ or utterances $\mathbf{a}_i^-$. The idea is that these pairs encourage the model to focus on keywords within utterances and images, rather than focusing on them as a whole. Based on these pairs, we use a contrastive loss [9]:

$$\ell = d\left(S(\mathbf{a},\mathbf{v}),100\right) + \sum_{i=1}^{N_{\text{neg}}} d\left(S(\mathbf{a}_i^-,\mathbf{v}),0\right) + \sum_{i=1}^{N_{\text{neg}}} d\left(S(\mathbf{a},\mathbf{v}_i^-),0\right) + \sum_{i=1}^{N_{\text{pos}}} d\left(S(\mathbf{a},\mathbf{v}_i^+),100\right) + \sum_{i=1}^{N_{\text{pos}}} d\left(S(\mathbf{a}_i^+,\mathbf{v}),100\right) \tag{1}$$

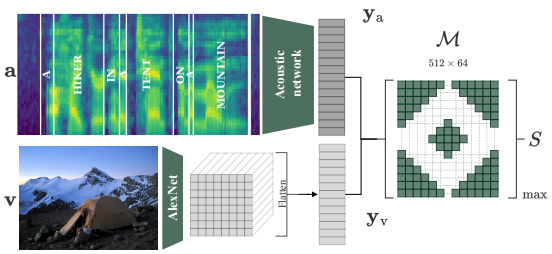

Fig. 2. LocAttNet consists of a vision and an audio branch connected through a localisation attention mechanism.

where d is the squared Euclidean distance. This loss attempts to make the similarity scores S close to 100 for positive pairs while negative pairs are pushed to have scores S close to zero.

B. Few-shot pair mining

For a low-resource language, we do not have access to the transcriptions required to sample the positive and negative pairs for (1) above. To get these pairs, we turn to the few-shot pair mining approach of [9]. We start by collecting a small number (K) of isolated speech examples for each of the keywords that we want to detect and localise. These are combined into a support set. For each keyword, we then use these spoken support set examples to automatically find utterances containing instances of the same keyword (e.g. "a group of boys playing soccer" and "the boys are climbing a tree"). Since utterances are also paired with images in our dataset, images can automatically be labelled with the predicted keywords of its paired utterance. In this way we obtain positive and negative utterances and images for each keyword. And all this without any transcriptions!

How do we use an example in the support set to find utterances containing the word? The support set word examples are used as queries in a query-by-example search approach called QbERT. This method uses HuBERT [12] to encode speech as a set of discrete units that approximate phones. Each query is then scored against each utterance in the dataset using a noisy string matching algorithm [13]. We take the mean score across the K word examples per keyword class for each utterance. The utterances are then ranked from highest to lowest for each keyword and the top n utterances are predicted to contain the keyword.

III. English experiments

To analyse our model and to compare it to previous work, we first perform VPKL experiments on English.

A. Experimental setup

Data. We train an English LocAttNet on the Flickr8k Audio Captions Corpus (FAAC) [15], which consists of 8k images each paired with five spoken English captions. The dataset is split into 30k, 5k and 5k utterances for train, development and test sets. For the support set, we sample $K = 10$ examples per keyword from the training and validation sets. Using forced alignments, we isolate the keywords. To mine pairs, we use the remainder of the training and development sets as the unlabelled speech dataset and predict that the top $n = 200$ samples per keyword class contain the keyword. Utterances are parametrised as mel-spectrograms with a hop length of 10 ms, a window of 25 ms and 40 mel bins. These are truncated or zero-padded to 1024 frames. Images are resized to 224×224 pixels and normalised with means and variances calculated on ImageNet [16].

Evaluation. We follow exactly the same protocol as in [8]. For each of the 34 keywords, 10 images from the Flickr8k test split were manually cropped to serve as image queries. The cropped images mostly contain the region corresponding to the keyword, but in some cases also include adjacent objects. In testing, the similarity score S for an utterance and an image query is calculated. If S is above a threshold α, the keyword depicted in the image query is predicted to be in the utterance. The α for each model is tuned on the development set. If a keyword is detected, the frame where the maximum attention score occurs is predicted as the keyword's position. We use the ground truth alignments to evaluate the predictions: a true positive is taken when the predicted frame falls within the ground-truth time-span of the keyword. It is counted as a mistake if a keyword is falsely detected or the prediction falls outside the time-span. Each model is trained three times to get mean scores and standard deviations.

Our model. The image branch of LocAttNet is initialised with the convolutional encoder of AlexNet [10], pretrained on ImageNet [16]. For the audio branch, we use an acoustic network pretrained using a self-supervised CPC task [17] on LibriSpeech [18] and the multilingual (English and Hindi) Places dataset [19]. We take $N_{\text{neg}} = N_{\text{pos}} = 4$ in (1), based on development experiments. The model is trained for 100 epochs using Adam [20]. A validation task is used for early stopping, with pair mining again used for constructing validation pairs (so transcriptions are never used).

Baselines. We compare our approach to that of [8]. Instead of pair mining, this model uses an external visual tagger to automatically label training images and then use these predicted tags for getting positive and negative pairs in a contrastive loss. The study [8] also has a topline model that uses transcriptions to get perfect pairs. We also compare to the visual bag-of-words (BoW) method of [14], which is queried with written keywords instead of images. This model is also trained using a visual tagger to generate textual BoW labels for training images. These labels are then used to train a model that takes speech as input and predicts the location of written keywords as output. While the task is somewhat different to ours (queries are text instead of images), we can still compare to how well a given keyword is detected and localised.

B. Results

English VPKL results are given in Table I. Line 3 gives the results of our approach. Compared to the model of [8] trained without transcriptions on visual tags (line 2), our new few-shot mining approach is consistently better. Additionally, our few-shot mining approach outperforms even the unsupervised textual keyword localisation method of [14] in line 1. This is noteworthy given that a written keyword arguably gives a stronger and less variable query signal than an image. The left part of Fig. 3 shows qualitative examples of the few-shot model detecting and localising image queries within utterances. We see for the keyword "soccer", the system makes a localisation error, but this is reasonable given the ambiguity in the visual query.

To establish the best possible results we could get from our approach, we train a LocAttNet model using ground truth pairs obtained from transcriptions instead of few-shot QbERT-mined pairs (Sec. II-B). By comparing lines 3 and 5,

TABLE I
VISUALLY PROMPTED KEYWORD DETECTION AND LOCALISATION RESULTS (%) ON ENGLISH. TOPLINE MODELS ARE SHOWN IN GREY.

	Model	Detection				Localisation			
		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
1	Visually grounded BoW [14]	**42.29**	36.32	39.08	36.63	33.39	31.02	32.17	28.50
2	Nortje et al. [8] (visual tagger)	31.02	31.83	31.42	32.02	23.20	25.75	24.21	23.57
3	LOCATTNET (few-shot mined pairs)	36.94±2.1	**48.80**±1.6	**42.03**±1.8	**49.16**±1.7	**33.72**±1.3	**46.52**±1.1	**39.09**±1.1	**44.21**±0.5
4	Nortje et al. [8] (ground truth pairs)	48.40	55.85	51.86	56.20	44.43	53.79	48.66	50.98
5	LOCATTNET (ground truth pairs)	63.18±2.1	45.12±3.7	52.62±3.0	45.54±3.6	58.98±0.5	43.61±3.1	50.11±2.1	42.84±2.9

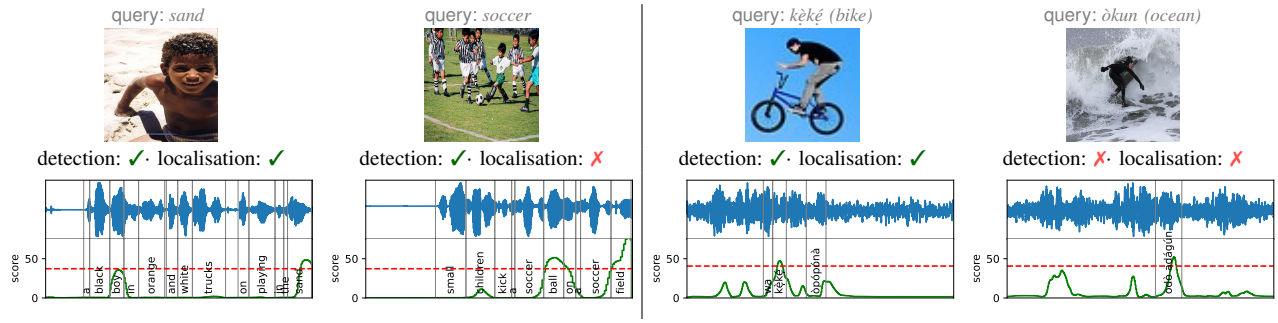

Fig. 3. Qualitative samples on English (left) and Yorùbá (right). Given a query image, we show the top detected audio sample and the scores for localisation. We include the corresponding keyword for reference, but this is not seen by the model. The red dotted line denotes the similarity score threshold α, which determines whether the query image is detected in the audio.

we see that both detection and localisation F1 drop by around 11% when moving from the ideal to the sampled pairs (e.g. localisation F1 goes from 50.1% to 39.1%). So there is still room for improvement by getting better positive and negative pairs.

We mostly followed the model architecture of [8], but proposed to simplify the method for getting a similarity score (see Sec. II-A). To see what the influence of the architectural change is, we compare our LOCATTNET topline model (line 5) to the topline of [8] (line 4). While accuracies are somewhat better with the more complex attention mechanism, detection and localisation F1 is better with the simpler model proposed here.

IV. LOW-RESOURCE EXPERIMENTS: YORÙBÁ

We train a Yorùbá LOCATTNET to detect and localise an image query depicting a keyword in a Yorùbá spoken utterance.

A. Experimental setup

Data. For the Yorùbá experiments, we use the Yorùbá version of the FAAC dataset, called YFACC [1]. This is a single-speaker dataset containing a single spoken Yorùbá caption for each of the 8k Flickr images. The dataset has 7k, 500 and 500 utterances in its train, development and test sets, respectively. We manually isolate $K = 5$ spoken examples for each of the 34 keywords from the training and validation sets to obtain the support set. We use the remainder of the train and development sets as the unlabelled speech dataset for pair mining. Because this dataset is much more limited than the English case, to set n, we use the actual number of samples in the training and validation sets in which the keyword occurs.

Models. There are a few changes in the Yorùbá model compared to the English one (Sec. III-A). First, for pair mining (Sec. II-B) we replace the English HuBERT in QbERT with a multilingual HuBERT trained on English, French and Spanish [21]. The idea is that multilingual representations would be more robust on the unseen language. To tailor the representations to Yorùbá even more, we train the clustering model on background Yorùbá data consisting of 51 hours of Bible recordings [22], [23]. This model gives the discrete units for pair mining. To initialise the audio branch of the Yorùbá model, we also use the Yorùbá Bible data to train the CPC model (Sec. II-A).

Evaluation. We use the same image queries for the 34 keywords used in the English task. The only difference here is that instead of the English utterances, we use the Yorùbá utterances from the YFACC test set as search utterances.

B. Yorùbá VPKL results

Table II reports the Yorùbá VPKL scores. Line 2 shows the scores achieved by our Yorùbá few-shot LOCATTNET model. This is the first time VPKL is performed on an actual low-resource language. This is also only the second time that keyword localisation is performed on a low-resource language with a visually grounded model, with the first being the model in line 1 (which takes text queries instead of images). The detection recall and accuracy scores of the few-shot LOCATTNET (line 2) are competitive to the visual BoW model (line 1). However, the detection precision and localisation scores are lower. To investigate why this happens, we look at a Yorùbá LOCATTNET model trained on ground truth pairs (line 4). This topline model outperforms the BoW model with roughly 4–13%

TABLE II
KEYWORD DETECTION AND LOCALISATION RESULTS (%) ON YORÙBÁ. TOPLINE MODELS ARE SHOWN IN GREY.

	Model	Detection				Localisation			
		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
1	Visually grounded BoW [1]	**38.55**	45.39	**41.69**	46.29	**22.75**	**32.89**	**26.90**	**26.33**
2	LOCATTNET (few-shot mined pairs)	7.62±0.3	**46.62**±2.1	13.10±0.4	**46.62**±2.1	2.43±0.4	21.60±1.1	4.36±0.7	14.73±1.5
3	LOCATTNET (ground truth pairs & no CPC)	8.43±1.1	12.00±2.9	14.40±1.5	50.12±2.9	2.68±1.0	24.74±3.3	5.08±1.6	16.49±3.0
4	LOCATTNET (ground truth pairs & CPC)	59.96±2.3	50.42±1.4	54.74±0.2	50.42±1.4	45.08±2.9	43.31±0.7	44.13±1.1	37.86±0.0

on detection and 10–18% on localisation. In terms of precision, LOCATTNET outperforms the visual BoW model by roughly 21% on detection and 22% on localisation. It therefore seems that the pair mining (Sec. II-B) is responsible for the poorer scores in line 2, and in particular for the worse precision. To support this further, we found that the accuracy of the Yorùbá mined pairs is 37% whereas the English mined pairs are 70%. Improving the mined pairs could therefore lead to a very accurate VPKL model. The major difference in the pair mining implementation in Yorùbá is that the HuBERT model has not been seen any Yorùbá data. This seems to be crucial for accurate mining.

To further show the importance of the representations being tailored to the target language beforehand, we investigate the contribution of the Yorùbá CPC initialisation. In line 3, we retrain the ground truth LOCATTNET model from a random initialisation without warm-starting from a Yorùbá CPC model. Comparing this model to the ground truth model in line 4, we see that CPC initialisation on the target language is essential. This highlights the advantage of using large unlabelled data to improve low-resource models through self-supervised learning.

The right part of Fig. 3 shows qualitative examples of the Yorùbá few-shot model performing VPKL. In the "òkun" ("ocean") example, the wrong keyword is detected and localised, "odò adágún" ("pool"), which is reasonable given the query.

V. DISCUSSION

The visually prompted keyword localisation (VPKL) task is not a typical mainstream task. Its formulation involves some nuanced assumptions that are worth further discussion.

Visual queries. Is VPKL really useful in low-resource settings? E.g. query-by-example search could be done using spoken queries rather than images. Or if we want to search speech in a low-resource language, we could use a BoW-based visually grounded speech model [24], e.g. allowing Yorùbá speech to be searched with English written keywords. We respond that a visual query is more flexible than either a textual or a spoken query: it can allow a user to search for words that they do not know or, compared to BoW-based approaches, to search for words outside of the vocabulary of the visual tagger that is used for supervision.

Multiple query objects. The VPKL task implicitly assumes that the query image refers to a single keyword. In our approach we try to achieve this by cropping the most relevant region, but this is not always perfect; as seen in Fig. 1, multiple objects may appear in an image. An interesting future direction would be to extend the framework to support multi-object settings, enabling the system to distinguish and localize several keyword referents within the input image.

Speech representations. Our approach relies on pre-trained HuBERT representations, which are not optimized for Yorùbá or other low-resource languages. Although these features transfer reasonably well, they may miss language-specific acoustic properties, which are important for keyword localization. A potential solution is to train a Yorùbá-specific HuBERT on unlabeled speech to obtain more tailored representations and improve performance in cross-lingual low-resource retrieval.

Few-shot samples. A more important limitation of our approach (and one that we agree should be addressed) is that we rely on a few-shot support set containing the keywords we would want to search for. This makes the approach applicable in low-resource settings, but it means that the vocabulary is constrained. Future work will look at removing the support set by adapting QbERT to compare whole utterances in a fully unsupervised mining approach, thereby enabling search for arbitrary words.

VI. CONCLUSIONS

We performed visually prompted keyword localisation (VPKL) – detecting and localising an image query depicting a keyword in spoken utterances – in a low-resource setting. We did this by building on previous work that followed an idealised scenario on English data. To make VPKL applicable in real low-resource settings, we proposed a few-shot approach to automatically mine positive and negative pairs in a contrastive loss for training a visually grounded speech model. The few-shot method relies on a small set of isolated examples for the keywords of interest. Coupled with a simpler attention mechanism than in previous work [8], we showed that this real low-resource approach is effective in VPKL experiments on English and Yorùbá. Future work directions include adding support for multiple visual queries, adapting the speech representations to the target language, and removing the need of few-shot support set.

REFERENCES

[1] K. Olaleye, D. Oneață, and H. Kamper, "YFACC: A Yorùbá speech-image dataset for cross-lingual keyword localisation through visual grounding," in *Proc. SLT*, 2023.

[2] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proc. ACL*, 2017.

[3] G. Chrupała, "Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques," *J. Artif. Intell. Res.*, 2022.

[4] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux, "Speech technology for unwritten languages," *IEEE/ACM TASLP*, 2020.

[5] S. Scholten, D. Merkx, and O. Scharenborg, "Learning to recognise words using visually grounded speech," in *Proc. ISCAS*, 2021.

[6] P. Peng and D. Harwath, "Fast-slow transformer for visually grounding speech," in *Proc. ICASSP*, 2022.

[7] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. ECCV*, 2018.

[8] L. Nortje and H. Kamper, "Towards visually prompted keyword localisation for zero-resource spoken languages," in *Proc. SLT*, 2023.

[9] L. Nortje, D. Oneață, and H. Kamper, "Visually grounded few-shot word learning in low-resource settings," *arXiv preprint arXiv:2306.11371*, 2023.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *ACM*, 2017.

[11] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2019.

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *ACM*, 2021.

[13] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, 1970.

[14] K. Olaleye and H. Kamper, "Attention-based keyword localisation in speech using visual grounding," in *Proc. Interspeech*, 2021.

[15] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. ASRU*, 2015.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[17] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," in *Proc. Interspeech*, 2020.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[19] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *Proc. ICASSP*, 2018.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[21] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," in *Proc. NAACL*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., 2022.

[22] J. Meyer, D. I. Adelani, E. Casanova, A. Öktem, D. W. J. Weber, S. Kabongo, E. Salesky, I. Orife, C. Leong, P. Ogayo, C. Emezue, J. Mukiibi, S. Osei, A. Agbolo, V. Akinode, B. Opoku, S. Olanrewaju, J. Alabi, and S. Muhammad, "BibleTTS: A large, high-fidelity, multilingual, and uniquely African speech corpus," in *Proc. Interspeech*, 2022.

[23] P. Ogayo, G. Neubig, and A. W. Black, "Building African voices," in *Proc. Interspeech*, 2022.

[24] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," in *Proc. SLTU*, 2018.