Now you're speaking my language: Visual language identification

Triantafyllos Afouras[1], *Joon Son Chung*[1,2], *Andrew Zisserman*[1]

[1] Visual Geometry Group, Department of Engineering Science, University of Oxford
[2] Naver Corporation

{afourast,joon,az}@robots.ox.ac.uk

Abstract

The goal of this work is to train models that can identify a spoken language just by interpreting the speaker's lip movements. Our contributions are the following: (i) we show that models can learn to discriminate among 14 different languages using only visual speech information; (ii) we compare different designs in sequence modelling and utterance-level aggregation in order to determine the best architecture for this task; (iii) we investigate the factors that contribute discriminative cues and show that our model indeed solves the problem by finding temporal patterns in mouth movements and not by exploiting spurious correlations. We demonstrate this further by evaluating our models on challenging examples from bilingual speakers.
Index Terms: language identification, language recognition.

1. Introduction

Language identification from audio is a relatively easy task for humans. Indeed we can distinguish between languages that we do not speak or understand [1]. Moreover, automatic language identification (LID) from audio speech, is a well studied problem [2, 3, 4, 5], and determining the spoken language is often a first step for multilingual speech recognition [6, 7].

But is it possible to infer the language spoken by only *looking* at the speaker's lip movements, without the audio? There is evidence that humans can infer the spoken language by observing the lip movements of the speaker [8, 9, 10]. Moreover, Newman and Cox [11, 12] have shown that, under controlled visual conditions, visual language identification can also be automated.

Our objective in this paper is visual language identification '*in the wild*' – speaker independent, and text (content) independent identification. To this end, we train and evaluate visual language identification (VLID) models on a large multilingual audio-visual speech dataset, composed of public datasets of TEDx talks. We show that VLID can be accomplished under more general conditions, with good accuracy and for a large number of languages. To ensure that the models are indeed distinguishing between languages by finding patterns in the mouth movement, and not instead using other factors (e.g. inferring ethnicity from appearance cues) or spurious correlations, we compare with a face recognition baseline and also evaluate the models on a dataset from a different domain, VoxCeleb2 [13].

VLID opens up a host of interesting applications such as automatically recognising the language in silent films, automatically detecting dubbing in films, or recognising the spoken language from a distance. Most importantly, from a practical perspective, it can be used to pre-condition lip reading models, which are highly dependent on context, and to make audio-based language identification more robust in noisy environments. Please see our website `http://www.robots.ox.ac.uk/~vgg/vlid` for video examples.

1.1. Related Work

Audio language identification. Research in audio language identification has a long history, and the performance given reasonably long speech segments is very high. The architectures, aggregation methods and loss functions used in the LID task are similar to those in speaker recognition. For example, Geng et al. [14] investigate the use of RNNs for temporal aggregation in language identification. Cai et al. [15] explore the encoder and loss function for LID and propose some efficient temporal aggregation strategies, while Chen et al. [16] use NetVLAD [17] for temporal aggregation. In more recent work [18] use a 2D CNN as feature extractor with a BLSTM backend for temporal modelling and a self-attentive pooling layer for utterance level aggregation. The experiments show that decision-level fusion of different architectures yields the best results. Miao et al. [19] propose the use of a CNN-LSTM-TDNN encoder in combination with attention mechanisms in both time and frequency. Padi et al. [20] use a BLSTM-based attention model, obtaining state-of-the-art results on the NRE17 dataset. Wan et al. [21] and Mazzawi et al. [22] also investigate LSTM based architectures for this dataset. Titus et al. [23] explore the effect of accent in language identification performance and train models robust to accented speech.

Visual language identification. The ability of humans to recognize languages by observing the lip movements of the speaker has been researched in psycholinguistics. Soto et al. [8] first report that facial speech information alone is sufficient for language identification. Weikum et al. [9] study visual speech identification in infants, while Ronquest et al. [10] investigate if humans are able to distinguish between English and Spanish based on visual speech.

However, there is limited research in using the visual modality to automatically identify the spoken language. Previous works by Newman and Cox [11, 12] are of closest relevance to ours: they introduce visual language identification as a classification problem, and show that languages can be classified by using only lip motion. However, the videos used are constrained to studio conditions, with a small number of subjects reading a set text, and their method does not use deep learning methods. Also related is [24] that identifies language in music videos by using both audio and video cues, while [25] used facial landmarks to classify between two languages, English and French. Brahme et al. [26] use constrained local models to the solve same task.

Lip reading. The methods used in visual language identification are closely related to those used for lip reading. There has been significant progress in the recent years, mainly due to the advances in deep learning and the creation of large scale datasets. While earlier work in the field used neural networks to predict phonemes [27] or words [28, 29], it has been proven

Table 1: *Statistics of audio-visual datasets used for training and evaluating our VLID models and baselines.* **# videos:** *Number of original YouTube videos.* **# hours:** *Total number of hours* **# clips:**. *Number of clips (each video is separated into multiple clips). For each statistic, we shown the minimum per language in parenthesis.*

dataset	# hours	# videos	# clips
LRS3-Lang+ (dev)	1,707 (38)	19,300 (342)	683k
LRS3-Lang+ (test)	166 (0.9)	1,816 (30)	59k
VoxCeleb2-Lang	9 (0.8)	1,595 (98)	8.8k
VoxCeleb2-Biling	20.7 (0.7)	921 (26)	15k

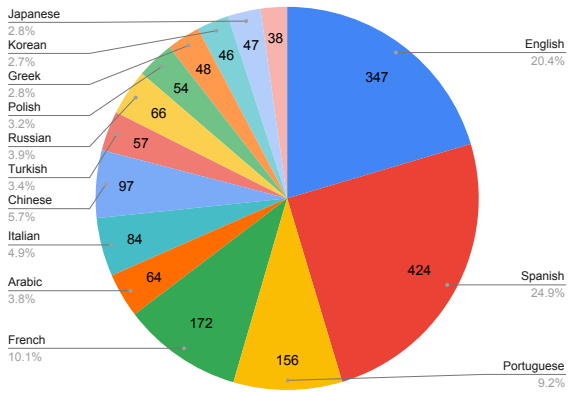

Figure 1: *Language distribution of the* `LRS3-Lang+` *dataset in number of hours.*

more recently that automatic lip reading can be generalised to continuous speech in unconstrained domains [30, 31, 32, 33, 34]. Recent works have shown that lip reading models trained on very large datasets can achieve word error rates as low as 33% on a real-world dataset, far exceeding the performance of professional lip readers [35].

2. Datasets

For training and evaluation, we use the `LRS3-Lang` [36] and `LRS3` [37] datasets, as well as `VoxCeleb2` [13] as a second multilingual test set. We show aggregate statistics of all datasets used in Table 1.

2.1. LRS3-Lang+

`LRS3-Lang` [36] is a multilingual audio-visual dataset based on videos collected from TEDx talks. The dataset covers 13 different (non-English) languages with a total of over 1,300 hours of video. For English we use the "pretrain" set of `LRS3` [37], where the videos come from the same domain (TED(x) talks) and the exact same process has been followed to collect the data. The test set of `LRS3` is small and contains short segments of no more than 6 seconds long. Therefore we re-split the "pretrain" set into a development and test set containing disjoint speakers. We incorporate this new split into `LRS3-Lang` as the English part to create a composite multilingual dataset of 14 languages, which we call `LRS3-Lang+`. The relative distribution of languages in our composite dataset are shown in Figure 1.

2.2. VoxCeleb2

`VoxCeleb2` is an audio-visual speech dataset which consists of 5,994 speakers with a total of 1,092,009 clips in the development set, and 118 speakers with 36,237 clips in the test set. To assess the cross-domain generalization capabilities of the models and baselines (trained on `LRS3-Lang+`), we create two subsets from the development set of `VoxCeleb2`, which we use as test sets.

VoxCeleb2-Lang. `VoxCeleb2` contains no language labels, however the identity of the speakers and their nationality are known. We therefore obtain language labels from two sources. The first is training an audio-only model on `LRS3-Lang+` (details in Section 3) and using it to classify the audio of the speakers in `VoxCeleb2`. The second source is using the nationality of the speakers: each language is assigned a list of nationalities – i.e. countries where the language is predominantly spoken. For example, English is associated with American, British, Australian, and Scottish nationalities; Spanish is associated with Spanish, Mexican, and Argentinean nationalities etc. For every speaker, we then use their nationality to list a set of possible languages. This narrows down the search space for each language considerably. The final language pseudo-labels are obtained by exploiting the redundancy between these two sources: For a given video, we only assign a language label when the audio-only model predicts one of the languages associated with the nationality of the speaker with a probability higher than a strict threshold (90%). This process gives us very accurate pseudo-labels, however leaves very few samples (less than 0.5 hour in total) for Japanese, Arabic and Greek. We therefore exclude these languages during evaluation on this dataset. The above procedure results in 11 languages, each containing material from at least 98 original YouTube videos each (see Table 1).

VoxCeleb2-Biling. To assess our models on bilingual speakers, we isolate individual speakers in `VoxCeleb2-Lang` who, across multiple videos, appear to be speaking both in English and in a non-English language with a high confidence, as determined by the audio model prediction. This is common due to the Celebrity content of the `VoxCeleb2` dataset (international actors, football players, politicians etc). We then create pairs of mother-tongue and English clips for those speakers. We refer to the resulting split as `VoxCeleb2-Biling`.

3. Architecture

We implement two types of models: an audio baseline, using audio features for LID, and our lip models using video features for VLID.

3.1. Input representation

Audio features. The input to the audio LID network is 80-dimensional log-mel spectrograms, extracted at every 10ms with 25ms frame length.

Video features. We extract embeddings modelling the lip movement with a spatio-temporal (3D/2D) ResNet18 network [38, 29] pretrained on word-level lip reading in English [31]. The model ingests a sequence of video frames (converted to grayscale) and outputs 512-dimensional visual features densely, one for every input frame.

3.2. Sequence modeling

We consider variations of Time-Delay Neural Networks (TDNN) and BLSTM [39, 40] encoders for the back-end. Those models ingest the visual features and convert them to representations more discriminative for the language recognition task, whilst potentially modelling longer term temporal dependencies. We experiment with 3 different encoder architectures.

TDNN model. This is a 10-layer residual temporal (1D) convolutional network. We use depth-wise separable convolutions [41] which we find to train faster and overfit less. The kernel width is set to 5, the number of channels to 512, and the temporal stride to 1 for all the layers.

TDNN + BLSTM. This model uses a TDNN as described above, followed by a bi-directional LSTM (BLSTM) with a cell dimension of 512.

3×BLSTM. This model, inspired by [22], uses a stack of 3 BLSTMs with cell size 512.

Utterance level aggregation. In line with the common practices in the audio LID literature, we also experiment with 3 different utterance-level aggregation techniques.

Temporal average pooling (TAP). The TAP layer simply takes the mean of the features along the time domain.

Self-attentive pooling (SAP). Unlike the TAP layer that equally pools the features over time, [15] introduces a self-attentive pooling layer that pays attention to the frames that are more informative for utterance-level speaker recognition.

NetVLAD. We also consider NetVLAD[17], which has been successfully used for temporally aggregating features in speech models for LID [16] and speaker verification [42]. NetVLAD mimics the BoW-derived VLAD[43] descriptor by learning a feature vocabulary from the input representations, then soft-quantising them over this dictionary and finally aggregating the results (in our case temporally).

3.3. Face recognition ablation

In order to assess to what extent our models learn to distinguish between spoken languages and are not using other appearance cues that are strongly correlated (e.g. ethnicity), we also consider the following baseline: We take a ResNet50 convolutional network [38] pretrained for face recognition on the VGGFace2 dataset [44] and fine-tune it on the on the VLID task. We consider 2 versions: (i) the model is trained end-to-end; (ii) the model is frozen at the penultimate residual block, i.e. only the last residual block and classification layers are fine-tuned.

4. Experimental Setting

Training. All models are trained only on `LRS3-Lang+`. We train the LID and VLID models by randomly sampling a segment of T contiguous frames from a given training clip. To accelerate training for all models we use a curriculum, first setting $T = 64$ and then increasing it to 128 and 256 frames (2.5s, 5s and 10s). During training the batches are balanced for languages. For languages with more samples available, the same frames are seen less often. To run inference with the RNN-based models on sequences longer than 256 frames (max seen during training), we split the sequence into 128 frame segments with 50% overlap and then average the predictions [21].

The face recognition baselines are trained by feeding one random frame from a clip at a time, with a batch of 32. For a fair comparison with our models, during inference we feed the face recognition models with all the frames of each test clip (e.g. 125 frames for the 5 seconds ones). The prediction is then obtained by averaging the model logits for all the frames.

Table 2: *Language Identification performance on the test set of `LRS3-Lang+` and the `VoxCeleb2-Lang` split. The average class accuracy is reported everywhere (higher is better). For all lip-reading models, a 3D/2D ResNet18 frontend is implied, and only the sequence-processing backend is varied and listed for comparison.* ***Mod.:*** *Input modality; A: Audio; L: Lips; F: Face.* ***Agg.:*** *Temporal aggregation strategy; NV.: NetVLAD. For `LRS3-Lang+` we report the average over all 14 languages (chance = 7%), while for `VoxCeleb2-Lang`, the average over 11 languages (excluding Japanese, Korean and Greek, chance = 9%). As the audio model is used to generate the pseudo-labels for the `VoxCeleb2-Lang` dataset, we don't report it's accuracy on this test set.*

			LRS3-Lang+			**VoxCeleb2**	
Model	**Mod.**	**Agg.**	**5s**	**10s**	**30s**	**5s**	**10s**
TDNN + BLSTM	A	TAP	95.6	96.6	97.3	-	-
ResNet50	F	AP	66.0	67.0	67.5	16.3	20.6
ResNet50 frozen	F	AP	39.9	40.8	41.2	24.5	27.4
TDNN	L	TAP	**67.2**	**76.3**	81.8	56.0	64.8
TDNN	L	SAP	66.4	74.2	76.8	52.9	62.0
TDNN	L	NV	66.3	74.0	75.8	46.4	59.8
TDNN + BLSTM	L	TAP	64.0	75.5	79.1	52.4	61.5
TDNN + BLSTM	L	SAP	65.4	75.2	79.2	52.1	61.1
3×BLSTM	L	TAP	64.8	75.5	82.0	**59.5**	**67.4**
3×BLSTM	L	SAP	64.5	76.0	**84.0**	58.5	66.7

Evaluation protocol. We evaluate on sequences of 5, 10 and 30 seconds long. As continuous clips of 30 seconds are very scarce in the datasets, we synthesize those by merging smaller clips from the same video together. For all experiments, the metric that we report is the average class language identification accuracy. We evaluate our models and baselines on the test set of `LRS3-Lang+`, on `VoxCeleb2-Lang`, and on `VoxCeleb2-Biling`.

5. Results

We summarize the results of our experiments on `LRS3-Lang+` and `VoxCeleb2-Lang` in Table 2.

As expected, the audio LID model achieves a very high accuracy. The visual VLID models also perform well. In both cases the model's performance improves as more temporal input is available. Indeed, when the visual models are supplied with 30 seconds of input the accuracy rises as high as 84%.

In terms of architectures, all options that we examine perform reasonably. The simplest of the models, *TDNN* performs best on `LRS3-Lang+`, except for the 30s case where the *3×BLSTM* model achieves marginally better results. When evaluating the models on the different domain of `VoxCeleb2-Lang`, the advantage of using the *3×BLSTM* is more apparent. Adding a BLSTM layer on top of the *TDNN* model impairs performance. In terms of utterance-level aggregation, neither SAP or NetVLAD clearly outperform simple temporal pooling. We conjecture that these results are due to

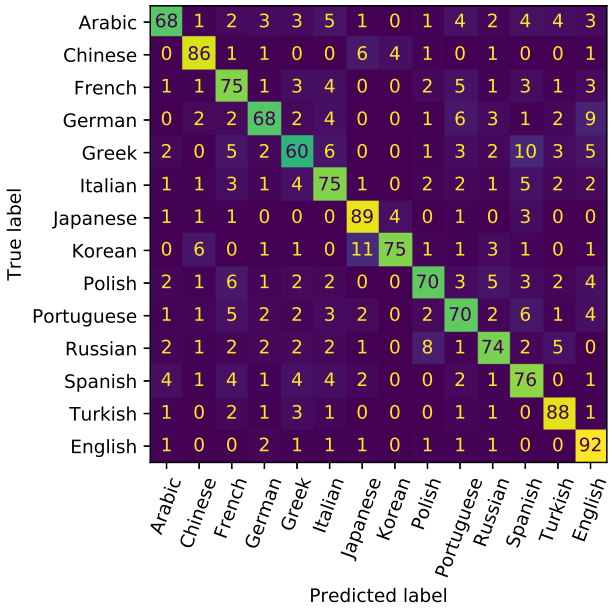

Figure 2: *Confusion matrix for predictions of 3×BLSTM-SAP model on the test set of* `LRS3-Lang+` *(10 seconds experiment).*

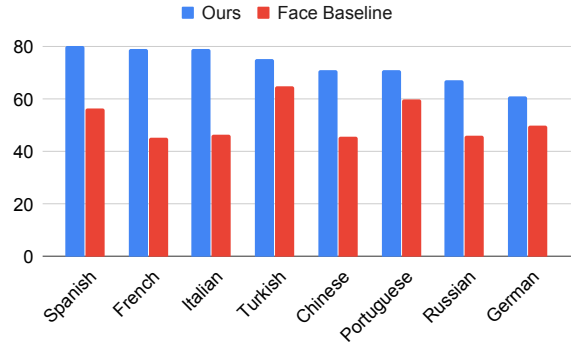

Figure 3: *Visual language identification accuracy on bilingual test set (*`VoxCeleb2-Biling`*). The model is tasked with discriminating between each language and English. Utterances of length 5 seconds are used. Chance accuracy is 50%.*

overfitting in the more complicated models.

We show the confusion matrix for the predictions of the *3×BLSTM-SAP* model in Figure 2. We note that the languages that are most commonly confused have phonetic similarities (e.g. German-English, Greek-Spanish, Korean-Japanese, Russian-Polish).

We next turn to the question of whether the visual model is indeed modelling the temporal mouth patterns to recognize the language or is just relying on appearance cues, such as face shape or skin tone. It is worth noting that (i) the visual features only use monochrome (not RGB) inputs, and (ii) they are trained on a word-level lip reading task on videos from British television and then frozen. This limits the extent of the information that they can access from the raw frames. In contrast, the baselines have a varying degree of access to the raw frames – and it can be seen that they can exploit this in solving the task. Examining the performance of the ResNet50-based face models, we notice that the model trained end-to-end obtains good results on `LRS3-Lang+`. However, when evaluated on `VoxCeleb2-Lang` the same model performs very poorly.

Figure 4: *Challenging examples from* `VoxCeleb2-Biling` *for which our VLID model correctly predicts the spoken language (indicated by the flag). Modelling of the lip movements is essential to solve this task.*

On the other hand, the evaluations of the model based on the frozen ResNet50, pretrained on face recognition, shows relatively worse performance on `LRS3-Lang+`, but its generalization on `VoxCeleb2-Lang` is better. The above suggest that the end-to-end model finds some shortcut which leads it to greatly overfitting the dataset. We conjecture that this might be due to background landmarks or camera artefacts correlated with the location of shooting of the TEDx events.

VoxCeleb2-Lang. We note that there is a significant domain shift between `LRS3-Lang+`, where the models have been trained, and `VoxCeleb2` as well as that the speaker identities between the two datasets are disjoint. As can be seen, the VLID models exhibit strong performance despite this domain shift. The face baselines, in contrast and as discussed above, drop in performance to near chance level. This demonstrates again that the VLID models are indeed using the mouth shape (visemes) and temporal changes for LID, and not employing shortcuts from the face and raw frames.

Bilingual speakers. On figure 3 we show results on bilingual speakers from VoxCeleb2. As expected, the accuracy of the face baseline fluctuates around the random performance (50%), as inferring the spoken language given the same face is very hard without any lip movement modelling. Our model significantly outperforms the baseline, reaching 80% accuracy for Spanish.

We show some qualitative examples of clips of bilingual speakers that our model predicts correctly in Figure 4. Please refer to our website for video examples.

6. Conclusion

We can give a qualified answer to the question posed in the introduction: Yes, it is possible to infer the spoken language only by observing the speaker's lips, and to a remarkably good accuracy. Our experiments have shown that using lip movements for this task exceeds using appearance cues captured by face embeddings. Finally, by performing analysis on bilingual speakers we demonstrated that our trained models can even distinguish between different languages spoken by the same person.

In future work we plan to investigate which lip movements provide the most discriminative cues, as well as explore the visual similarities and differences between languages – e.g. determine if certain viseme combinations are more prominent for some groups of languages than in others.

7. Acknowledgements

Funding for this research is provided by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems, the Oxford-Google DeepMind Graduate Scholarship, and the EPSRC Programme Grant Seebibyte EP/M013774/1.

8. References

[1] R. Lass, *Phonology: An introduction to basic concepts*. Cambridge University Press, 1984.

[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[3] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Interspeech*, 2011.

[4] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Interspeech*, 2011.

[5] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[6] J. Gonzalez-Dominguez, D. Eustis, I. Lopez-Moreno, A. Senior, F. Beaufays, and P. J. Moreno, "A real-time end-to-end multilingual speech recognition architecture," *IEEE Journal of selected topics in signal processing*, vol. 9, no. 4, pp. 749–759, 2014.

[7] M. Mller, S. Stker, and A. Waibel, "Neural codes to factor language in multilingual speech recognition," in *Proc. ICASSP*, 2019.

[8] S. Soto-Faraco, J. Navarra, W. M. Weikum, A. Vouloumanos, N. Sebastián-Gallés, and J. F. Werker, "Discriminating languages by speech-reading," *Perception & Psychophysics*, vol. 69, no. 2, pp. 218–231, 2007.

[9] W. M. Weikum, A. Vouloumanos, J. Navarra, S. Soto-Faraco, N. Sebastián-Gallés, and J. F. Werker, "Visual language discrimination in infancy," *Science*, vol. 316, no. 5828, pp. 1159–1159, 2007.

[10] R. E. Ronquest, S. V. Levi, and D. B. Pisoni, "Language identification from visual-only speech signals," *Attention, Perception, & Psychophysics*, vol. 72, no. 6, pp. 1601–1613, 2010.

[11] J. Newman and S. Cox, "Speaker independent visual-only language identification," in *Proc. ICASSP*, 01 2010, pp. 5026–5029.

[12] J. L. Newman and S. J. Cox, "Language identification using visual features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1936–1947, 2012.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[14] W. Geng, W. Wang, Y. Zhao, X. Cai, B. Xu, C. Xinyuan *et al.*, "End-to-end language identification using attention-based recurrent neural networks," *INTERSPEECH*, 2016.

[15] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Speaker Odyssey*, 2018.

[16] J. Chen, W. Cai, D. Cai, Z. Cai, H. Zhong, and M. Li, "End-to-end Language Identification using NetFV and NetVLAD," in *International Symposium on Chinese Spoken Language Processing*. IEEE, 2018, pp. 319–323.

[17] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. CVPR*, 2016.

[18] W. Cai, C. Danwei, S. Huang, and M. Li, "Utterance-level end-to-end language identification using attention-based cnn-blstm," in *ICASSP*, 05 2019, pp. 5991–5995.

[19] X. Miao, I. McLoughlin, and Y. Yan, "A new time-frequency attention mechanism for tdnn and cnn-lstm-tdnn, with application to language identification," *Interspeech*, pp. 4080–4084, 2019.

[20] B. Padi, A. Mohan, and S. Ganapathy, "Towards relevance and sequence modeling in language recognition," *arXiv preprint arXiv:2004.01221*, 2020.

[21] L. Wan, P. Sridhar, Y. Yu, Q. Wang, and I. L. Moreno, "Tuplemax loss for language identification," in *ICASSP*. IEEE, 2019, pp. 5976–5980.

[22] H. Mazzawi, X. Gonzalvo, A. Kracun, P. Sridhar, N. Subrahmanya, I. L. Moreno, H. J. Park, and P. Violette, "Improving keyword spotting and language identification via neural architecture search at scale," in *INTERSPEECH*, 2019.

[23] A. Titus, J. Silovsky, N. Chen, R. Hsiao, M. Young, and A. Ghoshal, "Improving language identification for multilingual speakers," *arXiv preprint arXiv:2001.11019*, 2020.

[24] V. Chandrasekhar, M. Emre Sargin, and D. A. Ross, "Automatic language identification in music videos with low level audio and visual features," in *Proc. ICASSP*, 2011.

[25] R. Špetlík, J. Čech, V. Franc, and J. Matas, "Visual language identification from facial landmarks," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 389–400.

[26] A. Brahme and U. Bhadade, "Lip detection and lip geometric feature extraction using constrained local model for spoken language identification using visual speech recognition," *Indian Journal of Science and Technology*, vol. 9, no. 32, pp. 1–7, 2016.

[27] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network." in *INTERSPEECH*, 2014, pp. 1149–1153.

[28] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2016.

[29] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Interspeech*, 2017.

[30] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017.

[31] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," in *INTERSPEECH*, 2018.

[32] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 513–520.

[33] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. de Freitas, "Large-Scale Visual Speech Recognition," *INTERSPEECH*, 2019.

[34] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[35] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2019.

[36] T. Afouras, L. Momeni, J. S. Chung, and A. Zisserman, "LRS3-Lang: a large-scale audio-visual dataset for multilingual visual speech recognition and language identification," in *arXiv*, 2020.

[37] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Sig. Proc.*, p. 26732681, Nov 1997.

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017.

[42] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP*, 2019.

[43] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the VLAD image representation," in *Proc. ACMM*, 2013.

[44] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.