INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

VoxCeleb: a large-scale speaker identification dataset

Arsha Nagrani[†], *Joon Son Chung*[†], *Andrew Zisserman*

Visual Geometry Group, Department of Engineering Science,
University of Oxford, UK
{arsha,joon,az}@robots.ox.ac.uk

Abstract

Most existing datasets for speaker identification contain samples obtained under quite constrained conditions, and are usually hand-annotated, hence limited in size. The goal of this paper is to generate a large scale text-independent speaker identification dataset collected 'in the wild'.

We make two contributions. First, we propose a fully automated pipeline based on computer vision techniques to create the dataset from open-source media. Our pipeline involves obtaining videos from YouTube; performing active speaker verification using a two-stream synchronization Convolutional Neural Network (CNN), and confirming the identity of the speaker using CNN based facial recognition. We use this pipeline to curate `VoxCeleb` which contains hundreds of thousands of 'real world' utterances for over 1,000 celebrities.

Our second contribution is to apply and compare various state of the art speaker identification techniques on our dataset to establish baseline performance. We show that a CNN based architecture obtains the best performance for both identification and verification.

Index Terms: speaker identification, speaker verification, large-scale, dataset, convolutional neural network

1. Introduction

Speaker recognition under noisy and unconstrained conditions is an extremely challenging topic. Applications of speaker recognition are many and varied, ranging from authentication in high-security systems and forensic tests, to searching for persons in large corpora of speech data. All such tasks require high speaker recognition performance under 'real world' conditions. This is an extremely difficult task due to both extrinsic and intrinsic variations; extrinsic variations include background chatter and music, laughter, reverberation, channel and microphone effects; while intrinsic variations are factors inherent to the speaker themself such as age, accent, emotion, intonation and manner of speaking, amongst others [1].

Deep Convolutional Neural Networks (CNNs) have given rise to substantial improvements in speech recognition, computer vision and related fields due to their ability to deal with real world, noisy datasets without the need for handcrafted features [2, 3, 4]. One of the most important ingredients for the success of such methods, however, is the availability of large training datasets.

Unfortunately, large-scale public datasets in the field of speaker identification with unconstrained speech samples are lacking. While large-scale evaluations are held regularly by the National Institute of Standards in Technology (NIST), these datasets are not freely available to the research community. The only freely available dataset curated from multimedia is the Speakers in the Wild (SITW) dataset [5], which contains speech samples of 299 speakers across unconstrained or 'wild' conditions. This is a valuable dataset, but to create it the speech samples have been hand-annotated. Scaling it further, for example to thousands of speakers across tens of thousands of utterances, would require the use of a service such as Amazon Mechanical Turk (AMT). In the computer vision community AMT like services have been used to produce very large-scale datasets, such as ImageNet [6].

This paper has two goals. The first is to propose a fully automated and scalable pipeline for creating a large-scale 'real world' speaker identification dataset. By using visual active speaker identification and face verification, our method circumvents the need for human annotation completely. We use this method to curate `VoxCeleb`, a large-scale dataset with hundreds of utterances for over a thousand speakers. The second goal is to investigate different architectures and techniques for training deep CNNs on spectrograms extracted directly from the raw audio files with very little pre-processing, and compare our results on this new dataset with more traditional state-of-the-art methods.

`VoxCeleb` can be used for both speaker identification and verification. Speaker identification involves determining which speaker has produced a given utterance, if this is performed for a closed set of speakers then the task is similar to that of multi-class classification. Speaker verification on the other hand involves determining whether there is a match between a given utterance and a target model. We provide baselines for both tasks.

The dataset can be downloaded from `http://www.robots.ox.ac.uk/~vgg/data/voxceleb`.

2. Related Works

For a long time, speaker identification was the domain of Gaussian Mixture Models (GMMs) trained on low dimensional feature vectors [7, 8]. The state of the art in more recent times involves both the use of joint factor analysis (JFA) based methods which model speaker and channel subspaces separately [9], and i-vectors which attempt to model both subspaces into a single compact, low-dimensional space [10]. Although state of the art in speaker recognition tasks, these methods all have one thing in common – they rely on a low dimensional representation of the audio input, such as Mel Frequency Cepstrum Coefficients (MFCCs). However, not only does the performance of MFCCs degrade rapidly in real world noise [11, 12], but by focusing only on the overall spectral envelope of short frames, MFCCs may be lacking in speaker-discriminating features (such as pitch information). This has led to a very recent shift from handcrafted features to the domain of deep CNNs which can be applied to higher dimensional inputs [13, 14] and for speaker identification [15]. Essential to this task however, is a large dataset obtained under real world conditions.

[†] These authors contributed equally to this work.

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-950

Many existing datasets are obtained under controlled conditions, for example: forensic data intercepted by police officials [16], data from telephone calls [17], speech recorded live in high quality environments such as acoustic laboratories [18, 19], or speech recorded from mobile devices [20, 21]. [22] consists of more natural speech but has been manually processed to remove extraneous noises and crosstalk. All the above datasets are also obtained from single-speaker environments, and are free from audience noise and overlapping speech.

Datasets obtained from multi-speaker environments include those from recorded meeting data [23, 24], or from audio broadcasts [25]. These datasets usually contain audio samples under less controlled conditions. Some datasets contain artificial degradation in an attempt to mimic real world noise, such as those developed using the TIMIT dataset [19]: NTIMIT, (transmitting TIMIT recordings through a telephone handset) and CTIMIT, (passing TIMIT files through cellular telephone circuits).

Table 1 summarises existing speaker identification datasets. Besides lacking real world conditions, to the best of our knowledge, most of these datasets have been collected with great manual effort, other than [25] which was obtained by mapping subtitles and transcripts to broadcast data.

Table 1: *Comparison of existing speaker identification datasets.* **Cond.:** *Acoustic conditions;* **POI:** *Person of Interest;* **Utter.:** *Approximate number of utterances. †And its derivatives. ‡Number of telephone calls. ∗ varies by year.*

Name	**Cond.**	**Free**	**# POI**	**# Utter.**
ELSDSR [26]	Clean Speech	✓	22	198
MIT Mobile [21]	Mobile Devices	-	88	7,884
SWB [27]	Telephony	-	3,114	33,039
POLYCOST [17]	Telephony	-	133	1,285‡
ICSI Meeting Corpus [23]	Meetings	-	53	922
Forensic Comparison [22]	Telephony	✓	552	1,264
ANDOSL [18]	Clean speech	-	204	33,900
TIMIT [28]†	Clean speech	-	630	6,300
SITW [5]	Multi-media	✓	299	2,800
NIST SRE [29]	Clean speech	-	2,000+	∗
VoxCeleb	Multi-media	✓	**1,251**	**145,000**

3. Dataset Description

VoxCeleb contains over 100,000 utterances for 1,251 celebrities, extracted from videos uploaded to YouTube. The dataset is gender balanced, with 55% of the speakers male. The speakers span a wide range of different ethnicities, accents, professions and ages.

Videos included in the dataset are shot in a large number of challenging multi-speaker acoustic environments. These include red carpet, outdoor stadium, quiet studio interviews, speeches given to large audiences, excerpts from professionally shot multimedia, and videos shot on hand-held devices. Crucially, all are degraded with real world noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and there is a range in the quality of recording equipment and channel noise. Unlike the SITW dataset, both audio and video for each speaker is released. Table 2 gives the dataset statistics.

4. Dataset Collection Pipeline

This section describes our multi-stage approach for collecting a large speaker recognition dataset, starting from YouTube videos. Using this fully automated pipeline, we have obtained hundreds of utterances for over a thousand different Persons of Interest (POIs). The pipeline is summarised in Figure 1 left, and key stages are discussed in the following paragraphs:

Table 2: VoxCeleb *dataset statistics. Where there are three entries in a field, numbers refer to the maximum / average / minimum.*

# of POIs	1,251
# of male POIs	690
# of videos per POI	36 / 18 / 8
# of utterances per POI	250 / 116 / 45
Length of utterances (s)	145.0 / 8.2 / 4.0

Stage 1. Candidate list of POIs. The first stage is to obtain a list of POIs. We start from the list of people that appear in the VGG Face dataset [30] , which is based on an intersection of the most searched names in the Freebase knowledge graph, and the Internet Movie Data Base (IMDB). This list contains 2,622 identities, ranging from actors and sportspeople to entrepreneurs, of which approximately half are male and the other half female.

Stage 2. Downloading videos from YouTube. The top 50 videos for each of the 2,622 POIs are automatically downloaded using YouTube search. The word ‘interview’ is appended to the name of the POI in search queries to increase the likelihood that the videos contain an instance of the POI speaking, and to filter out sports or music videos. No other filtering is done at this stage.

Stage 3. Face tracking. The HOG-based face detector [32] is used to detect the faces in every frame of the video. Facial landmark positions are detected for each face detection using the regression tree based method of [33]. The shot boundaries are detected by comparing colour histograms across consecutive frames. Within each detected shot, face detections are grouped together into face tracks using a position-based tracker. This stage is closely related to the tracking pipeline of [34, 35], but optimised to reduce run-time given the very large number of videos to process.

Stage 4. Active speaker verification. The goal of this stage is to determine the audio-video synchronisation between mouth motion and speech in a video in order to determine which (if any) visible face is the speaker. This is done by using ‘SyncNet’, a two-stream CNN described in [36] which estimates the correlation between the audio track and the mouth motion of the video. This method is able to reject the clips that contain dubbing or voice-over.

Stage 5. Face verification. Active speaker face tracks are then classified into whether they are of the POI or not using the VGG Face CNN. This classification network is based on the VGG-16 CNN [3] trained on the VGG Face dataset (which is a filtered collection of Google Image Search results for the POI name). Verification is done by directly using this classification score with a high threshold.

Discussion. In order to ensure that our system is extremely confident that a person is speaking (Stage 4), and that they have been correctly identified (Stage 5) without any manual interference, we set very conservative thresholds in order to minimise the number of false positives. Precision-recall curves for both tasks on their respective benchmark datasets [30, 31] are shown in Figure 1 right, and the values at the operating point are given in Table 3. Employing these thresholds ensures that although we discard a lot of the downloaded videos, we can be reasonably certain that the dataset has few labelling errors.

This ensures a completely automatic pipeline that can be scaled up to any number of speakers and utterances (if available) as required.

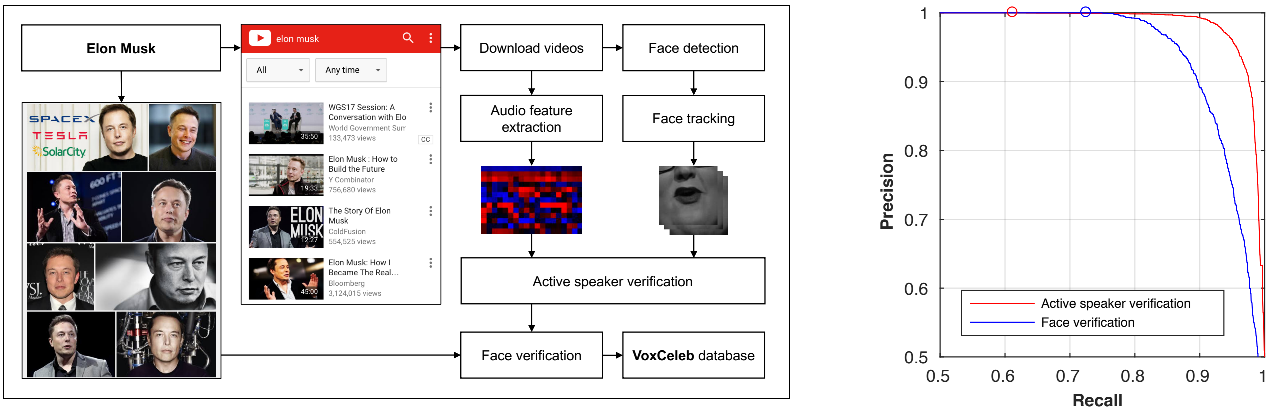

Figure 1: **Left:** *Data processing pipeline;* **Right:** *Precision-recall curves for the active speaker verification (using a 25-frame window) and the face verification steps, tested on standard benchmark datasets [30, 31]. Operating points are shown in circles.*

Table 3: *Precision-recall values at the chosen operating points.*

Task	**Dataset**	**Precision**	**Recall**
Active speaker verification	[31]	1.000	0.613
Face verification	[30]	1.000	0.726

5. CNN Design and Architecture

Our aim is to move from techniques that require traditional hand-crafted features, to a CNN architecture that can choose the features required for the task of speaker recognition. This allows us to minimise the pre-processing of the audio data and hence avoid losing valuable information in the process.

Input features. All audio is first converted to single-channel, 16-bit streams at a 16kHz sampling rate for consistency. Spectrograms are then generated in a sliding window fashion using a hamming window of width 25ms, step 10ms and 1024-point FFT. This gives spectrograms of size 512 x 300 for 3 seconds of speech. Mean and variance normalisation is performed on every frequency bin of the spectrum. This normalisation is crucial, leading to an almost 10% increase in classification accuracy, as shown in Table 7. No other speech-specific preprocessing (e.g. silence removal, voice activity detection, or removal of unvoiced speech) is used. These short time magnitude spectrograms are then used as input to the CNN.

Architecture. Since speaker identification under a closed set can be treated as a multiple-class classification problem, we base our architecture on the VGG-M [37] CNN, known for good classification performance on image data, with modifications to adapt to the spectrogram input. The fully connected *fc6* layer of dimension 9 × 8 (support in both dimensions) is replaced by two layers – a fully connected layer of 9 × 1 (support in the frequency domain) and an average pool layer with support 1 × n, where n depends on the length of the input speech segment (for example for a 3 second segment, $n = 8$). This makes the network invariant to temporal position but *not* frequency, and at the same time keeps the output dimensions the same as those of the original fully connected layer. This also reduces the number of parameters from 319M in VGG-M to 67M in our network, which helps avoid overfitting. The complete CNN architecture is specified in Table 4.

Identification. Since identification is treated as a simple classification task, the output of the last layer is fed into a 1,251-way softmax in order to produce a distribution over the 1,251 different speakers.

Verification. For verification, feature vectors can be obtained from the classification network using the 1024 dimension fc7 vectors, and a cosine distance can be used to compare vectors. However, it is better to learn an *embedding* by training a Siamese network with a contrastive loss [38]. This is better suited to the verification task as the network learns to optimize similarity directly, rather than indirectly via a classification loss. For the embedding network, the last fully connected layer (*fc8*) is modified so that the output size is 256 instead of the number of classes. We compare both methods in the experiments.

Testing. A traditional approach to handling variable length utterances at test time is to break them up into fixed length segments (e.g. 3 seconds) and average the results on each segment to give a final class prediction. Average pooling, however allows the network to accommodate variable length inputs at test time, as the entire test utterance can be evaluated at once by changing the size of the *apool6* layer. Not only is this more elegant, it also leads to an increase in classification accuracy, as shown in Table 7.

Table 4: *CNN architecture. The data size up to fc6 is for a 3-second input, but the network is able to accept inputs of variable lengths.*

Layer	**Support**	**Filt dim.**	**# filts.**	**Stride**	**Data size**
conv1	7×7	1	96	2×2	254×148
mpool1	3×3	-	-	2×2	126×73
conv2	5×5	96	256	2×2	62×36
mpool2	3×3	-	-	2×2	30×17
conv3	3×3	256	256	1×1	30×17
conv4	3×3	256	256	1×1	30×17
conv5	3×3	256	256	1×1	30×17
mpool5	5×3	-	-	3×2	9×8
fc6	9×1	256	4096	1×1	1×8
apool6	1×n	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Implementation details and training. Our implementation is based on the deep learning toolbox MatConvNet [39] and trained on a NVIDIA TITAN X GPU. The network is trained using batch normalisation [40] and all hyper-parameters (e.g. weight decay, learning rates) use the default values provided with the toolbox. To reduce overfitting, we augment the data by taking random 3-second crops in the time domain during training. Using a fixed input length is also more efficient. For verification, the network is first trained for classification (excluding the test POIs for the verification task, see Section 6), and then all filter weights are frozen except for the modified last layer and the Siamese network trained with contrastive loss. Choosing good pairs for training is very important in metric learning.

We randomly select half of the negative examples, and the other half using Hard Negative Mining, where we only sample from the hardest 10% of all negatives.

6. Experiments

This section describes the experimental setup for both speaker identification and verification, and compares the performance of our devised CNN baseline to a number of traditional state of the art methods on VoxCeleb.

6.1. Experimental setup

Speaker identification. For identification, the training and the testing are performed on the same POIs. From each POI, we reserve the speech segments from one video for test. The test video contains at least 5 non-overlapping segments of speech. For identification, we report *top-1* and *top-5* accuracies. The statistics are given in Table 5.

Speaker verification. For verification, all POIs whose name starts with an 'E' are reserved for testing, since this gives a good balance of male and female speakers. These POIs are not used for training the network, and are only used at test time. The statistics are given in Table 6.

Two key performance metrics are used to evaluate system performance for the verification task. The metrics are similar to those used by existing datasets and challenges, such as NIST SRE12 [29] and SITW [5]. The primary metric is based on the cost function C_{det}

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (1)$$

where we assume a prior target probability P_{tar} of 0.01 and equal weights of 1.0 between misses C_{miss} and false alarms C_{fa}. The primary metric, C_{det}^{min}, is the minimum value of C_{det} for the range of thresholds. The alternative performance measure used here is the Equal Error Rate (EER) which is the rate at which both acceptance and rejection errors are equal. This measure is commonly used for identity verification systems.

Table 5: *Development and test set statistics for identification.*

Set	# POIs	# Vid. / POI	# Utterances
Dev	1,251	17.0	139,124
Test	1,251	1.0	6,255

Table 6: *Development and test set statistics for verification.*

Set	# POIs	# Vid. / POI	# Utterances
Dev	1,211	18.0	140,664
Test	40	17.4	4,715

6.2. Baselines

GMM-UBM. The GMM-UBM system uses MFCCs of dimension 13 as input. Cepstral mean and variance normalisation (CMVN) is applied on the features. Using the conventional GMM-UBM framework, a single speaker-independent universal background model (UBM) of 1024 mixture components is trained for 10 iterations from the training data.

I-vectors/PLDA. Gender independent i-vector extractors [10] are trained on the VoxCeleb dataset to produce 400-dimensional i-vectors. Probabilistic LDA (PLDA) [41] is then used to reduce the dimension of the i-vectors to 200.

Inference. For identification, a one-vs-rest binary SVM classifier is trained for each speaker m ($m \in 1...K$). All feature inputs to the SVM are L2 normalised and a held out validation set is used to determine the C parameter (determines trade off between maximising the margin and penalising training errors). Classification during test time is done by choosing the speaker corresponding to the highest SVM score. The PLDA scoring function [41] is used for verification.

6.3. Results

Results are given in Tables 7 and 8. For both speaker recognition tasks, the CNN provides superior performance to the traditional state-of-the-art baselines.

For identification we achieve an 80.5% *top-1* classification accuracy over 1,251 different classes, almost 20% higher than traditional state of the art baselines. The CNN architecture uses the average pooling layer for variable length test data. We also compare to two variants: 'CNN-fc-3s', this architecture has a fully connected fc6 layer, and divides the test data into 3s segments and averages the scores. As is evident there is a considerable drop in performance compared to the average pooling original – partly due to the increased number of parameters that must be learnt; 'CNN-fc-3s no var. norm.', this is the CNN-fc-3s architecture without the variance normalization pre-processing of the input (the input is still mean normalized). The difference in performance between the two shows the importance of variance normalization for this data.

For verification, the margin over the baselines is narrower, but still a significant improvement, with the embedding being the crucial step.

Table 7: *Results for identification on* VoxCeleb *(higher is better). The different CNN architectures are described in Section 5.*

Accuracy	Top-1 (%)	Top-5 (%)
I-vectors + SVM	49.0	56.6
I-vectors + PLDA + SVM	60.8	75.6
CNN-fc-3s no var. norm.	63.5	80.3
CNN-fc-3s	72.4	87.4
CNN	**80.5**	**92.1**

Table 8: *Results for verification on* VoxCeleb *(lower is better).*

Metrics	C_{det}^{min}	EER (%)
GMM-UBM	0.80	15.0
I-vectors + PLDA	0.73	8.8
CNN-1024D	0.75	10.2
CNN-256D Embedding	**0.71**	**7.8**

7. Conclusions

We provide a fully automated and scalable pipeline for audio data collection and use it to create a large-scale speaker identification dataset called VoxCeleb, with 1,251 speakers and over 100,000 utterances. In order to establish benchmark performance, we develop a novel CNN architecture with the ability to deal with variable length audio inputs, which outperforms traditional state-of-the-art methods for both speaker identification and verification on this dataset.

8. Acknowledgements

Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1 and IARPA grant JANUS. We would like to thank Andrew Senior for helpful comments.

9. References

[1] L. L. Stoll, "Finding difficult speakers in automatic speaker recognition," *Technical Report No. UCB/EECS-2011-152*, 2011.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1106–1114, 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[5] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," *INTERSPEECH*, vol. 2016, 2016.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, S. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, 2015.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[9] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal, CRIM-06/08-13*, 2005.

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] U. H. Yapanel, X. Zhang, and J. H. Hansen, "High performance digit recognition in real car environments.," in *INTERSPEECH*, 2002.

[12] J. H. Hansen, R. Sarikaya, U. H. Yapanel, and B. L. Pellom, "Robust speech recognition in noise: an evaluation using the spine corpus.," in *INTERSPEECH*, pp. 905–908, 2001.

[13] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, pp. 1–5, 2015.

[14] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," *arXiv preprint arXiv:1609.09430*, 2016.

[15] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2016.

[16] D. van der Vloed, J. Bouten, and D. A. van Leeuwen, "NFI-FRITS: a forensic speaker recognition database and some first experiments," in *The Speaker and Language Recognition Workshop*, 2014.

[17] J. Hennebert, H. Melin, D. Petrovska, and D. Genoud, "POLYCOST: a telephone-speech database for speaker recognition," *Speech communication*, vol. 31, no. 2, pp. 265–270, 2000.

[18] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The Australian national database of spoken language," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. I–97, IEEE, 1994.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report*, vol. 93, 1993.

[20] C. McCool and S. Marcel, "Mobio database for the ICPR 2010 face and speech competition," tech. rep., IDIAP, 2009.

[21] R. Woo, A. Park, and T. J. Hazen, "The MIT Mobile Device Speaker Verification Corpus: Data collection and preliminary experiments," *The Speaker and Language Recognition Workshop*, 2006.

[22] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ Australian English speakers," *URL: http://databases.forensic-voice-comparison.net*, 2015.

[23] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, *et al.*, "The ICSI meeting corpus," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, IEEE, 2003.

[24] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, *et al.*, "The AMI meeting corpus," in *International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[25] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, *et al.*, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 687–693, IEEE, 2015.

[26] L. Feng and L. K. Hansen, "A new database for speaker recognition," tech. rep., 2005.

[27] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 517–520, IEEE, 1992.

[28] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, pp. 93–99, 1986.

[29] C. S. Greenberg, "The NIST year 2012 speaker recognition evaluation plan," *NIST, Technical Report*, 2012.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, 2015.

[31] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," *arXiv preprint arXiv:1603.08907*, 2016.

[32] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[33] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[34] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proceedings of the Asian Conference on Computer Vision*, 2016.

[35] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automatic naming of characters in TV video," *Image and Vision Computing*, vol. 27, no. 5, 2009.

[36] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference*, 2014.

[38] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 539–546, IEEE, 2005.

[39] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for MATLAB," *CoRR*, vol. abs/1412.4564, 2014.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[41] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the European Conference on Computer Vision*, pp. 531–542, Springer, 2006.