AUDIO-VISUAL PERSON RECOGNITION IN MULTIMEDIA DATA FROM THE IARPA JANUS PROGRAM

Gregory Sell, Kevin Duh, David Snyder, Dave Etter, Daniel Garcia-Romero*

Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

ABSTRACT

Currently, datasets that support audio-visual recognition of people in videos are scarce and limited. In this paper, we introduce an expansion of video data from the IARPA Janus program to support this research area. We refer to the expanded set, which adds labels for voice to the already-existing face labels, as the Janus Multimedia dataset. We first describe the speaker labeling process, which involved a combination of automatic and manual criteria. We then discuss two evaluation settings for this data. In the core condition, the voice and face of the labeled individual are present in every video. In the full condition, no such guarantee is made. The power of audio-visual fusion is then shown using these publicly-available videos and labels, showing significant improvement over only recognizing voice or face alone. In addition to this work, several other possible paths for future research with this dataset are discussed.

Index Terms— multimodal, audio-visual, speaker recognition, face recognition, multimedia

1. INTRODUCTION

The importance of video data in areas of audio, speech, and image processing has been steadily increasing in recent years. While interest has existed at some level for decades, the prevalence of video-sharing websites like YouTube as well as an ever-growing list of video-hosting social media sites has significantly increased the need to automatically process this type of data. As a result of this trend, using multiple modalities in video processing has drawn attention in numerous areas of research, such as diarization [1, 2], event detection [3], handwritten character recognition [4], and speech recognition [5, 6]. In the work to follow, we will explore the problem of person recognition in multimedia data.

Person recognition is the general task that includes speaker recognition and face recognition. The overarching paradigm of these tasks is that there is some labeled enrollment data that is used to define a model for a given person. That model is then compared against unknown data to determine the presence or absence of the enrolled people. In speaker recognition, a person is identified according to their voice, and in face recognition, a person is identified according to their face, but person recognition uses any information source available, which could include voice or face or both.

The potential power of combining face and speaker technologies was previously explored for automatic tagging of web videos [7]. The degree of success of that combination is difficult to assess, as the evaluation is performed on an internal set of exclusively automatically labeled data, and the fusion performance is not compared to any baselines for reference, but feasibility of the combination is clearly demonstrated.

A similarly-named task to person recognition was explored in the MediaEval challenge [8, 9]. Participants were asked to perform person discovery across multiple videos, which differs from recognition in that there is no enrollment data. And so, despite the similar terminology, this task is more like multimodal diarization.

In the work that follows, we will review existing datasets for speaker and face recognition in videos. We will then discuss our process for expanding one of those sets, the IARPA Janus video data, to include labels for speaker as well as face. Using that data, which we call the Janus Multimedia dataset, we benchmark speaker and face recognition systems, and then demonstrate that the fusion of the two systems results in large performance gains over either system alone. Finally, we discuss other areas of future work that the Janus Multimedia dataset can support.

2. VIDEO DATA FOR SPEAKER AND FACE RECOGNITION

As interest has expanded into the processing of multimedia data, datasets that originate from videos have become more prevalent in speaker and face recognition. However, though they are sourced from multimedia data, the unused modality is often thrown away prior to distribution. As a result, resources for research into multimodal recognition algorithms are currently limited.

2.1. Speaker Recognition Corpora

The recently released Speakers in the Wild (SITW) dataset [10] addresses speaker recognition in diverse video recordings. This corpus provides challenging conditions for speaker recognition, but, as of yet, the image portion of the videos has not been released. As a result, while this is an interesting and difficult dataset for speaker recognition, it does not support multimodal processing.

Even more recently, the VoxCeleb database [11] was released. This corpus is distributed as unique identifiers to online videos, and so the visual elements of the videos could be utilized as well. However, since the purpose of the dataset is speaker recognition, it was collected with automatic labeling of speakers based on face recognition and lip reading. As a result, the labels are likely strongly biased towards face recognition performance.

A collection of public speaking videos was released several years ago related to a demonstration of fast and efficient speaker identification with locality-sensitive hashing [12]. This dataset is potentially valuable for multimodal recognition, as the videos include both visual and audio representations of the individuals. However, despite the number of videos being reasonably large (1,111 videos), the number of speakers with multiple videos is somewhat small (74

*Work done during the HLTCOE SCALE workshop

speakers), significantly limiting experiment size. Also, since the videos are all of technical talks, the conditions are not as diverse as those in a generic video collection.

2.2. Face Recognition Corpora

The face recognition community has also shown strong interest in video data, though often these datasets are distributed as short segments or keyframes with no corresponding audio track. YouTube-Faces [13], for example, is drawn from YouTube videos that include individuals from the Labeled Faces in the Wild dataset [14], but is distributed only in the form of key frames drawn from the original videos.

Another type of face recognition corpus drawn from videos utilizes annotations of professionally produced video content, such as the Movie Trailer Face Dataset [15], which is built from annotated movie trailers found on YouTube. Other similar efforts have annotated sections of television episodes (i.e. *Buffy the Vampire Slayer* [16]) or films (i.e. *Hannah and Her Sisters* [17]). In most cases, these datasets can support audio-visual research (and some sets, such as *Hannah and Her Sisters*, are already even labeled for voice). However, the videos themselves here are professionally produced and often limited in the number of unique identities, and so they are highly domain-specific. Though there are numerous other potential research areas for which these are very well suited, their small size and specific domain potentially limit their generalizability.

In the last few years, the IARPA Janus program has been exploring face recognition in challenging conditions, including in videos[1]. While the videos are only labeled for face recognition, the original audio is available in many cases, allowing for the possibility of multimodal recognition. The videos include a challenging and diverse set of conditions and channels, and the data is multilingual, adding another unique challenge. Futhermore, the data is openly distributed under a Creative Commons license via NIST, making it ideal for open research. In the next section, we will discuss our process for determining speaker labels on the IARPA Janus video data in order to create a dataset for multimodal recognition.

It should also be noted that the recent video addition to the UMDFaces dataset [18] is also a compelling possibility for multimodal recognition, and labeling this dataset for speakers is a potential area for future work.

3. LABELING IARPA JANUS DATA FOR SPEAKER

The video portion of the IARPA Janus Benchmark-B (IJB-B) dataset [19] includes 7,011 videos. Manually inspecting all videos for the presence of voice would require a massive effort, and so we instead utilized several automatic measures to narrow down the videos which required manual labels.

First, we limited the video set to only those videos with an audio track, which immediately trimmed the list of videos to 2,312. We refer to this subset of the videos as the Janus Multimedia dataset, as this is the group of videos with both audio and visual elements. However, we still require labels for whether or not the labeled individual's voice is in the audio.

To lessen the manual annotation effort, we first ran an automatic process to identify easy cases of same or different speakers. For more reliable results, we first removed videos with less than five seconds of speech according to automatic speech activity detection. The videos that were found to have sufficient speech were then scored against each other with an i-vector speaker recognition system [20]. In this case, the i-vector extraction was trained using mel-frequency cepstral coefficients (MFCCs) computed on Fisher English data. The probabilistic linear discriminant analysis (PLDA) models [21, 22] were subsequently trained with telephony and microphone data from NIST Speaker Recognition Evaluations and Mixer6 that was also augmented with reverberation and added noise. This system was measured to yield a minimum detection cost function (mDCF) [23] of 0.67 ($P_{target} = 0.01$) on the SITW evaluation, which is a reasonable single-system performance on the task [24].

The scores of the audio tracks of the Janus videos were used to score every video pair for the likelihood of the presence of the same speaker. These scores were then used to identify videos that yielded large-magnitude negative scores in all trials where they are labeled for the same individual in the original face labels, suggesting that the videos do not have the same voice even though they do have the same face. Videos that were flagged according to these scores were assumed to be missing the voice matching the face label.

Anomalous cases that required manual inspection were identified visually in the score matrix sorted to group identities together, which created a block diagonal structure of high scores within label groups. Groups with any anomalous scores were flagged for manual annotation. Manual confirmation of the borderline trials is essential, because automatic labeling would necessarily remove all the difficult trials. Here, instead, the difficult trials were simply flagged for human annotation.

This alternatively means that any videos with high speaker recognition scores for all target trials are not guaranteed to have been manually checked for presence of voice. Many of these videos did receive subsequent confirmation of their labels during the selection of enrollment data (described below) or experimental error analysis, but only those with anomalous scores are guaranteed to have been checked manually.

The final step was to determine the set of individuals with sufficient videos to be included in the enrollment set. First, a list was assembled of individuals with at least two videos. The videos for these individuals were then manually inspected for suitability as enrollment data, meaning that the video is sufficiently long (at least five seconds of the correct voice confirmed by a human) and that the video is not a segment drawn from the same recording as any of the test trials. Some enrollment videos include multiple speakers, and so an approach inspired by the "assist" condition in SITW was adopted in which five seconds of the labeled speaker's speech was marked in all enrollment videos for system's to use as needed.

In the end, this process yielded a total of 1,593 videos that were determined to have both the face and voice of the labeled individual, drawn from the original 2,312 videos with an audio track. Furthermore, an enrollment set of 362 individuals was built, with exactly one video per person used for enrollment data. The enrollment list was split into a development (dev) list and evaluation (eval) list. The remaining videos were also split to ensure no overlapping speakers between dev and eval and then added to the test sets. We refer to the reduced set of 1,593 videos as the core subset, because these are the videos where the audio and visual recognition systems should agree. The full 2,312 videos provide a wilder situation where it is unknown if the audio and visual systems should agree or not. The importance of this full condition in application settings was emphasized in [7] for the purpose of automatic tagging of web videos. The statistics of the dev and eval lists for both conditions can be seen in Table 1.

[1]This product contains or makes use of the following data made available by the Intelligence Advanced Research Projects Activity (IARPA): IARPA Janus Benchmark B (IJB-B) data detailed at http://www.nist.gov/itl/iad/ig/facechallenges.cfm

List		Videos		Trials	
Condition	Split	Enroll	Test	Target	Nontarget
CORE	Dev	102	319	244	32,294
	Eval	258	914	681	235,131
FULL	Dev	102	436	317	44,155
	Eval	258	1,516	1,005	390,123

Table 1. Statistics of the Janus Multimedia dataset. The core subset includes only videos with both face and voice of the labeled person, while the full set includes all videos with audio.

4. FUSION OF SPEAKER AND FACE RECOGNITION

As a first experiment with this expanded set of labels, we explored the combination of speaker and face recognition systems for multimodal (audio-visual) person recognition.

4.1. Speaker Systems

We considered two different speaker systems in these experiments. The first is the i-vector system described above, which uses MFCC features and augmented telephone and microphone data for PLDA training, and was measured to perform reasonably well on SITW, a video-based speaker recognition corpus.

The second system uses deep neural network (DNN) embeddings [25] denoted x-vectors, but still utilizes a PLDA back-end trained with augmented telephone and microphone data. The DNN that produces the x-vectors is built as described in [25], with several time-delay layers [26] followed by a temporal pooling to aggregate statistics across a wide context. The output of this pooling is passed through several densely-connected layers, finally outputting through a soft-max classification layer trained to identify the speakers in the training set. While the system is trained for this direct identification, the network is used at test-time to extract the x-vectors from the first layer after the temporal pooling. This system is available for download through the Kaldi distribution[2].

Neither speaker system utilized diarization, either for enrollment or test videos, instead building representations from all available speech. These experiments were intended as an initial exploration into multimodal recognition, and so the effects of additions like diarization were left for future work.

4.2. Face System

Our face processing system used an open-source deep learning package available online[3]. This system uses a Multi-task Cascaded Convolutional Network [27] for detection, and FaceNet [28] trained on a subset of MS-Celeb-1M for recognition (model 20170512-110547).

Frames for processing were uniformly extracted every second. For enrollment data, we averaged the representation of all faces detected in the video's sampled frames in order to yield the face model for that individual. Then, the maximum cosine score (minimum cosine distance) between the model and each detected face in a test video was used as the score for the entire video with that model. In this way, the face recognition system was able to utilize the most basic version of diarization, though more sophisticated methods like face tracking were left for future work.

[2]http://kaldi-asr.org/models.html

[3]https://github.com/davidsandberg/facenet

		CORE		FULL	
	System	EER	mDCF	EER	mDCF
Audio	i-vector (s1)	13.0	0.663	23.0	0.762
	x-vector (s2)	11.5	0.534	21.9	0.675
	s1+s2	9.8	0.561	21.1	0.681
Visual	FaceNet (f1)	6.0	0.372	8.1	0.474
AV	s1+f1	3.7	0.265	8.2	0.410
	s2+f1	3.8	0.243	7.6	0.394
	s1+s2+f1	4.0	0.249	9.0	0.417

Table 2. Equal error rate (EER) and minimum detection cost function (mDCF) for all systems on the Janus Multimedia eval dataset. The fusion of audio and visual systems yields much better performance than any system alone, especially on the core subset.

4.3. Fusion

The fusion of the systems was performed by converting the output scores into log-likelihood ratios using calibration [29] trained on the dev set. Once the scores are in this form, fusion can be accomplished by summing the log-likelihood ratios from the individual systems. This form of fusion assumes that the systems are independent draws from the same label, which is a valid assumption for the core subset but not for the full list.

4.4. Results

Performance results on eval for each system can be seen in Table 2 in the form of equal error rates (EERs) and mDCFs ($P_{target} = 0.01$), metrics commonly used in speaker recognition. The full detection error tradeoff (DET) curves [30] for a subset of the systems on the core subset can also be seen in Fig. 1.

A first observation is that the i-vector performance metrics are very similar for the core subset of Janus Multimedia as for the SITW evaluation (mDCF $= 0.67$). While these numbers are not directly comparable, it is reassuring that the similar datasets yield similar results. This also demonstrates that, like SITW, the audio conditions in this dataset are very difficult and challenging, which is not surprising considering the diversity of conditions.

It is worth noting that the x-vector system comfortably outperforms the i-vector system. It is also clear in Table 2 that the face recognition system yields better performance than either speaker recognition system.

However, the most noteworthy result is that the fusion of audio and visual systems yields huge improvements in performance, especially on the core subset. The inclusion of the audio scores with the visual scores reduces the metrics by roughly a third (relative) from those of face alone, and the DET curve in Fig. 1 confirms that these gains are consistent across all operating points. This gain from multimodal fusion is massive compared to the smaller and operating-point-dependent improvement seen from combining the two speaker systems. So, the improvements from audio-visual fusion truly appear to come from the complementarity of the modalities rather than simply from the benefit of multiple systems.

However, the results on the full dataset also demonstrate that these gains are dependent on the correctness of the assumption that all scores are drawn from the same label. When that assumption no longer applies, relative improvement from fusion is reduced by half. The assumption fails because these trials in which the face is present

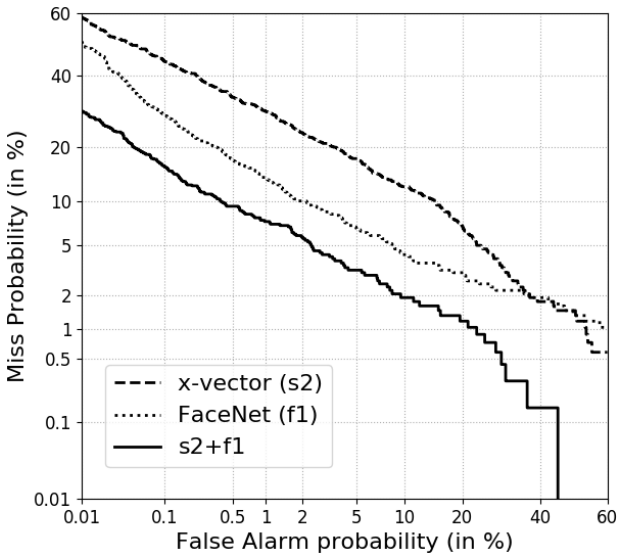
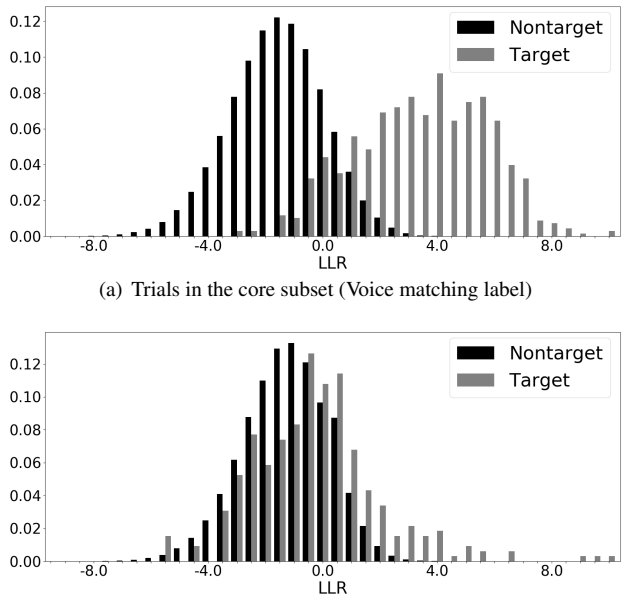


Fig. 1. DET curves on the core subset for the x-vectors (s2 from Table 2), FaceNet (f1), and their fusion, which show the gains from fusion are consistent across all operating points.



(a) Trials in the core subset (Voice matching label)

(b) Remaining trials in the full set (Voice likely not matching label)

Fig. 2. Score histograms from the x-vector system for target and nontarget trials (normalized for each class). In trials with the labeled person's voice (a) the two distributions are reasonably well separated. But in trials without the labeled person's voice (b) the distributions are essentially indistinguishable, showing why fusion is not as effective.

without the corresponding voice are drawn from the target distribution for face scores, but from the non-target distribution for speaker scores (since the labeled person is not speaking). This effect is especially clear in the score histograms in Fig. 2 which show that target and nontarget scores are generated from different distributions for trials in the core subset, but they are largely generated from the same nontarget distribution in the remainder of the trials. This indicates that an alternative approach is likely necessary for fusion, which is an area for future work.

5. FUTURE WORK

Several areas of future work were identified above, such as incorporation of speaker diarization, face tracking, or alternative fusion methods for improved performance when both face and voice are not reliably present, but there are several other potentially interesting directions for research that the Janus Multimedia dataset can support.

Many videos include multiple faces and/or voices, and multimodal diarization could be a valuable improvement over speaker diarization or face tracking alone. Additionally, matching faces to voices in the video through diarization could be one potential solution to the fusion problem, in which the speaker and face scores are only fused if they are from the same person. Similarly, syncronizing the speech with lip movement [31] could serve the same purpose.

Many videos also include written text, either in subtitles or in background images. Optical character recognition (OCR) could be an interesting contribution to determining the identity of individuals in the video. The work above focused on audio-visual recognition, but other modes like written text could easily contribute as well.

The IJB-B release also includes names of the individuals in the videos (all are famous public figures), which opens up interesting opportunities as well. One such possibility is to utilize a knowledge base with information about the enrolled individuals. In this approach, additional characteristics such as age, spoken language, or topic of spoken content could be utilized to further refine the likelihoods. Incorporating this type of information with automatic speech recognition, OCR, or scene analysis is another potential area of future work.

6. CONCLUSION

In this work, we reduced the IARPA Janus video data to a subset with both audio and visual elements, called the Janus Multimedia dataset. This set of videos was further reduced to those with both the face and voice of the labeled individual, called the core subset. The validity of these datasets and the power of the multimodal approach for recognition was shown in preliminary experiments, where the fusion of the two improved performance by 30% relative or more.

In addition to these initial experiments, several areas of potential future work were discussed, showing that this dataset has the potential to support a diverse set of future research. In order to facilitate this work throughout the community, labels for both conditions of the Janus Multimedia dataset will be included in future releases of the IARPA Janus data through the NIST website[4].

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of the IARPA Janus program in providing the data and in making the expanded audio labels available in future releases.

[4]https://www.nist.gov/programs-projects/face-challenges

8. REFERENCES

[1] Athanasios Noulas, Gwenn Englebienne, and Ben J. A. Kröse, "Multimodal Speaker Diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 79–93, January 2012.

[2] Ken Hoover, Sourish Chaudhuri, Caroline Pantofaru, Malcolm Slaney, and Ian Sturdy, "Putting a Face to the Voice: Fusing Audio and Visual Signals Across a Video to Determine Speakers," *arXiv preprint arVix:1706.00079*, 2017.

[3] George Awad, Jonathan Fiscus, David Joy, and Martial Michel, "TRECVID 2016: Evaluation Video Search, Video Event Detection, Localization, and Hyperlinking," in *Proceedings of TRECVID 2016*, 2016.

[4] Smita Vemulapalli and Monson Hayes, "Audio-video based character recognition for handwritten mathematical content in classroom videos," *Integrated Computer-Aided Engineering*, vol. 21, pp. 219–34, 2014.

[5] Jing Huang and Brian Kingsbury, "Audio-Visual Deep Learning for Noise Robust Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[6] Yaji Miao and Florian Metze, "Open-Domain Audio-Visual Speech Recognition: A Deep Learning Approach," in *Proceedings of Interspeech*, 2016.

[7] Mehmet Emre Sargin, Hrishikesh Aradhye, Pedro J. Moreno, and Ming Zhao, "Audiovisual Celebrity Recognition in Unconstrained Web Videos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[8] Johann Poignant, Hervé Bredin, and Claude Barras, "Multimodeal Person Discovery in Broadcast TV at MediaEval 2015," in *MediaEval Workshop*, 2015.

[9] Hervé Bredin, Claude Barras, and Camille Guinaudeau, "Multimodal Person Discovery in Broadcast TV at MediaEval 2016," in *MediaEval Workshop*, 2016.

[10] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The Speakers in the Wild (SITW) Speaker Recognition Database," in *Proceedings of Interspeech*, 2016.

[11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proceedings of Interspeech*, 2017.

[12] Ludwig Schmidt, Matthew Sharifi, and Ignacio Lopez Moreno, "Large-Scale Speaker Identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[13] Lior Wolf, Tal Hassner, and Itay Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," in *Proceedings of CVPR*, 2011.

[14] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition," Tech. Rep. 07-49, University of Massachusetts, Amherst, 2007.

[15] Enrique G. Ortiz, Alan Wright, and Mubarak Shah, "Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-based Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[16] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–59, April 2009.

[17] Alexey Ozerov, Jean-Ronan Vigouroux, Louis Chevallier, and Patrick Perez, "On evaluating face tracks in movies," in *Proceedings of the IEEE International Conference on Image Processing*, 2013.

[18] Ankan Bansal, Carlos Castillo, Rajeev Ranjan, and Rama Chellapa, "The Do's and Don'ts for CNN-based Face Verification," *arXiv preprint arXiv:1705.07426*, 2017.

[19] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother, "IARPA Janus Benchmark-B Face Dataset," in *Proceedings of CVPR*, 2017.

[20] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–98, May 2011.

[21] Sergey Ioffe, "Probabilistic Linear Discriminant Analysis," in *ECCV*, A. Leonardis, H. Bischof, and A. Pinz, Eds., pp. 531–42. Springer-Verlag, 2006.

[22] Simon J. D. Prince and James H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[23] "The NIST Year 2010 Speaker Recognition Evaluation Plan," (Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST\_SRE10\_evalplan.r6.pdf`), 2010.

[24] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The 2016 Speakers in the Wild Speaker Recognition Evaluation," in *Proceedings of Interspeech*, 2016.

[25] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, 2017.

[26] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[27] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, 2015.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a Unified Embedding for Face Recognition and Clustering," in *Proceedings of CVPR*, 2015.

[29] Niko Brümmer and David A. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey*, 2006.

[30] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, pp. 1895–8.

[31] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.