



TELVID: A Multilingual Multi-modal Corpus for Speaker Recognition

Karen Jones¹, Kevin Walker¹, Christopher Caruso¹, Elliot Singer²^{}, Trang Nguyen²^{*},
Robert Dunn²^{*}, Stephanie Strassel¹*

¹Linguistic Data Consortium, University of Pennsylvania, USA

²MIT Lincoln Laboratory, MA, USA

karj@ldc.upenn.edu, walker@ldc.upenn.edu, carusocr@ldc.upenn.edu, strassel@ldc.upenn.edu,
es@ll.mit.edu, trang.nguyen@ll.mit.edu, rbd@ll.mit.edu

Abstract

The TELVID corpus is a new multi-language, multi-modal resource for speaker recognition, comprising multiple conversational telephone speech and video recordings from each of 300 multilingual speakers. Consented subjects contributed recordings in a wide variety of recording conditions, with a minimum of 11 calls, 10 videos and one selfie image per person. Every speaker made recordings in Tunisian Arabic, North African French and/or English, along with two “freestyle” recordings that utilize the speaker’s choice of any language, dialect or mix of varieties. Recordings were audited to verify quality and speaker identity and portions of the data were selected for test data for the NIST 2024 Speaker Recognition Evaluation. We developed audio and visual baseline systems and measured baseline system performance. The TELVID corpus will be published in the Linguistic Data Consortium Catalog, making it broadly available for language-related research, education and technology development.

Index Terms: speaker recognition, speech database, video, telephony, corpus, Tunisian Arabic

1. Introduction

The TELVID Corpus is a new multilingual, multi-modal corpus consisting of conversational telephone speech (CTS) and audio from video (AfV) recordings made by speakers of Tunisian Arabic who are also fluent in North African French and/or English. The corpus is designed to support the development and evaluation of speaker recognition technologies, with a particular focus on speaker detection in cross-lingual and cross-modal trials. Consented, enrolled speakers submitted recordings that vary across multiple factors including modality, language, noise condition, handset, location, physical

appearance and camera position. Each speaker was required to contribute 10 or more telephone calls lasting at least 8 minutes, 10 or more videos lasting at least 1 minute, and one selfie image consisting of a close-up photograph of the speaker’s face. Telephone calls were made and recorded in Tunis in 2022, utilizing a custom-built telephone collection platform operated by local collection partners, while video recordings and selfie images produced by each speaker were uploaded to a custom website. Collected data was manually audited to verify recording quality and speaker identity, and multi-speaker videos were manually diarized to label regions where the target speaker was audible. Over 300 speakers contributed data to the corpus, with 291 speakers completing all data requirements.

Portions of the corpus provided development and test data for the National Institute of Standards and Technology (NIST) 2024 Speaker Recognition Evaluation (SRE24). To help inform test set selection, we developed audio and visual baseline systems and measured baseline performance.

The sections that follow describe the TELVID corpus design and implementation in detail. After an overview of related work in Section 2, we discuss speaker and language requirements in Section 3, methods used to collect the CTS and AfV data in Sections 4-5, data validation, auditing and video diarization procedures in Section 6, baseline system development and baseline results in Section 7, and finally a summary of corpus properties in Section 8 with conclusions in Section 9.

2. Related Work

Comparatively few other Tunisian language datasets have been specifically developed to support speaker recognition research. One exception is Call My Net 2 [1] which consisted of Tunisian Arabic CTS and was used in the 2018 and 2019 NIST Speaker

¹*DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

© 2024 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

Recognition Evaluations [2,3]. While Call My Net 2 consisted of CTS only, the TELVID corpus is multi-modal, containing both telephone and video data recorded under a variety of conditions.

In terms of language and CTS genre overlap, two datasets comparable to TELVID are the OrienTel Tunisia Modern Colloquial Arabic database [4] and the Maghrebi Language Identification Corpus [5], which was used to provide a subset of test and development segments for the 2022 NIST Language Recognition Evaluation [6].

Other Tunisian datasets, mostly created to support ASR activities, include: the Spoken Tunisian Arabic Corpus consisting of radio broadcast [7]; the Tunisian Arabic Railway Interactive Corpus [8]; TunSpeech (consisting of parliamentary recordings and audio of people reading books in Tunisian Arabic [9]; TunSwitch TO (consisting of sentences read in Tunisian only) and TunSwitch CS (broadcast/podcasts with various degrees of code switching [10]; and the English language Tunisian Lecture Corpus [11].

The TELVID collection followed similar design principles and protocols as the WeCanTalk corpus [12] used in SRE21 [13], which was also a multilingual, multi-modal corpus where speakers in Hong Kong, speaking Cantonese and/or Mandarin or English each contributed several telephone calls and videos.

3. Speaker and Language Requirements

3.1. Speaker Recruitment and Enrollment

The TELVID corpus was collected with oversight from the University of Pennsylvania's Institutional Review Board. We adopted a claque-based protocol in which individual target speakers (clagues) were recruited by study personnel to make calls and videos; clagues were then responsible for inviting their own friends, family members and acquaintances to participate as anonymous speaking partners in calls and videos. Both clagues and speaking partners provided consent prior to each recording. Upon enrollment via the project website, clagues provided basic demographic information (sex, primary language and year of birth) and informed consent. Clagues were compensated for each recording that satisfied requirements and earned a bonus upon completion of 11 calls, 10 videos and one selfie image. Anonymous speaking partners were not compensated. Over 400 clagues were recruited to support a goal of 300 speakers with a complete set of telephone and video of recordings. Over-recruitment was necessary to counteract the reality, generally true for most speech collections, that some enrolled participants fail to finish the required number of recordings or even start making any recordings at all.

3.2. Speaker Requirements

To be eligible to participate in the study, clagues were required to be at least 18 years old and be willing to provide basic demographic information upon enrollment via the study website as described above. Additionally, all clagues had to be multilingual with native or near-native fluency in Tunisian Arabic and fluency in either North African French or English or both. Each claque was assigned a unique and persistent ID that was associated with every call, video or selfie completed by this claque. Speaking partners were also assigned a speaker ID for each recording, but this ID is not guaranteed to be unique and persistent since speaking partners participated anonymously, possibly appearing in multiple calls, videos and/or claque speaker networks.

3.3. Language Requirements

Clagues were required to produce calls and videos with a specific language distribution, and anonymous speaking partners were required to speak the same language as clagues throughout the recording. Although clagues were necessarily multilingual and used to code-switching in everyday discourse, the requirements for the NIST SRE24 evaluation were that evaluation test segments consist of unambiguously monolingual speech. For this reason, clagues and speaking partners were instructed to avoid code-switching with the exception of designated “freestyle” calls and videos in which any language or mixture of languages could be used. (Freestyle recordings were not selected for use in SRE24 but remain part of the TELVID corpus.) The language distribution requirements are summarized in Table 1.

Table 1: *TELVID Language Requirements*

| Language | Calls | Videos |
|----------------------------------|-------|--------|
| Tunisian Arabic Monolingual | 5 | 5 |
| English or NA French Monolingual | 5 | 4 |
| Freestyle | 1 | 1 |
| Total | 11 | 10 |

4. CTS Collection

4.1. Call Requirements

Calls could be made by the claque no more than once per day and they needed to contain at least 3 minutes of speech on their own call side. To meet this goal, clagues received specific instructions to do at least half of the talking in calls at least 8 minutes long. Any topic of conversation was permissible though speakers were instructed to avoid discussing sensitive topics and mentioning personal identifying information such as full names. At least 25% of calls in the collection had to be made in a noisy environment, so the collection strategy was to instruct clagues to ensure half of their calls were made in noisy conditions. Clagues were allowed to call the same person more than once, but were instructed to ensure that they made calls to at least three unique individuals across their set of 11 calls. Calls were to be made using either cellphones or landlines and using a variety of devices across the 11 calls was encouraged. For all devices, whether the claque’s or their call partner’s, the phone numbers were anonymized in the collection metadata in line with privacy protection requirements.

4.2. Collection Platform and Participant Experience

The telephone collection platform consists of a control computer, custom Interactive Voice Recording software that triggered North African French language prompts providing participants with instructions, an Asterisk dialplan for routing calls programmatically, database servers in Tunis and Philadelphia for compilation of speaker metadata, and VPNs for storing recordings securely at the Tunis site before transfer to project servers. The system utilizes an E1 trunk service for voice traffic and two gateways: a GSM to VoIP gateway for access to the cellular network and an ISDN/PRI to VoIP gateway for access to the traditional phone network.

To initiate a call, clagues dialed a dedicated phone number and used their phone keypad to respond to prompts asking for consent to be recorded plus information about the call including language, phone and microphone type, noise level and whether

their call partner was a repeat callee. The claque then entered their speaking partner’s phone number, prompting the system to place the call. The speaking partner then heard a recorded greeting and was prompted to provide consent. After the platform bridged both sides of the call, recording began. Call recording terminated automatically after 10 minutes.

5. AfV Collection

5.1. Video Requirements

Each claque needed to contribute at least 10 video recordings where they were both visible and audible, and one selfie image to be used for reference. Variation within each claque’s set of videos was an important requirement of the corpus. To support this goal, claques were asked to provide a) one primary monolog video in which they appeared alone speaking only Tunisian Arabic for 3-6 minutes, facing the camera in a well-lit and quiet setting; and b) nine secondary videos lasting at least 1 minute, satisfying the language requirements specified in Table 1. Across the secondary videos, at least 5 had to involve another speaker, and the videos taken as a whole were required to show variation in terms of recording location, lighting and noise conditions, camera orientation and claque appearance. Use of filters and special effects was prohibited in the video collection.

5.2. Collection Platform Participant Experience

Claques logged in to the study website to upload their selfie and videos; if suitable claque videos were already present online they could instead enter the video URL for subsequent harvesting. For each video submitted, claques answered a series of questions about the languages used, recording date, and who was audible and visible in the recording (self only, self + 1 or self + multiple). Claques also indicated whether their appearance and location were distinct relative to other recordings, and answered questions about camera orientation and noise conditions. Claques provided click-through consent prior to completing the upload of each video or selfie. A modified open-source audio/video downloader (yt-dlp) was used to collect the highest resolution video version available.

6. Data Validation and Annotation

6.1. Automatic Validation Checks

A range of automatic checks on incoming recordings were set up to identify specific problems as early as possible. For calls, automated validation included checking that the recording duration and the amount of speech on the claque call side met requirements, as determined by our Broad Phonetic Class Speech Activity Detector [14]. Videos were automatically checked for duration and whether either the video or the extracted audio were duplicates of any previously collected files. Recordings failing any of these validation steps were excluded from the corpus.

6.2. Quality Auditing

Manual auditing began with a general quality audit. Video recordings were audited in their entirety, while pre-defined segments were extracted from calls for auditing. Call audit segments were extracted from the start, middle and end of the claque side of each call. The first 15 seconds was extracted to use as a reference segment, a procedure that is especially useful

for checking the accuracy of speaker labels since speakers often use the same greeting style in their calls. The next 15 seconds were skipped, then the remainder of the call was divided into thirds and the most speech dense 60 second stretch was identified via SAD from each third for extraction as an audit segment. During call auditing, auditors listened to these pre-selected segments and answered questions about language, speaker and overall quality using a web-based user interface. Video quality auditing followed a similar procedure, but auditors were instructed to watch the entire video rather than pre-selected segments, and also answered questions about who was visible and audible in the video.

6.3. Video Diarization

Whereas claques were required to be the sole speaker on their side of each call, six of each claque’s 10 video recordings involved other speakers. Consequently, it was necessary to do manual diarization on multi-speaker videos to establish when the claque was speaking. Diarization annotators first familiarized themselves with the claque’s voice and appearance by paying close attention to the claque’s primary video in which they are the sole person who is audible and visible. Next they watched and listened to each of the claque’s multi-speaker videos and marked the segment boundaries where the claque was speaking. Diarization did not attempt to label where claques were visible in the recording, nor did it attempt to label when other speakers may be audible during claque speech segments.

6.4. Speaker Auditing

Since the TELVID corpus was created specifically to support a speaker recognition evaluation, a robust process for ensuring that all the calls and videos with the same speaker ID were accurately labeled was crucial and so speaker auditing examined the entire set of recordings associated with a claque speaker ID. The first call that the claque made was designated a reference call. The auditor listened to all the other calls and for each one made a yes/no decision about whether the speaker was the same as in the reference call. Auditors then watched a sufficient amount of each video to again ensure that the same speaker as in the reference call was audible and that the person visible in each video was the same person in the selfie.

7. Baseline System Development and Results

7.1. Audio System

For the audio baseline system, the incoming signal was converted to 16 kHz, either by upsampling (CTS) or downsampling (AfV). Speaker diarization was performed using the marks provided with the corpus. The data was augmented with noise from the MUSAN corpus [15] and room impulse responses (RIRs) [16]. Kaldi energy-based voice activity detection (VAD) was applied, and utterances shorter than five seconds were discarded. A 80-dimensional mel-scale filter bank with a frequency range from 20 Hz to 7600 Hz was used for feature extraction.

We used a Res2Net50 model with additive angular margin loss ($m = 0.3$ and $s = 20$) for embedding extraction. This network was configured with a width of 26 and scale of 8 and trained on 4 second chunks with a batch size of 512. It used an Adam optimizer with a starting learning rate of 0.02, a warm-up of 1k steps, hold of 40k steps, and then a learning rate decay

by half every 10k steps. For fine-tuning, the network was trained on 10 second chunks with a batch size of 128. A SGD optimizer with a learning rate of 0.01 and momentum of 0.9 was used. After embeddings were extracted, they were centered, whitened, and unit-length normalized, and linear discriminant analysis (LDA) was used for dimension reduction to 200. Backend scoring was done using a simplified PLDA (SPLDA) model. The flow diagram of this baseline is shown in Figure 1.

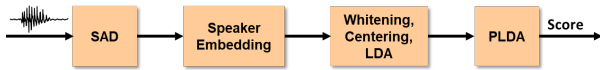


Figure 1: Diagram of audio baseline for speaker recognition

7.2. Visual System

For the visual baseline system, we used the RetinaFace [17] for face detection and InsightFace [18] to extract face embeddings. RetinaFace is a single-stage dense face localization network trained on a MobileNet-0.25 backbone using the WIDER Face data set [19]. InsightFace is a 101-layer ResNet trained with additive angular margin loss, on the Microsoft Celeb data set [20].

The flow diagram of the visual baseline is shown in Figure 2. First, an enrollment embedding is obtained from a selfie image. Then a test video is processed frame-by-frame as a series of still images. RetinaFace detects any faces that are present in each frame and InsightFace extracts an embedding for each detected face. An agglomerative clustering process operates across all frames to create clusters of similar face embeddings. The cosine score between the enrollment embedding and the mean of each cluster is computed and the maximum score from among all the clusters is taken. This maximum score is S-Normed, using the development set from the Janus Multi-Media corpus [21] as a cohort, to generate the final score.

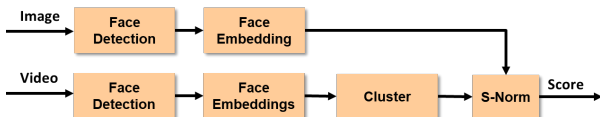


Figure 2: Diagram of visual baseline for face recognition

7.3. Baseline Experiment and Results

The results of the baseline system are shown in Table 2. For the Audio-Visual task, the sum of the trial scores from the audio and visual baseline systems was used.

Table 2: Baseline system performance

| Set | Task | EER | min C |
|------|--------------|-------|-------|
| Dev | Audio | 12.45 | 0.696 |
| Dev | Visual | 2.0 | 0.069 |
| Dev | Audio-Visual | 1.53 | 0.108 |
| Eval | Audio | 12.75 | 0.729 |
| Eval | Visual | 2.56 | 0.166 |
| Eval | Audio-Visual | 2.23 | 0.153 |

8. Corpus Results

In total, 3331 videos and 3574 calls from 311 unique claques passed quality and speaker audits. The corpus is roughly evenly divided by speaker sex, with 45% male and 55% female speakers. Language distribution across calls and videos demonstrated the variety expected from the corpus design, with roughly half containing Tunisian Arabic, a third containing North African French with smaller amounts of English, other languages and code mixing; these results are shown in Table 3.

Table 3: Language

| Language | Call Count | Video Count |
|-----------------|------------|-------------|
| Tunisian Arabic | 1716 | 1784 |
| NA French | 1417 | 1136 |
| English | 263 | 228 |
| Other | 171 | 12 |
| Mix | 7 | 171 |

Over half of the video recordings were made in a noisy recording environment, while only 20% of calls were noisy. Virtually all calls were made from a mobile phone, and while 71% of claques used a single phone number to complete their calls, 87% used multiple devices (e.g. the same phone both with and without headphones).

A major design focus of the TELVID corpus was ensuring variety in the video data, both in terms of the recording environment and the speaker features. Roughly half of the videos contained at least one audible speaker in addition to the claque, and 34% contained at least one additional visible speaker. While nearly all of the videos were made with the claque facing the camera and at a close distance, there was good variation in other features based on claque-reported metadata. Claques were stationary relative to the camera in 75% of the data but were always fully visible to the camera in under 60% of the videos. Over 40% of the videos were recorded outdoors, with fewer than 17% recorded in a repeat location. Finally, claques varied their physical appearance across the data, with over 77% of the data featuring a change in clothing or in other appearance features like the addition of accessories like hats, glasses or scarves.

The final TELVID corpus comprises 8KHz alaw CTS recordings, with one channel per speaker, plus video recordings in the original formats submitted by claques. The corpus also contains recording metadata, speaker demographic information and associated video diarizations and audit judgments.

9. Conclusions

The TELVID corpus is a new multi-modal corpus consisting of speech and video from multilingual speakers with data in three languages: Tunisian Arabic, North African French and English. The corpus was designed to support speaker recognition research and was successfully used in the NIST SRE24 evaluation, providing challenging new data. The TELVID corpus will be published in the LDC Catalog after the data is authorized for public release, making it broadly available for speaker recognition and other language-related research, education and technology development.

10. Acknowledgements

The authors would like to thank Craig Greenberg, Audrey Tong and Lukas Diduch from NIST for their collaboration and

invaluable feedback on collection design. Additionally, we are grateful to Dr. Mohamed Maamouri, the TELVID study participants and participant recruiters for their hard work, dedication and contributions to the collection.

11. References

- [1] K. Jones, S. Strassel, K. Walker and J. Wright, "Call My Net 2: A New Resource for Speaker Recognition," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6621–6626, Marseille, France, May 2020.
- [2] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason and J. Hernandez-Cordero, "The 2018 NIST Speaker Recognition Evaluation," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pp. 1483-1487, Graz, Austria, Sep. 2019.
- [3] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason and J. Hernandez-Cordero, "The 2019 NIST Speaker Recognition Evaluation CTS Challenge," *The Speaker and Language Recognition Workshop: Odyssey 2020*, Tokyo, Japan [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=929506 (Accessed January 29, 2025)
- [4] D. Iskra, R. Siemund, J. Borno, A. Moreno, O. Emam, K. Choukri, O. Gedge, H. Tropsf, A. Nogueiras, I. Zitouni, A. Tsopanoglou and N. Fakotakis, "OrienTel -Telephony databases across Northern Africa and the Middle East," *Proceedings of the Fourth Language and Resources Evaluation Conference*, pp. 552, Lisbon, Portugal, May 2004. European Language Resource Association (ELRA)
- [5] K. Jones, K. Walker, C. Caruso and S. Strassel, "MAGLIC: The Maghrebi Language Identification Corpus," *Proc. The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp.86-90, Quebec, Canada, June 2024.
- [6] Y. Lee, C. Greenberg, E. Godard, A. Butt, E. Singer, T. Nguyen, L. Mason and D. Reynolds, "The 2022 NIST Language Recognition Evaluation," *Proc. INTERSPEECH 2023*, pp.1928-1932, doi:10.21437/Interspeech.2023-241
- [7] I. Zribi, M. Ellouze, L. Belguith and P. Blache, "Spoken Tunisian Arabic Corpus STAC," *Transcription and Annotation. Research in Computing Science*. 90.10.13053/rcs-90-1-9. 2015.
- [8] A. Masmoudi, M.E. Khmekhem, E. Yannick, L.H Belguith and N. Habash, "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 306-310, Reykjavik, Iceland, May 2014. European Language Resource Association (ELRA)
- [9] A. Messaoudi, H. Haddad, C. Fourati, M.B.H. Hmida, A.B.E. Mabrouk and M. Graiet, "Tunisian Dialectal End-to-end Speech Recognition based on Deep Speech," *Procedia Computer Science, Volume 189*, 2021, pp. 183-190, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2021.05.082>.
- [10] A. A. B. Abdallah, A. Kabboudi, A. Kanoun and S. Zaiem, "Leveraging Data Collection and Unsupervised Learning for Code-Switched Tunisian Arabic Automatic Speech Recognition," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 12607-12611, doi: 10.1109/ICASSP48485.2024.10445734.
- [11] B. Bouziri, "A corpus-assisted genre analysis of the Tunisian Lecture Corpus: An exploratory study," *Research in Corpus Linguistics*, 8(2), pp. 103-132., 2020 <https://doi.org/10.32714/ricl.08.02.06>
- [12] K. Jones, K. Walker, C. Caruso, J. Wright, S. Strassel, "WeCanTalk: A New Multi-language, Multi-modal Resource for Speaker Recognition." *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 3451–3456, Marseille, France, June 2022, European Language Resources Association (ELRA)
- [13] O. Sadjadi, C. Greenberg, E. Singer, L. Mason and D. Reynolds, "The 2021 NIST Speaker Recognition Evaluation," *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, pp.322-329, Beijing, China, June 2022.
- [14] N. Ryant, "Linguistic Data Consortium Broad Phonetic Class Speech Activity Detector (ldc-bpcsd)," *Linguistic Data Consortium*, 2023. <https://github.com/Linguistic-Data-Consortium/ldc-bpcsd>
- [15] D. Snyder, MUSAN: A Music, Speech, and Noise Corpus, "JOUR", October 2015
- [16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5220-5224, doi: 10.1109/ICASSP.2017.7953152.
- [17] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia and S. Zafeiriou, "RetinaFace: Single-stage Dense Face Localisation in the Wild", May 2019, <https://doi.org/10.48550/arXiv.1905.00641>
- [18] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," arXiv:1801.07698, 2018.
- [19] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A Face Detection Benchmark," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 5525-5533, doi: 10.1109/CVPR.2016.596.
- [20] Y. Guo, L. Zhang, Y. Hu, X. He and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," Microsoft Research, 2016, <https://doi.org/10.48550/arXiv.1607.08221>
- [21] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the IARPA Janus program," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.