Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2024
A Regional Conference of the Society for Industrial and Systems Engineering

Multimodal Authentication

Tobias Hild[1], Patrick Moore[2], and Mike Powell[1]

[1] Department of Mathematical Sciences,
United States Military Academy,
West Point, NY

[2] Striveworks, Inc.,
Austin, TX

Corresponding author's email: mike.powell@westpoint.edu

Author Note: CDT Tobias Hild is a cadet at the United States Military Academy majoring in Applied Statistics and Data Science. Patrick Moore is a data scientist at Striveworks, Inc. LTC Mike Powell teaches in the Department of Mathematical Sciences at the United States Military Academy.

Abstract: Multimodal authentication offers the possibility of improvement in security and accuracy compared to single-modality authentication. Currently implemented approaches relying on facial identification alone are vulnerable to presentation and replay attacks. We propose a multimodal authentication system that utilizes facial recognition, automatic speech verification and recognition, and a comparison of the audio signal to face movement. This system demonstrates resilience against presentation and replay attacks by requiring a new, randomly generated password for each authentication attempt as well as facial verification, password recognition, voice verification, and synthetic media detection, which together prevent a video-only or audio-only presentation or replay attack. Using the GRID audio-visual speech corpus, we find that this system provides improved security against these types of attacks.

Keywords: Authentication, Multimodal Machine Learning, Biometrics

1. Introduction

Authentication can be accomplished with three different factors: something you know, something you have, or something you are. Something you know refers to knowledge-based authentication, such as passwords, PINs, or security questions. Something you have refers to possession-based authentication, such as a physical token, smart card, or mobile device. Something you are refers to biometric authentication, which includes unique physical or behavioral traits such as fingerprints, facial recognition, iris scans, or voice recognition. Combinations of these factors are commonly used in multi-factor authentication to enhance the security of systems and accounts. In this paper we will demonstrate the concept of multimodal authentication to enhance security for biometric authentication by combining different modes of biometric data. We will train a prototype authentication system on three modes of authentication: facial identification (FID), automatic speech verification (ASV), and automatic speech recognition (ASR). We propose a fourth modality to detect synthetic media by computing a similarity metric between the lip movements of the user and the audio of the user's authentication sample. We combine the outputs of these modules using logistic regression to reach an accept/reject decision for a collection of video samples. For all these modes we incorporate pretrained, open-source models to varying degrees to decrease the training time and computational expense of the system.

1.1 Literature Review

Facial recognition. Facial recognition and identification necessitate converting an image into a numeric encoding of facial geometry. These encodings can be compared numerically to a known facial geometry to determine whether the faces match. Modern approaches generally utilize recurrent neural networks to convert an image into an encoding, achieving human-level performance, or above 98% accuracy (Taigman et al., 2014) on the Labeled Faces in the Wild dataset, which is considered the benchmark dataset for FID. More recent research has focused on improving loss functions and training methods (Deng et al., 2022).

Speech recognition. The wav2vec automatic speech recognition (ASR) system used in this paper utilizes self-supervised learning to develop a numerical representation of an audio signal of speech (Baevski et al., 2020). This system is

trained on a large body of unlabeled speech data. During the self-supervised phase, the model learns to estimate future audio samples from preceding ones, enabling it to capture meaningful speech representations. Using a comparatively small amount of labeled data, a quantization model converts these representations learned on the unlabeled data to a vector of words with a word error rate (Klakow & Peters, 2002) of 1.8/3.3 on the Librispeech dataset (Panayotov et al., 2015). This approach is flexible for different languages and makes very efficient use of labeled data.

Audio-visual speech recognition. Recent innovations in ASR demonstrate that including video data can improve the accuracy of speech-to-text models, especially in environments where the audio signal is noisy or there are multiple, simultaneous speakers. There are generally two approaches which demonstrate this result: either analyzing audio and video components separately and combining the processed outputs into a single prediction of the spoken words (Afouras et al., 2022) or combining the audio and video into a single input vector before processing (Chao et al., 2016). When analyzing components separately, some approaches use the video signal to determine which parts of the audio signal belong to a given speaker by analyzing lip movements (Yu et al., 2020). Both approaches demonstrate significant reductions in word error rate compared to audio-only ASR in clean and noisy environments.

Automatic speech verification uses audio data to determine the speaker's identity. Deep learning models can be applied to preprocessed acoustic characteristics, the raw audio signal itself, or to speech representations learned through self-supervised models as seen with the wav2vec ASR model (Chen et al., 2022). A key difference from facial recognition is that audio data is presented as time-series data, so different model architectures such as time-delay neural networks are used (Desplanques et al., 2020). In some approaches, the audio signal is first converted into a spectrogram so that image analysis techniques can be applied (Park et al., 2019).

Lip-Based Biometric Authentication (LBBA). Our work shares similarities with lip-based biometric authentication (Koch & Grbić, 2024). This method utilizes a region of interest (ROI) centered around the lips of a user as input, processing it as a raw video signal to determine if two samples originate from the same speaker uttering the same phrase. This approach serves to authenticate a speaker using a predetermined password. Notably, their method distinguishes between error types, such as misidentifying the speaker or misinterpreting the spoken phrase, achieving false acceptance and false rejection rates of less than 4% on the GRID dataset. We intend to expand upon their methodology by incorporating multiple modalities, thus addressing a broader range of security vulnerabilities. Our approach differs from theirs in that we use a random password used only once rather than a predetermined password. We also use ASR to determine whether the password was said, and then we use FID and ASV to determine whether the correct person is speaking that password.

2. Methods

Our proposed system is initialized with several video samples from each speaker in our dataset, which are used to compute encodings for face and voice identification. In our proposed application, each user would be prompted to speak a randomly generated password. This generates a video sample which would be compared using FID, ASV, and ASR to an authorized user database and the prompted password. An overview of our approach is shown in Figure 1.

2.1 Data

We use the GRID dataset for training and validating all components of the authentication pipeline. This dataset consists of video recordings of 34 speakers with 1000 sentences per speaker. All sentences are six syllables long and syntactically identical, of the form "place green at B 4 now" (Cooke et al., 2006). For one speaker, only audio is available and is not usable for this project. Of the 33 speakers for which there is usable video data, 18 are male and 15 are female. The video recordings are 3 seconds long with 75 frames of video and 48000 samples of single channel audio; a transcript of the spoken sentence is attached to each video sample. While this dataset allows us to illustrate multimodal authentication using a straightforward, noise-free, and demographically homogeneous dataset, it also constrains the applicability of any system trained on it. Our results thus serve as an estimation of the upper bound of our system's effectiveness.

2.2 Facial Recognition

For every authorized user in the dataset, we initialize the facial recognition module with five images of the user, taken at random from the videos in the dataset. These videos are excluded from later training data. We encode each image using FaceNet-512 encoding (Schroff et al., 2015), which results in a vector of length 512 for each image. To compare this encoding with a video input, we take the encoding of the face in each frame of the video sample. We compare encodings from the input video frames to the encodings obtained during initialization by computing the mean cosine similarity.

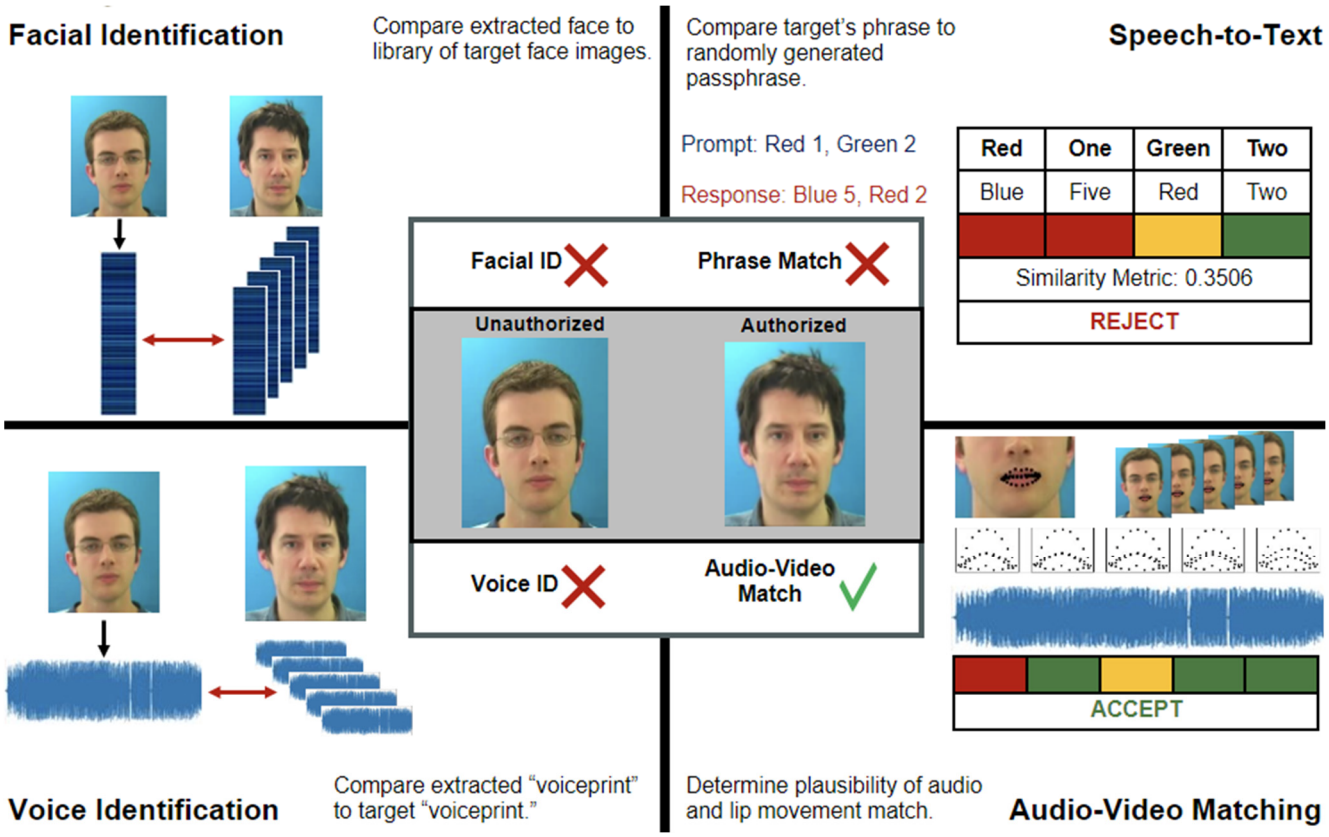

Figure 1: System Overview. The authentication system prompts the user to read a randomly generated passphrase and then evaluates the recorded video for face and voice identification, passphrase similarity, and audio/video compatibility.

2.3 Password Verification

To determine whether the user's video matches the prompted password, we compare the wav2vec transcript of the video to the password (Baevski et al., 2020). Since there are multiple ways in which a similar audio signal could be transcribed (e.g., maybe vs. may be vs. May bee), we utilize the double metaphone algorithm (Philips, 2000) to transcribe both the password and the video transcript to a common phonetic alphabet (metaphones). We measure the similarity between the password and the input metaphones using edit distance, Jaccard similarity, and cosine similarity.

We use a normalization of the Levenshtein distance (edit distance) metric, which is defined as the number of one-character deletions, insertions, and substitutions required to transform one string into another string. We normalize by dividing the edit distance by the length of the longer string to obtain a value in the range [0,1] with 1 representing identical strings and 0 representing two strings with no common characters (Yujian & Bo, 2007).

We then compute Jaccard similarity (Niwattanakul et al., 2013) by dividing the size of the intersection of the two sets of metaphones by the size of the union of the two sets of metaphones. This method does not account for the position of characters in the set; thus, anagrams have a Jaccard similarity of 1.

Finally, we also compute the cosine similarity between the metaphone transcriptions of the password and the input video. Both metaphone strings are vectorized using the frequency of metaphones (Wu et al., 2010) in the union of sets of the metaphones (bag-of-words model). We take the cosine similarity between these representations.

To make a probabilistic determination of password match, we fit a logistic regression using the above similarity metrics as explanatory variables. We fit this model on a dataset of password matches and non-matches. The non-matches in the logistic regression training data are generated by randomly choosing a password and a spoken phrase from the passwords in the GRID dataset.

2.4 Speaker Verification

For every authorized user in the dataset, we initialize the facial recognition module with five video samples of the user, corresponding to the images which were used to initialize the facial recognition module; these videos are excluded from later training data. We use a pretrained, open-source model from SpeechBrain for speaker verification (Ravanelli et al., 2020/2024). We use an ECAPA-TDNN model (Emphasized Channel Attention, Position-Aware Temporal Convolutional Neural Network) (Desplanques et al., 2020) that was trained on the VoxCeleb dataset (Nagrani et al., 2017). The system produces an encoding for each speaker corresponding to characteristics of speech such as intonation, pitch, and tempo. We calculate an overall similarity metric based on cosine similarity.

2.5 Audio-Video Comparison

We implement a simple check to determine whether a video matches its accompanying audio signal. For every frame of video, we extract a mediapipe representation of facial landmarks (Savin et al., 2021); this method creates a numerical representation of facial geometry, with every number corresponding to the relative position of a certain facial feature. We use an autoencoder to reduce the dimensionality of the lip-position portion of this representation to a single scalar for a frame of video. This encoding is generally on the range (-0.3, 0.3) with lower encoded values corresponding to a more closed-lip position.

Next, we train a linear feed-forward neural network to predict the lip position encoding using the audio signal of a frame of video as input. To determine whether a video is "live" (i.e., the audio and video signal align), we compare the vector of encodings of the video frames and the vector of encodings of the audio signal and take the mean squared error (MSE) of these vectors. Generally, these MSEs are in the range (0, 0.1), with larger MSEs corresponding to a higher probability of samples where the video and audio are from different samples.

2.6 Decision

To reach an accept/reject decision for each video sample, we use a logistic regression with the similarity scores from each modality as input. For similarity scores on the range (0,1), we can interpret the magnitude of the fitted coefficient for each modality as the importance of that modality. We fit the logistic regression on a training dataset sampled from a balanced dataset of control, replay, impersonation, and presentation samples, and it is important to note that these coefficients are dependent on the composition of the training set. For an application, it is therefore imperative that the training set is representative of the attacks which the application is expected to face. Our training dataset is likely not ideal for an implemented system since it seems unrealistic that any authentication system would face more attacks than legitimate authentication attempts and that these attacks would be balanced across several attack types; we are merely trying to demonstrate that a multimodal authentication system can provide security against these attacks.

The ease of interpretation of logistic regression coefficients helps an end user appropriately adjust the settings of the decision function; for example, if the system administrator believes that face presentation or replay attacks are unlikely, they might increase the importance of the facial identification module since this module's output is likely to correspond to the correct decision. In general, a multiple logistic regression, which combines the effects of the modality similarities, is desired since a low similarity score in any one modality will likely result in a rejection decision. This characteristic of the decision function provides security against attacks which appear legitimate in all but one modality. The ability to add terms for additional modalities is another advantage of logistic regression. In practice, this means that other authentication modalities with probabilistic outputs could be added to such a multimodal system (e.g., fingerprinting or retinal scans).

2.7 Evaluation

To test the system, we created four types of video samples, shown in Table 1. The first is a control sample, which is a sample that should be authenticated because the password, face, and voice match the authorized user. The second is an impersonation where a person claims to be authorized user but is not and speaks the prompted password; in this case, the password matches while the face and voice do not. To generate an impersonation sample, we start with a video and choose a different authorized user at random from the dataset. The third type of sample is a replay attack in which an old video of an authorized user is played during which the user is unlikely to state the correct password; we generate this sample by taking a video and testing it against a password taken at random from a different sample. It is possible that the randomly chosen password

is phonetically similar to the spoken password; this is a vulnerability that should be addressed before implementing any such system, for example by randomizing the length of passwords. The fourth type of sample is a presentation sample, where an old video of an authorized user is presented, and an unauthorized user speaks the correct password simultaneously. In this attack, the facial ID matches the authorized user, the voice ID does not match, the password is correct, and the audio and video are not aligned. To generate a sample for a presentation attack, we combine a video and a randomly chosen audio sample, and we set the password to be the password from the audio sample. For each video in the dataset, we generate a control sample and a sample of every type of attack.

Table 1. Tested Attack Types. Video samples for each attack type combine samples from the GRID corpus.

Attack Type	Facial ID Match	Voice ID Match	Password Match	Audio-Video Match
Control	Yes	Yes	Yes	Yes
Impersonation	No	No	Yes	Yes
Replay	Yes	Yes	No	Yes
Presentation	Yes	No	Yes	No

3. Results

3.1 Password Verification

Each of the password similarity metrics performs well on their own; however, each has certain limitations. For example, Jaccard and cosine similarity cannot distinguish between anagrams, which is reflected in a smaller difference of means in Table 2. However, by leveraging the strengths of each of these similarity metrics through a logistic regression, we can achieve a very accurate system when comparing wav2vec generated phonetic transcripts of GRID video samples to the known transcripts of the same video. The results in Table 2 can be interpreted as showing that these metrics can distinguish between cases where the correct password is spoken vs. random sentences of the same length.

Table 2. Password Similarity Metrics. Three metrics (edit, Jaccard, and cosine) serve as inputs to a logistic regression. Here we show the mean and standard deviation of similarity scores using each password similarity metric.

Metric	Match mean (SD)	Non-match mean (SD)	Difference of means	t-statistic (p-val.)	F1 Score
Edit similarity	0.911 (0.091)	0.330 (0.142)	0.581	267.89 (<0.001)	0.969
Jaccard similarity	0.951 (0.090)	0.603 (0.147)	0.348	157.76 (<0.001)	0.840
Cosine similarity	0.974 (0.040)	0.681 (0.146)	0.293	141.99 (<0.001)	0.926
Logistic regression	0.976 (0.092)	0.047 (0.142)	0.933	435.92 (<0.001)	0.972

3.2 System Performance

All the modules were tested on modality-specific outcomes as shown in Table 3. That is, we compare the distributions of password, facial identification, audio identification, and audio-video comparison scores for samples where each modality should support authentication and samples where each modality should not support authentication (e.g., facial identification similarity scores where the sample shows the correct face vs. samples where the face is incorrect, etc.).

Table 3. Module Performance. The performance of each of the four modules was evaluated by comparing mean similarities for matches vs. non-matches, as well as by computing an F1 score associated with modality-specific models.

Module*	Match mean (SD)	Non-match mean (SD)	Difference of means	t-statistic (p-val.)	F1 Score
Password	0.976 (0.092)	0.047 (0.142)	0.933	435.92 (<0.001)	0.972
Facial Identification	0.733 (0.147)	0.192 (0.161)	0.540	204.90 (<0.001)	0.979
Audio Identification	0.739 (0.072)	0.128 (0.155)	0.612	288.35 (<0.001)	0.991
Audio-Video Comp.	0.040 (0.020)	0.060 (0.020)	0.020	28.97 (<0.001)	0.251

ISBN: 97819384962-4-0

(MSE of encodings)					

*Note that each of these performance comparisons is for matches and non-matches for that modality only.

Although there is a statistically significant difference between the calculated MSEs of video samples where the audio and video signal align (see Table 3), this modality is not useful on its own to make an authentication decision as it does not correspond to one of the three factors of authentication. Additionally, the results of the audio-video comparison can be completely determined by the results of the facial identification and voice identification modules. If the outputs of these modules do not align (if the voice is correct but not the face, or vice versa), then the audio-video comparison should also return a decision not to authenticate. This module is intended to serve as a check to determine that an authentication attempt comes from a live individual; we can also fulfill the live-check function of the authentication system with the password module, since it requires a new password for every attempt. The utility of this module would come from detecting synthetic media (deepfakes); as we will discuss later, we believe that this function could be better accomplished by a generative adversarial network. Therefore, we exclude this module from our further analysis.

As expected, the modalities which incorporate open-source, pretrained models perform very well at their intended tasks. Our experiment serves to combine the different modalities to provide security against attacks which exploit a weakness in a single-modality authentication system. As shown in Table 4, single-modality authentication systems appear vulnerable to various attack methods. However, by combining modality similarity scores in a logistic regression, we can achieve a relatively high success rate on all types of generated attacks.

Table 4. Attacks Tested. The results for three attack types and the control setting show the ability of this multimodal authentication approach to correctly accept authorized users while correctly rejecting various types of attacks.

Attack Type	FID Only F1 Score	ASV Only F1 Score	PWD Only F1 Score	System Accept rate	System Reject rate	System F1 Score
Control	0.468*	0.729*	0.994*	98.52%	1.48%	*0.996
Impersonation	0.998	1	0.020	0.16%	99.84%	1
Replay	0.817	0.616	0.986	2.49%	97.51%	0.998
Presentation	0.903	0.999	0.030	0.11%	99.89%	0.993

*Note that to compute the F1 scores for control samples, we re-leveled our response variable to be positive when the sample should be rejected.

We hypothesize that the lower reject rate for replay attacks comes from the fact the some of the passwords in our dataset are similar. Therefore, there is a random chance that the replayed password is similar to the prompted passwords; in other words, our random-password-used-once system should be improved by making passwords less similar.

4. Discussion

4.1 Known Limitations

Before implementing this kind of system for any operational purpose, each module and the combined system should be trained on a dataset that is representative of the population that will be interacting with the system. We trained our proof-of-concept system on a small sample of British university students; thus, it would be inappropriate to deploy it before fine-tuning on a general population. The training population should be representative of any potential target population in terms of attributes that could impact the function of any module of the final system. These attributes should certainly include sex, age, ethnicity, language ability, and other demographic attributes.

The training dataset should also be representative of the expected attacks the system could face. As discussed in section 2.6, the composition of the training dataset will impact the final parameters of the decision function and should therefore be designed with the specific use case in mind. The audio and video noise levels will affect the performance of the individual modules, and by training these modules in the presence of real-world noise, the decision function would likely require less certainty from each module, leading to changes in overall performance requiring further study. Therefore, with a

specific use-case in mind (e.g., a physical gate at the entrance of a compound), the training dataset should be adjusted to reflect the expected noise levels.

The accuracy of the audio-video comparison module is also unsatisfactory for implementation as a production authentication system; we discuss various methods of further addressing the problem of dealing with synthetic media below. One of the most significant problems with this module is that its effectiveness differs significantly by speaker. We hypothesize that this difference comes from the fact that different speakers vary their lip positions to different extents when speaking, and therefore without finetuning on individual users, our module cannot differentiate between a speaker with small lip moment and a video where the audio signal and lip movement do not align.

4.2 Future Directions of Research

There are several alternative methods that could be used to compare the audio and video components of an input video. A frame-by-frame comparison between audio and video could be implemented using spectrograms instead of raw audio input. Considering multiple frames of audio and video would likely also result in a more accurate check.

Comparing audio to video helps determine whether a video is synthetic or an authentic recording. A generative adversarial network (GAN) would likely be useful to address this problem (Aldausari et al., 2022). Methods of implementing a GAN might include a model which is trained to distinguish between synthetic and live videos; however, this would be computationally expensive and difficult to implement. A simpler GAN implementation could be used just to validate lip positions against an audio signal.

Koch et al. introduce the value of hard-negative mining in their work on lip-based biometric authentication (Koch & Grbić, 2024). As an example of hard-negative mining, we could examine the success of the password module on samples where the spoken phrase is an incorrect but similar password to the prompted password. We could define a difficulty metric for every authentication sample and train the model on progressively more difficult samples. We could also use a GAN to deliberately create more difficult samples. The next iteration of this experiment should include samples that are more intentionally constructed rather than randomly generated.

5. Conclusion

The system outlined above demonstrates that adding authentication modalities can decrease the error rate of an authentication system against certain attacks. Multimodal authentication can also protect systems against attacks which might only appear in certain modalities; for example, a video replay attack might only be preventable by requiring a password module, and an image presentation attack might be prevented by a voice identification module. There are significant obstacles which must be overcome before implementing such a system; these obstacles vary by use case and the selected modalities, but at a minimum developing any system of this type requires preparing a training dataset which is demographically representative of the user population and contains samples that adequately represent the variety of attacks the system is likely to face.

A continuing concern for authentication systems is synthetic media. We demonstrated a very rudimentary module for comparing lip movement to audio signal to detect synthetic media where the audio and video components do not align. However, more advanced synthetic media, already publicly available to some extent, will be undetectable by this method. Future work should extend the capabilities developed here with an emphasis on the growing threat of synthetic media.

6. References

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2022). Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8717–8727. https://doi.org/10.1109/TPAMI.2018.2889052

Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Video Generative Adversarial Networks: A Review. *ACM Computing Surveys*, *55*(2), 30:1-30:25. https://doi.org/10.1145/3487891

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Chao, G.-L., Chan, W., & Lane, I. (2016). Speaker-Targeted Audio-Visual Models for Speech Recognition in Cocktail-Party Environments. *Interspeech 2016*, 2120–2124. https://doi.org/10.21437/Interspeech.2016-599

Chen, Z., Chen, S., Wu, Y., Qian, Y., Wang, C., Liu, S., Qian, Y., & Zeng, M. (2022). Large-Scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6147–6151. https://doi.org/10.1109/ICASSP43922.2022.9747814

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5 Pt 1), 2421–2424. https://doi.org/10.1121/1.2229005

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2022). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 5962–5979. https://doi.org/10.1109/TPAMI.2021.3087709

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Interspeech 2020*, 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, *38*(1), 19–28. https://doi.org/10.1016/S0167-6393(01)00041-3

Koch, B., & Grbić, R. (2024). One-shot lip-based biometric authentication: Extending behavioral features with authentication phrase information. *Image and Vision Computing*, *142*, 104900. https://doi.org/10.1016/j.imavis.2024.104900

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. *Interspeech 2017*, 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard Coefficient for Keywords Similarity. *Hong Kong*.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, *18*(6), 38–43.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2024). *SpeechBrain* [Python]. https://github.com/speechbrain/speechbrain/ (Original work published 2020)

Savin, A. V., Sablina, V. A., & Nikiforov, M. B. (2021). Comparison of Facial Landmark Detection Methods for Micro-Expressions Analysis. *2021 10th Mediterranean Conference on Embedded Computing (MECO)*, 1–4. https://doi.org/10.1109/MECO52532.2021.9460191

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Wu, L., Hoi, S. C. H., & Yu, N. (2010). Semantics-Preserving Bag-of-Words Models and Applications. *IEEE Transactions on Image Processing*, *19*(7), 1908–1920. https://doi.org/10.1109/TIP.2010.2045169

Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., & Yu, D. (2020). Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6984–6988. https://doi.org/10.1109/ICASSP40776.2020.9054127

Yujian, L., & Bo, L. (2007). A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1091–1095. https://doi.org/10.1109/TPAMI.2007.1078