Fig. 1: The architecture of the speaker identification systems tested in our comparative analysis where several combinations of feature extractions techniques and models were compared.

Fig. 2: Dataset preprocessing levels that we considered for audio recordings.

Fig. 3: The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction process.

Fig. 4: The sliding window approach for real-time speaker identification.

Fig. 5: The CNN architecture used for training spectrograms of one-channel and three-channel.

Fig. 6: The CNN architecture used for training spectrograms of three-channel with VGG16 as a baseline model.

Fig. 7: The CNN architecture used for training MFCC coefficients as a vector of frames.

Fig. 8: The CNN architecture used for training MFCC coefficients as a mean vector.

Fig. 9: A confusion matrix that shows 100% accuracy in the same-settings testing using GMM and MFCC.

Fig. 10: The relationship between training dataset size and the GMM model’s accuracy when trained and tested with 22 MFCCs that are extracted from 3-second utterances recorded in both same-settings and different-settings recordings.

Fig. A.1: The tab of speaker enrollment from a dataset of speakers of our Speaker Identification Application.

Fig. A.2: The tab of training speakers on the fly of our Speaker Identification Application.

Fig. A.3: A screenshot from the demo showing the tab of real-time speaker identification of our Speaker Identification Application while recognizing the speaker.

Fig. A.4: The format conversion tab of our developed Audio Processing Toolkit.

Fig. A.5: The silence removal tab of our developed Audio Processing Toolkit.

Fig. A.6: The segmentation tab of our developed Audio Processing Toolkit.

Fig. A.7: The trimmer tab of our developed Audio Processing Toolkit.

Fig. A.8: The noise reduction tab of our developed Audio Processing Toolkit.

Fig. A.9: The audio normalization tab of our developed Audio Processing Toolkit.