C3 A Text-Independent, Automatic Speaker Recognition System Evaluation With Males Speaking Both Arabic and English

Safi S. Alamri, MS, PO Box 91904, Riyadh 11634, SAUDI ARABIA*

After attending this presentation, attendees will better understand how to compare text-independent samples of speakers using different languages against a single-language reference population.

This presentation will impact the forensic science community by illustrating how conducting a study to better understand a design may be beneficial in further developing software that can complete accurate, text-independent, automatic speaker recognition for bilingual speakers against a single-reference population. All samples were taken from a text-independent speaker recognition system and enhanced to optimal performance. The data obtained were processed by a BATVOX 4.1, which deploys the Mel Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Model (GMM) methods of speaker recognition and identification. The results of testing through BATVOX 4.1 was a likelihood ratio for each sampled voice that was evaluated and the problems experienced. This presentation provides a brief overview of the area of text-independent, automatic speaker recognition system evaluation with males speaking both Arabic and English. The purpose is to compare text-independent samples of speakers using different languages against a single language reference population.

Automatic speaker recognition can be classified into speaker identification and speaker verification. Speaker identification deals with determining who the speaker is in the provided sample. The speaker usually lacks identity, so it is assumed the unknown speaker must come from a set of known speakers fixed from the system. Speaker verification deals with determining whether or not the speaker is the claimed person. In this research, the emphasis is on text-independent automatic speaker recognition; however, the process can be classified as either text-dependent or text-independent. But it depends on the cooperation of the involved parties and the available data. The text-dependent application is designed to identify the speaker through the recognition systems regardless of what they say.[1]

Automatic speaker recognition has several applications, both commercially and forensically. Some of the commercial applications entail telephone banking, voice mail, prison call monitoring, voice dialing, and biometric authentication.[2] The emphasis on the current study is on forensic applications, which includes systems such as BATVOX. It can be applied to both investigative and evidential purposes. The systems have two main processes, which are feature extraction and classification. Feature extraction takes small portions of samples that will be stored and used later for speaker identification. The most common technique for feature extraction is the MFCCs.[3] Classification is a two-phase process that starts with speaker modeling, which is the features of a new speaker, then uses speaker matching, which includes the features saved in a database.[4]

For text-independent applications, there must be a speaker model in place. The speaker model is a recognition system that has trained speaker samples stored in a database that used acoustic feature vectors extracted from each trained samples comparison to any given sample. This is what allows the text-independent application to have no restrictions on the words the speaker can use, but also makes it a more challenging method of automatic speaker recognition because of differences in linguistic content and potential phonetic mismatch.[5] Many models are available, such as Hidden Markov Model, MFCCs, Vector Quantization, GMM, Neural Networks, and Radial Basic Functions.[6]

In this study, the MFCCs and GMM-Universal Background Model (UMB) models are used. Nonetheless, MMFCs are the most notable features in automatic speaker recognition. The goal of the MFCC is to model the vocal tract's spectral envelope, consisting of the formants and a smooth curve connecting them, and using them as an identifier. This takes place by taking the spectral envelope and applying a filter based on human perception experiments, which applies filters to the spectral envelope and creates the spectrum known as the Mel-Spectrum. Cepstral transformation is then performed on the Mel-Spectrum, and the outputs are the MFCCs and speech then represented as a sequence of cepstral vectors.[7] The GMM is a weighted cumulative of the features observed from a sample when compared to the trained model, the outcome being the Log Likelihood (LL). The higher the value of the LL the higher probability that the mode and evidence are the same speakers. The GMM is a representation of the cumulative observed features from the speaker taken from the underlying model. Forensic automation speaker recognition was created to make it easier and more accurate to conduct speaker recognition. This involved creating an algorithm that then makes a quantitative analysis of the speech signal.[8]

Reference(s):

1. Reynolds D. (2002) An overview of automatic speaker recognition. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (S. 4072-4075).
2. El-Samie A., Fathi E. (2011). *Information Security for Automatic Speaker Identification*. Springer New York.
3. Drygajlo A. (2012). Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. *Forensic Speaker Recognition*. Springer New York. 21-39.
4. El-Samie A., Fathi E. (2011). *Information security for automatic speaker identification*. Springer New York.
5. Drygajlo A. (2012) Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. *Forensic Speaker Recognition.* Springer New York. 21-39.
6. Drygajlo A. (2012) Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. *Forensic Speaker Recognition*. Springer New York. 21-39.
7. Huang X., Acero A., Hon H.W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR.
8. Drygajlo A. (2012). Automatic Speaker Recognition for Forensic Case Assessment and Interpretation. *Forensic Speaker Recognition*. Springer New York. 21-39.

Automatic Speaker Recognition, Likelihood Ratio, Arabic and English