DOI:10.34918/87370

Adam DUSTOR[1]

10. VOICE BIOMETRIC SYSTEMS IN SMART CITIES

10.1. Introduction

Security and related issues are extremely important in modern information systems. Therefore, there is a huge demand for effective and reliable methods of identification and identity verification. The existing methods of verification based on specific knowledge (password, PIN number) or possessed equipment (bank card, key) are highly unreliable. The first requires memorizing and is susceptible to eavesdropping or guessing while the second must be properly stored and is vulnerable to loss, destruction, copying or theft. The third and most modern method is verification based on biometric features of the human body. Among the biometric features used today we can distinguish physical and behavioral ones. The most important physical characteristics are fingerprints, iris pattern, facial image, retina pattern, hand shape. Behavioral features are mainly voice and handwritten signature. There is no biometric feature that can provide flawless identity verification, therefore biometric methods are usually complementary to traditional methods.

Security systems based on physical features are usually more reliable and stable in time. Iris pattern or fingerprint does not change over time while behavioral biometric like human voice is severely affected by the speaker's emotional state or illness. As a result, the accuracy of voice based biometric systems is significantly lower than iris-based systems.

In near future, biometric systems will be very common in all information and security systems that are an inherent part of each smart city. In some countries, especially in PRC, such systems are almost everywhere. Face recognition implemented in majority of street cameras is used to recognize criminals that are sought by the police.

[1] Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science, Department of Telecommunications and Teleinformatics, e-mail: adam.dustor@polsl.pl.

Human voice as a security key

Automatic speaker recognition is a process that implements a series of decision rules on measurable features of speech signal to determine whether a given utterance belongs to a specific speaker or set of speakers. The need for voice-based person recognition occurs where no other way of checking or verifying the identity of people is possible, e.g., when communicating with a computer using speech, telephone, or radio communications. Other examples of automatic speaker recognition include data protection systems against unauthorized access, control of devices by voice by authorized persons, telephone shopping, access to computers, information services or databases. Automatic speaker recognition is also widely used in forensics and judiciary.

Automatic speaker recognition is usually divided into two tasks, namely automatic speaker identification and automatic speaker verification. Speaker identification is the process of assigning an unknown speaker utterance to a given speaker class from a specified set of M speakers. Voice verification is the process of confirming or rejecting the claimed identity based on a sample utterance of the speaker being verified. The basis for the acceptance or rejection decision is the comparison of the similarity function between the voice pattern and the recognized sample with a fixed threshold value. The issue of speaker verification can be treated as a recognition process with two classes: YES – belonging to the set, and NO – not belonging to the set of speakers. Speaker verification can also be considered as a special case of speaker identification in an open set of speakers for M = 1.

Speaker recognition systems can be divided also into two important classes, namely text-dependent and text-independent systems. In text-dependent systems, it is required that the linguistic content of the test utterance is the same as that previously recorded during system training. The consequence of this is the necessity to utter specific passwords during the recognition procedure and their frequent repetition in case of difficulties with user recognition. Due to the type of password, they can be divided into systems with personal password and common password. Despite being more reliable and simpler than text-independent systems, these systems have very serious disadvantage of not being immune to an impostor's reconstruction of a speaker's previously recorded password. Since people recognize speakers independently of their utterances, text--independent systems, in which the test utterance may differ from the training utterance and may be prompted by the system, are of greater interest to researchers.

Text-independent systems are divided into:

- fixed vocabulary – during training, users must utter a fixed set of words, the order of which randomly changes during each user testing;

- event-dependent – the test utterance contains a linguistic event that is extracted from the speaker utterance;
- unrestricted – there are no restrictions on the test utterance.
- Regarding the method of communication with the user, we can divide them into:
- systems with text prompting – the system displays a text on the screen that the user has to read;
- systems with voice prompting – the user has to repeat the statement heard;
- systems without prompting – using spontaneous speech.

Because prompted systems require the user to repeat a different utterance each time, they are very effective in preventing the possibility of replaying a previously recorded user voice. However, in the case of these systems, it is required to check whether the user said what was asked, so they must also implement speech recognition.

10.2.1. Speaker recognition – a general idea

The block diagram of the automatic voice identification system is shown in Figure 10.1, while the verification system is shown in Figure 10.2. In both cases, the decision-making process uses rules based on the calculation of similarity function values to patterns taken from the pattern bank. In the case of identification, the recognition result is the class for which the calculated similarity value to a given pattern is the highest. In the case of verification, the similarity is calculated only for one pattern (the one whose identity is declared) and an additional decision is required to accept or reject the speaker.

In speaker identification, the only error that can occur is the misidentification of the speaker voice. The parameter characterizing the accuracy of recognition is the identification rate in %. As the number of speakers registered in the system increases, the probability of wrong identification also increases. This phenomenon is a serious obstacle on the way of building systems with many registered users.

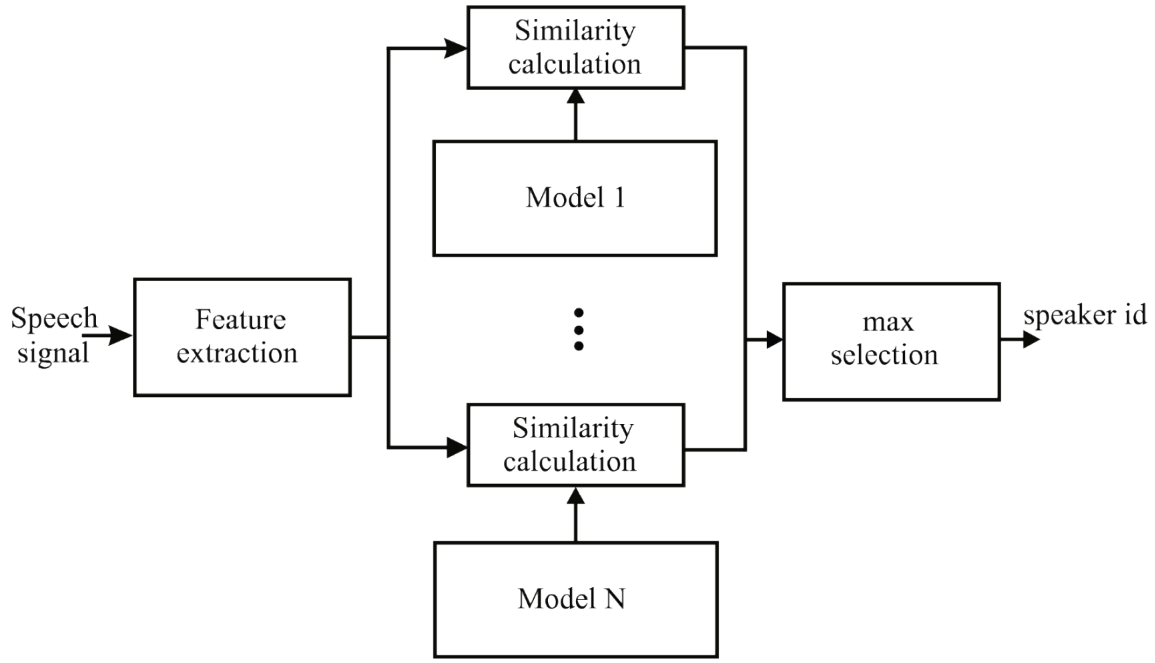

Fig. 10.1. Block diagram of speaker identification system
Rys. 10.1. Schemat blokowy systemu identyfikacji mówcy
Source: Own research.

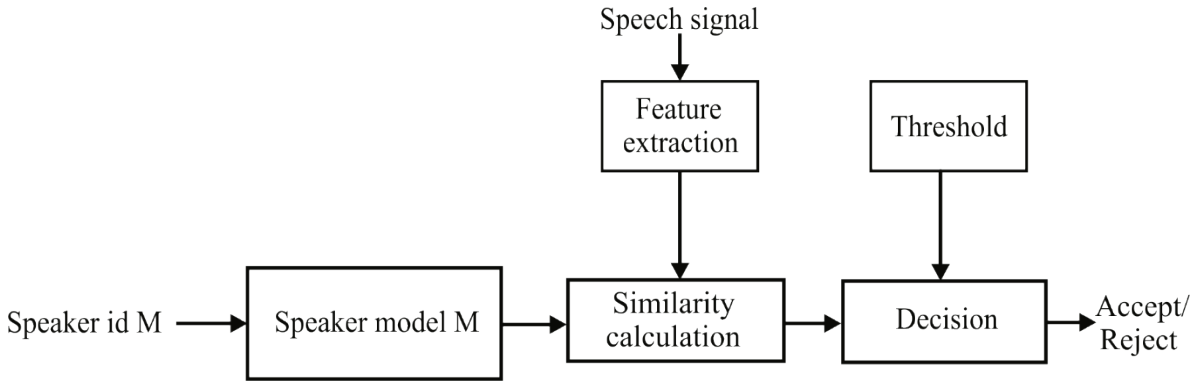

Fig. 10.2. Block diagram of speaker verification system
Rys. 10.2. Schemat blokowy systemu weryfikacji mówcy
Source: Own research.

In the process of verification, a test utterance is compared with only one model, the one whose identity is declared. As a result of such comparison, the utterance may be falsely rejected by the system – False Rejection error, or falsely accepted – False Acceptance error. The quality of the system is characterized by the probabilities of False Acceptance Rate (FAR) and False Rejection Rate (FRR). These errors depend on the value of the decision threshold[2]. Depending on the application of the speaker recognition system, it is advantageous to vary these probabilities and thus regulate the

[2] Dustor A.: Problematyka błędów w biometrycznych systemach rozpoznawania głosu, [in:] IC-SPETO 2011, Ustroń 2011, pp. 133–134.

security of the system. For example, in applications related to remote banking services, it is necessary to ensure a very low probability of accepting a fraudster even at the cost of a longer verification procedure and a higher probability of false rejection.

A full evaluation of the verification system is possible with the help of Detection Error Tradeoff (DET) curves[3,4] which provide a full description of all possible system operating conditions. With their help it is possible to determine the value of FAR if FRR is known, the value of FRR if FAR is known and the EER value (Equal Error Rate) corresponding to the threshold for which the false acceptance rate is equal to the false rejection rate[5].

10.2.2. Feature extraction

A fundamental issue that affects the effectiveness of a voice recognition system is the choice of the best measurable physical parameters of the speech signal associated with the recognized classes. The effectiveness of the recognition system will depend largely on to what extent the physical parameters of the speech signal capture the speaker's personal characteristics. The most used physical parameters of the speech signal in speaker recognition include:

- parameters determined directly from the time course such as relative utterance lengths of individual phonetic elements, time envelope of the amplitude or intensity of the sound, zero crossing rate of the speech signal, distribution of time intervals;
- parameters determined from the spectrum of the speech signal such as averaged amplitude spectrum, power spectrum, fundamental frequency, frequencies and amplitude ratios of formants, short-term spectrum, spectral moments;
- linear predictive coding parameters and their derivatives such as Linear Prediction Cepstrum Coefficien (LPCC) cepstral parameters and reflection coefficients.

The ideal parameters used for voice recognition should have high inter-speaker and low intra-speaker variability. Over the years, it has been shown that the lowest recognition error rates are provided by the means of Mel Frequency Cepstral Coefficients (MFCC) parameters[6] which take into account the nonlinear characteristics of human hearing.

---

[3] Dustor A.: Problematyka błędów w biometrycznych systemach rozpoznawania głosu, [in:] IC-SPETO 2011, Ustroń 2011, pp. 133–134.
[4] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M.: The DET curve in assesment of detection task performance, [in:] Proceedings of the EUROSPEECH 1997, Rhodes 1997, pp. 1895–1898.
[5] Dustor A.: Problematyka błędów w biometrycznych systemach rozpoznawania głosu, [in:] IC-SPETO 2011, Ustroń 2011, pp. 133–134.
[6] Rabiner L.R., Juang B.H.: Fundamentals of speech recognition, Prentice Hall, 1993.

Studies on the influence of parameters and their dimensionality on error rates include works[7,8]. The parameters LPC, LPCC as well as reflection coefficients k and MFCC were investigated. The ROBOT and TIMIT speech resources were used for the study. By far the lowest error rates were obtained for cepstral MFCC parameters, not much worse for LPCC parameters and by far the worst for LPC prediction coefficients.

Using GMM (Gaussian Mixture Models), 100% correct identifications is possible for a small set of speakers, it is only necessary to provide sufficiently long training and test utterances and to use MFCC parameters with sufficient dimensionality.

The paper[9] also investigated the effect of the speech corpora used on error rates. It turned out that this parameter is crucial when testing systems with low error rates. The results for the TIMIT[10] database (630 speakers) were almost four times worse than for the ROBOT[11] database (30 speakers). Reliable voice biometrics research requires large datasets to get results statistically significant.

10.2.3. Speaker models

The problem of voice modelling, similarly, to feature extraction, is a key stage on the way to building a low-error voice recognition system. As a result of many years of research on this issue, we have a very rich set of algorithms based both on generative approach (estimation of probability distribution) and discriminative approach (minimization of classification error on training set). Among the generative algorithms, the simplest approach is to use only the nearest neighbor algorithm. The speaker model then consists of all learning vectors. The similarity of a test sequence to such a model is calculated sequentially by finding for each test vector the nearest neighbor from the set of all learning vectors of a given speaker and calculating the distance between these vectors.

As the number of learning vectors can be large, for long learning sequences a more common approach is to use data clustering based on the k-means or LBG (Linde – Buzo – Gray) algorithm. This allows all learning vectors to be replaced by some small set of centroids. The vector quantization approach produces satisfactory results provided that

---

[7] Dustor A.: Speaker verification with TIMIT corpus – some remarks on classical methods, [in:] Proceedings of the 24th International Conference on Signal Processing Algorithms, Architectures, Arrangements and Applications (SPA), Poznań, Poland 2020.

[8] Dustor A., Kłosowski P., Izydorczyk J.: Influence of Feature Dimensionality and Model Complexity on Speaker Verification Performance, [in:] Communications in Computer and Information Science, Springer-Verlag, Berlin, Germany 2014, Vol. 431, p. 177–186.

[9] Dustor A., Kłosowski P., Izydorczyk J., Kopański R.: Influence of corpus size on speaker verification, [in:] Communications In Computer and Information Science, Springer-Verlag, Berlin Heidelberg, Germany 2015, Vol. 522, p. 242-249.

[10] TIMIT LDC93S1: https://catalog.ldc.upenn.edu/LDC93S1.

[11] Adamczyk B., Adamczyk K., Trawiński K.: Zasób mowy ROBOT, [in:] Biuletyn Instytutu Automatyki i Robotyki, WAT, 2000, Vol. 12, p. 179–192.

the speech quality is relatively high (no noise) and the number of speakers is small. Studies on the application of data clustering in voice biometrics include works[12] in which the effect of the number of centroids on the verification error rate was investigated.

A more complicated but also more effective approach (assuming there is enough learning material) is the use of statistical modelling where the voice model is a certain set of parameters. In the simplest version, these are parameters characterizing a single multivariate normal distribution. In a more complex case, the speaker is represented by a linear combination of M normal distributions, each of which is described by a vector of mean values μi, a covariance matrix Ci and a certain weighting factor pi. Such a model, called Gaussian mixture GMM, and its modifications have been for many years the so-called gold standard in speaker modelling. The disadvantages of GMM models include the necessity of estimation many parameters, especially in the case of models with full covariance matrices, and thus the frequent phenomenon of overfitting. Of the most important modifications, the GMM-UBM (Gaussian Mixture Model – Universal Background Model) algorithm, in which individual models are created from one reference model, should be mentioned first [13]. Another, even more advanced approach is to combine GMM models with a Support Vector Machine (SVM)[14].

Studies on voice recognition using GMM models include publications[15]. The research was performed for the TIMIT resource. It turned out that using MFCC cepstral parameters and GMM models with diagonal covariance matrices, it is possible to obtain EER error values for verification of the order of one percent or even below.

Another approach to voice modelling is the use of discriminative approaches. Examples of classifiers that have been applied in voice recognition are the SVM[16] and

---

[12] Dustor A., Kłosowski P.: Biometric voice identification based on fuzzy kernel classifier, [in:] Communications in Computer and Information Science, Vol. 370, pp. 456–465, Springer-Verlag, Berlin Heidelberg, Germany 2013; Dustor A., Kukiełka A.: Detekcja sygnału mowy w systemach rozpoznawania głosu, [in:] IC-SPETO 2012, Ustroń 2012, pp. 79–80.

[13] Reynolds D.A., Quatieri T.F., Dunn R. B.: Speaker verification using adapted gaussian mixture models, [in:] Digital Signal Processing, 2000, Vol. 10, pp. 19–41.

[14] Vapnik V.: The nature of statistical learning theory, Springer, New York 1999.

[15] Dustor A.: Speaker verification with TIMIT corpus – some remarks on classical methods, [in:] Proceedings of the 24th International Conference on Signal Processing Algorithms, Architectures, Arrangements and Applications (SPA), Poznań, Poland 2020; Kłosowski P., Dustor A., Izydorczyk J.: Speaker verification performance evaluation based on open-source speech processing software and timit speech corpus, [in:] Communications in Computer and Information Science, Springer-Verlag, Berlin Heidelberg, Germany 2015, Vol. 522, pp. 400–409.

[16] Dustor A., Bąk M.: Wykorzystanie maszyny wektorów podpierających w weryfikacji mówcy, [in:] Współczesne Aspekty Sieci Komputerowych, Tom 1, Wydawnictwa Komunikacji i Łączności, Warszawa 2008, pp. 299–308.

also KHK which is a kernel version of the classical Ho-Kashyap (HK) classifier[17]. The nonlinearity introduced with a Gaussian kernel function has allowed for very promising results when there is little learning data.

Fuzzy set theory can also be used for voice modelling. Fuzzy Ho-Kashyap (FHK) classifier[18] which is a kind of combination of linear HK classifiers based on fuzzy inference system allowed with the use of Kleene-Dienes implication to obtain lower EER error value than for GMM approach[19]. It is also possible to combine fuzzy approach with the kernel approach (kernel function). Example results for the Fuzzy Kernel Ho-Kashyap (FKHK) classifier with linguistic interpretation of the kernel matrix are given in[20].

Recent modelling approaches are based on the use of Joint Factor Analysis (JFA) and the so-called i-vector[21]. Deep neural networks[22] and the so-called x-vectors[23] based on them are also used in modelling. However, these methods require very large learning data sets to correctly estimate the parameters of these models.

10.2.4. Speaker verification system – implementation

In this chapter, an example of real speaker verification system is presented. Implementation of all signal processing procedures, speaker training and recognition was done with the help of an open-source software such as Sidekit Python package[24] and Anaconda[25] with separate environment with all required by Sidekit packages.

---

[17] Dustor A., Kłosowski P., Izydorczyk J.: Speaker recognition system with good generalization properties, [in:] 2014 International Conference on Multimedia Computing and Systems (ICMCS), Marrakech 2014, pp. 206–210; Dustor A.: Voice verification based on nonlinear ho-kashyap classifier, [in:] 2008 IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering, Novosibirsk 2008, pp. 296–300.

[18] Dustor A.: Speaker verification based on fuzzy classifier, [in:] Man-Machine Interactions, AISC 59, Springer--Verlag, Berlin Heidelberg 2009, pp. 389–397.

[19] Dustor A.: Speaker verification based on fuzzy classifier, [in:] Man-Machine Interactions, AISC 59, Springer--Verlag, Berlin Heidelberg 2009, pp. 389–397.

[20] Dustor A., Kłosowski P.: Biometric voice identification based on fuzzy kernel classifier, [in:] Communications in Computer and Information Science, Vol. 370, pp. 456–465, Springer-Verlag, Berlin Heidelberg, Germany 2013; Dustor A.: Application of fuzzy kernel ho-kashyap classifier to speaker verification, [in:] Proceedings of the 18th International Conference Mixed Design of Integrated Circuits and Systems – MIXDES 2011, Gliwice 2011, pp. 581–586.

[21] Dehak N., Kenny P.J., Dehak R., Dumouchel P., Ouellet P.: Frontend factor analysis for speaker verification, [in:] IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, 2011.

[22] Sztahó D., Szaszák G., Beke A.: Deep learning methods in speaker recognition: a review, [in:] ArXiv, https://arxiv.org/ftp/arxiv/papers/1911/1911.06615.pdf, 2019.

[23] Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S.: Xvectors: Robust dnn embeddings for speaker recognition, [in:] 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329–5333.

[24] Larcher A., Lee K.A., Meignier S.: An extensible speaker identification SIDEKIT in Python, [in:] International Conference on Audio Speech and Signal Processing ICASSP 2016.

[25] Anaconda software – www page: https://www.anaconda.com.

For the research on speaker verification, TIMIT[26] corpus was applied. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT has resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency – Information Science and Technology Office (DARPA-ISTO). Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and has been maintained, verified, and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. It may be obtained from Linguistic Data Consortium.

For the research, whole data of TIMIT corpus were used. Universal background model was obtained from the speech of the 130 speakers. Speaker models for the rest 500 speakers were adapted from the UBM via MAP (Maximum A Posteriori) procedure[27]. During training speaker's GMM[28] only mean supervector was adapted. As there are 10 utterances for each of 630 speakers, 6 were used for training and the remaining 4 for testing. As a result, there were 2000 (500·4) target trials and 998000 (499·4·500) nontarget trials for each combination of feature dimensionality and number of mixtures. Number of mixtures in GMM-UBM models was changed from 16 to 512 (diagonal covariance matrix). As a feature vector MFCC[29] parameters with supplemental delta features were used. Performance was examined for vectors with 5, 10, 15 and 20 features. All tests were conducted without voice activity detection. During training and testing the same signal processing procedure was used. Speech files were pre-emphasized and segmented into 25ms pieces. Verification performance was characterized in terms of the two error measures, namely the false acceptance rate FAR and false rejection rate FRR. These measures correspond to the probability of acceptance an impostor as a valid user (acceptance of the non-target) and the probability of rejection of a valid user (rejection of target). Varying the decision level, Detection Error Tradeoff (DET) curves, which show dependence between FRR and FAR, may be plotted. Another performance measure is an EER which corresponds to

[26] TIMIT LDC93S1: https://catalog.ldc.upenn.edu/LDC93S1.
[27] Reynolds D.A., Quatieri T F., Dunn R.B.: Speaker verification using adapted gaussian mixture models, [in:] Digital Signal Processing, 2000, Vol. 10, pp. 19–41.
[28] Reynolds D.A., Quatieri T.F., Dunn R.B.: Speaker verification using adapted gaussian mixture models, [in:] Digital Signal Processing, 2000, Vol. 10, pp. 19–41.
[29] Rabiner L.R., Juang B.H.: Fundamentals of speech recognition, Prentice Hall, 1993.

error rate achieved for the decision threshold for which FRR = FAR. In other words, EER is just given by the intersection point of the main diagonal of DET plot with DET curves.

Performance of the speaker verification system as a function of complexity of the speaker model is shown in Figure 10.3, Figure 10.4, Figure 10.5, Figure 10.6, for the feature vector of dimensionality 5, 10, 15 and 20 respectively. Achieved EER values for each model are depicted.

It may be observed that error rates of the speaker verification system are strongly dependent on the number of features extracted from each segment of speaker utterance and on the complexity of the voice model. For 5 features per frame of speech, the lowest EER value was approximately 5.28% for the GMM-UBM with 512 mixtures (Fig. 10.3) while for 20 features per segment of speech and 512 mixtures achieved EER was 1.04% (Fig. 10.6). Although these error rates are relatively high for the standalone security system, such biometric system may be implemented as an additional security level accompanying traditional system based on knowledge (PIN number) or equipment (bank card). After careful selection of decision threshold in such biometric system, it is possible for example in Figure 10.5 to achieve false acceptance rate 0.2% when FRR = 10% is acceptable. In other words, 1 out of 10 valid users may require re-verification but only 1 out of 500 impostors would be falsely accepted by the system.

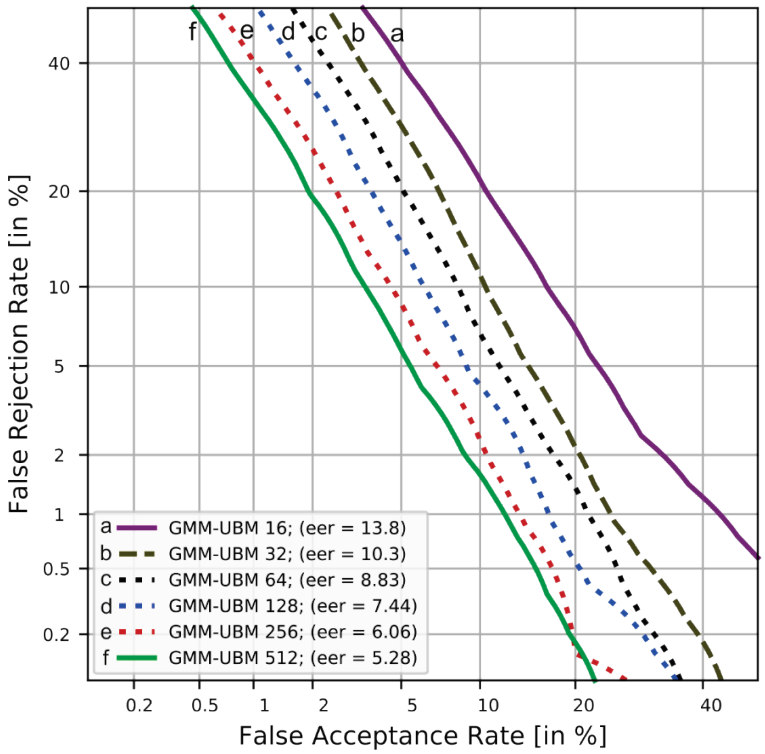

Fig. 10.3. Influence of model size on speaker verification (5 features)
Rys. 10.3. Wpływ rozmiaru modelu (5 parametrów na segment) na proces weryfikacji
Source: Own research.

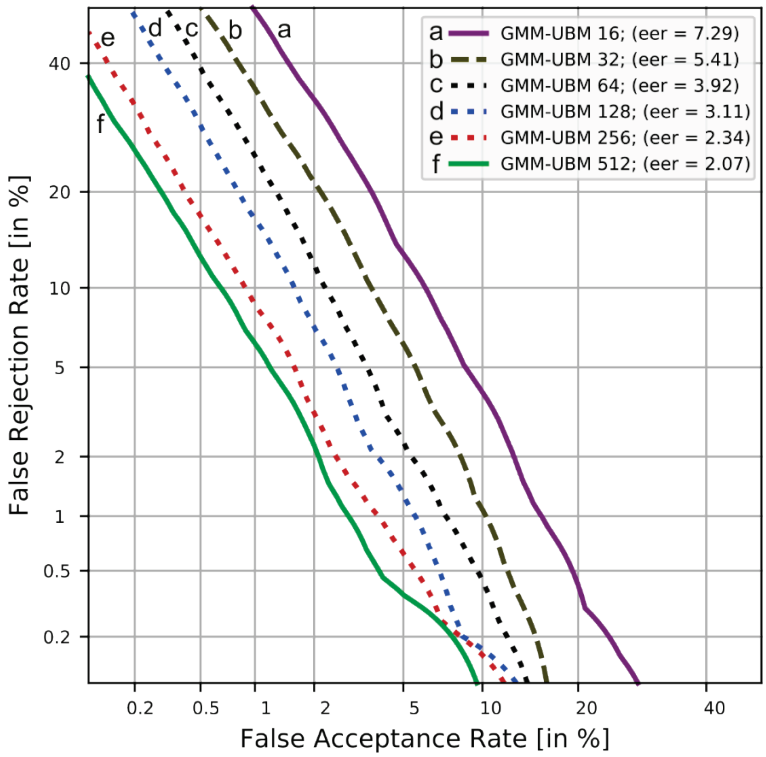

Fig. 10.4. Influence of model size on speaker verification (10 features)
Rys. 10.4. Wpływ rozmiaru modelu (10 parametrów na segment) na proces weryfikacji
Source: Own research.

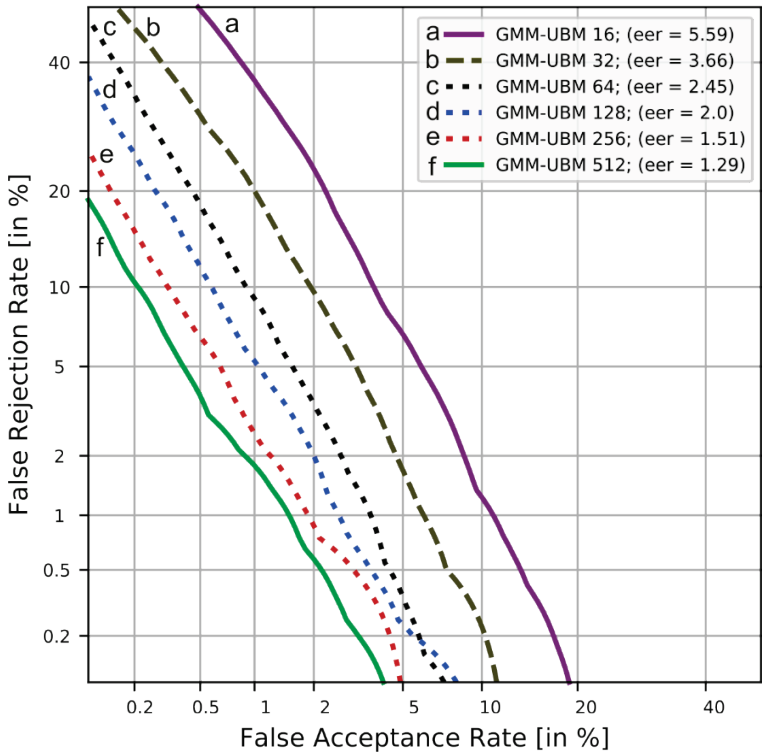

Fig. 10.5. Influence of model size on speaker verification (15 features)
Rys. 10.5. Wpływ rozmiaru modelu (15 parametrów na segment) na proces weryfikacji
Source: Own research.

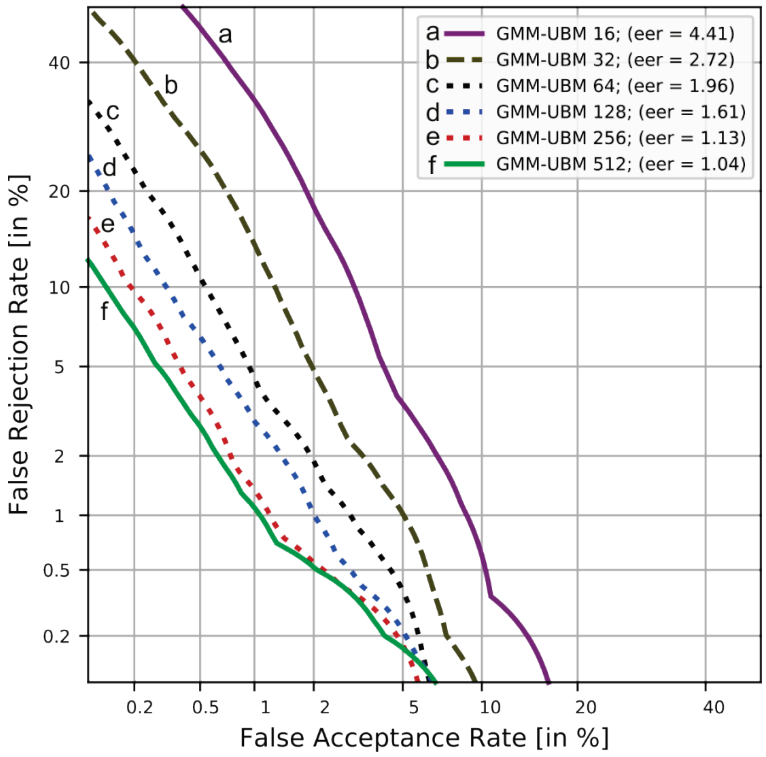

Fig. 10.6. Influence of model size on speaker verification (20 features)
Rys. 10.6. Wpływ rozmiaru modelu (20 parametrów na segment) na proces weryfikacji
Source: Own research.

10.3. Summary

In this study fundamentals of voice biometric systems were presented. The most important functional blocks of the recognition system, namely feature extraction and voice model construction were discussed. An example of real speaker verification system implemented in Python and its performance was also covered. It was demonstrated that careful selection of feature dimensionality and number of mixtures in speaker model may lead to significant reduction of verification errors, leading in the best scenario to EER value of 1.04%.

Additional security level provided by such biometric system may be implemented in many smart applications which would require user personalization or identity recognition. In many information systems, very common in future smart cities, such additional level of security will be of great importance.