DOI: 10.1109/CIVEMSA.2017.7995297

Text-Independent Speaker Recognition for Ambient Intelligence Applications by Using Information Set Features

Abhinav Anand*, Ruggero Donida Labati*, Madasu Hanmandlu†, Vincenzo Piuri*, Fabio Scotti*
*Department of Computer Science, University of Milan, Crema, Italy
Email: {abhinav.anand, ruggero.donida, vincenzo.piuri, fabio.scotti}@unimi.it
†Department of Computer Science and Engineering, MVSR Engineering College, Hyderabad, India
Email: mhmandlu@gmail.com

***Abstract*—Biometric systems are enabling technologies for a wide set of applications in Ambient Intelligence (AmI) environments. In this context, speaker recognition techniques are of paramount importance due to their high user acceptance and low required cooperation. Typical applications of biometric recognition in AmI environments are identification techniques designed to recognize individuals in small datasets. Biometric recognition methods are frequently deployed on embedded hardware and therefore need to be optimized in terms of computational time as well as used memory. This paper presents a text-independent speaker recognition method particularly suitable for identification in AmI environments. The proposed method first computes the Mel Frequency Cepstral Coefficients (MFCC) and then creates Information Set Features (ISF) by applying a fuzzy logic approach. Finally, it estimates the user's identity by using a hierarchical classification technique based on computational intelligence. We evaluated the performance of the speaker recognition method using signals belonging to the NIST-2003 switchboard speaker database. The achieved results showed that the proposed method reduced the size of the template with respect to traditional approaches based on Gaussian Mixture Models (GMM) and achieved better identification accuracy.**

***Index Terms*—Biometrics, Human-Computer Interaction, Speaker recognition, Text-independent, Computational Intelligence, Ambient Intelligence**

I. Introduction

Most of current biometric systems are designed for security applications, like automated border control [1]–[4]. A growing research area consists of designing biometric technologies to improve the Human-Computer Interaction (HCI) in Ambient Intelligence (AmI) environments. These technologies should be based on less-constrained technologies with respect to traditional biometric systems [5]–[8].

Voice recognition provides a true unobtrusive HCI method. Voice recognition applications can be divided in two categories, namely: speaker recognition (which aims to recognize the user based on her voice) and speech recognition (which aims to recognize what is said). Speech recognition technologies are widely used in HCI for AmI environments [9], [10]. On the other hand, speaker recognition is mostly used for authentication purposes [11]. Studies in the literature use the speech-based interaction between human and computers to facilitate the users inside their home [12]–[15]. The work in [13] presents an AmI environment based on speech and speaker recognition. The authors deployed the proposed technologies in the domotic system named STARHome, which is a functional prototype. Many commercial applications use speech recognition for HCI technologies designed for AmI environments, such as Apple Siri, Ubi and Amazon Echo, etc. The work proposed in [16] presents a list of commercial applications using speech for HCI in AmI environments.

Speaker recognition systems can be classified into text-dependent and text-independent [17]–[19]. The first class requires that the user enunciates a specific set of words, while the second class does not impose this limitation. Text-independent speaker recognition systems are usually less accurate than text-dependent systems. Nevertheless, text-independent systems are based on a natural and unconstrained HCI modality and are therefore suitable for user-friendly application scenarios.

In AmI environments, it is frequently needed to identify the users in relatively small closed-sets by using biometrics. There are two main categories of closed set identification systems: systems performing multiple identity comparisons, and systems that search the identity of the user by classifying a single biometric template. Fig. 1 shows the schema of the two types of identification systems. The first category has the advantage of being more scalable since the enrollment of new users does not require to train any classifier. Nonetheless, having a biometric database of N identities, these systems need to compute N identity comparison to estimate the identity corresponding to the fresh template. Differently, the second category of identification systems estimates the final result by performing a single classification, thus requiring less computational time and resources. However, enrolling a new user requires to re-train the classifier.

Biometric recognition algorithms for AmI environments are frequently deployed in embedded systems, characterized by reduced computational resources with respect to general purpose architectures. Therefore, these algorithms should be optimized in terms of computational time and memory. Specifically, identification applications should be based on fast feature extraction algorithms and use templates of limited size. The computational limitations also justify the choice of

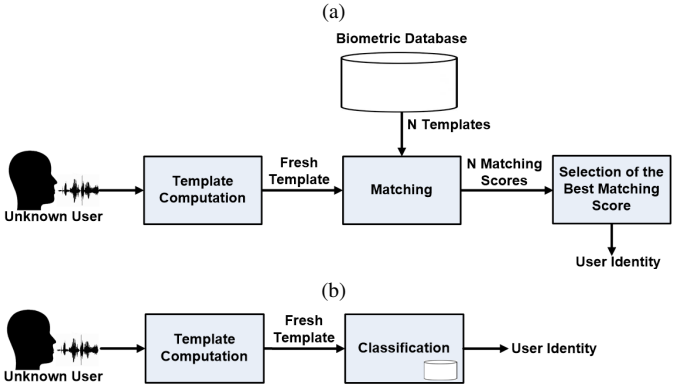

Fig. 1: Schemas of biometric identification systems: (a) systems performing multiple identity comparison, and (b) systems that search the identity of the user by classifying a single biometric template.

identification systems based on classifiers for a wide range of applications.

This paper presents a novel text-independent speaker identification method particularly suitable for AmI environments. The method requires limited computational resources since it uses a classification-based approach and computes templates composed of only 12 floating-point numbers.

Our method first computes the Mel Frequency Cepstral Coefficients (MFCC) [17], [19]–[21] and then creates Information Set Features (ISF) [22]–[24] by applying a fuzzy logic approach. It estimates the identity corresponding to the computed template by using a hierarchical trained classifier based on computational intelligence techniques.

The main novelty of this paper consists of the proposed method for computing ISF templates for speaker recognition, which are composed by only 12 floating-point numbers, representing the first 12 cepstral coefficients. Compared to the size of the templates used by other methods in the literature, the size of the proposed templates can be considered as particularly small.

We evaluated the performance of our speaker recognition method using signals from NIST-2003 switchboard speaker database [25] and compared the achieved accuracy with that of traditional approaches based on Gaussian Mixture Models (GMM). Our approach obtained better performance by using smaller templates.

This paper is organized as follows. Section II briefly overviews the state of the art for speaker recognition. Section III describes the theory on IFS and presents the proposed speaker recognition method. Section IV analyzes the performed experiments and the achieved results. Finally, Section V concludes the work.

II. State of the Art

Speaker recognition systems estimate the identity of a person based on her speaking utterances [11], [19]. In text-dependent speaker recognition systems, the phrases spoken is matched with the same enrolled phrase. These systems consider the feature dynamics of the words for identification. The most common modeling techniques for text-dependent speaker recognition are the Hidden Markov Models (HMM) [26] and Dynamic Time Warping (DTW) [27].

Text-independent systems pose no restrictions on the phrases spoken, hence these systems do not consider the feature dynamics and process the feature vector as a bag of symbols. In this kind of systems, the speakers are frequently modeled by using the Gaussian Mixture Model (GMM) [17], [21] or Vector Quantization (VQ) [28]. GMM requires a large amount of training data to create the speaker model and to estimate a set of distinctive parameters (mean, variance, and weights related to each speaker). VQ clusters the speaker data by using k-means clustering. Each cluster is represented by a code that denotes the centroid of the clusters. The set of codes are known as codebooks, which are used to model the individuals. Universal Background model (UBM) [29], [30] is an another technique to model the speaker distribution, which is generally used for verification purpose. Usually, it uses a very large GMM trained to represent speaker-independent datasets. Other approaches in literature use Support Vector Machines based on GMM [31] or Artificial Neural Networks [18].

The features corresponding to the speech signal represent differences between the vocal traits of sets of individuals, which are frequently described using the frequency spectrum of the signal [17], [21]. The Mel-Frequency Cepstral Coefficient (MFCC) is one of the mostly used feature extraction techniques [17], [19]–[21]. MFCC is a filterbank-based approach designed to resemble the human auditory frequency perception. Other feature extraction methods are: delta-MFCC and delta-delta MFCC [32], linear predictive cepstral coefficients [33], perceptual linear prediction [34], coefficients cepstral mean and variance normalization [35], relative spectral transform filtering [36], feature warping [27], i-vectors and super-vectors [26].

Speaker recognition systems based on deep learning have recently been proposed [37], [38]. The advantage of deep learning is that the system can learn discriminative features from the raw input signal. Studies have shown that deep learning can obtain better accuracy with respect to MFCC and GMM features [39], [40].

Methods in literature for text-independent speaker recognition frequently suffer from some drawbacks. The GMM method with MFCC features [17], [21] provides very reliable accuracy for speaker recognition, but it requires templates composed of a big number of features, which are difficult to store in low-cost hardware architectures. Deep learning shows improved accuracy, but it requires large amounts of training data, which increases the training time of identification applications based on single classifiers.

In AmI environments, it is frequently required to identify the users in small closed-sets using limited hardware resources. In this regard, the proposed system uses low dimensional speaker templates, consisting only of 12 floating-point numbers. More-

over, it performs fast identifications by using a classification technique based on computational intelligence, which can be trained using relatively small datasets and in a reasonable time.

III. Proposed Method

The proposed speaker recognition method is designed to perform closed-set identifications by using a limited amount of computational resources and templates of small size, thus allowing for its use in embedded architectures for AmI environments. Our method can be divided into three main steps (Fig. 2): computation of MFCC features, ISF computation, and hierarchical classification.

A. MFCC Feature Extraction

MFCC features are widely used in the literature for text-independent speaker recognition [17], [19]. The computation of these features can be divided into the following tasks:

- Framing and windowing: papers in the literature show that the speaker signal in small time duration windows is stationary and it is possible to extract reliable features in these windows. Hence, the signal is divided into windows of 20ms. The signal extracted from each window is called frame.
- Computation of the DFT: to extract the spectral information from the signal of each window, we compute the energy available for each frequency band. Therefore, we compute the Discrete Fourier Transform (DFT), as follows:

$$S_i(k) = \sum_{f=1}^{F} S_i(n)h(n)e^{-(j2\pi kn)/F}, 1 \leq k \leq K, \quad (1)$$

  where $h(n)$ is a sample analysis Hamming window, and K is the length of the DFT.
- Computation of the Mel Filter Banks: to estimate the energy in different frequency regions, we use Mel Filter Banks [20]. These are triangular filter banks, non-linearly placed throughout the bandwidth and Mel scale. The Mel-spaced scale changes the signal from the frequency domain to the Mel-scale as follows:

$$m = 2595 \log_{10}(1 + f/700) \quad (2)$$

  This bank of filters estimates the energy of every frequency band.
- Computation of the Logarithm: after computing the Mel filter banks, we compute the logarithm of each filter bank. This task allows us to use the cepstral mean subtraction, which is a channel normalization technique.
- Computation of the DCT: finally, we compute the Discrete Cosine Transform (DCT) of the filtered signal to estimate the cepstral coefficients. The resulting features are called Mel Frequency Cepstral Coefficients. Similar to most of the methods in the literature, we use only the first 12 MFCC coefficients. Studies demonstrated that these coefficients represent the information solely about the vocal tract filter, cleanly separated from information about the glottal source [41].

B. ISF computation

The concept of information set has been introduced in [24] to enlarge the scope of a fuzzy set using the Hanman-Anirban entropy function [23]. The fuzzy set theory considers only the value obtained by applying a membership function to a property, without taking into account the value of the property itself. Differently, an information set connects the attribute values and the fuzzyfied values by using empowered membership functions [42]. Feature extraction approaches based on the information set theory have been applied in biometric systems based on face [22] and ear [43].

Let us suppose a collection of values of an attribute $\Phi = \{\Phi_1, \Phi_2, ...\Phi_n\}$, an empowered membership function is defined as follows:

$$I_\Phi = \sum_i X_\Phi(\varphi_i) G_\Phi(\varphi_i), \quad (3)$$

where $G_\Phi(\varphi_i)$ is a gain function. $G_\Phi(\varphi_i)$ is computed as follows:

$$\begin{aligned} G_\Phi(\varphi_i) = \; & e^{-}[a_\Phi(x_\Phi(\varphi_i))^3 + b_\Phi(x_\Phi(\varphi_i))^2 \\ & + c_\Phi(x_\Phi(\varphi_i)) + d_\Phi]^{\beta_\Phi}, \end{aligned} \quad (4)$$

where the parameters ($a_\Phi, b_\Phi, c_\Phi, d_\Phi, \beta_\Phi$) are the real valued variables.

This formulation of entropy function can be modulated by selecting a suitable choice of parameters ($a_\Phi, b_\Phi, c_\Phi, d_\Phi, \beta_\Phi$). As an example, using the variables ($a_\Phi = b_\Phi = 0$, $c_\Phi = 1/2\sigma_j$, $d_\Phi = -\mu_j/2\sigma_j$), we get the following function:

$$G_\Phi(\varphi_i) = e^{-[(x_\Phi(\varphi_i) - \mu_j)/2\sigma_j]^{\beta_\Phi}} \quad (5)$$

In this work, we apply the information set theory to reduce the size of the feature set. ISF enables to extract the cepstral as well the temporal possibilistic uncertainties from the MFCC features.

The MFCC feature matrix X is of dimension $(d \times m)$, where d is the number of cepstral coefficients and m is the number of frames. From each cepstral coefficient j of X, the proposed algorithm extracts the first and second order moments, creates a gain function G_j according to the extracted information, and computes ISF value. The number of features composing the final ISF vector Y is equal to number of cepstral coefficients d.

We compute every gain function as follows:

$$G_j = e^{-1/2[(X_{ij} - \mu_j)/\sigma_j]^2}, j = 1, 2, ...d, \quad (6)$$

where, μ_j and σ_j are the mean and variance of the cepstral coefficient j.

The ISF value for the cepstral coefficient j is then computed using the concept of empowered membership function, as follows:

$$\mathrm{Y}_j = \sum_{i=1}^{m} (X_{ij} \cdot G_j). \quad (7)$$

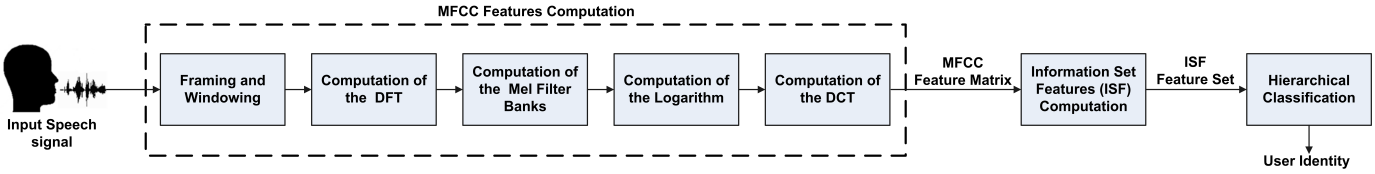

Fig. 2: Schema of the proposed biometric identification method.

The resulting ISF features are composed of 12 floating-point numbers, equal to the number of cepstral coefficients d of the MFCC feature matrix X.

C. Hierarchical classification

We use a hierarchical classification strategy to estimate the identity corresponding to the fresh template Y. Single classifiers may obtain unsatisfactory accuracy for problems with high numbers of classes involved [44]. To achieve higher accuracy, many studies in the literature use hierarchical classification approaches based on pools of classifiers. There are different categories of strategies, including the flat classification approach, the local classifier approach, and the global classifier approach. In this paper, we use a flat classification approach since it is one of the simplest and mostly used techniques in the literature.

Considering a biometric database composed of N enrolled identities, our method uses a pool of N binary classifiers and a score fusion strategy. Each classifier C_i considers the identity i as the positive class and returns a score value $s_i \in [0, 1]$. We use the following strategy:

$$\text{Identity} = \underset{i=1...N}{\operatorname{argmax}}(s_i). \tag{8}$$

We consider different types of classifiers: k-Nearest Neighbors (kNN) [45], Feed-forward Neural Networks (FFNN) [46], and Support Vector Machines (SVM) [47]. More details on the learning strategies and configurations of the single classifiers will be provided in Section IV.

IV. Experimental results

A. Database description

We evaluated the proposed method using a set of signals belonging to the Switchboard NIST 2003 SRE speaker database [25]. This database consists of 356 voice signals recorded on telephone for a duration of 2 minutes, with a sampling rate of 8kHz at 16 bit. We extracted the speech signals of the 149 males of the training set. We selected this set of signals to easily compare the performance of our method with that of other techniques in the literature. In fact, most of the studies in the literature using the NIST 2003 SRE speaker database only consider this subset of the signals.

To create the samples for our tests, we divided the 2 minutes audio signals into five samples of 24 seconds each. In this manner, we obtained 745 samples (5 samples per individual).

B. Evaluated methods and configurations

We used GMM with MFCC features (MFCC+GMM) as a baseline since it is a widely used method in the literature. To learn the parameters of GMM, we tested different numbers of mixtures (in the range of [1,...,32]) and we found the best configuration by using 16 Gaussian mixtures.

To evaluate the performance of proposed feature set (ISF), we used hierarchical classifiers based on three computational intelligence techniques, namely: kNN, FFNN, and SVM.

We tested kNN classifiers with k = 1, 3, and 5, and achieved the best results using k = 1 and Euclidean distance.

The considered configurations of FFNN are designed as follows. We used a single linear node for the output layer of the neural network. We tested different numbers of nodes with tan-sigmoidal transfer functions for the hidden layer (in the range of [1,...,100]) . We trained the neural networks using the Levenberg-Marquardt back-propagation algorithm with 500 epochs. We obtained the best results using 80 nodes in the hidden layer.

We tested three variants of SVM kernels: the linear kernel and two non-linear kernels: Gaussian kernel, and polynomial kernel of order 2. To learn the parameters of non-liner kernels, we optimized the value of σ in the range [0.1,...,3]. We achieved the best results using a Gaussian kernel with σ = 1.70.

C. Analysis of the identification performance

To evaluate the performance of the speaker recognition methods, we adopted an iterative-validation strategy. We performed 5 iterations. For each iteration, we randomly selected 4 samples per user to create the training set. The validation set was composed using the remaining sample for each individual.

We compared the results obtained using the proposed ISF templates and different hierarchical classifiers with that achieved using MFCC features. Table I summarizes the results of the evaluated methods. This table shows that the proposed method achieved better accuracy with respect to the compared speaker recognition systems. The proposed feature extraction in combination with a hierarchical classifier based on SVMs (ISF+SVM) achieved best result on the considered dataset, with Rank-1 identification accuracy of 85.64%.

Fig. 3 shows the Cumulative Match Characteristic (CMC) curves obtained by comparing the baseline method (MFCC+GMM) and the proposed method in its best configuration (ISF+SVM). Notably, our method achieved higher identification accuracy for all the considered ranks.

TABLE I: Rank-1 Identification Accuracy achieved by the baseline method (MFCC+GMM) and the proposed method in different configurations

Methods	Rank-1 Accuracy (%)
ISF+KNN	64.43
ISF+FFNN	83.22
ISF+SVM	85.64
MFCC+GMM	78.66

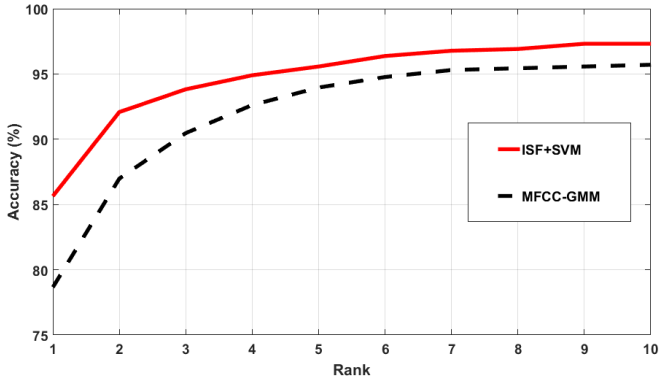

Fig. 3: CMC curve achieved by the proposed method in its best configuration (ISF+SVM) and by the baseline method (MFCC+GMM). Our method achieved better accuracy for each considered rank.

D. Performance with reduced number of enrolled samples

To investigate the application of the proposed method with less enrolled samples, we trained the hierarchical classifiers with reduced numbers of enrolled samples per user.

We adopted a 2-fold validation strategy. The first n samples per individual were selected as training set. We tested four scenarios in which the classifiers were trained with 1 enrolled sample per user, 2 enrolled samples per user, 3 enrolled samples per user, and 4 enrolled samples per user respectively. The validation set was composed using the remaining samples per individual.

Table II reports the results obtained. In each evaluated scenario, the proposed method achieved the best results. Our method (ISF+SVM) achieved Rank-1 identification accuracy of 84.56% for 4 enrolled samples per user, 80.54% for 3 enrolled samples per user, 74.72% for 2 enrolled samples per user, and 65.94% for 1 enrolled sample per user. These results show that our method achieved better accuracy with respect to the baseline method also with a limited number of training samples.

V. Conclusion

This paper presented a text-independent speaker recognition method particularly suitable for identification in Ambient Intelligence (AmI) environments. Our method first extracts MFCC features from the raw signal and then creates Information Set Features (ISF) by applying a fuzzy logic approach. ISF features reduce the size of the MFCC features and computes templates composed of only 12 floating-point numbers. The proposed biometric recognition method estimates the user's identity by applying a hierarchical classification technique based on computational intelligence.

TABLE II: Rank-1 Identification Accuracy achieved using different number of enrolled samples per user

	Rank-1 Accuracy (%)			
Methods	1 enrolled	2 enrolled	3 enrolled	4 enrolled
ISF+KNN	42.62	51.90	59.06	59.06
ISF+FFNN	48.20	69.60	71.80	80.50
ISF+SVM	65.94	74.72	80.54	84.56
MFCC+GMM	66.95	67.79	73.49	76.51

We evaluated the performance of our speaker recognition method using signals from the NIST-2003 switchboard speaker database and compared the achieved accuracy with that of traditional approaches based on Gaussian Mixture Models (GMM). The obtained results demonstrated that the proposed method reduced the size of the template with respect to traditional approaches based on GMM and achieved better identification accuracy.

We also evaluated the performance of the proposed speaker recognition method using reduced numbers of enrolled samples per individual. The obtained results showed that our method is capable of achieving better accuracy than the baseline. For future work, we should evaluate the performance of the proposed method on larger datasets acquired in less-constrained conditions and including males as well as females.

VI. Acknowledgments

This work was supported in part by: the EC within the 7FP under grant agreement 312797 (ABC4EU); the EC within the H2020 program under grant agreement 644597 (ESCUDO-CLOUD); and the Italian Ministry of Research within PRIN 2015 project COSMOS (201548C5NT).

References

[1] A. Anand, R. Donida Labati, A. Genovese, E. Muoz, V. Piuri, F. Scotti, and G. Sforza, "Enhancing fingerprint biometrics in automated border control with adaptive cohorts," December 2016.

[2] A. Anand, R. Donida Labati, A. Genovese, E. Muoz, V. Piuri, F. Scotti, and G. Sforza, "Enhancing the performance of multimodal automated border control systems," September 2016.

[3] R. Donida Labati, A. Genovese, E. Muoz, V. Piuri, F. Scotti, and G. Sforza, "Biometric recognition in automated border control: a survey," *ACM Computing Surveys*, vol. 49, pp. 24:1–24:39, June 2016.

[4] R. Donida Labati, A. Genovese, E. Muoz, V. Piuri, F. Scotti, and G. Sforza, "Emerging biometric technologies for automated border control gates," October 2016.

[5] R. Donida Labati, V. Piuri, and F. Scotti, *Touchless fingerprint biometrics*. CRC Press, 2015.

[6] R. Donida Labati, A. Genovese, V. Piuri, and F. Scotti, "Toward unconstrained fingerprint recognition: a fully-touchless 3-d system based on two views on the move," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 46, pp. 202–219, February 2016.

[7] R. Donida Labati and F. Scotti, "Noisy iris segmentation with boundary regularization and reflections removal," *Image and Vision Computing, Iris Images Segmentation Special Issue*, vol. 28, pp. 270 – 277, February 2010.

[8] A. Genovese, V. Piuri, and F. Scotti, *Touchless palmprint recognition systems*, vol. 60. Springer, 2014.

[9] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[10] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Toward conversational human-computer interaction," *Artificial Intelligence magazine*, vol. 22, no. 4, pp. 27–37, 2001.

[11] G. R. Doddington, "Speaker recognitionidentifying people by their voices," *Proc. of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[12] C. I. Nass and S. Brave, *Wired for speech: How voice activates and advances the human-computer relationship*, vol. 9. MIT press, 2005.

[13] K.-A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Proc. of the 12th Annual Conf. of the Int. Speech Communication Association*, pp. 3317–3318, 2011.

[14] W. Wahlster, "Smartkom: Fusion and fission of speech, gestures, and facial expressions," in *Proc. of the 1st Int. Workshop on Man-Machine Symbiotic Systems*, pp. 213–225, 2002.

[15] J. Kleindienst, T. Macek, L. Serédi, and J. Sedivy, "Vision-enhanced multi-modal interactions in domotic environments," *IBM Tecnologías de Voz y Sistemas. República Checa*, pp. 1059–1064, 2004.

[16] M. Vacher, B. Lecouteux, J. S. Romero, M. Ajili, F. Portet, and S. Rossato, "Speech and speaker recognition for home automation: Preliminary results," in *Proc. of the 8th Int. Conf. on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, IEEE, 2015.

[17] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[18] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. on speech and audio processing*, vol. 2, no. 1, pp. 194–205, 1994.

[19] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[21] D. Reynolds, "An overview of automatic speaker recognition," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075)*, 2002.

[22] F. Sayeed and M. Hanmandlu, "Three information set-based feature types for the recognition of faces," *Signal, Image and Video Processing*, vol. 10, no. 2, pp. 327–334, 2016.

[23] F. Sayeed and M. Hanmandlu, "Properties of information sets and information processing with an application to face recognition," *Knowledge and Information Systems*, pp. 1–23, 2017.

[24] M. Hanmandlu, "Information sets and information processing," *Defence Science Journal*, vol. 61, no. 5, p. 405, 2011.

[25] M. Przybocki and A. Martin, "The nist year 2003 speaker recognition evaluation plan," 2003.

[26] H. Zeinali, H. Sameti, L. Burget, J. Černockỳ, N. Maghsoodi, and P. Matějka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge," *Interspeech 2016*, pp. 440–444, 2016.

[27] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[28] J. Pelecanos, S. Myers, S. Sridharan, and V. Chandran, "Vector quantization based gaussian modeling for speaker verification," in *in Proc. Of 15th Int. Conf. on Pattern Recognition, 2000*, vol. 3, pp. 294–297, IEEE, 2000.

[29] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1890–1899, 2011.

[30] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 108–121, 2012.

[31] A. O. Hatch, A. Stolcke, and B. Peskin, "Combining feature sets with support vector machines: Application to speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 75–79, IEEE, 2005.

[32] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4787, IEEE, 2011.

[33] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[34] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[35] D. A. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami, "The 2004 mit lincoln laboratory speaker recognition system," in *in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'05)*, vol. 1, pp. I–177, IEEE, 2005.

[36] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[37] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[38] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056, IEEE, 2014.

[39] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm," in *in IEEE Workshop on Spoken Language Technology Workshop (SLT)*, pp. 131–136, IEEE, 2012.

[40] L. Li, Y. Lin, Z. Zhang, and D. Wang, "Improved deep speaker feature learning for text-dependent speaker recognition," in *in Annual Summit and Conf, on Signal and Information Processing Association (APSIPA)*, pp. 426–429, IEEE, 2015.

[41] D. Jurafsky and J. H. Martin, *Speech and language processing*, vol. 3. Pearson, 2014.

[42] M. Aggarwal and M. Hanmandlu, "Representing uncertainty with information sets," *IEEE Trans. on Fuzzy Systems*, vol. 24, no. 1, pp. 1–15, 2016.

[43] M. Hanmandlu *et al.*, "Robust ear based authentication using local principal independent components," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6478–6490, 2013.

[44] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[45] K. C. A. R.Webb, "Statistical pattern recognition, 3rd edition."

[46] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[47] L. Wang, *Support vector machines: theory and applications*, vol. 177. Springer Science & Business Media, 2005.