Chapter 1
Introduction

Abstract This chapter introduces the concept of speaker recognition (SR) and its applications. It emphasizes on explaining the requirement of developing SR technologies that are robust towards background environments. The intermediate sections provide broad overviews of various stages associated in developing a SR system and different categories of SR. The later sections highlight the issues addressed in the book and its contributions.

1.1 Introduction

Telecommunication networking has made a pervasive impact in the human society in the last few decades. Much of our personal information today, is shared over the Internet or exchanged through hand-held devices. This obviously drives the demand for technology that secures human access to confidential data. Recent developments in the area of remote transactions such as telebanking, e-commerce, online railway or airline reservations etc., have made individual authentication a crucial factor. Traditional modes of security such as passwords and personal identification numbers (credit/debit cards) are often vulnerable since they can be easily forgotten, misplaced or stolen. A feasible alternative is the use of biometric authentication i.e., identifying individuals by their physical traits, which are least susceptible to physical misuse and impersonation. However, practical use of common biometric techniques like iris, face and fingerprint recognition is constrained by factors like close proximity/direct contact with individuals or requirement of costly sensors, which thereby limits their application in remote operations.

Speaker recognition (SR) is the task of recognizing individuals using their speech. As the most common mode of human communication, speech is readily available, can be easily recorded by inexpensive devices and transmitted over long-distance telecommunication channels. This is evident from the wide range of voice communication applications available over the Internet e.g., Skype, Google talk, Google voice search etc. As such, speaker recognition also provides an

K.S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-07130-5_1, © The Author(s) 2014

attractive biometric alternative to its sophisticated counterparts. Speaker recognition technologies are being readily deployed today in three major areas of applications i.e., security, surveillance and forensics [1].

The key applications that demand biometric security based SR technology are tele-commerce and forensics [1] where the objective is to automatically authenticate speakers of interest using his/her conversation over a voice channel (telephone or wireless phone). In forensics (e.g., criminal investigation), the speakers can be considered non-cooperative as they do not specifically wish to be recognized. On the other hand, in telephone-based services and access control, the users are considered to be cooperative. With the ever increasing popularity in multimedia web-portals (e.g., Facebook and Youtube), large repositories of archived spoken documents such as TV broadcasts, teleconference meetings, and personal video clips can be accessed through the Internet. Searching for topic of discussion, participant names and genders from these multimedia documents would require automated technology like speaker verification and recognition.

While the SR technologies promise an additional biometric layer of security to protect the user, the practical implementation of such systems faces many challenges. For example, a handheld-device based recognition system needs to be robust to noisy environments, such as office, street or car environments, which are subject to unpredictable and unknown sources of noise (e.g., abrupt interference, sudden environmental change, etc.).

1.2 Speaker Recognition

Human beings can reliably recognize known voices by barely hearing a few seconds of speech. The uniqueness of one's voice can be attributed to both physical and acquired characteristics of a person. Physical differences occur largely due to the distinct shapes and sizes of the voice producing organs (e.g., vocal folds, vocal tract, larynx, etc.) and partly due to the articulators (e.g., tongue, teeth, lip etc.). Apart from these anatomical properties, individuals can also be distinguished by their accent, vocabulary, speaking rate and other personal mannerisms that are acquired over a period of time. State-of-the-art speaker recognition systems exploit these properties in parallel to achieve high recognition accuracy [2, 3]. While subjective tests have revealed that humans often show superior performance in recognizing familiar [4] or disguised voices [5], machines outperform humans when it comes to recognition on a large scale [6] especially for non-cooperative speakers. Automatic speaker recognition (ASR) systems would ideally imitate the human voice recognition process which in turn is dependent on a complex auditory perception mechanism. Human beings are inherently capable of integrating a wide range of knowledge sources in speech signals at various levels (e.g., acoustic, articulatory, syntactic etc.). However, the exact nature of speech comprehension or segregation of speaker information at the cognitive or neurobiological level is still largely unknown. Thus, the general approach is to enumerate perceptual cues used by humans at various levels and estimate their patterns for later classification. The broad stages of the ASR process are briefly outlined in the following paragraphs.

- **Preprocessing:** This stage corresponds to the acquisition of a speech signal for the recognition process. The analog speech signal is digitized by sampling it at a desired frequency. The digital speech is usually 'pre-emphasized' using a high pass filter which emphasizes higher frequency components and compensates for the human speech production mechanism which tends to attenuate them. For several ASR tasks, a 'voiced activity detection' (VAD) stage is often used to separate speech segments from a given audio signal. It is often challenging to implement VAD that works consistently across various background environments especially for short-duration utterances [2].
- **Feature Extraction:** This stage corresponds to the enumeration of knowledge sources in a speech signal. The raw speech signal is reduced to a set of parameters in which speaker-discriminative properties are emphasized and redundant information is suppressed. The vast numbers of features explored for ASR tasks can be broadly categorized as spectral, source, prosodic and high-level features. The first two categories, often collectively termed as 'low-level' features, convey physiological information about the speaker (e.g., size of vocal folds, structure of vocal tract etc.). The latter two categories comprise high-level features which reflect acquired behavioral aspects of a speaker (e.g., temperament, accent, vocabulary etc.). Selection of appropriate features for ASR is usually based on certain criterion [7]. An ideal feature is expected to have high inter-speaker variability, low intra-speaker variability, natural occurrence in speech, robustness towards noise/channel-distortion, immunity towards a speaker's health/mood fluctuations and ease of extraction. Apart from these, the features should have a compact representation to avoid requirement of a large amount of training data. Though short-term spectral features (e.g., MFCC) [8] are often preferred for ASR tasks due to their high accuracy and real-time extraction, they are susceptible to noise degradation [9]. High-level features improve noise/channel-robustness at the cost of a difficult extraction procedure and high amount of training data. Feature selection is thus a tradeoff between speaker-discrimination, robustness and practical application.
- **Acoustic Speaker Modeling:** In this stage various statistical modeling techniques are employed to capture the distribution of features extracted from individual speakers. The feature extraction and speaker modeling stage jointly represent the training or enrollment phase of ASR in which speakers register/enrol for the SR system. The goal of this stage is to build unique templates or models for each enrolled speaker. Standard speaker modeling techniques can be categorized in different ways. Depending on the nature of modeling the feature distribution, they may be either *parametric* or *non-parametric*. *Parametric* models assume a fixed probability density of the feature distribution (e.g., Gaussian Mixture Models (GMMs) [10, 11], Hidden Markov Models (HMMs) [12]) whereas *non-parametric* models use non-stochastic template-based modeling (e.g., Vector Quantization (VQ) [13], Dynamic Time Warping [14]). On the

basis of their training paradigm, speaker models are classified as *generative* and *discriminative*. The *generative* models individually estimate feature distribution within each speaker (class) (e.g., GMMs, HMMs, VQ) while *discriminative* models are based on learning the differences between enrolled speakers (classes) (e.g., Support Vector Machines (SVMs) [15], Neural Networks (NNs) [16]). Recent research trends have also focussed on combining generative and discriminative models for improved ASR tasks [15, 17, 18].

- **Pattern Matching and Classification:** In this stage an unknown (test) utterance based on its statistical similarities with a known speaker model. The pattern matching and classification stage is collectively termed as the testing/evaluation phase in which the ASR system is evaluated on the basis of its classification accuracy. Pattern matching is entirely dependent on the nature of the acoustic speaker models. In case of stochastic generative models, matches are quantified in the form of log-likelihood scores whereas for parametric ones they might be simple distance metrics (e.g., Euclidean distance for VQ). For discriminative models, scores may be based on the distance from the decision boundary of two classes (speakers) (e.g., SVMs) or the difference between the actual and predicted class (e.g., NNs). A decision is taken based on the scores obtained i.e., the test utterance is classified as the speaker (model) producing the highest score.

1.3 Types of Speaker Recognition

Speaker Recognition can be broadly categorized into two types i.e., Speaker Identification (SI) [10] and Speaker Verification (SV) [11].

1.3.1 Speaker Identification

Closed-set speaker identification (SI) is the task of detecting a unique speaker responsible for producing a test utterance, out of a closed-set of enrolled speakers. In case the test utterance doesn't belong to any member of the closed-set, the task is an 'Open-set' SI. Considering each speaker model as a class, the SI task is basically a multi-class classification problem in which an unknown test utterance is assigned to a particular class. Figure 1.1 shows the block-diagram of a SI system. The training phase shows the estimation of acoustic models from individual speakers. This is usually time-consuming and hence performed offline. The evaluation phase, performed online requires fast identification of a known speaker. However, since the unknown utterance has to be compared against all enrolled speaker models, increase in the number of speakers in the set causes performance degradation (in terms of both accuracy and computational burden).

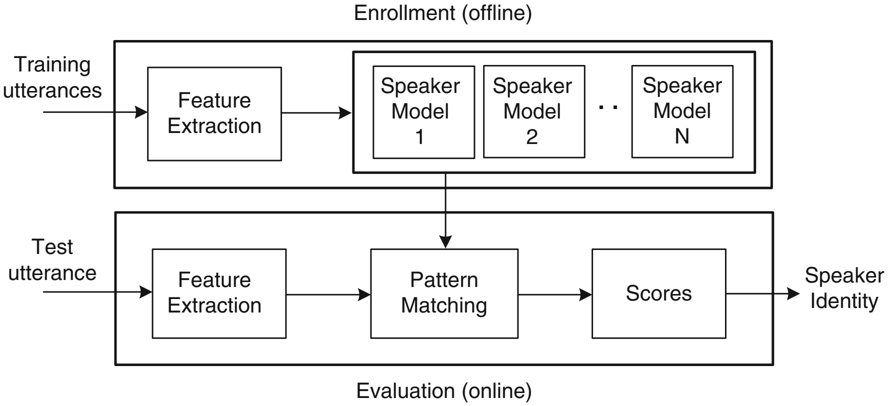

Fig. 1.1 Block diagram of a speaker identification system

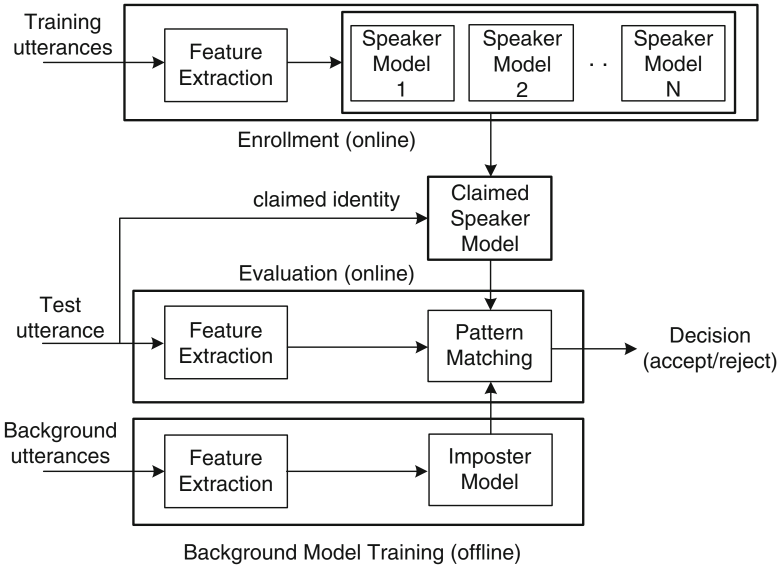

Fig. 1.2 Block diagram of a speaker verification system

1.3.2 Speaker Verification

Speaker verification (SV) is the task of validating the claimed identity of a speaker. It is a binary classification problem in which the claim is either accepted or rejected based on the statistical similarities of a test utterance with the claimed speaker model (true class) and a selected background/impostor model (false class). Figure 1.2 shows the block-diagram of a typical SV system. A number of differences can be observed in contrast to SI. Firstly, a fixed pool of background speakers are required for offline training of the impostor model. The background speakers can be used as negative examples for training a discriminative model [15] or used to train a 'Universal Background Model' (UBM) [11] for GMM-based SV. In the latter case the enrolled speaker models are obtained online by adapting the UBM using a speaker's training data. Secondly, the pattern matching stage in SV requires comparison of the unknown utterance against a single claimed model and another imposter model, which makes it much faster and unaffected by the number of speakers enrolled for the SV system. The ratio of scores obtained against either model is compared with a threshold for final decision. Furthermore, SV is able to reject speech from arbitrary speakers (i.e., the open-set case) which is not true for speaker identification. Applications of ASR involving surveillance and monitoring usually require identification rather than verification. However, most online applications and security based transactions (e.g., online reservation, telebanking) require an individual to be verified rather than identified (i.e., authenticity of a claimed identity is judged irrespective of the actual identity of the speaker).

Both the above types of ASR systems may further be 'text-dependent' [19] or 'text-independent' [20]. In text-dependent systems (suitable for cooperative users) [20], the recognition phrases are fixed, or known in advance. Such systems additionally require a speech recognizer in the front-end causing more accurate but costly applications. In text-independent systems, there are no constraints on the words which the speakers are allowed to speak. Thus, the reference (what are spoken in training) and the test (what are uttered in actual use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account. Text-independent speaker recognition is thus much more challenging of the two tasks.

1.4 Challenging Issues in Speaker Recognition

A number of very common yet challenging issues concerning ASR, especially speaker verification has been highlighted in this section.

- **Mismatched training and test conditions:** This refers to the family of problems that arise primarily due to the differences (mismatch) in recording devices, channel, background etc., during the enrollment and evaluation phase of ASR. A typical example scenario is the development of recognition models using enrollment data acquired over the Internet and acquiring the speech data via a mobile phone during verification or testing. The medium of data acquisition or transfer seemingly encodes new information into the speech signal which largely affects the feature extraction, speaker modeling and pattern matching stages. These problems, often collectively termed as '*session variability*', has been identified as one of the most challenging issue in the field of ASR and a major source of verification errors [21,22]. The problem has been addressed over the last few decades starting with primitive methods [6] and gradually advancing into more recent techniques [21,22].

- **Intra-Speaker Variability:** While 'mismatch' occurs primarily due to extraneous factors (e.g., recording devices, background etc.), it is not solely restricted to them. Fluctuations in intrinsic/personal factors of a speaker (e.g., health, emotion, mood etc.) are also reflected across different sessions causing poor recognition. For text-independent SV systems, lack of constraints in the form of utterances spoken during training and evaluation may additionally lead to a phonetic mismatch. In general text-independent systems are more affected due to intra-speaker variability compared to text-dependent ones [23].
- **Background Noise:** Background noise is a prominent factor responsible for the loss of performance accuracy in generalized speech-based recognition tasks. Noise can be severely detrimental for ASR in both matched and mismatched conditions, the latter usually being the worse case [24]. The problem of noise or environmental degradation had been studied in past primarily in the context of speech recognition [25, 26]. A number of techniques developed for 'noise suppression' or 'noise compensation' since then, can be interchangeably applied for speaker recognition tasks. The discussion on SV for background noise shall be continued in Sect. 1.5 in more details.
- **Limited Enrollment Data:** The availability of data is a critical factor for training acoustic speaker models. The generative speaker models which are most commonly used for ASR, especially demand a high amount of training data. Usually, the required amount of training data increases proportionally with the dimension of the features extracted. This phenomenon is often termed as 'curse of dimensionality' [27]. The problem of limited data arises particularly for real-time ASR applications such as hand-held devices or in non-cooperative scenarios where speakers purposely avoid enroling for longer durations. The problem is usually tackled using statistical adaptation techniques where an already built model is modified using the acquired data [11,28,29].

1.5 Issue Addressed in Book

The book addresses the issue of speaker verification in noisy background environment. Substantial number of studies have been previously carried out in the area of robust speech recognition [25, 26]. Due to the advent of online transaction processing and the large-scale deployment of ASR technologies in hand-held devices in recent times, robustness for ASR systems has received a renewed interest [30]. In systems deployed for telephony applications the main form of degradation is due to channel variabilities induced by the handset and/or microphone. However, for speaker recognition carried out in far field applications environmental or background distortions are also of concern. As an example we may consider the typical scenario where a user enrolls for a SV system through his mobile phone while walking on a busy street. During his next access to the SV system for verification, he may be present in a secluded environment (e.g., car interior, room, office etc.). Three facts can be observed. Firstly, the background keeps changing

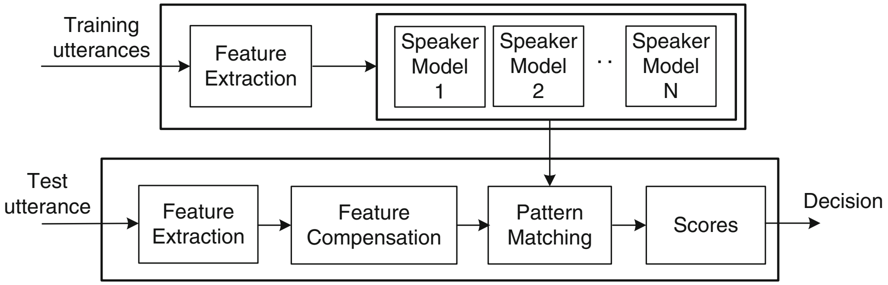

Fig. 1.3 Block diagram of the feature compensation process

during enrollment where the user may even enter a totally unknown environment. Secondly, the obvious environmental mismatch that occurs during verification. Thirdly, there might even be a handset/channel mismatch if a different device is used during testing. In fact, in most cases especially for mismatched conditions one can expect a combined impact of both channel and background.

Background noise, in general considered additive in nature, primarily affects the spectral properties of a signal. Handling noise distortions is a challenge due to a number of reasons. Firstly, it is very difficult to quantify the effect of noise in speech primarily due to its random nature. More specifically, a clean speech segment exposed to a particular noisy environment in different intervals of time may yield noisy signals with different spectral properties. Such problems increase manifold if the noise is non-stationary i.e., its statistical properties change over time. Secondly, addition of noise results in arbitrary distortion of the feature distribution causing loss of discriminative information. This is indirectly reflected in each distinct stage of the ASR process discussed in Sect. 1.2.

The present study shall emphasize on the impact of noise in the feature-level and acoustic model-level, respectively. Noise-robustness obtained via the aforementioned stages has two broad interpretations. Firstly, the features extracted or the classifiers trained in the modeling stage may themselves be relatively immune towards the effects of channel distortions or background noise, by design. Secondly, the features and models used for generic recognition tasks in one environment may be modified or 'adapted' in another environment, to suppress the effect of mismatch. The former category comprises the group of robust features and robust speaker models while the later category comprises the family of 'compensation' or 'adaptation' techniques.

Feature compensation techniques aim to transform the features extracted during the evaluation phase such that they reflect the environmental conditions present during the training phase. Figure 1.3 shows a simplified block diagram of the feature compensation process. This is particularly applicable but not restricted to scenarios where a person enrols in a clean environment but verifies himself in a noisy one.

Despite much research for developing robust features [9], feature compensation techniques are often preferred due to the implementation costs associated with the

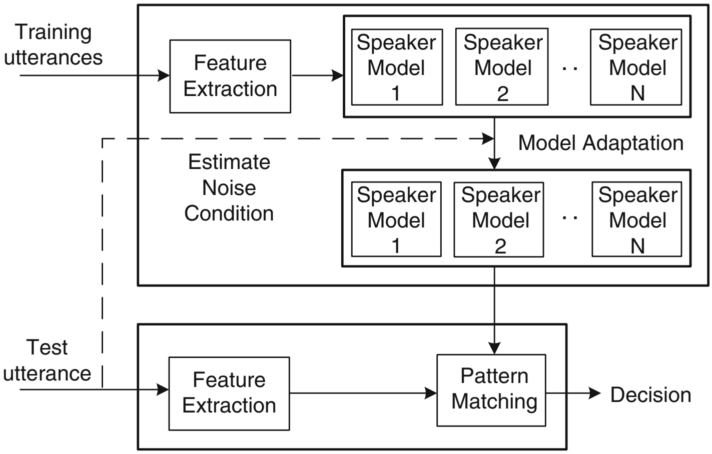

Fig. 1.4 Block diagram of the model compensation process

former and the considerable performance improvement obtained in the latter [30]. A discussion about the various kinds of feature compensation techniques have been provided in the next chapter.

Model compensation/adaptation techniques (shown in Fig. 1.4) alters the acoustic modeling and pattern matching stages in order to account for the interfering noise. The model parameters learnt during the training phase are modified to reflect the new/mismatched environment of the evaluation phase. The traditional model compensation methods mostly rely on a priori knowledge about the test environment to adapt clean speaker models. They may be either (i) 'data-driven' in which available noisy adaptation data is used to alter pre-estimated speaker models or (ii) 'analytical' in which a mathematical structure of noise corruption is used to synthesize noisy speaker models from clean speaker models and noise models [30]. The 'data-driven' methods are usually more preferred for practical SV applications due to their low data-requirements in comparison to the 'analytical' ones which require high amount of training data. Though these methods perform significantly well (often better than feature compensation techniques), prior knowledge of test environment is sometimes considered as a major drawback for real-life scenarios. Robust speaker modeling techniques are alternatively explored as a tradeoff between accuracy and practical applications [24]. Detailed discussion about robust speaker modeling and model compensation approaches have been provided in the next chapter.

1.6 Objective and Scope of Work

The book aims to study alternative methods for developing ASR systems that are robust towards environmental noise. Specific focus is laid on text-independent speaker verification (SV) rather than speaker identification, since the former has a greater range of biometric applications especially in hand-held devices and online transactions.

Amongst various available strategies, the present work explores data-driven stochastic feature compensation (SFC) and robust speaker modeling methods. Two distinct categories of SFC methods based on (i) independent probability models and (ii) joint probability models, are explored. Amongst robust speaker modeling methods, the significance of supervector-based approaches in a discriminative framework for SV in noisy environment, is explored. Certain drawbacks concerning the conventional speaker modeling framework are highlighted and addressed. A boosting algorithm is proposed to combine robust discriminative classifiers for enhanced SV in degraded environments. Significance of all the methods explored in the present work is analyzed on the basis of their effectiveness and computational costs.

1.7 Organization of the Book

- Chapter 1 provides a brief introduction to the concept of automatic speaker recognition, its stages, categories and modern applications. A number of challenging issues in the field of ASR are highlighted. A brief discussion of the issue addressed in the book is provided followed by the objective and scope of work.
- Chapter 2 provides an overview of various feature and model-based approaches developed in past for robust speaker recognition. The advantages and disadvantages of some standard methods applied for robust SV tasks have been highlighted.
- Chapter 3 discusses baseline SV systems developed using the GMM-UBM framework in noisy environments. A feature mapping technique using multiple background model framework has been explored for robust SV in time-varying noisy environments.
- Chapter 4 explores the impact of standard stereo-based stochastic feature compensation (SFC) methods for robust speaker verification in uniform noisy environments. Integration of a SFC stage in the GMM-UBM framework is proposed for SV evaluation under mismatched conditions.
- Chapter 5 explores robust speaker-modeling methods for SV in noisy environments. Specifically, the combined GMM-SVM and SVM-i vector approaches are used for developing SV systems and evaluating them in matched conditions
- Chapter 6 provides a brief summary and conclusion of the Book.

1.8 Contribution of the Book

The contribution of the book lies in exploring feature compensation and robust speaker modeling methods, the impact of which have not been erstwhile studied explicitly for speaker verification in noisy environments. The major contributions can be broadly summarized under the following points

- A class of data-driven stochastic feature compensation methods has been explored for robust speaker verification (SV) in noisy background environments.
- The robustness of some state-of-the-art speaker modeling methods (e.g., GMM supervector, i-vector) in a discriminative framework using SVM classifiers, has been explored for SV in noisy environments.
- A novel boosting algorithm is proposed for combining robust SVM classifiers for improving SV performance.

References

1. J.P. Campbell, W. Shen, W.M. Campbell, R. Schwartz, J.F. Bonastre, D. Matrouf, Forensic speaker recognition. IEEE Signal Process. Mag. **26**(2), 95–103 (2009)
2. B.G.B. Fauve, D. Matrouf, N. Scheffer, J.F. Bonastre, J.S.D. Mason, State-of-the-art performance in text-independent speaker verification through open-source software. IEEE Trans. Audio Speech Lang. Process. **15**(7), 1960–1968 (2007)
3. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**, 12–40 (2010)
4. D.V. Lancker, J. Kreiman, K. Emmorey, Familiar voice recognition: patterns and parameters Part I: recognition of backward voices. J. Phon. **13**, 19–38 (1985)
5. A. Reich, J. Duke, Effects of selected vocal disguises on speaker identification by listening. J. Acoust. Soc. Am. **66**(4), 1023–1029 (1979)
6. G. Doddington, Speaker recognition—identifying people by their voices. Proc. IEEE **73**(11), 1651–1664 (1985)
7. J. Wolf, Efficient acoustic parameters for speaker recognition J. Acoust. Soc. Am. **6**(51), 2044–2056 (1972)
8. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
9. D.A. Reynolds, Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech Audio Process. **2**(4), 639–643 (1994)
10. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Acoust. Speech Signal Process. **3**(1), 72–83 (1995)
11. D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. **10**(1), 19–41 (2000)
12. M. BenZeghiba, H. Bourland, On the combination of speech and speaker recognition, in *Proceedings of the European Conference of Speech Communication and Technology (EUROSPEECH '03)*, Geneva, 2003
13. D. Burton, Text-dependent speaker verification using vector quantization source coding. IEEE Trans. Acoust. Speech Signal Process. **35**(2), 133–143 (1987)
14. L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, 1st edn. (Prentice-Hall, Englewood Cliffs, 1993)

15. W. Campbell, J. Campbell, D. Reynolds, E. Singer, P. Carrasquillo, Support vector machines for speaker and language recognition. Comput. Speech Lang. **20**, 210–229 (2006)
16. K. Farrell, R. Mammone, K. Assaleh, Speaker recognition using neural networks and conventional classifiers. IEEE Trans. Speech Audio Process. **2**(1), 195–204 (1994)
17. W. Campbell, J. Campbell, D. Reynolds, Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
18. B. Yegnanarayana, S.P. Kishore, AANN: an alternative to GMM for pattern recognition. Neural Netw. **15**, 456–469 (2002)
19. F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, D.A. Reynolds, A tutorial on text-independent speaker verification. EURASIP J. Adv. Signal Process. (Spec. Issue Biom. Signal Process.) **4**(4), 430–451 (2004)
20. M. Hebert, Text-dependent speaker recognition, in *Springer Handbook of Speech Processing* (Springer, Berlin/Hiedelberg, 2008), pp. 743–762
21. P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio Speech Lang. Process. **15**(4), 1448–1460 (2007)
22. R. Vogt, S. Sridharan, Explicit modeling of session variability for speaker verification. Speech Commun. **1**(22), 17–38 (2008)
23. J. Kahn, N. Audibert, S. Rossato, J.F. Bonastre, Intra-speaker variability effects on speaker verification performance, in *The Speaker and Language Recognition Workshop (Odyssey '10)*, Brno, 2010
24. J. Ming, T.J. Hazen, J.R. Glass, D. Reynolds, Robust speaker recognition in noisy conditions. IEEE Trans. Audio Speech Lang. Process. **15**(5), 1711–1723 (2007)
25. A. Acero, Acoustical and environmental robustness in automatic speech recognition. PhD thesis, Carnegie Mellon University, Sept 1990
26. P. Moreno, Speech recognition in noisy environments. PhD thesis, Electrical & Computer Engineering Department, Carnegie Mellon University, Pittsburgh, 1996
27. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006)
28. J. Gauvain, C. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. **2**(2), 291–298 (1994)
29. V. Hautamaki, T. Kinnunen, I. Karkkainen, M. Tuononen, J. Saastamoinen, P. Franti, Maximum a posteriori adaptation of the centroid model for speaker verification. IEEE Signal Process. Lett. **15**, 162–165 (2008)
30. R. Togneri, D. Pullella, An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. **11**(2), 23–61 (2011)