Chapter 2
Robust Speaker Verification: A Review

Abstract This chapter 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 speaker verification tasks have been highlighted. The main focus is to summarily introduce popular state-of-the-art techniques adopted for enhancing speaker verification performance in noisy conditions.

This chapter provides a broad overview of research methods developed for robust speaker recognition tasks in past. The focus is to summarily introduce popular state-of-the-art techniques adopted for enhancing speaker verification performance in noisy conditions, especially those within the current scope of work. The chapter mainly emphasizes in the feature extraction and statistical modeling stages of speaker recognition. The merits and de-merits of some of these techniques are discussed in the purview of the book. It is to be noted that many of these methods have primarily been applied for robust speech recognition in noisy environment. Since some of the intermediate stages of speaker verification are similar to that of speech recognition, they may be interchangeably used for the former. The readers are encouraged to follow the references for detailed description of the methods discussed especially notable reviews such as [1–4] or recent research works [5]. Concise overviews of methods adopted for feature compensation, feature extraction, model compensation and robust speaker modeling are briefly presented in different sections of this chapter. The role of each of these stages has been discussed in the first chapter. The final section briefly describes the motivation of carrying out the present research work.

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_2, © The Author(s) 2014

2.1 Feature Compensation

Ever since parameterization of raw speech signal was first studied [6], the motivation was to discover speaker-discriminative features for generalized recognition tasks [7]. The significance of cepstral features [8], especially the mel-cepstrum [9] for speaker recognition (SR) was established during the contemporary period. However, there were practical limitations of the use of cepstral features due to arbitrary modification of the cepstral distribution in the presence of channel distortions or background noise. A series of feature compensation techniques were proposed during the early 1990s as a refinement of the common feature extraction process [10–12]. The motivation was to make real-life applications of SR or speech recognition which countered channel-induced distortions and handset mismatches over telephonic conversations [13]. The class of feature compensation methods developed since then, may be broadly categorized into three groups i.e., filtering-based compensation, noise model-based compensation and empirical compensation. Apart from the conventional compensation techniques, there exists a group of feature transformation methods which are often used in conjunction with the former. In [14], neural network models are used as mapping functions for transforming the emotion-specific features to emotion-independent features for developing robust speaker recognition system.

The filtering techniques aim to denoise or suppress the effect of noise in the extracted features. They exploit the fact that convolutive channel or environmental distortions become additive in the log-spectral and cepstral domain. It was studied in [15] that slow variations in the channel appear as an offset of individual co-efficients of a cepstral vector. Cepstral Mean Subtraction (CMS) [15] suppresses the channel effects by subtracting the mean of cepstral co-efficients extracted from short-term frames, from the individual coefficients. The removal of the average spectrum also suppress inter-session variabilities to certain extent [11]. Apart from simple mean-removal as in CMS the variance of the cepstral vectors are often scaled to unity. Relative Spectra (RASTA) [16], principally similar to CMS, was proposed to compensate for rapidly varying channel conditions. Instead of uniform mean subtraction over the entire cepstra, a moving average filter was employed for an exponentially decaying mean subtraction. CMS and RASTA are commonly applied for front-end compensation in SR tasks due to the simplicity of implementation. A set of more sophisticated '*kernel filtering*' methods [17] were later developed which captured the non-linear features of speech by fitting a higher dimensional mapping function and eventually projecting the features to a lower dimensional manifold. However later studies had promptly revealed that these techniques are not much effective for channel mismatches and additive background noise.

The model-based feature compensation methods assume a priori knowledge of the noise spectrum. An estimate of the clean speech parameters is made using either a noise model or representation of the effects of noise in speech. The primitive methods in this group include Spectral Equalization [18] and Spectral Subtraction (SS) [19]. In SS, the clean speech spectra is estimated by subtracting the mean magnitude of an approximate noise spectra from that of the noisy speech spectra. These methods relied on the stationary assumption of noise and independence of spectral estimates across frequencies explicitly. To overcome this limitation, some of the methods developed later were based on the minimum mean squared error (MMSE) predictor [20] which modeled the correlation of frequency components e.g., MMSE log spectral amplitude estimator [21]. During the early 1990s, stereo-data based compensation techniques were first introduced [10]. Cepstral compensation vectors were derived from a stereo database and applied to the training data to adapt to environmental changes. The compensation could also be in the form of affine transformations learned from stereo data [12]. Popular examples are Codeword Dependent Cepstral Normalization (CDCN) [22] and its variants like Fast CDCN (FCDCN) [10]. Other methods relied on a mathematical model of the environmental mismatch due to noise. The parameters of the model were estimated and applied to the appropriate inverse operation to compensate the test signal e.g., feature-level Vector Taylor Series [23].

The third group of feature compensation techniques are entirely data-driven and are stochastic in nature. They are ‘blind’ towards the nature of the corrupting process and are based on empirical compensation methods that use direct spectral comparison. Prior work shows that they often outperform the previous two approaches for feature enhancement [24]. During the training phase, some transformations are estimated by computing the frame-by-frame differences between the vectors representing speech in the clean and noisy environments (stereo data). The differences between clean and noisy feature vectors are modeled by training additive bias vectors on the mean and covariance of either of the two (clean or noisy) probability distributions. During evaluation phase, the bias vectors are used to transform noisy test feature vectors to their clean feature equivalent based on the MMSE predictor. Previous MMSE-based methods like CDCN [22] and FCDCN [10], used vector quantization (VQ) codebooks to represent the distribution of clean feature vectors. Due to their quantization-based framework, these algorithms were unable to learn the variance of a distribution and were later replaced by the more flexible Gaussian Mixture Model (GMM)-based normalization techniques e.g., Multivariate Gaussian-based Cepstral Normalization (RATZ) [25]. Although the RATZ family of algorithms approximated the normalized features, the posterior probability of clean GMM components with respect to the noisy test feature vectors were usually distorted, causing poor MMSE estimates. To suppress these distortions, the Stereo-based Piecewise Linear CompEnsation for Environments (SPLICE) algorithm proposed in [26] modeled the noisy feature space using GMMs instead. This produced significantly better result in robust speech recognition tasks compared to its predecessors [27]. The effectiveness of SPLICE framework has since then encouraged its extended applications e.g., speech recognition in non-stationary noisy environments within cars using the Multi Environment Model-based Linear Normalization (MEMLIN) algorithm [28] and word recognition using Noise Adaptive Training [27]. The more recently proposed Stereo-based Stochastic Mapping (SSM) [29] is principally a more accurate version of SPLICE based on joint probability modeling of the noisy and clean feature spaces using GMMs.

2.2 Robust Feature Extraction

The conventional features used for SR tasks can be broadly categorized as spectral, prosodic and high-level features. In this section we briefly discuss each.

Prosody is a collective term for certain aspects manifested in long term speech segments e.g., stress, intonation pattern, rhythm etc. The most significant amongst these is intonation which is characterized by the fundamental frequency contour (F_o). F_o contour and energy (stress) were effectively used for speaker recognition in [30]. A few other significant applications of prosodic features for SR include combination of energy trajectory with F_o [31] and construction of SVM speaker models using pitch, duration and pause features [32]. In [33], temporal variations in speaker-specific prosodic parameters are proposed in addition to conventional spectral features for improving the speaker recognition accuracy in presence of noisy background environments. A comparative study about the significance of the various prosodic features for SR tasks can be found in [34]. Modeling the different levels of prosodic information (instantaneous, long-term) for speaker discrimination is considered to be a difficult task. At the same time, it is desired that the features are free from the effects that a speaker can voluntarily control. Due to these complications, prosodic features haven't been much used for robust SR tasks.

High-level features exploit speaker's choice of words or vocabulary for recognizing them. The term 'high-level' refers to modeling speech utterances using the sequence of 'tokens' present in them. The co-occurrence pattern in the tokens, often termed as 'idiolect' [35], characterizes speaker differences. The tokens that are commonly used for speaker recognition may be in the form of phones [36], words [35], prosodic gestures [31, 34] or even articulatory movements [37]. Significant applications of these features for SR include [36, 38], where GMMs trained using individual sets of extracted tokens are used in parallel for classification. Due to their nature, high-level features can often be interchangebly used for speaker and language recognition [39]. Other approaches share similarities with the common prosodic features [31]. A study on the joint application of prosodic and high-level features for robust SR tasks can be found in [40]. However, high-level features are not a very attractive group to work with, due to the computational complexity involved in recognizing tokens.

The most common features for generalized speech related tasks as well as speaker recognition, are the family of spectral or spectro-temporal features. These features are extracted from short overlapping frames (10–25 ms) which are pre-emphasized and smoothed. Based on their interpretation they can be categorized as temporal, spectral or spectro-temporal. Popular examples of cepstral features are the Linear Prediction Cepstral Coefficient (LPCC) [8], Perceptual Linear Prediction (PLP) [41] coefficients and Mel Frequency Cepstral Coefficient (MFCC) [9], etc.

LPCCs are based on the principle of correlation of a sample with its adjacent ones. An instantaneous sample is approximated in terms of its neighborhood samples weighted with a set of predictor coefficients. The error in estimation is often termed as LP residual. The frequency domain equivalent of this representation is that of an all-pole filter with the same set of LP coefficients. The coefficients are determined by minimizing the residual energy using the Levinson Durbin algorithm [42]. The prediction coefficients instead of being used by themselves are transformed into a set of robust, less correlated features like LPCCs, PLP [41], Line Spectral Frequencies (LSF) [43], formant frequencies and bandwidth etc. [42].

The MFCCs [9] are the most successful and extensively used features for speaker recognition. MFCCs were psychoacoustically motivated in the sense that they were found to mimic the human auditory perception. MFCCs are extracted by a non-linear filter-bank analysis of the Discrete Fourier Transform (DFT) magnitude spectrum of short-term frames. The filterbank usually consist of a set of triangular band-pass filters, which are spaced according to the 'mel' scale. The log-magnitude of the filtered spectra is subjected to a Discrete Cosine Transform (DCT) for obtaining the cepstral features. MFCCs have arguably shown the best results compared to contemporary features like LPCC, PLP, LSF etc., in several prior works in SR using clean speech [10–12, 44]. Thus they are considered to be the default features for several speech related tasks including SR.

However the presence of background noise or channel effects inhibit the performance of MFCCs significantly primarily due to the distortion in the feature distribution [25]. The default use of MFCCs in most baseline SR systems necessitated the development of feature compensation methods as discussed in Sect. 2.1. However quite recently, researchers have focussed on alternative ways of modifying the cepstral feature extraction process for resistance towards ambient noise. Amongst several others, some notable features are Mean Hilbert Envelope Coefficient (MHEC) [45], Power Normalized Cepstral Coefficient (PNCC) [46] and Normalized Modulation Cepstral Coefficient (NMCC) [47]. Instead of modifying the features for compensating the effect of noise, features can be extracted from selective-regions of speech. Even in presence of noise also, glottal closure region in each pitch cycle and steady vowel regions contain high signal to noise ratio, and hence, features extracted from these regions are more robust compared to other regions of speech. In [48–50], features extracted from above mentioned regions are explored for robust speaker and language recognition tasks.

In MHEC extraction, the pre-emphasized speech is first decomposed into a number of spectral subbands using a gammatone filter constrained in the telephonic bandwidth of 300–3400 Hz. Unlike MFCC, the filters are uniformly spaced on an equivalent rectangular bandwidth (ERB) scale. The temporal envelope (Hilbert envelope) of each subband is estimated by using the Hilbert transform of the subband signal followed by low-pass filtering. The smoothed envelope is then used for deriving the required cepstral features. In PNCC extraction, the pre-emphasized signal is analyzed using short overlapping frames. A short-time Fourier analysis is performed over the Hamming windowed data, followed by frequency domain filtering using a gammatone filterbank constrained in 133 and 4,000 Hz, where the center frequencies of the gammatone bank are spaced equally in the ERB scale. The NMCCs are similar to PNCC except that amplitude modulation (AM) signals are estimated from the gammatone filtered sub-band signals using a Teager non-linear energy operator. The resulting signal is power normalized followed by DCT transform to obtain the cepstral features.

2.3 Model Compensation

Though feature-level compensation techniques are often applied as a front-end denoising process due to their low computational complexity and independence of any recognition model, they have certain limitations. In most cases, the feature compensation techniques produce point estimates of clean speech features. Due to this, they are unable to capture the uncertainty of observations which is represented as the variance of the conditional distribution of noisy speech given clean speech [51]. An alternative is to alter the statistical parameters of the acoustic model learned during the training phase to compensate for the channel or environmental mismatch of the evaluation phase. Since the evolution of statistical models for speech recognition, much research has been devoted in exploring model compensation issues in parallel [52, 53].

The earlier methods focussed on rendering the speaker models ineffective towards channel mismatches or handset variations [54]. In most cases, the mismatch would be caused due to unseen channel data during the evaluation phase. Unlike speech recognition tasks where multiple channel adaptation data could be obtained by pooling all speaker data over individual channels, SR required speaker-specific enrollment speech over multiple channels which could be later used for verification. This was unfavourable for practical SR applications. Alternate methods would cluster the data from a single conversation into multiple channel types to meet data requirements. Synthetic variance distribution [55] used an auxillary database of stereo recordings to artificially construct a global distribution of variances. Transformations derived from this distribution were used to modify the variance of individual speaker models. Speaker Model Synthesis (SMS) [56] learned speaker-independent transformations between different channels and applied it to synthesize speaker models under unseen enrollment conditions. The transformations were learned in the form of mean shift, weight scaling and variance scaling of GMM model parameters trained across various channel conditions.

In contrast to model-based channel compensation schemes, model-based environment adaptation methods developed during the contemporary period, modify speaker model parameters to reflect the acoustic environment of the evaluation phase. Two most popular data-driven environmental adaptation techniques initially proposed for robust speech recognition are *Maximum aPosteriori* (MAP)[57] and Maximum Likelihood Linear Regression (MLLR) [58]. The successful application of GMMs in the field of speaker recognition [59] has since then encouraged their usage in robust speaker verification (SV) tasks [60]. Both these methods use adaptation data to build speaker-specific models from a speaker independent background model constructed offline. MAP is a two stage process in which Bayesian statistics estimated using the training/adaptation data in the first stage, are used to update the ‘a priori’ available background model parameters (mean, covariances and weights) in the second stage. The Speaker Model Synthesis [56] method was based on deriving individual channel dependent GMMs by MAP-adaptation of a channel-independent background model. In another application, MAP was jointly used for model adaptation as well as feature transformation [61]. The advantage of MAP adaptation is its close approximation to the ideal maximum-likelihood estimates given sufficient enrollment data. However in situations where training data is sparse, MAP would only update a fractional number of GMM components. The MLLR adaptation technique transform the background GMM means and covariance matrices (optionally) by an affine transformation aiming at maximizing the likelihood function given new adaptation data. The parameters of the transformation are derived by iteratively using the Expectation Maximization (EM) algorithm [62]. Unlike MAP, all the GMM components are updated with limited amount of enrollment data. Other variants of MLLR like constrained MLLR (CMLLR) [63] are often used for online model adaptation [64]. However the performance improvement in MLLR-based methods saturates with increasing adaptation data and at a certain stage they are outperformed by MAP. A comparison of MLLR and Neural Network based environmental techniques was made in [65].

Apart from the traditional data driven methods that are dependent on adaptation data representing acoustic conditions of the evaluation phase, another approach is to exploit a priori information about the test environment. Popular state-of-the-art techniques in this category are Parallel Model Combination (PMC) [66, 67] and Vector Taylor Series (VTS) [68]. PMC relies on an available statistical noise model of the recognition phase and clean speaker GMMs trained during enrollment. The aim is to obtain noise-corrupted model for pattern matching, by combining the clean speech and noise models. This is done in two stages. Firstly, clean speaker models (GMMs/HMMs) and a simplified noise model (GMM) are built independently from clean training data and a noise signal, respectively. Secondly, the effect of additive noise on clean speech in the cepstral domain is analysed by using a function of noise corruption. This function is then extended to the parametric space to estimate the corrupted model parameters (mean and variances) from the clean and noise model parameters, respectively. Prior work shows that PMC model parameter estimation gets increasingly complex for dynamic and acceleration coefficients of MFCC. A recent state-of-the-art technique [69] addresses this problem by exploiting the relation between static and dynamic coefficients. The VTS method [70] uses a similar mathematical structure to represent the noise corruption process. However, unlike PMC the noise and channel statistics are obtained via an approximate taylor series expansion of the function around the mean of GMM components. This method is relatively much simpler compared to PMC and the tradeoff in terms of accuracy is not significant.

Though the model-compensation techniques perform better than their feature-level counterparts, they are computationally intensive and often require substantial amount of training data. Apart from the two broad types of compensation techniques discussed in Sects. 2.1 and 2.3, there exists hybrid approaches which can be termed as a combination of the two methods. Examples include Stochastic Matching [71] and Joint Uncertainty Decoding [72]. These methods account for the imperfections in feature enhancement process by approximating the marginal distribution of noisy features. In realistic situations, it may also turn out that the verification environment is entirely unknown [73]. In such scenarios, one might not expect availability of adaptation data or stereo training data. Quite recently, researchers have addressed this issue [74] by combining ‘missing feature theory’ based techniques [75] to subdue noise variation outside training conditions. The ‘posterior union model’ in [74], require detection and exclusion of the heavily mismatched subbands of the speech spectra. However, the improvement in performance accuracy of all these methods is usually associated with increased computational load and dependency on numerical approximations.

2.4 Robust Speaker Modeling

Speaker modeling techniques have been extensively explored in the past few decades of SR research. The scope of applying diverse pattern recognition techniques for classification and clustering of features makes this field an exciting area to work with. The broad classes of modeling techniques that are used in practice can be broadly categorized as generative models (GMM, VQ, Joint Factor Analysis (JFA)) or discriminative models (Neural Networks (NN) and Support Vector Machines (SVMs)). A family of hybrid modeling techniques also exist which are a combination of both e.g., GMM-SVM, SVM-JFA, etc. In this subsection we shall briefly discuss each.

Vector Quantization, introduced in the late 1980s [76] is one the most primitive form of SR model. Based on the principle of K-means [62], the set of feature vectors extracted from a speaker’s training utterance are grouped into a number of non-overlapping clusters. Individual speaker models are represented by the stack of cluster centroids often termed as codebook. Classification of a test utterance is based on minimization of a distortion measure commonly given by the average Euclidean distance of a vector from each codebook. Despite its crude form of clustering, VQ is often used for computational speedup required for real-time SR applications [77].

The Gaussian Mixture Models (GMMs) introduced in the mid-1990s [59] is widely considered to be a benchmark for modern text-independent Speaker Recognition. In contrast to VQ, a number of overlapping multivariate Gaussian functions are used to cluster the feature space. The GMMs are able to characterize general properties like multi-modal feature distribution, speaker-dependent spectral shapes etc. Unlike VQ, GMMs are able to capture the variance of feature distribution. In contrast to the naive K-means, GMM training is based on a more rigorous approach of maximizing the likelihood of a given speaker’s data. The parameters are estimated iteratively using the Expectation Maximization (EM) algorithm [62]. Classification of test utterances are done on the basis of log likelihood scores obtained from the sequence of test vectors. Though speaker-specific GMMs performed reasonably well for SR given clean speech, a good amount of data was required for parameter estimation. Besides, a more generalized approach was required for unifying model-compensation techniques with the GMM framework. A novel GMM-based approach was proposed in [60], where a single speaker-independent GMM (Universal Background Model (UBM)) trained using multiple speaker data across various channels and sessions, was used as a common impostor model for speaker verification. HMM-based speaker models were derived using MAP and MLLR adaptation of the UBM using the speaker's training data. Besides reducing data requirements, these techniques provided scope for model adaptation as discussed in Sect. 2.3. Comparative studies of alternate adaptation techniques were made in [78]. Efforts were also made to approximate the common MAP adaptation process in terms of a VQ model [79]. However in the context of environmental robustness, GMMs often provide limited performance improvement despite model-adaptation. This problem received a new direction with the introduction of Joint Factor Analysis and its variants [80].

Prior to the introduction of GMMs, role of Neural Networks (NN) for text-independent SR was first studied in [81]. An advantage of NNs is its ability to perform feature transformation and speaker modeling simultaneously [82]. In a later study, Auto Associative Neural Networks (AANNs) were introduced for speaker modeling [83]. Since GMMs relied on first and second order statistics, it was hypothesized in [83] that they fail to capture feature distribution based on higher order statistics. AANNs were found to be effective for SR tasks where distribution of data is highly non-linear [83]. However, NNs have not been used much in practice primarily due to the heavy computational costs involved in training them. Besides, prior determination of the appropriate structure for NNs (number of neurons in each layer) is a non-trivial task.

Support Vector Machines (SVMs) have emerged as a powerful discriminative classifier in the field of robust SR in the last decade [32, 36, 84]. A SVM is a binary classifier which distinguishes between two classes (true speaker and impostor) by learning a decision hyperplane which separates them in some higher dimensional feature space [62]. SVMs have been initially used to model individual speakers using high-level [36] and prosodic features [32]. However the real significance of SVMs in robust SV tasks was found in its effective combination with the traditional GMM classifier [85]. A novel method of representing variable length training utterances using fixed-length vectors was discovered contemporarily. The mean vectors of MAP-adapted speaker GMMs were stacked together to produce a high dimensional vector commonly termed as a 'supervector'. The labelled supervectors were used as input the SVMs. This led to the scope of exploring various 'sequence kernels' or non-linear mappings for transforming features to high dimensional spaces [85–87]. Several normalization techniques for minimizing inter-session and intra-speaker variabilities in the supervector space have been introduced since then. Common examples are Nuisance Attribute Projection (NAP) [88], Within Class Covariance Normalization (WCCN) [89] and Linear Discriminant Analysis (LDA) [62]. The GMM-SVM approach is often considered as a effective alternative of the GMM-UBM method.

Supervector-based speaker recognition opened an interesting new direction for compensating channel and session variabilities. It was thought that channel variations in recorded training utterances might lead to the problem of mismatch in the supervector space. A feasible alternative was to explicitly model the channel variability by representing the supervector space as a combination of statistically independent channel and speaker subspaces. This approach was named Joint Factor Analysis (JFA) [80] where the term ‘factor’ denotes the low-dimensional projection of the speaker or channel supervectors in their corresponding spaces. JFA as a new research trend has been extensively studied for robust SR tasks since the late 2000s [80, 90]. However it was later argued in [91], that instead of two distinct subspaces a single ‘total variability’ space could in fact be useful for simultaneously representing both speaker and channel variabilities. A low-dimensional projection of the supervectors in the total variability space, commonly known as ‘i-vectors’ has since then been considered as the modern state-of-the-art in robust speaker verification. Various studies have since then been conducted to combine JFA and SVM based methods with appropriate normalization techniques [92]. Quite recently, i-vector based studies have conducted for robust speaker recognition tasks where authors have proposed alternative methods of projecting the i-vectors into a subspace for improved speaker discrimination and suppression of channel-effects [93].

2.5 Motivation for the Present Work

Robust speaker recognition in noisy environments till date remains an open issue despite the diverse array of methods developed to address it in past. The ever increasing usage of hand-held devices in the modern era has driven new demand for robust speaker recognition applications. Despite being well explored in past, new methods keep unfolding in this field which are either suggested improvements or alternatives of the existing ones. This makes robust SR a very challenging and yet an interesting area to work in.

Despite the availability of robust features as discussed in Sect. 2.2, feature compensation techniques play a crucial role for SV applications that demand noise-robustness without compromising on speaker-discriminative power [94]. An interesting fact to notice about the state-of-the art data-driven feature compensation methods discussed in Sect. 2.1, is that their application has mostly been restricted to robust speech recognition tasks but rarely studied for robust SV tasks. The brief discussion about model compensation techniques in Sect. 2.3 reveal some of their vulnerabilities. They either rely explicitly on an available clean speaker model (e.g., PMC [66], VTS [70]) or a priori knowledge about the noisy environment (e.g., noise model for PMC, adaptation data for MAP [57], MLLR [58]). These drawbacks suggest the use of robust speaker modeling methods as an alternative for practical scenarios (e.g., unknown noisy environment, unavailable clean speaker models). In a similar context it can be argued that the state-of-the-art robust speaker modeling methods (e.g., GMM supervectors [85], i-vectors [95] etc.) have mostly been applied to counter channel/handset mismatches but not additive background noise specifically.

Summarily, the above two points motivates us to propose new studies in which we explore the application of feature enhancement techniques for speaker verification in additive background noise. Studies are also conducted to demonstrate the effectiveness of supervector-based approaches and its state-of-the-art variants (e.g., i-vectors) for robust speaker verification in noisy environments.

References

1. J. Campbell, Speaker recognition: a tutorial. Proc. IEEE **85**(9), 1437–1462 (1997)
2. 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)
3. 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)
4. T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. **52**, 12–40 (2010)
5. S. Sarkar, Robust speaker recognition in noisy environments. Master's thesis, School of Information Technology, Indian Institute of Technology Kharagpur, Mar 2014
6. R. Schafer, L. Rabiner, Digital representations of speech signals. Proc. IEEE **63**(4), 662–677 (1975)
7. B. Atal, Automatic recognition of speakers from their voices. Proc. IEEE **64**(4), 460–475 (1976)
8. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
9. 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)
10. A. Acero, Acoustical and environmental robustness in automatic speech recognition. PhD thesis, Carnegie Mellon University, Sept 1990
11. D.A. Reynolds, Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech Audio Process. **2**(4), 639–643 (1994)
12. R. Mammone, X. Zhang, R. Ramachandran, Robust speaker recognition: a feature-based approach. IEEE Signal Process. Mag. **13**(5), 58–71 (1996)
13. D. Reynolds, The effects of handset variability on speaker recognition performance: experiments on the Switchboard corpus, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, 1996, vol. 1, pp. 113–116
14. K.S. Rao, J. Yadav, S. Sarkar, S.G. Koolagudi, A.K. Vuppala, Neural network based feature transformation for emotion independent speaker identification. Int. J. Speech Technol. (Springer) **15**(3), 335–349 (2012)
15. S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**(2), 254–272 (1981)
16. H. Hermansky, N. Morgan, RASTA processing of speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994)
17. A. Kocsor, L. Toth, Kernel-based feature extraction with a speech technology application. IEEE Trans. Signal Process. **52**(8), 2250–2263 (2004)
18. T.G. Stockham, T.M. Cannon, R.B. Ingebretsen, Blind deconvolution through digital signal processing. Proc. IEEE **63**(4), 678–692 (1975)
19. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
20. A. Erell, M. Weintraub, Spectral estimation for noise robust speech recognition, in *Proceedings of DARPA Speech and Natural Language Workshop*, Philadelphia, 1989

21. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
22. A. Acero, R.M. Stern, Environmental robustness in automatic speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '90)*, Albuquerque, 1990, vol. 2, pp. 849–852
23. S. Suhadi, S. Stan, T. Fingscheidt, C. Beaugeant, An evaluation of VTS and IMM for speaker verification in noise, in *Proceedings of 4th Annual Conference of the International Speech Communication Association (INTERSPEECH '03)*, Geneva, 2003, pp. 1669–1672
24. L. Deng, J. Droppo, A. Acero, Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition. IEEE Trans. Speech Audio Process. **11**(6), 568–580 (2003)
25. P.J. Moreno, B. Raj, R.M. Stern, Data-driven environmental compensation for speech recognition: a unified approach. Speech Commun. **24**(4), 267–285 (1998)
26. L. Deng, A. Acero, M. Plumpe, X. Huang, Large-vocabulary speech recognition under adverse acoustic environments, in *Proceedings of the International Conference of Spoken Language Processing (ICSLP '00)*, Beijing, 2000, pp. 806–809
27. L. Deng, A. Acero, L. Jiang, J. Droppo, X. Huang, High-performance robust speech recognition using stereo training data, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, 2001, vol. 1, pp. 301–304
28. L. Buera, E. Lleida, A. Miguel, A. Ortega, Multi-environment models based linear normalization for speech recognition in car conditions, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, Montreal, 2004
29. M. Afify, X. Cui, Y. Gao, Stereo-based stochastic mapping for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. **17**(7), 1325–1334 (2009)
30. L. Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition. Speech Commun. **50**, 782–796 (2008)
31. A.G. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey, Modeling prosodic dynamics for speaker recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, Hong Kong, 2003
32. L. Ferrer, E. Shriberg, S. Kajarekar, K. Sonmez, Parameterization of prosodic feature distributions for SVM modeling in speaker recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, Honolulu, 2007, pp. 233–236
33. S.G. Koolkagudi, K.S. Rao, R. Reddy, A.K. Vuppala, S. Chakrabarti, Robust speaker recognition in noisy environments: using dynamics of speaker-specific prosody, in *Forensic Speaker Recognition* (Springer, New York, USA, 2013), pp. 183–204
34. E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, A. Stolckea, Modeling prosodic feature sequences for speaker recognition. Speech Commun. **46**, 455–472 (2005)
35. G. Doddington, Speaker recognition based on idiolectal differences between speakers, in *Proceedings of the European Conference of Speech Communication Technology (EUROSPEECH '01)*, Aalborg, 2001, pp. 2521–2524
36. W.M. Campbel, J.P. Campbell, D.A. Reynolds, D.A. Jones, T.R. Leek, Phonetic speaker recognition with support vector machines, in *Proceedings of the Neural Information Processing Systems Conference*, Vancouver, 2003, pp. 1377–1384
37. K. yee Leung, M. wai Mak, M. Siu, S. yuan Kung, Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. Speech Commun. **48**, 71–84 (2006)
38. B. Ma, D. Zhu, H. Li, R. Tong, Speaker cluster based GMM tokenization for speaker recognition, in *Proceeding of the 7th Annual Conference of the International Speech Communication Association (INTERSPEECH '06)*, Pittsburgh, 2006
39. B. Ma, H. Li, R. Tong, Spoken language recognition using ensemble classifiers. IEEE Trans. Audio Speech Lang. Process. **15**(7), 2053–2062 (2007)

40. D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adomi, Q. Jin, D. Kluracek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, S. Xiang', The supersid project: exploiting high-level information for high-accuracy speaker recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, Hong Kong, 2003
41. H. Hermansky, Perceptual linear prediction (PLP) analysis for speech. J. Acoust. Soc. Am. **87**, 1738–1752 (1990)
42. L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, 1st edn. (Prentice-Hall, Englewood Cliffs, 1993)
43. X. Huang, A. Acero, H. Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development* (Prentice Hall, Upper Saddle River, 2001)
44. S. Sarkar, K.S. Rao, D. Nandi, Multilingual speaker recognition on Indian languages, in *IEEE INDICON*, Mumbai (IIT Mumbai, Mumbai, 2013)
45. J.W. Suh, S.O. Sadjadi, G. Liu, T. Hasan, K.W. Godin, J.H. Hansen, Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA, in *Proceedings of NIST Speaker Recognition Evaluation Workshop*, Gaithersburg, USA, 2011
46. C. Kim, R.M. Stern, Power-normalized cepstral coefficients (PNCC) for robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, Kyoto, 2012
47. V. Mitra, H. Franco, M. Graciarena, A. Mandal, Normalized amplitude modulation features for large vocabulary noise-robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, Kyoto, 2012
48. A.K. Vuppala, K.S. Rao, Speaker identification under background noise using features extracted from steady vowel regions. Int. J. Adapt. Control Signal Process. **27**(9), 781–792 (2013). Wiley
49. A.K. Vuppala, K.S. Rao, S. Chakrabarti, Improved speaker identification in wireless environment. Int. J. Signal Imaging Syst. Eng. **6**(3), 130–137 (2013)
50. K.S. Rao, S. Maity, V.R. Reddy, Pitch synchronous and glottal closure based speech analysis for language recognition. Int. J. Speech Technol. **16**, 413–430 (2013). Springer
51. T. Kristjansson, B. Frey, Accounting for uncertainity in observations: a new paradigm for robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, Orlando, 2002, vol. 1, pp. 61–64
52. C.H. Lee, On stochastic feature and model compensation approaches to robust speech recognition. Speech Commun. **25**, 29–47 (1998)
53. C.H. Lee, Q. Huo, On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proc. IEEE **88**(8), 1241–1269 (2000)
54. T. Quatieri, D. Reynolds, G. O'Leary, Estimation of handset nonlinearity with application to speaker recognition. IEEE Trans. Speech Audio Process. **8**, 567–584 (2000)
55. H.A. Murthy, F. Beaufays, L.P. Heck, M. Weintraub, Robust text-independent speaker identification over telephone channels. IEEE Trans. Speech Audio Process. **7**(5), 554–568 (1999)
56. R. Teunen, B. Shahshahani, L. Heck, A model-based transformational approach to robust speaker recognition, in *Proceeding of the Annual Conference of the International Speech Communication Association (INTERSPEECH '00)*, Beijing, 2000, vol. 2, pp. 495–498
57. 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)
58. C. Leggetter, P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput. Speech Lang. **9**, 171–185 (1995)
59. 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)
60. D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. **10**(1), 19–41 (2000)
61. D. Zhu, B. Ma, H. Li, Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, Taipei, 2009, pp. 4045–4048

62. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006)
63. V. Digalakis, D. Rtischev, L. Neumeyer, E. Sa, Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Trans. Speech Audio Process. **3**(5), 357–366 (1995)
64. S. Kozat, K. Visweswariah, R. Gopinath, Feature adaptation based on Gaussian posteriors, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, 2006, pp. 221–224
65. K.K. Yiu, M.W. Mak, S.Y. Kung, Environment adaptation for robust speaker verification, in *Proceedings of the European Conference of Speech Communication and Technology (EUROSPEECH '03)*, Geneva, 2003, vol. 2, pp. 2973–2976
66. M.J.F. Gales, S.J. Young, Robust speech recognition in additive and convolutional noise using parallel model combination. Comput. Speech Lang. **9**, 289–307 (1995)
67. L.P. Wong, M. Russell, Text-dependent speaker verification under noisy conditions using parallel model combination, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'01)*, Salt Lake City, 2001, pp. 457–460
68. P. Moreno, Speech recognition in noisy environments. PhD thesis, Electrical & Computer Engineering Department, Carnegie Mellon University, Pittsburgh, 1996
69. K.C. Sim, M.T. Luong, A trajectory-based parallel model combination with a unified static and dynamic parameter compensation for noisy speech recognition, in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU '11)*, Waikoloa, Dec 2011, pp. 107–112
70. P.J. Moreno, B. Raj, R.M. Stern, A vector Taylor series approach for environment-independent speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, 1996, pp. 733–736
71. A. Sankar, C.H. Lee, Stochastic matching for robust speech recognition. IEEE Signal Process. Lett. **1**(8), 124–125 (1994)
72. H. Liao, M.J.F. Gales, Joint uncertainty decoding for noise robust speech recognition, in *Proceedings of 6th Annual Conference of the International Speech Communication Association (INTERSPEECH '05)*, Lisbon, 2005
73. J. Ming, D. Stewart, S. Vaseghi, Speaker identification in unknown noisy conditions – a universal compensation approach, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, Philadelphia, 2005
74. 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)
75. A. Drygajlo, M. El-Maliki, Speaker verification in noisy environment with combined spectral subtraction and missing data theory, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, Seattle, 1998
76. D. Burton, Text-dependent speaker verification using vector quantization source coding. IEEE Trans. Acoust. Speech Signal Process. **35**(2), 133–143 (1987)
77. T. Kinnunen, E. Karpov, P. Franti, Real-time speaker identification and verification. IEEE Trans. Audio Speech Lang. Process. **14**(1), 277–288 (2006)
78. M.W. Mak, R. Hsiao, B. Mak, A comparison of various adaptation methods for speaker verification with limited enrollment data, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, Toulouse, 2006, pp. 929–932
79. 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)
80. P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Factor analysis simplified, in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP '05)*, Philadelphia, 2005, vol. 1, pp. 637–640
81. K. Farrell, R. Mammone, K. Assaleh, Speaker recognition using neural networks and conventional classifiers. IEEE Trans. Speech Audio Process. **2**(1), 195–204 (1994)
82. L.P. Heck, Y. Konig, M. Sonmez, M. Weintraub, Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Commun. **31**, 181–192 (2000)

83. B. Yegnanarayana, S.P. Kishore, AANN: an alternative to GMM for pattern recognition. Neural Netw. **15**, 456–469 (2002)
84. 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)
85. W. Campbell, J. Campbell, D. Reynolds, Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
86. V. Wan, S. Renals, Speaker verification using sequence discriminant support vector machines. IEEE Trans. Acoust. Speech Audio Process. **13**(2), 203–210 (2005)
87. C.H. You, K.A. Lee, H. Li, An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. IEEE Signal Process. Lett. **16**(1), 49–52 (2009)
88. A. Solomonoff, C. Quillen, I. Boardman, Channel compensation for SVM speaker recognition, in *IEEE Workshop on Speaker and Language Recognition (Odyssey '04)*, Toledo, 2004, pp. 57–62
89. A.O. Hatch, S. Kajarekar, A. Stolcke, Within-class covariance normalization for SVM-based speaker recognition, in *Proceedings of the International Conference of Spoken Language Processing (ICSLP '05)*, Lisbon, Portugal, 2005
90. P. Kenny, G. Boulianne, P. Dumouchel, Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process. **13**(3), 345–354 (2005)
91. N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, P. Dumouchel, Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in *Proceeding of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, Brighton, 2009
92. N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, F. Castaldo, Support vector machines and joint factor analysis for speaker verification, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, Taipei, 2009, pp. 4237–4240
93. M. McLaren, D. van Leeuwen, Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources. IEEE Trans. Audio Speech Lang. Process. **20**(3), 755–766 (2012)
94. T. Kinnunen, Spectral features for automatic text-independent speaker recognition. PhD thesis, Department of Computer Science, University of Joensuu, 2004
95. N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)

Springer

http://www.springer.com/978-3-319-07129-9

Robust Speaker Recognition in Noisy Environments
Rao, K.S.; Sarkar, S.
2014, XII, 139 p. 31 illus., 25 illus. in color., Softcover
ISBN: 978-3-319-07129-9