An Overview of Speaker Identification: Accuracy and Robustness Issues

Roberto Togneri
and Daniel Pullella

Abstract

This paper presents the main paradigms for speaker identification, and recent work on missing data methods to increase robustness. The feature extraction, speaker modeling and system classification are discussed. Evaluations of speaker identification performance subject to environmental noise are presented. While performance is impressive in clean speech conditions, there is rapid degradation with mismatched additive noise. Missing data methods can compensate against arbitrary disturbances and remove environmental mismatches. An overview of missing data methods is provided and applications to robust speaker identification summarized. Finally combined approaches involving bottom-up estimation and top-down processing are reviewed, and their significance discussed.

I. Introduction

Biometric recognition systems are increasingly being deployed as a more "natural" means for the recognition of people. Instead of remembering passwords and PINs (which can be stolen or forgotten) or written signatures (which can be forged), biometric cues such as fingerprints, voice and face are specific to an individual (and hence cannot easily be stolen or forged) and characterizes that individual (and hence cannot be forgotten) [1]. The simplest to acquire, most used and pervasive in society, and least obtrusive biometric measure is that of human speech. Thus we refer to speaker recognition systems as those technologies which utilize human speech to recognize, identify or verify an individual [2]. Some of the key, early papers providing an overview of speaker recognition systems and the various paradigms popular at the time can be found in [3]–[5], for a more modern treatment [2], [6]–[8].

A. Speaker Identification and Verification Systems

Human society functions by communication between individuals. Language in both its written and spoken form underpin all aspects of human interactions. The spoken language is the most fundamental as this is how

Digital Object Identifier 10.1109/MCAS.2011.941079
Date of publication: 27 May 2011

Since spoken language is one of the easiest measures to acquire (all you need is a microphone), it comes as no surprise that speaker recognition is one of the key research areas in signal processing and pattern recognition.

individuals communicate with one another using only the human vocal apparatus. Consequently the acoustic signal of human speech characterizes not only what is being said but also embodies individual characteristics of the speaker, in particular individual pitch and vocal tract resonances as well as speaking styles and durations. Since spoken language is one of the easiest measures to acquire (all you need is a microphone), is used in a variety of transaction applications (e.g. telephone banking), and has the potential for security by surveillance (using eavesdropping technology) it comes as no surprise that speaker recognition is one of the key research areas in signal processing and pattern recognition. In security applications where a person has to be recognized there are two distinct modes of operation: identification and verification.

In *speaker identification* (see Fig. 1) human speech from an individual is used to identify who that individual is. There are two distinct operational phases. In *training* (also called *enrolment*) the speech from each known, verified speaker, for all speakers that need to identified, is acquired to build (train) the model for that speaker. Usually this is carried out off-line as part of the system configuration and before the system is deployed. In *testing* the true operation of the system is carried out where the speech from an unknown utterance is compared against each of the trained speaker models. In *closed-set identification* the unknown individual belongs to a pre-existing pool or database of speakers (speaker models) and the problem then becomes that of choosing which speaker from the pool the unknown speech is derived from. The main performance measure of such systems is the identification rate (percentage of correct identification averaged across all speakers in the pool). Closed-set identification is typical in departmental organizations where the group members are known, their speaker profiles can be acquired and stored in a database, and for which identification is internal to the department (i.e. there are no "outside" users). In *open-set identification* the unknown individual can come from the general population. However as identification is always carried out against a finite, known pool of individuals it is not possible to identify arbitrary people. Thus the first task of an open-set identification system is to detect whether the speaker belongs to the pool or database of known speakers, if not, that speaker is rejected, otherwise, closed-set identification is carried out. It is important in these systems to detect whether a speaker belongs to the pool, otherwise a random individual from the pool will always be identified.

In *speaker verification* (see Fig. 2) human speech from an individual is used to verify the claimed identity of that individual. As with speaker identification the initial configuration of the system is carried out during the training or enrolment when each speaker to be verified by the system has to provide samples of speech which are then used to train the model for that speaker. In testing the verification takes place when the individual has

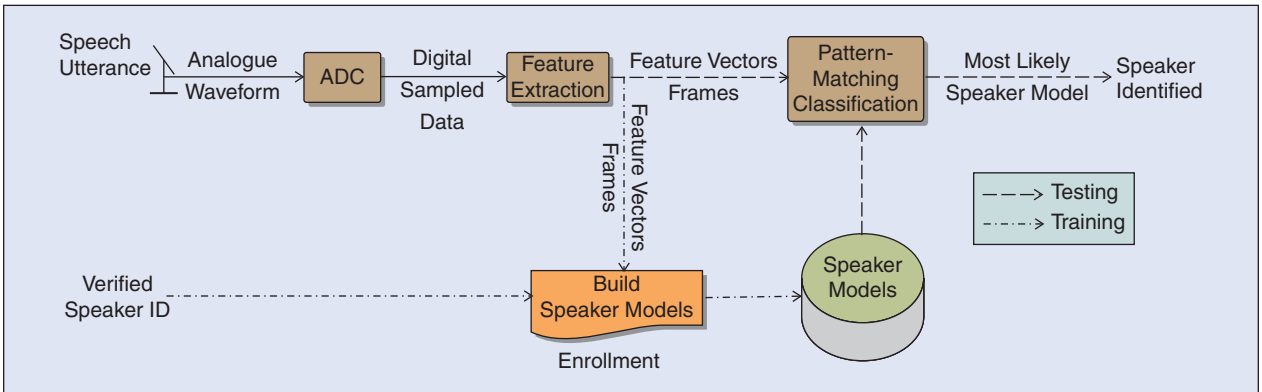

Figure 1. Block Diagram of a Speaker Identification System.

Roberto Togneri and Daniel Pullella are with The University of Western Australia. E-Mail: Roberto.Togneri@uwa.edu.au.
This work was supported in part by the Samaha Research Scholarship (F8046) of The University of Western Australia.

Most forms of security based transactions (e.g. telephone banking) require an individual to be verified rather than identified (i.e. a claim is made as to who the individual is supposed to be, and this has to be verified).

to make a claim as to who he/she is, and the system then proceeds to verify whether that claim is true or false. With speaker verification the speech of the unknown person is compared against both the claimed identity and against all other speakers (the imposter or background model(s)). The ratio of the two measures is then taken and compared to a threshold, if above the threshold the claim is accepted as true, if below, the claim is rejected as false. In verification systems two key performance measures are popular, the *false rejection rate* (FRR), the number of times the true speaker is incorrectly rejected, and *false acceptance rate* (FAR), the number of times an imposter speaker is incorrectly accepted. By varying the decision threshold the FAR and FRR will change in opposing directions. For example raising the threshold will lower FAR but increase the FRR as true claims will start to be rejected since the "bar" is raised, conversely if the threshold is lowered the FRR is reduced but FAR will increase since not only are all true claims now accepted but more false ones will as well. The typical operating point for the selection of the threshold is when FAR = FRR, termed the *equal error rate* (EER) condition.

Most forms of security based transactions (e.g. telephone banking) require an individual to be verified rather than identified (i.e. a claim is made as to who the individual is supposed to be, and this has to be verified). However applications of speaker recognition involving surveillance, monitoring and automated ID tagging will usually require identification rather than verification. In closed-set speaker identification an unknown utterance has to be compared against all speaker models in the pool and as the number of speakers in the pool increases performance is degraded (both in terms of accuracy and computational burden). However since in speaker verification the unknown utterance is only compared against the one claimed model and the other imposter model, the verification is faster and does not degrade with increasing number of speakers in the pool. Furthermore speaker verification is able to reject speech from arbitrary speakers (i.e. the open-set case) which is not true for speaker identification. Not surprisingly the majority of research in speaker recognition concentrates on speaker verification, given the practical importance of this to, especially, secure call telephony applications. The interested reader is referred to [2],[6],[7],[9],[10] for a review of the methods and technologies specifically directed at speaker verification.

Although human speech is used to identify or verify an individual it should be remembered that human speech is primarily used to convey meaning by words. Thus it is possible to strengthen speaker recognition by also performing speech recognition (what is being said). By restricting the speech to specific words or phrases (typically passwords or PINs, or even one's own name) *text-dependent* speaker recognition is possible. This

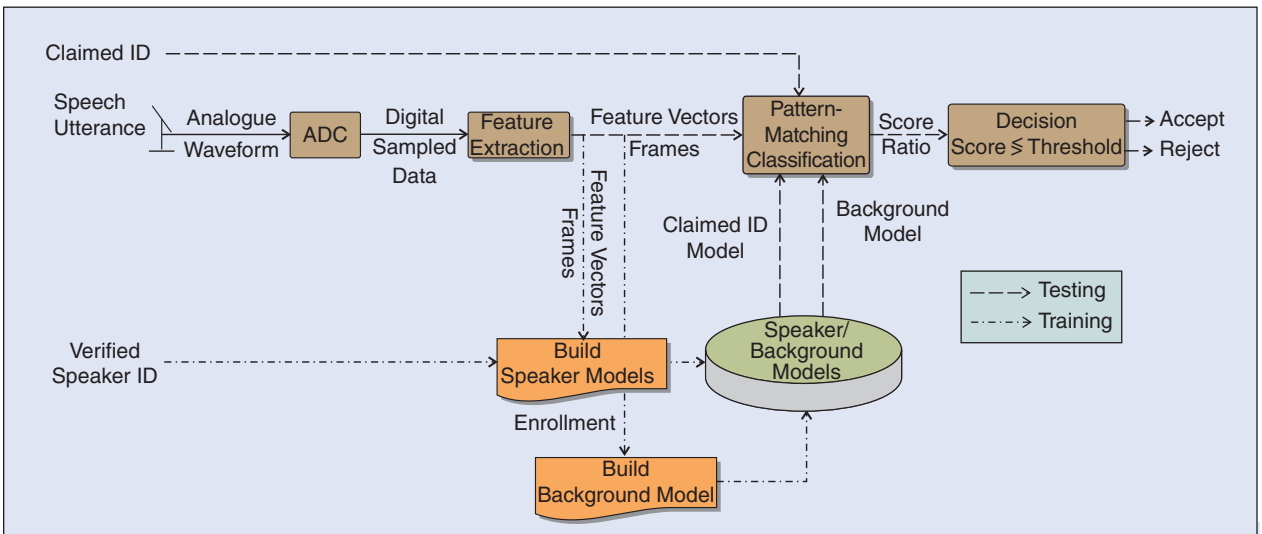

Figure 2. Block diagram of a speaker verification system.

requires both the words being spoken to be correctly identified and then identified as originating from the claimed speaker (or identifying who that speaker is). Although text-dependent speaker recognition can result in improved performance these systems are more complex (requiring the use of some form of speech recognition) and their practical application limited. Consequently there is greater interest in *text-independent* speaker recognition which recognizes an individual without any constraints on what the individual is saying (although it is assumed the individual is actually speaking and, in most cases, in the language of interest).

In this tutorial overview the key emphasis will be on text-independent, closed-set, speaker identification based on some of the key advances in the modeling and classification paradigms in the area. We commence our review with the key features extracted from the acoustic waveform in Section II, the main speaker modeling paradigms in Section III, the key classification paradigms in Section IV and then in Section V present some recognition results on a standard speech corpus both in clean and in noisy conditions. It should be remembered that the technologies used in speaker identification are interchangeably adopted by speaker verification systems as can be seen by the common processing blocks in Fig. 1 and Fig. 2.

B. Robustness of Speaker Identification Systems

Practical speaker recognition systems are often subject to noise or distortions within the input speech which degrades performance. In systems deployed for telephony applications and in office environments 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. Typical solutions to providing robustness in speaker recognition can be categorized as feature-based, score-based or model-based approaches. Feature-based methods remove the noise from the speaker characteristic information directly, and include methods such as cepstral mean normalization [11], RASTA processing [12], warping methods [13] and robust parameterizations [14], [15]. Score-based methods alter the classifier scores at the utterance or frame level, while model-based approaches (such as parallel model combination [16]) attempt to incorporate distortion characteristics into the speaker models themselves to achieve robustness.

Traditionally compensation against channel induced distortion was of primary concern due to the prevalence of telephone recorded speech, and the desire to perform robust speaker recognition on this speech (typically for security related applications). However, the recent interest in speaker recognition technology for far field type commercial applications has increased the need for environment distortion compensation. While effective for channel distortion compensation, the previously proposed feature, score and model based robustness methods are limited in their suitability for environmental disturbance compensation: typically these techniques require strict assumptions about the nature of the environmental disturbance (such as stationarity), or require explicit noise modeling (as in the model-based paradigm) resulting in poor performance in unseen noise conditions.

Missing data methods are capable of compensating against additive distortions of arbitrary type, and are thus naturally suited to the problem of environmental noise mismatch. These methods are based on the time-frequency representation of the speech signal and the labeling of each individual time-frequency point as speech or noise dominant ('reliable' or 'unreliable'). Constructing these labels in the form of a time-frequency *reliability mask* allows robust recognition to be performed via a reconstruction of the speech spectrogram, or by integrating over the noise dominated points. The effectiveness of these missing data approaches is critically dependent on the accuracy of the decisions within the reliability mask. In the case of a priori noise knowledge the ideal reliability mask can be constructed resulting in high robustness to extreme non-stationary noises. However, in practice a priori noise knowledge is not available and so the reliability masks must be estimated. Past research has concentrated on producing accurate mask estimates by utilizing the properties of the speech signal as well as auditory and perceptual principles. The weakness of these traditional *bottom-up* approaches is the lack of protection afforded to the recognizer against errors in the estimated reliability decisions. By utilizing *top-down* knowledge from the trained models a significant amount of these errors may be removed, thus providing a modification to the GMM speaker recognition paradigm which is extremely noise robust.

In the latter part of this tutorial we provide a review of the techniques which are beneficial for robust speaker recognition. This overview begins in Section VI with an examination of traditional feature space, score based and model based approaches to noise compensation. The missing data approach is introduced in Section VII which initially presents the fundamentals of spectrographic masking, reliability mask construction and also provides a brief summary of the different missing data recognition strategies. A review of bottom-up mask estimation methods for speaker

identification is then presented including SNR-based approaches, auditory and perceptual methods, and classifier-based techniques. Top-down only approaches for constructing reliability decisions are then discussed with emphasis on speaker recognition. Finally, motivated by the desire for efficient and robust recognition, we discuss approaches which combine bottom-up and top-down sources of information. Specifically, two distinct combined approaches are reviewed in the context of speech recognition, followed by a discussion of the potential of combining information sources for missing data speaker recognition tasks.

II. Feature Extraction

A. Frames

The most fundamental process common to all forms of speaker and speech recognition systems is that of extracting vectors of features uniformly spaced across time from the time-domain sampled acoustic waveform. Irrespective of the features derived from the waveform (of which there are many) the initial framing of the waveform, with reference to Fig. 3, proceeds as follows (the numerical parameter values mentioned are those typically adopted in practice):

a) ***Pre-emphasis:*** A high-pass filter is applied to the waveform. This emphasises the higher frequencies and compensates for the human speech production process which tends to attenuate high frequencies. A simple 1st order, high-pass filter is used, with a typical co-efficient value of 0.97 (i.e. the filter function is $y(t) = x(t) - 0.97x(t-1)$ where $x(t)$ is the input speech data and $y(t)$ is the output).

b) ***Framing:*** The time-domain waveform of the utterance under consideration is divided into overlapping fixed duration segments called *frames*. Typical duration values for frames are anywhere from 20 ms to 30 ms (usually 25 ms) and a frame is generated every 10 ms (thus consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).

c) ***Windowing:*** Each frame is multiplied by a window function. The window function is needed to smooth the effect of using a finite-sized segment for the subsequent feature extraction by tapering each frame at the beginning and end edges. As most features are spectral in nature the Fourier Transform is employed and the multiplicative effect of the window function in the time domain is convolutive in the spectral domain. A tapered window function creates a smoother and less distorted (by artefacts) spectrum. Without a specified window function the default arising from the framing operation is that of a rectangular window effect which will generate undesirable spectral artefacts. Any of the window functions used in FIR digital filter design can be deployed, with the Hamming window function being the most popular.

B. MFCC Features

For speaker recognition it is important to extract features from each frame which can capture the speaker-specific characteristics. Many such features have been investigated in the literature [17]. Linear Prediction Co-efficients (LPCs) have received special attention in this regard [18] as they are directly derived from the speaker's speech production model. So too Perceptual Linear Prediction (PLP) co-efficients [17], [19] as these

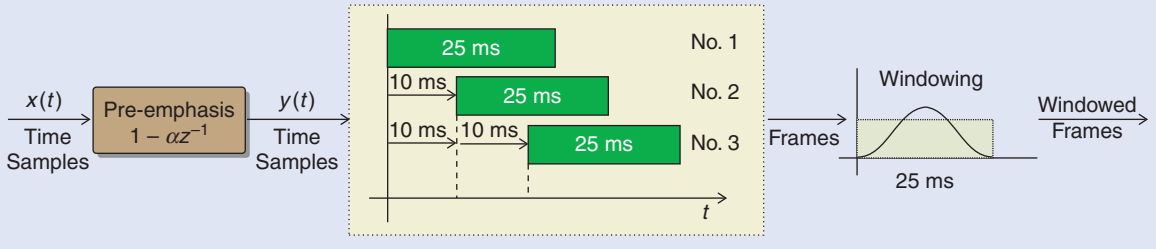

Figure 3. Analysis block diagram for framing.

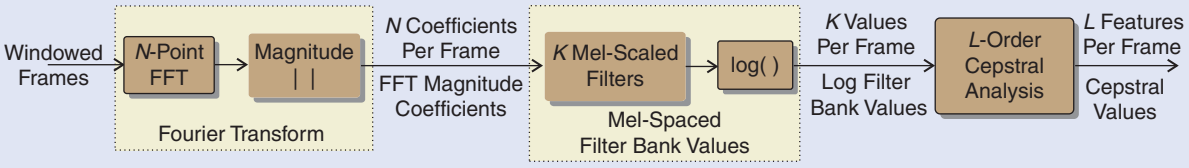

Figure 4. Analysis Block Diagram for MFCC feature vectors.

are based on human perceptual and auditory processing. However over the past two decades spectral based features, most typically derived by direct application of the Fourier Transform, have become popular. Investigations [19] have shown that the same features adopted in speech recognition are equally successful when applied to speaker recognition. These features are Mel-Frequency spaced Cepstral Co-efficients (MFCCs) and their success arises from the use of perceptually based Mel-spaced filter bank processing of the Fourier Transform and the particular robustness (to the environment) and flexibility that can be achieved using cepstral analysis. Referring to Fig. 4 MFCC features are derived as follows:

a) ***Fourier Transform:*** A Fast Fourier Transform (FFT) operation is applied to each frame to yield complex spectral values. If, say, a 512-point FFT is applied, then 256 complex spectral values uniformly spaced from 0 to $F_s/2$ (where F_s is the sampling frequency) are produced (ignoring the mirror values). In speech processing the phase information is ignored and only the FFT magnitude spectrum is considered.

b) ***Mel-spaced filter bank values:*** The N FFT magnitude co-efficients are converted to K filter bank values. This is necessary since $N = 256$ represents too much spectral detailed information and by smoothing the spectrum to only $K = 30$, or so, values per frame a more efficient representation is achieved. Furthermore this can be carried out in a perceptually meaningful way by smoothing logarithmically rather than linearly, specifically using a Mel or Bark scale. The filter bank values are derived by cross-wise multiplying the N FFT magnitude co-efficients by the K triangular filter bank weighting function from Fig. 5 and then accumulating or *binning* the results from each filter triangle. The centers of the triangle filter banks are spaced according to the Mel scale:

$$f_{\text{MEL}} = 2595 \log_{10}\left(1 + \frac{f_{\text{LIN}}}{700}\right). \tag{1}$$

Denote the accumulated output from the kth filter bank as S_k. As human hearing exhibits logarithmic compression in the dynamic range the log of the filter bank output, $\log(S_k)$, is usually taken to reflect this. Taking the logarithm is also beneficial in that it transforms multiplicative frequency filtering channel distortions into additive effects which can more easily be compensated for as shown later.

c) ***Cepstral analysis:*** The final step is to convert the K log filter bank spectral values, $\{\log(S_k)\}_{k=1}^{K}$, into L cepstral co-efficients using the Discrete Cosine Transform

$$c_n = \sum_{k=1}^{K} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, \ldots, L. \tag{2}$$

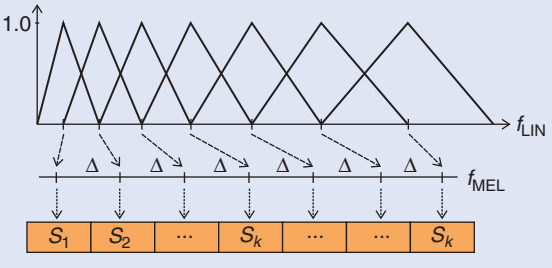

Figure 5. Mel-scale filter bank.

Unlike spectral features which are highly correlated, cepstral features yield a more decorrelated, compact representation. Typically only $L = 12$ MFCC co-efficients are extracted per frame (which comprises the feature vector for that frame). Additionally one can also include the special c_0 cepstral co-efficient which, by definition from (2) when $n = 0$, represents the average log-power of the frame. However as there is little speaker specific discriminant information provided by the average log-power it is not uncommon to exclude c_0, where in most cases of its inclusion this is by "default."

C. Environmental Compensation

A key advantage in transforming spectral magnitude features to log spectral cepstral features is that multiplicative channel and environment effects (especially arising from the use of different microphones between

the training and testing environments) become additive [20]. Assuming these effects are invariant for the duration of the utterance (and constant across all utterances in the testing or training session) a very simple and effective procedure is to subtract the utterance average MFCC feature vector from the individual utterance MFCC feature vectors [11],[21]. The idea is that this will remove out any time-invariant channel effects (as well as the average speech information) and retain only the important dynamic variations which characterise the speech of the speaker. These compensated MFCC features are derived by the process of *cepstral mean normalization* (CMN) as follows:

1) Calculate the mean over all the MFCC features of the utterance: $\vec{\mu}_T = \frac{1}{T}\sum_{t=1}^{T}\vec{c}_t$ where $\vec{c}_t = [c_1, c_2, \ldots, c_L]'_t$ is the MFCC feature vector at frame index t and there are T frames in the utterance.
2) Compensate the MFCC feature vector at each frame by: $\vec{c}_t^{\,c} = \vec{c}_t - \vec{\mu}_T$.

In speaker verification this front-end compensation is augmented by hand-set or channel score normalization strategies [6], [7], [22]. In speaker identification, as in speech recognition, other front-end environmental compensation can be implemented, especially if training data obtained from the different testing environments is available. Specifically CMN can be replaced by both mean and variance normalization: $\vec{c}_t^{\,c} = (\vec{c}_t - \vec{\mu}_e)/\vec{\sigma}_e$, where the mean, $\vec{\mu}_e$, and variance, $\vec{\sigma}_e$, are calculated from all of the speaker data derived from the testing environment, e [21]. Testing environments in this context include clean (no noise), different types of noise (e.g. white, babble, factory), and variable levels of relative noise power (the SNR).

D. Concatenation with Temporal Derivatives

In speech recognition an important feature processing attribute is to be able to capture and model the dynamic, temporal information between frames by concatenation of the compensated MFCC feature vectors with the first, second and even third order derivative approximations. Such dynamic information provides vital acoustic clues as to the nature of the speech being spoken. Similarly in speaker recognition the dynamic information also plays a role in helping to identify speaking styles and durations (albeit in a very simplistic fashion compared to prosodic cues). Define $\vec{c}_t^{\,c}$ as the compensated MFCC feature vector at frame time index t. The first-order derivative or "delta" feature is approximated by [23]:

$$\vec{d}_t = \frac{\sum_{p=1}^{P} p\left(\vec{c}_{t+p}^{\,c} - \vec{c}_{t-p}^{\,c}\right)}{2\sum_{p=1}^{P} p^2}, \tag{3}$$

where typically $P = 2$. By replacing $\vec{c}_t^{\,c}$ by $\vec{d}_t$ one can similarly derive the second-order delta-delta or "acceleration" parameters, $\vec{a}_t$. These temporal derivatives are concatenated with the original MFCC coefficients to yield an augmented feature vector. For example, with 13-dimensional MFCC feature vectors (L = 12 plus c_0) this gives 26-dimensional MFCC + delta features and 39-dimensional MFCC + delta + acceleration features. This feature parameterization is elucidated in Fig. 6.

III. Speaker Modeling with GMM and GMM-UBM Systems

A. Gaussian Mixture Models (GMM)

Assume an utterance of length T frames for speaker j and feature extraction yielding a D-dimensional feature vector for each frame, that is for each utterance we have: $\{\vec{x}_t \in \mathbb{R}^D : 1 \le t \le T\}$. How can we now build a model for speaker j such that for any utterance from that same speaker the T feature vectors will in, some sense, be represented by that speaker model and not by any other speaker models? If we can successfully do this and the unknown utterance can somehow be matched to the speaker model, we can recognize the speaker. Given that utterances from a speaker consist of random sequences of time samples covering all possible spoken words a statistical model will be the best candidate. In this regard the most generic modeling paradigm one can adopt is that of a Gaussian Mixture Model (GMM). A Gaussian model assumes the feature vectors follow a Gaussian

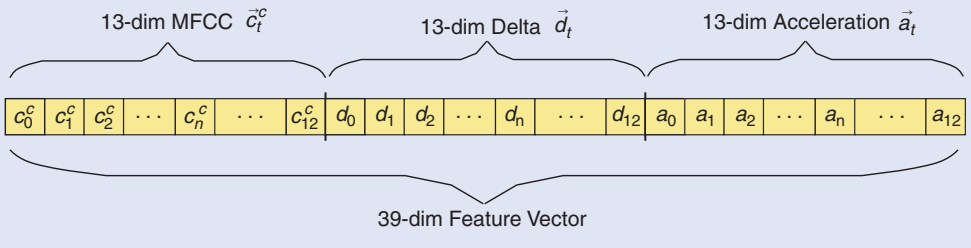

Figure 6. Concatenation of MFCC features with temporal derivatives.

distribution, characterized by a mean and a deviation about the mean. Furthermore by allowing a mixture of such Gaussians the distribution of the features from a particular speaker may be characterized. Indeed GMMs were one of the first modeling paradigms proposed for speaker recognition [24], [25] and with advances in the parameter estimation, computations and scoring of these models, they remain one of the most widely used to this day [22], [26].

The Gaussian mixture model for speaker j, λ_j, is a weighted sum of M component densities (as depicted by Fig. 7) governed by the output probability expression (for a given feature vector, $\vec{x}_t$):

$$p(\vec{x}_t|\lambda_j) = \sum_{i=1}^{M} g_i \mathcal{N}(\vec{x}_t; \vec{\mu}_i, \Sigma_i). \tag{4}$$

where g_i are the *mixture weights* satisfying $\sum_{i=1}^{M} g_i = 1$. The $\mathcal{N}(\vec{x}_t; \vec{\mu}_i, \Sigma_i)$ are the individual component densities, which for a D-variate Gaussian function are of the form:

$$\mathcal{N}(\vec{x}_t; \vec{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\vec{x}_t - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x}_t - \vec{\mu}_i)} \tag{5}$$

with *mean vector* $\vec{\mu}_i \in \mathbb{R}^D$ and *covariance matrix* $\Sigma_i \in \mathbb{R}^{D \times D}$. The GMM model for speaker j, λ_j, is parameterized by the mean vectors, covariance matrices and mixture weights from all M component densities:

$$\lambda_j = \{\vec{\mu}_i, \Sigma_i, g_i\}_j \quad i = 1, 2, \ldots, M. \tag{6}$$

Why are GMMs so successful in speaker recognition? This question needs to be asked since the GMM essentially "blindly" pools all of the speech data from a single speaker and hence has the difficult task of modeling all possible acoustic variations from anything the speaker can say (since the system is text-independent). Nevertheless with enough mixtures (on the order of 64 or more) the component densities may be able to represent the individual speaker's broad phonetic class distribution. Indeed given English speech has no more than 45 or so distinct phones one can surmise that with at least as many mixtures all of the possible distinct ways a speaker can speak are modelled. And under the GMM framework the overall model provides a smooth transition from one acoustic class (or mixture) to the other via the linear weighting function defined by (4) thus making the system text-independent in nature.

Given a collection of utterances from speaker j, how do we train the model, that is, estimate all of the speaker model parameters in (6)? Yet another reason for the success of the GMM is the availability of a powerful and versatile parameter estimation paradigm: the *expectation-maximization* (EM) algorithm [27]. The interested reader is referred to the many texts on the parameter estimation of the GMM for speaker recognition,

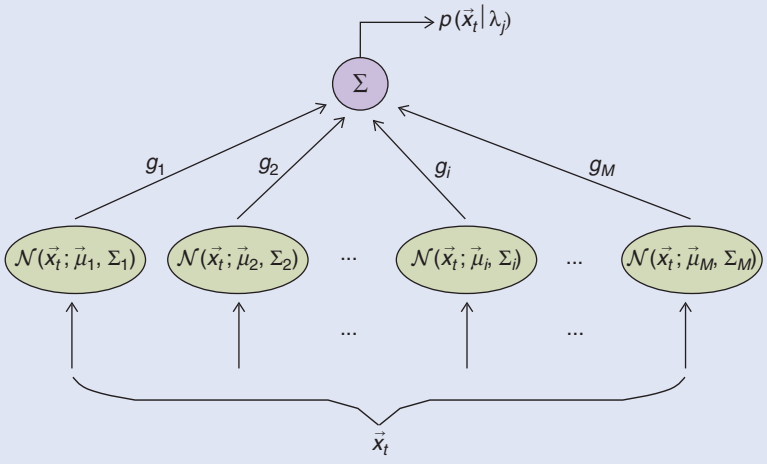

Figure 7. Depiction of the GMM mixtures.

in particular [24]. A key feature of the EM algorithm is that it can guarantee monotonic convergence to the set of optimal parameters (in the *maximum-likelihood* (ML) sense) in only a few (5 or so) iterations.

Although GMMs are quite powerful they do suffer from two important drawbacks. One is the need for enough training data to properly estimate the model parameters. A common trick is to use diagonal covariance matrices (i.e. $\Sigma_i = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2)_i \in \mathbb{R}^D$) rather than "full" covariance matrices which is possible given that MFCC features vectors are decorrelated (i.e. exhibit low correlation values for the off-diagonal covariance matrix elements). Not only are computational resources reduced using diagonal covariance matrices but investigations have shown [22], [24] that performance is unimpaired (and in most cases even improved). Even so with a typical 128 mixture GMM modeling 39 dimension feature vectors, a total of $128 \times (39 + 39) + 128 = 10112$ floating-point parameters have to be estimated and depending on the amount of initial training data for each speaker there will be an upper limit to the number of mixtures before performance degrades due to unreliable estimation of too many model parameters. The second problem with a GMM, as with any generative modeling paradigm, is that data unseen in the training which appears in the test data will trigger a low score on that data and degrade the overall system performance. For example, a particular speaker specific phone feature distribution not present in the training data will not be captured by the GMM and when this appears in the test data it will generate a low likelihood score. The solution is simply more and varied training data, but in practical speaker recognition this may be hard to come by. In section III-B we discuss a clever solution to both of these problems.

The GMM framework just described for speaker identification is depicted in Fig. 8 where the GMM ML scoring is discussed in Section IV.

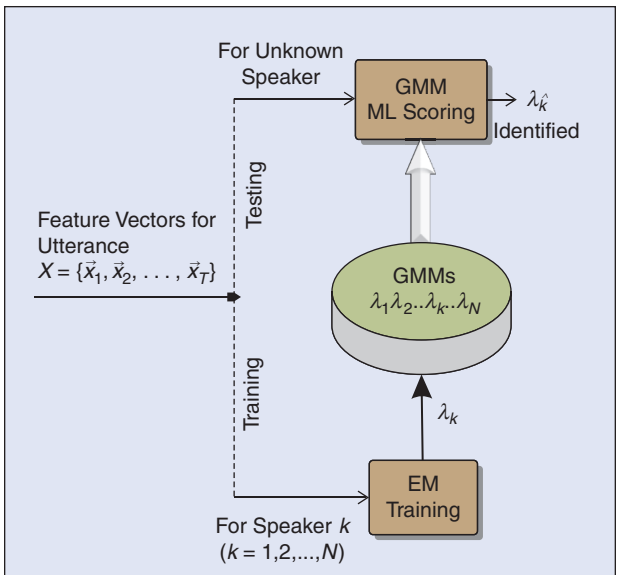

Figure 8. GMM training and testing framework assuming N speakers.

B. Universal Background Model (UBM) and the GMM-UBM

In speaker verification the claimed identity of the speaker is checked by scoring the unknown utterance against the claimed speaker model and comparing this to the score against the imposter model. The imposter model is a GMM which models all speakers other than the claimed speaker and is sometimes referred to as the Universal Background Model (UBM). In the strictest sense for verification the GMM for the imposter model should be trained by pooling all of the speaker data with the exception of the claimed speaker in question. In practice the GMM for a UBM is trained by pooling all of the speaker data (including the claimed speaker) assuming the available speaker data is balanced across all subgroups (e.g. different gender). This has the advantage that the same UBM can be used for speaker verification for any claimed speaker identity. Since the amount of training data is greatly increased for a UBM, the GMM parameters are reliably estimated and a larger number of mixtures (on the order of 256 or more) is not uncommon. Conceptually the UBM represents the speaker-independent distribution of features across all speaker data.

Although a UBM can be used in open-set speaker identification for the detection of unknown speakers, in closed-set speaker identification there is no direct need for a UBM since the individual speaker GMMs are sufficient to carry out the identification process. However as the UBM is more reliably trained than any one speaker GMM and is also more accurate in modeling all of the feature space across all speaker data it does not suffer from the problems of insufficient training data and unseen data. Statistical models like the GMM are not only able to be estimated directly using a powerful technique like the EM algorithm, but with small amounts of data the parameters can be further *adapted* to the new data using either Maximum Likelihood Linear Regression (MLLR) or Maximum A-Posteriori (MAP) adaptation [23]. Thus an alternative to the individual speaker GMMs is to train a UBM and then form the speaker GMMs by adaptation of the UBM using the individual speaker data as the adaptation data [22]. The training and test using this GMM-UBM framework is depicted in Fig. 9.

The governing equation for generic MAP adaptation is:

$$\hat{\vec{\mu}}_m = \frac{N_m}{N_m + \tau}\bar{\vec{\mu}}_m + \frac{\tau}{N_m + \tau}\vec{\mu}_m, \quad (7)$$

where, $\hat{\vec{\mu}}_m$, is the adapted mean for mixture m, τ is a weighting parameter of the a priori knowledge, N_m is the occupation likelihood of the adaptation data (individual speaker) data, $\vec{\mu}_m$, is the speaker-independent (UBM) mean and, $\bar{\vec{\mu}}_m$ is the mean of the observed adaptation (individual speaker) data.

As the UBM is initially trained over all of the available data, parameters are estimated reliably as there is sufficient data, and even with small amounts of individual speaker adaptation data the resultant GMM-UBM will be equally reliable, much more so than using the same small amounts of data to train a GMM directly. Due to the increased training data the number of mixtures in a UBM is more than an individually trained GMM can reliably estimate and the adapted GMM-UBM has the same number of mixtures as the original UBM. Thus a GMM-UBM is better able to handle unseen data as it inherits the modeling power of the underlying UBM. According to (7) the adaptation algorithm will only adapt (i.e. modify) mixture parameters for which observations exist from the available individual speaker data (large N_m), where this data is non-existent (unseen) (i.e. small or zero N_m) the original UBM mixtures are copied over, thereby mitigating against low scores with unseen data.

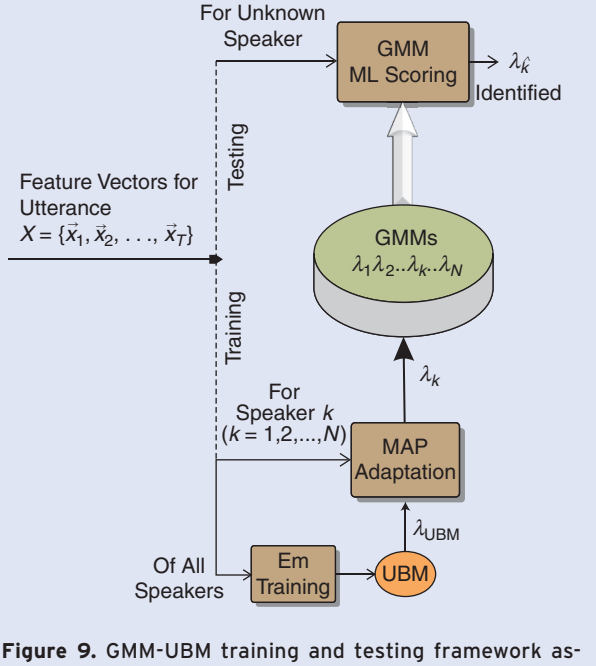

Figure 9. GMM-UBM training and testing framework assuming N speakers. Note that MAP adaptation can only be applied after the EM training produces the UBM from all of the speaker data.

C. The Problem with Silence

The GMM estimates the parameters and scores the models based on all of the utterance data from a speaker. As long as the utterance data contains acoustic data from the speaker then valid data is being used to estimate or score the model. However in practice when speaker data is acquired an uncertain amount of initial and end silence is present. With a push-to-talk system there could be a large silence gap between the activation and the actual speech. With an automated system using a Voice Activity Detector (VAD) this can be controlled but a small amount of silence is always inevitable. As silence data contains no speaker (or speech) specific information inclusion of such data will degrade the training of the speaker GMMs (in proportion to the amount of silence versus speech). With only small amounts of beginning and end silence a GMM may be able to cope by sacrificing some of the component mixtures to model the silence class. Otherwise as with speech recognition, special silence models can be specifically trained, however this will then require a Hidden Markov Model (HMM) training and scoring approach (e.g. as provided by the HMM ToolKit (HTK) [23]). With this solution utterances of a speaker are used to train the HMM sequence: *sil*+*speaker_ID*+*sil* instead of the GMM *speaker_ID* and in scoring a language modeling network: [*sil*] (*speaker models*) [*sil*] is adopted rather than the GMM scoring of (*speaker models*). Since a GMM is just a single-state HMM, one can indeed do this using, say, HTK or equivalent toolkits. Alternatively if one is able to manually segment the training data into the speech regions and silence regions the silence models can be directly trained and then frames can be simultaneously scored against the silence and speaker models, and any silence scores ignored [7].

D. Other Modeling Paradigms

Although the GMM is the mainstay of speaker modeling, it is not the only modeling paradigm to have received attention. Here we briefly mention two alternative paradigms which have also been considered: eigen-voices and utterance covarance matrices.

In the face recognition community, the idea of face-spaces and eigen-faces has been popularized from the view that the pixel face images can be considered as derived from an underlying linear subspace. A similar idea can be developed by considering the collection of GMM means as representing an underlying linear sub-space, where each GMM mean is transformed to a lower dimension *eigen-voice* by application of Principle Components Analysis (PCA) [28], [29]. Unknown test utterances are

then projected onto this *eigen-space* (formed by the linear combination of the eigenvoices as the basis vectors) and the nearest eigen-voice is the identified speaker.

The GMM is a model of the speaker derived from all of the utterances of that speaker. An alternative view is to consider the sampled covariance calculated directly from the collection of utterance frames. Thus the speaker model is not a GMM but a covariance matrix. Unlike the GMM which maps the position of the speaker's feature space the covariance matrix captures the long-term distribution or shape of the feature space as a whole. Thus the covariance will not be sensitive to channel biasses nor the short-time speech characteristics [30]. Furthermore by considering utterance level covariances (covariance matrix calculated per utterance) one can then consider utterance level scoring instead of the usual frame level scoring.

IV. Speaker Classification with GMM and GMM-SVM Systems

A. GMM Maximum-Likelihood (ML) Scoring

In testing, a sequence of T feature vectors are extracted from the unknown speaker utterance. Given a set of N GMM (or GMM-UBM) speaker models how do we identify (or classify) the sequence of feature vectors? For minimum classification error the optimal classifier is given by the speaker model which exhibits the maximum a-posterior probability [25]:

$$\hat{j} = \underset{1 \le j \le N}{\operatorname{argmax}} P(\lambda_j | X), \tag{8}$$

where $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ is the utterance feature vector sequence and $\hat{j}$ is the identified speaker (classification decision). It can be shown this reduces to Maximum-Likelihood (ML) scoring of the log likelihoods:

$$\hat{j} = \underset{1 \le j \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t | \lambda_j) \tag{9}$$

from which we calculate $p(\vec{x}_t | \lambda_j)$ using (4). Speaker identification using GMM ML scoring is depicted in Fig. 10.

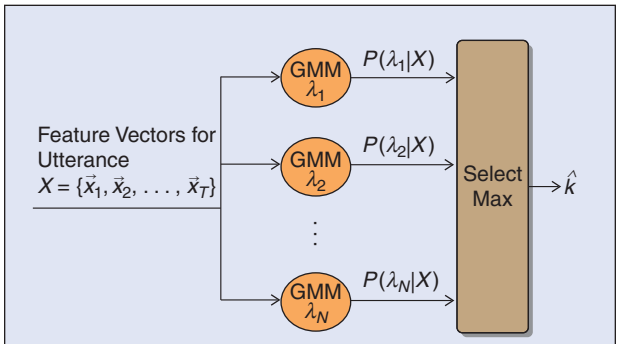

Figure 10. GMM ML Scoring.

B. Support Vector Machines (SVM) for Speaker Recognition

For classification problems most paradigms can be described as falling in one of two families: *generative models* (like GMMs) which only require training data samples from the class or target speaker and build a statistical model that describes the target speaker distribution, or *discriminative classifiers* which require training data for both the target and non-target (imposter) speakers and derive an optimal separation between the different speakers. Popular in the latter category for speaker recognition has been the increasing adoption of Support Vector Machines (SVMs). A key feature is that SVMs can achieve comparable or superior performance to GMMs with much less training data.

An SVM is basically a two-class classifier that fits a separating hyperplane between the two classes (assuming linear separability). The optimal hyperplane is chosen according to a *maximum margin criterion*. That is, the optimal hyperplane is chosen such that it maximizes the Euclidean distance to the nearest data points on each side of the plane. The nearest data points on each of the separating hyperplane are known as the *support vectors*. The basic concept behind this is highlighted by Fig. 11. A good tutorial discussion of SVMs can be found in [31].

Application of SVMs to speaker recognition requires the following considerations:

- Speakers are not linearly separable and the basic SVM has to be augmented by the use of *slack variables* and a *kernel function* that projects the

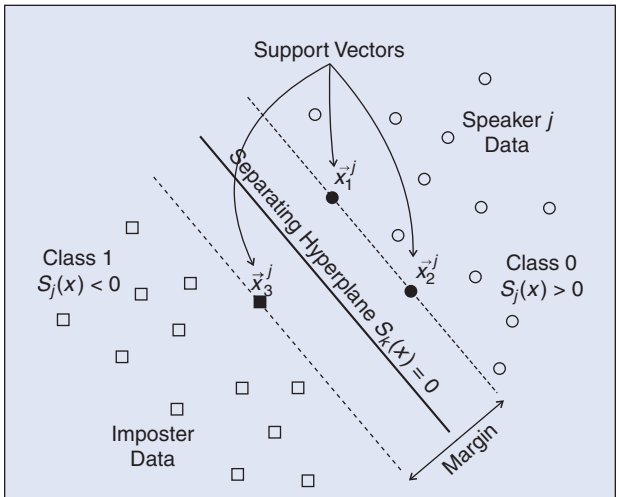

Figure 11. Basic principle of the SVM separating hyperplane.

non-linearly separable data to a linearly separable higher dimension [31], [32].

- Standard SVMs as described in, say, [31], are defined only for static data vectors. However when recognizing speakers, an utterance will generate a sequence of feature vectors rather than the one data vector. How can one transform the sequence of feature vectors to a single data vector suitable to be classified by an SVM? Possible solutions include the use of a *polynomial classifier* [32], *sequence kernel* functions [33] and *GMM supervectors* [26].
- As SVMs are inherently two-class classifiers how can one extend these to multiclass speaker identification? This can be achieved by designing a *one-against-all* (OAA) SVM classifier for each of the N speakers. The SVM classifier for speaker j is a two-class system where class 0 is the training data from speaker j and class 1 represents all the other speaker data (all speaker data except speaker j). Identification is then carried out by performing N SVM classification operations and selecting the SVM with the maximum decision function value.

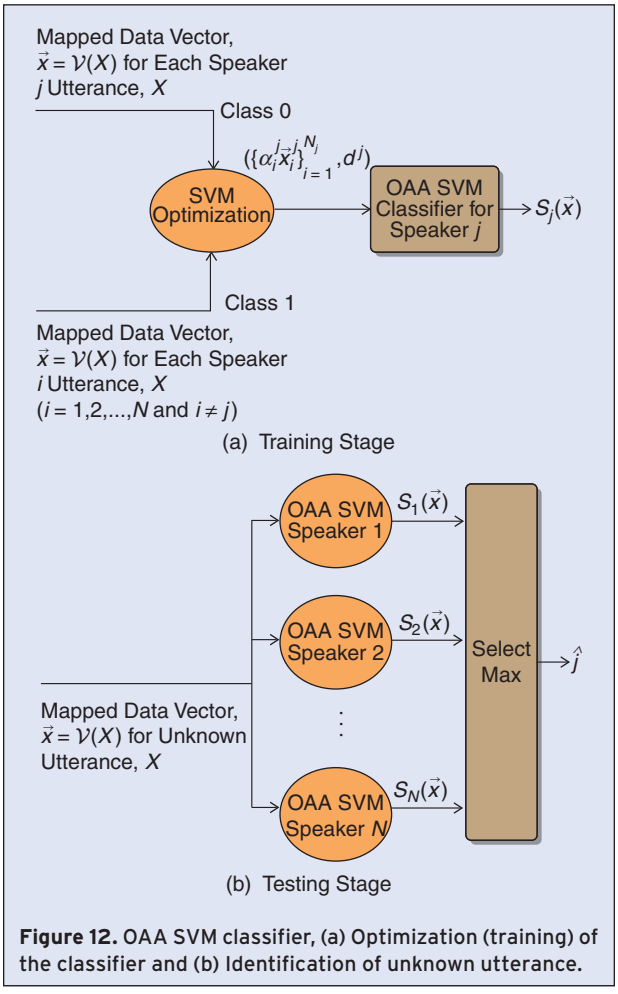

Figure 12. OAA SVM classifier, (a) Optimization (training) of the classifier and (b) Identification of unknown utterance.

The functional operation of the OAA SVM classifier for a target j can be described as the sum of the kernel functions:

$$S_j(\vec{x}) = \sum_{i=1}^{N_j} \alpha_i^j t_i^j K(\vec{x}, \vec{x}_i^j) + d^j, \tag{10}$$

where α_i^j are the Lagrangian multipliers, d^j is a learned constant, $\vec{x}_i^j$ are the support vectors and N_j are the number of support vectors.

The $(\{\alpha_i^j, \vec{x}_i^j\}_{i=1}^{N}, d^j)$ are obtained from the SVM optimization algorithm [31], [32], [33] given the t_i^j ideal outputs ($t_i^j = 1$ for class 0 and $t_i^j = -1$ for class 1), the class 0 (target j) and class 1 (all other) training data, and the kernel function $K(\vec{x}, \vec{y})$. The kernel function is constrained to be of the form $K(\vec{x}, \vec{y}) = B(\vec{x})'B(\vec{y})$ where $B(.)$ is the mapping from the input space to the higher dimensional separating space.

For speaker identification the sequence of feature vectors for the unknown utterance, $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, is first mapped to a single data point in the higher dimension separating space, $\vec{x} = \mathcal{V}(X)$, by one of the methods previously described (e.g. a GMM supervector, see Section IV-C). Identification is then performed by determining the SVM classifier with the maximum decision function value:

$$\hat{j} = \underset{1 \leq j \leq N}{\operatorname{argmax}} S_j(\vec{x}). \tag{11}$$

The OAA SVM classifier for speaker identification is depicted in Fig. 12.

C. GMM-SVM Speaker Identification System

The face image used in face recognition is the equivalent to the speaker utterance. However in face recognition the face image can be mapped to a vector of pixel values, thus creating a single template vector for the person. Thus application of pattern classifiers, such as SVMs, in face recognition is relatively straightforward since there is a defined template vector or data point for each (known or unknown) person. In speaker recognition, of course, one needs to deal with an utterance of arbitrary length T making the application of SVMs and similar pattern classifiers more complex.

Campbell [26] proposed an intriguing application of the GMM-UBM to the generation of a speaker template which can then be directly used with an SVM classifier. The system is depicted in Fig. 13. In the GMM-UBM system described in Section III-B the UBM is adapted using the training data for speaker j to form the GMM for that speaker. But by only adapting the UBM to a single utterance one can instead form the GMM for that utterance. Although the GMM of an utterance is in itself of no practical use, by concatenating the means of the

GMM components, $\vec{\mu}_i$, formed by adapting the means of the UBM using only the feature vectors from the utterance X the so-called *GMM supervector* can be formulated:

$$\vec{\mu}_X = [\vec{\mu}_1' | \vec{\mu}_2' | \vec{\mu}_3' | \cdots | \vec{\mu}_M']', \tag{12}$$

where $\vec{\mu}_i$ is a column vector, $\vec{\mu}_i'$ is a row vector and hence $\vec{\mu}_X$ is a column vector.

The ordering of the GMM component means is not important as long as the same ordering is adopted for all utterances. This is guaranteed by the adaptation of the same UBM for all speaker utterances. Furthermore adaptation of UBM ensures that even for a single utterance a supervector in a sufficiently high dimensional space results (enough mixtures M). For example, with, say, a 512-component GMM and 39-dimension feature vectors a $512 \times 39 = 19968$ dimension supervector results. Thus in the application of SVMs a simple, linear kernel and trivial kernel mapping is possible. From [26] one solution is to use a linear kernel directly on the data, that is, $K(\vec{x}, \vec{y}) = \vec{x}'\vec{y}$ where the supervectors are given by

$$\vec{x} = \tilde{\vec{\mu}}_X = \left[\sqrt{g_1}\Sigma_1^{-(1/2)}\vec{\mu}_1' | \sqrt{g_2}\,\Sigma_2^{-(1/2)}\vec{\mu}_2' | \sqrt{g_3}\,\Sigma_3^{-(1/2)}\vec{\mu}_3' | \cdots | \sqrt{g_M}\Sigma_M^{-(1/2)}\vec{\mu}_M'\right]', \tag{13}$$

which have been scaled by the respective UBM component mixture weights and diagonal covariances.

The concept of a speaker template derived from an utterance in this way opens up the possibility of directly applying other classification and modeling paradigms from pattern recognition, especially those popular in face recognition, for example, eigenspace analysis [28].

V. Speaker Identification Experiments and Evaluations

A. Experimental Setup

We now examine the performance of speaker identification described previously by carrying out experimental evaluations as follows. The three systems for speaker identification, GMM, GMM-UBM and GMM-SVM were implemented and evaluated under different scenarios: clean matched conditions, with additive white noise at different signal-to-noise or SNR values, and mismatched handset or telephone channels. All evaluations were carried out using the widely available TIMIT speech corpus [34]. The TIMIT corpus is a collection of phonetically balanced sentences, 10 sentence utterances from 630 speakers across 8 dialect regions in the USA. The data was acquired in a clean, studio environment and sampled at 16 kHz (8 kHz bandwidth). Although the TIMIT corpus is not recommended for speaker recognition evaluations due to the ideal acquisition environment [35], it is a widely available and still popular speech corpus and quite suitable for the tutorial, exploratory nature of this investigation. The following software tools were used:

- For the implementation of the GMM and GMM-UBM systems the Hidden Markov Model ToolKit (HTK version 3.4.1) [23], was configured to model a single-state HMM with the standard MLLR and MAP adaptation scripts to adapt the UBM accordingly for the GMM-UBM models and GMM-SVM supervectors.
- For the GMM-SVM the SVM-KM toolbox [36] was used to implement the one-against-all SVM classifer.
- Both additive noise and mismatched channel effects were created making use of the FaNT software [37], especially for the implementation of the G.712 and MIRS channel characteristics.

In all experiments a pre-emphasis filter with co-efficient 0.97 was applied to the sampled waveform and features were extracted from each 25 ms frame and generated every 10 ms. All frames were windowed using the Hamming window function. Unless otherwise specified, 13 MFCC features (including c_0) were extracted together with the delta and acceleration temporal features, yielding a 39-dimensional feature vector. Environmental compensation by cepstral mean normalization (CMN) was also applied to each sequence of MFCC features generated from an utterance.

For all evaluations 64 speakers from the TIMIT corpus were selected for closed-set speaker identification. The speakers were evenly balanced between the 8 different dialects and gender (i.e. 32 male and 32 female speakers with 4 male and 4 female speakers from each dialect region). Following the protocol suggested in [25] the 10 utterances per speaker were divided into 8 utterances for training (two SA, three SX and three

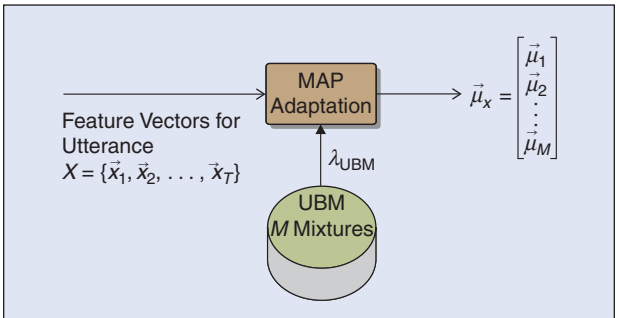

Figure 13. The GMM-UBM supervector.

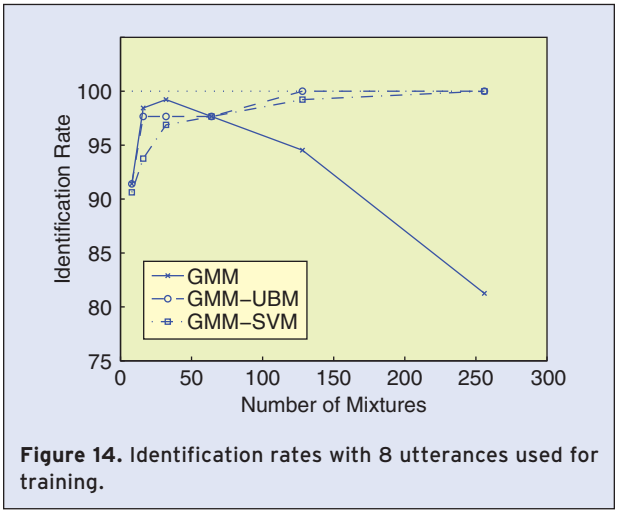

Figure 14. Identification rates with 8 utterances used for training.

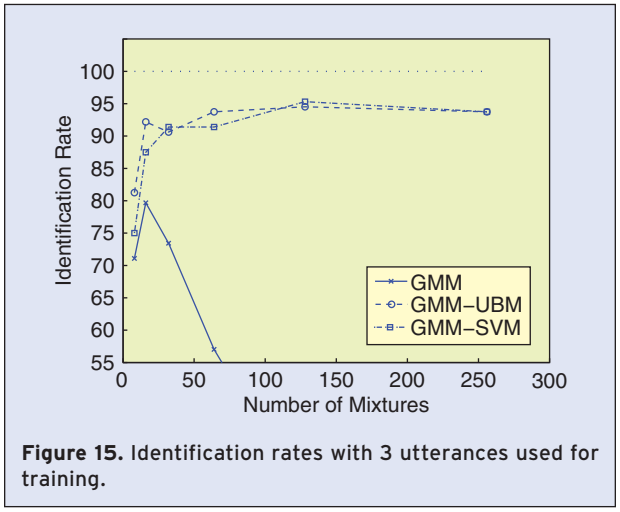

Figure 15. Identification rates with 3 utterances used for training.

SI sentences) and 2 utterances (remaining two SX sentences) for testing. In evaluations where only 3 utterances for training were used these comprised the three SI sentences.

The key investigations that were undertaken were the recognition peformance of the three systems measured against the number of component mixtures for the underlying GMM and the amount of training data, and effect on the GMM-UBM performance in additive noise at different SNR levels and mis-matched channel effects.

B. Comparison of GMM, GMM-UBM and GMM-SVM systems

The number of mixtures for the GMM, GMM-UBM and GMM-SVM systems were varied from 8 to 256 mixtures and the identification rates are plotted in Fig. 14. As expected when the number of mixtures increase the GMM's performance fails as there are too many parameters that need to be estimated given the limited amount of training. The best performance for the GMM was 99.2% identfication with 32 mixtures, with performance rapidly deteriorating for more component mixtures. However for both the GMM-UBM and GMM-SVM the performance improved with increasing number of mixtures and achieved 100% with 256 mixtures. The results clearly indicate the importance of using a GMM-UBM approach when confronted with limited amounts of training data (which is typical of most speaker recognition applications).

Table 1.
Identification rates using original features and reduced features (without the temporal derivatives) based on 3 utterances per speaker for training.

	GMM-UBM		GMM-SVM	
Mixtures	Original	Reduced	Original	Reduced
64	93.8	92.1	91.4	95.3
128	94.5	93.0	95.3	96.0
256	93.8	89.8	93.4	92.1

The effect of the limited training can be further examined by reducing the number of training utterances from 8 per speaker to 3 per speaker. The results are plotted in Fig. 15. It is even more evident with these results how sensitive the GMM is to the amount of training data and the number of mixture components. The best result achieved with the GMM was only 79.7% with 16 mixtures, and rapid deterioration was observed when more mixture components were added. The GMM-UBM and GMM-SVM systems, on the other hand, achieved a much better performance of around 95% with 128 mixtures. To put these results in context commercial applications of biometric identification require error rates to be no more than 2% (ideally 1% or better) depending on the number of speakers enrolled in the system [1].

From the results presented so far the GMM-UBM and GMM-SVM achieved similar performance even though these represent different classifier paradigms. One possible explanation is that the supervector used by the SVM is based on the same UBM used by the GMM-UBM. Another possible reason is that the data set used did not sufficiently exploit any of the advantages of a discriminant SVM classifier over the generative GMM based system. To examine the differences in more detail both the original feature sets and a reduced feature set without the inclusion of the temporal derivatives (i.e. only a 13-dimension, rather 39-dimension, MFCC feature vector) were evaluated on the limited training data set for 64, 128 and 256 mixtures and the results presented in Table 1. For this case the GMM-SVM achieves

Table 2.
Effect of matched (G.712) and mismatched (MIRS) training and testing environments with and without cepstral mean normalization (CMN) for the 128 mixture GMM-UBM system.

	G.712	MIRS
with CMN	94.5	94.5
without CMN	99.2	78.9

Table 3.
Effect of additive white noise (at different SNRs) and matched/mismatched channels on the 128 mixture GMM-UBM system and the standard features (with CMN).

	G.712	MIRS
clean	94.5	94.5
30 dB	74.2	75.8
20 dB	42.2	39.8
10 dB	10.9	7.8
5 dB	3.1	2.3

superior recognition over the GMM-UBM of around 3% for all cases when both the amount of training data and the number of features is reduced, as compared to when only the training data is limited. The results highlight the potential superiority of an SVM classifier when presented with limited amounts of training data and/or reduced feature sets.

C. Effect of Mismatched Channels and Additive Noise

For the remaining experiments we examine the performance of speaker identification in the presence of additive noise and mismatched channel effects. Results are presented only for the 128 mixture GMM-UBM system. Similar experiments on the GMM and GMM-SVM confirm identical findings, but are not reported here for brevity. The feature set used was the 39-dimension MFCC + delta + acceleration vector with and without CMN as described below. In all cases the system was trained with 8 utterances per speaker filtered by the G.712 channel filter characteristic [38] and tested on the 2 utterances filtered using either the same (matched) G.712 characteristic or the different (mis-matched) MIRS characteristic. The G.712 applies a flat passband response between 300 Hz and 3400 Hz whereas the modified impulse response system (MIRS) exhibits a rising passband response thus representing different telephone or handset characteristics.

To examine the effect of channel mis-match and the importance of CMN in compensating for channel mis-match we first present results in clean (no additive noise) using both matched and mis-matched testing conditions, with and without CMN. These findings are presented in Table 2. The robustness of CMN compensation to channel mis-match is evident in the identical recognition rate of 94.5% in matched (G.712) and mis-matched (MIRS) conditions. In contrast to this without CMN there is a significant degradation in the mis-matched case, from 99.2% in the matched case to only 78.9%. However it should be noted that a side-effect of CMN is that it can reduce performance in ideal, matched conditions (from 99.2% down to 94.5% in this case), but as practical scenarios rarely involve ideal conditions the benefits of CMN are apparent.

Speaker identification is a statistical pattern classification task and as such is very sensitive to environmental changes between the training and test conditions which can introduce spectral biases and distortions. The performance under additive white noise at different SNR levels on the test data, with training data not subject to any noise (clean) is examined for both matched and mis-matched channel conditions. The results are presented in Table 3. From the results obtained, as features were compensated by CMN, there is little performance difference between the channel mis-match, but there is significant degradation in the presence of additive noise. Even under the mild noisy condition of 30 dB SNR the recognition performance has dropped by about 20% (from 94.5% down to 74.2%). It is quite obvious that the standard features and modeling paradigms described are insufficient to deal with the severe consequences of environmental mismatch due to additive noise. In the next section we discuss how to embed robustness in speaker identification systems, especially in regards to additive noise.

VI. Robustness for GMM Speaker Identification

Despite the high recognition accuracies produced by Gaussian Mixture Model based identification systems in clean conditions, in practice noise distortion often affects the input test speech leading to a dramatic degradation in performance. These distortions are typically categorized as channel effects (which result from differences in the characteristics of the handsets or microphones used to capture training or testing speech), or environmental effects such as background noise. Both of these effects produce a mismatch between the speaker dependent information extracted from the input speech utterance, and the information contained within the trained model corresponding to the true speaker. Achieving robustness to these effects requires a modification of the standard GMM speaker identification framework.

In this section the effect of training-testing mismatch on GMM based speaker identification is outlined, and a

brief review is provided for some of the common noise robustness techniques which are applicable to speaker recognition systems.

A. Speech Distortion and GMM Mismatch

Consider a recognition system where each speaker is represented by a clean speech trained GMM. In a practical test environment the classification of a sampled speech utterance signal $s(z)$ will be performed in the presence of a noise distortion signal $n(z)$. Therefore, at the speaker recognizer input the received test signal $x(z)$ is a function of both the speech signal and the noise distortion:

$$x(z) = \mathcal{F}(s(z), n(z)), \tag{14}$$

where z is the sample index and the functional relationship $\mathcal{F}$ is dependent on the type of distortion experienced. In the case of channel distortion the mismatch is often approximated by a convolutional relationship between the time domain speech and noise signals, where $n(z)$ represents the impulse response of the filter approximating the channel:

$$x(z) = \mathcal{F}_{\text{channel}}(s(z), n(z)) = s(z) * n(z). \tag{15}$$

In the spectral energy domain the received signal is obtained by multiplication of the speech signal spectra $S(f)$ and the channel frequency response $N(f)$. Following the filterbank and log magnitude feature extraction operations (as described in Section II-B), the received signal's log-spectral value for filterbank component f processed from time frame t is given by:

$$\log X_t(f) = \log S_t(f) + \log N_t(f). \tag{16}$$

For environmental distortion there is an additive relationship between the sampled time domain speech and noise signals:

$$x(z) = \mathcal{F}_{\text{additive}}(s(z), n(z)) = s(z) + n(z). \tag{17}$$

The corresponding relationship in the log-spectral domain is non-linear:

$$\log X_t(f) = \log [S_t(f) + N_t(f)]. \tag{18}$$

Conversion of the log-filterbank values to cepstral coefficients requires the application of the Discrete Cosine Transform (DCT) (see Section II-B). Due to the linear nature of the DCT, the relationship between the speech and noise spectra remains linear in the case of convolutional distortion and non-linear for additive effects. Therefore, the presence of a noise signal with sufficiently large energy compared to the speech signal will produce variation in the feature values extracted, regardless of whether such features are cepstral or spectral based. Within the statistical GMM classification framework, this mismatch between the clean trained component distributions in the true speaker model λ and the feature observation vector $\vec{x}_t$ extracted from the noisy input speech signal results in a decrease in the frame likelihood scores $p(\vec{x}_t | \lambda)$ (as in (4)). When the scores are accumulated over all frames in the utterance (as in (9)) an incorrect ML speaker decision is often produced (see Fig. 16). Typically the extent of the feature mismatch is dependent on the relative speech-to-noise energy (the SNR) of the input utterance, and the time-varying nature of the noise process.

Traditionally speaker and speech recognition systems have utilized telephone speech where channel effects (such as handset mismatch) are the dominant distortion source [39]. Previous approaches to providing robustness against these effects can be categorized as feature

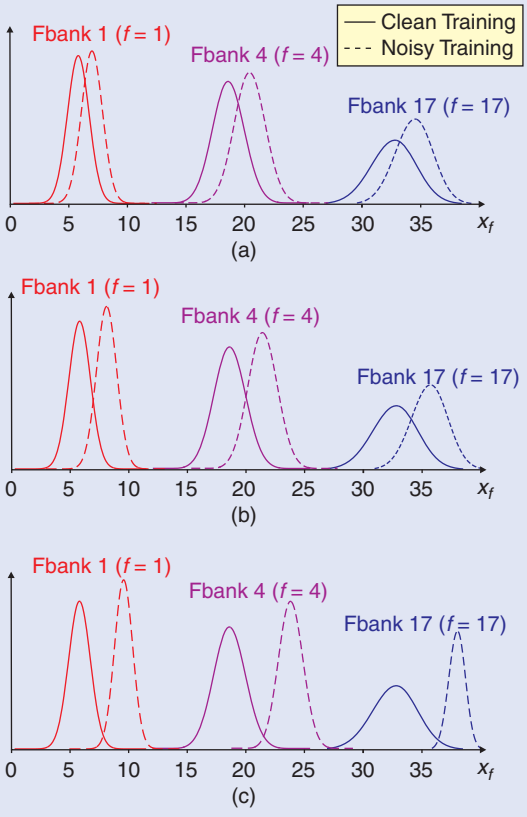

Figure 16. The maximal mixture Gaussian for log-filterbank feature components 1, 4 and 17 of TIMIT speaker fadg0's GMM. Distributions are shown from the training data in clean conditions (solid) and the same data in 20 dB (a), 10 dB (b) and 0 dB (c) additive white noise conditions (dashed).

based compensation, which is concerned with removing the effect of the noise within the feature parameterization, score based compensation which attempts to remove model score biases and shifts due to acoustic variabilities, and model based compensation which alters the learned speaker models in an attempt to sensitize against the distortion and thus reduce mismatch.

B. Feature Compensation Approaches

The focus on feature compensation approaches for noise robustness in past work is due to the importance of the feature extraction stage in the speaker recognition process. The ideal feature set should provide discrimination between speakers while exhibiting invariance to non-speaker characteristic information within the input speech. In this way the design of noise robust features has the advantage of generality in that a single technique may show robustness to multiple types of distortion. This is illustrated in Fig. 17, where a feature compensation stage may be inserted following extraction and pre-recognition.

Motivated by the ability of cepstral features to produce high speech and speaker recognition rates in clean conditions, compensation approaches which apply normalization to these standard features are common. Cepstral Mean Normalization (CMN) compensates against linear filtering effects (such as channel distortion) by taking advantage of the additive nature of the distortion in the log-spectral domain [18], [40], [11]. If in the cepstral domain the clean speech contribution has zero mean, subtracting the time averaged value from each cepstral coefficient is effective in removing the contribution of the channel. Evaluations have shown that the removal of global cepstral means prior to training and recognition improves robustness to intersession variability [19], [11]. Despite its benefits for mismatched channel compensation, when applied to matched clean speaker recognition tasks cepstral mean normalization causes performance loss (see Tab. II). This is due to the assumption that the cepstral time average of the corrupted received speech approximates that of the channel distortion, however this is only valid given sufficient phonetic balance within the speech utterance such that the speech cepstral mean is zero [17]. Methods such as augmented CMN [41] have been recently proposed to address this, where the probable noise and speech regions are normalized with different values leading to a reduction in error rate over the standard algorithm. Other modifications attempt to improve CMN in additive noise conditions including Fixed Codeword Dependent Cepstral Normalization (FCDCN) and variants which aim to reduce environmental dependence [42].

The relative spectral (RASTA) approach [12] also aims to provide compensation against channel distortion. The method is based on the observation that the rate of

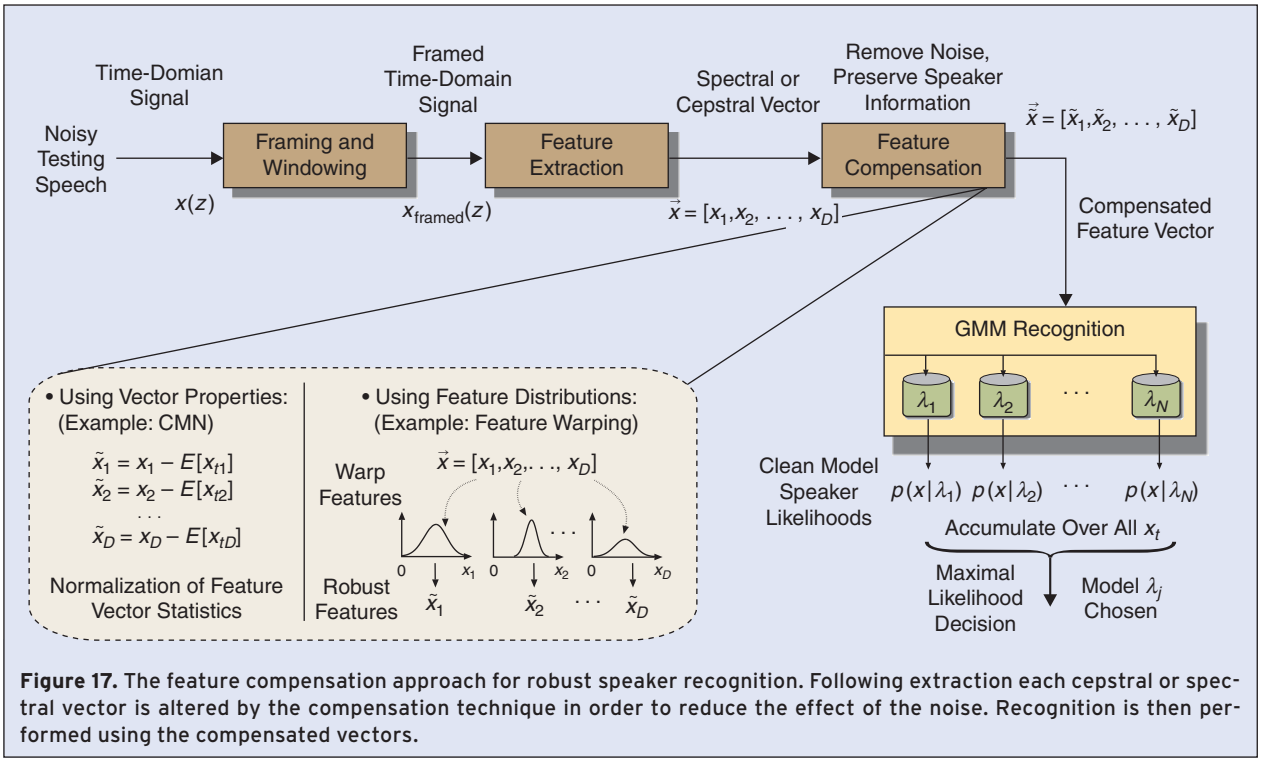

Figure 17. The feature compensation approach for robust speaker recognition. Following extraction each cepstral or spectral vector is altered by the compensation technique in order to reduce the effect of the noise. Recognition is then performed using the compensated vectors.

change of distortion effects within a corrupted speech mixture will significantly differ compared to the change of the vocal characteristics. Thus, by suppressing slowly varying components over all frequencies in the log-spectral domain, those constant components which represent the channel are removed [20], [43]. Compared to CMN, RASTA has an advantage in that its suppression of components is dynamic and as a result the application of standard RASTA compensation to the matched clean case produces negligible decrease in performance.

In addition to techniques which remove or suppress channel contributions, research has focused on modification of the power spectral representation technique itself to enhance robustness. Linear predictive (LP) analysis features are based on an all-pole model which captures vocal tract information in the magnitude spectrum of the LP filter response, and pitch information within the LP residual [44]. Unlike filterbank based features where spectral detail over frequency is fixed (MFCC, LFCC), LP cepstral coefficients are adaptive in that the detail in the representation of the spectral peaks is limited only by the model order. The clear separation of vocal tract and pitch information is advantageous for robust speaker identification due to the ability to characterize the vocal tract properties in the presence of some types of noise [17], [45]. Standard LP cepstral analysis has been augmented with weighting schemes to improve robustness. Basic liftering prioritizes low order coefficients due to their sensitivity to spectral slope while minimizing the noise sensitive higher order coefficients. Adaptive component weighted cepstrum (ACW) normalizes residues within the LP transfer function such that formants are emphasized while variations due to channel distortions are suppressed [46]. Post-filter cepstrum is conceptually similar in that it seeks to emphasize formant regions on the assumption that the effect of noise is lessened in these regions [47].

The performance of the traditional cepstral features and channel compensation methods was compared for speaker identification [19]. The results showed that recognition performance using both the filterbank based (MFCC, LFCC) and linear predictive based cepstral features (LPCC, PLPCC) degraded significantly in the presence of mismatched telephone channel effects. Of the channel compensation methods tested, CMN was found to be superior to RASTA in most cases where Reynolds [19] cites the short time window used to calculate the RASTA channel means as a possible explanation. The study concludes that for robust recognition with channel distortion the compensation method applied has greater importance than the base cepstral feature utilized.

Feature transformation approaches have been previously utilized to remove channel variability and improve speaker recognition robustness. In the cepstral domain, affine transforms have been proposed to correct feature distortion due to noise [17]. By modeling the effect of the noise distortion on the speech as an affine transformation compensation can be achieved by applying the appropriate inverse transform. Subject to correct selection of the transform parameters, affine transforms can model composite channel distortion and stationary additive noise effects leading to improved speaker recognition rates. A more recent approach performs MAP adaptation from a channel independent model to a set of channel dependent models, and utilizing the mapping parameters in the feature domain [48]. The approach shows a significant reduction in error rate compared to the uncompensated system when evaluated on verification tasks with Switchboard-II landline and cellular speech data. While similar to model synthesis approaches (discussed in detail later) utilizing the map information in the feature domain provides generality and enables the use of the technique for speech recognition. A feature warping approach has also been proposed which involves altering the distribution of a cepstral feature stream over a certain time interval to match a target distribution [13]. By conforming the distribution of the individual feature components to a particular form the variation experienced by the distribution across environments is reduced. Speaker verification experimentation on the NIST 1999 database shows that the warping approach is at least comparable to other common compensation techniques (such as mean normalization) in all cases [13].

Reducing telephone handset mismatch was the specific focus of the feature compensation method developed by Quatieri and colleagues [49]. By matching the spectral magnitude of the distorted signal to the output of a reference driven channel model, a handset mapper is designed which improves consistency between high quality training data and low quality testing data. Using HTIMIT [50] to train the mapper, recognition tests on NIST 1997 and NIST 1996 data showed improvement over the uncompensated baseline. Heck et al. use a discriminative feature design approach to further increase speaker recognition system robustness to telephone handset mismatch [51]. A non-linear artificial neural network is discriminatively trained for speaker recognition rate maximization under handset mismatch, and is employed to transform standard cepstral features. The method significantly outperforms a baseline mel-cepstral mean normalized system while not requiring stereo training data.

Despite the success of feature space techniques in reducing channel distortion effects, applying compensation

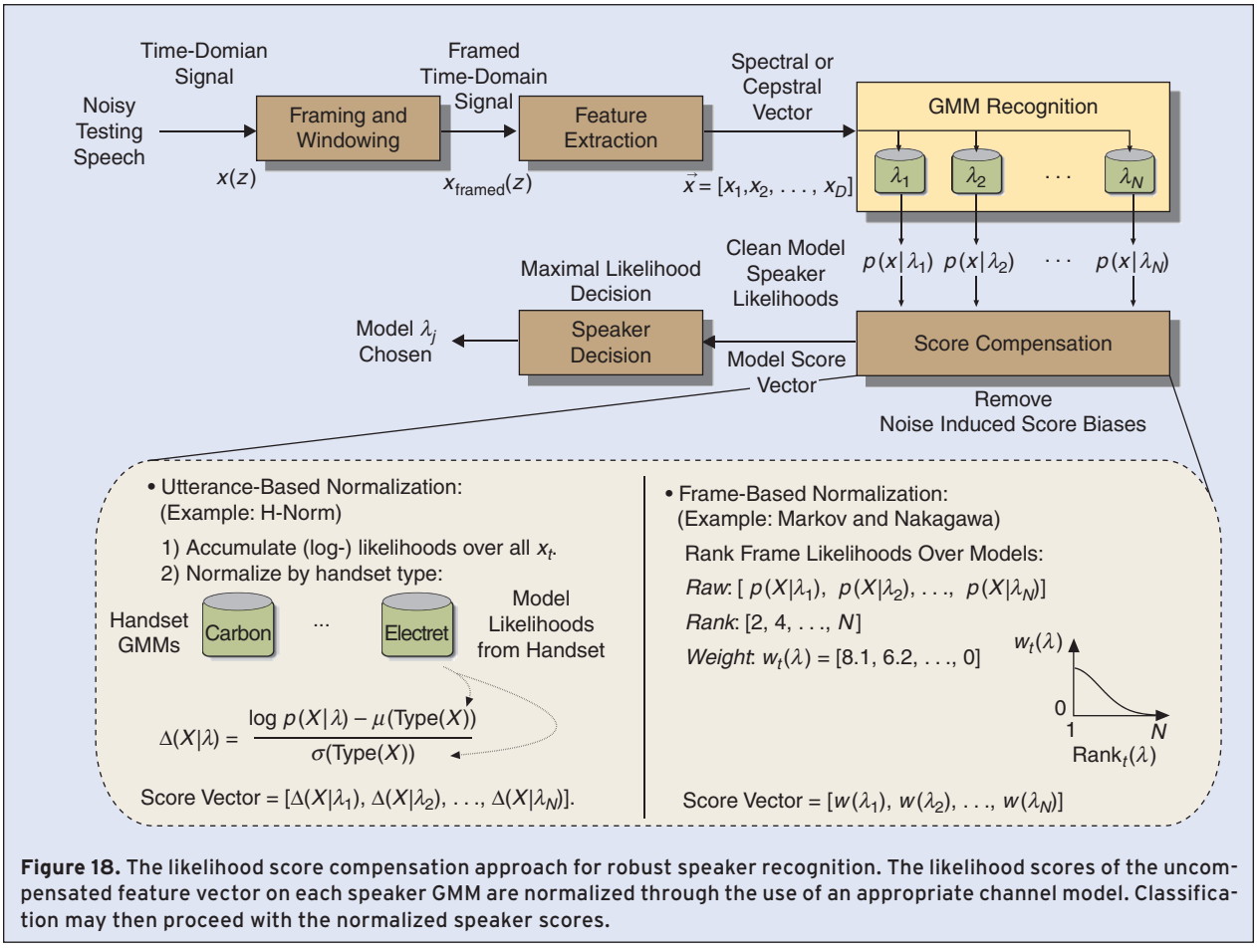

Figure 18. The likelihood score compensation approach for robust speaker recognition. The likelihood scores of the uncompensated feature vector on each speaker GMM are normalized through the use of an appropriate channel model. Classification may then proceed with the normalized speaker scores.

to standard cepstral or spectral features is less effective for eliminating environmental noise. Some past research has attempted to address this by examining alternate feature parameterizations which inherently provide more robustness to additive artifacts. Higher order spectral domains are advantageous for robust feature extraction due to the preservation of phase relations and information about Gaussianity which is not contained in the power spectrum [14]. Specifically bispectral magnitude based features have been proposed to suppress the effects of additive Gaussian noise on input speech signals for speaker identification [15]. An integrated bispectral phase feature has also been developed to perform identification under clean and mismatched conditions [52], [53]. Evaluation results suggest the feature can approximate MFCC performance for clean speech testing, and provide large improvements over the cepstral feature in low SNR additive white noise distortion. The disadvantage of higher order spectral based features is the large computation time required for spectral estimation, and this has largely prevented their use in the feature extraction stage of practical real-time systems.

C. Score Compensation Approaches

Although feature compensation approaches are effective in reducing linear channel effects, practical telephone handsets may induce other non-linear degradations on the input speech signal [39]. Score compensation methods attempt to eliminate the effects of these distortions on log-likelihood scores produced by the GMM recognizer (as shown in Fig. 18). To achieve this the handset dependent score normalization technique (H-norm) was developed which involves the construction of GMMs to model non-linear uncompensated channel effects within each of the relevant conditions [22]. During recognition the test segment is assigned a handset type classification based on the handset GMMs, and the speaker GMM likelihood is modified by normalization with the handset model parameters. Results on the NIST 1998 and NIST 1999 databases showed that H-norm improved robustness against channel mismatch for verification tasks, but is also applicable to speaker identification.

Other score normalization methods previously utilized in recognition tasks with channel variability include Z-norm [54] and T-norm [55]. Z-norm has an advantage in that it does not involve the explicit labeling

of each test utterance according to its channel type, and estimation of the normalization parameters can be performed offline. The T-norm approach extends this by scaling the score distribution with the variance of the imposter scores. Since the mean and variance normalization parameters are estimated from the test utterance, T-norm avoids the test-to-normalization mismatches which are possible in Z-norm [55]. An experimental evaluation comparing T-norm and Z-norm approaches was conducted on cellular data from the NIST 2001 database, and the results showed T-norm produced superior verification rates particularly at low EERs [56]. Other evaluations suggest that for data which has significant handset degradation the combination of H-norm and T-norm approaches (HT-norm) provides improved performance [55].

Score normalization approaches have also been previously applied directly to speaker identification tasks. In early work Gish and Schmidt combine multiple model training with segmentation and normalization to limit the effect of mismatched conditions [4]. Scoring using multiple GMMs trained on different conditions, the best likelihood over all models is considered for each speaker and normalization is performed to allow segment comparisons. A frame level normalization approach is presented by Markov and Nakagawa [57] where for each frame the speaker model likelihoods are ranked allowing model weights to be assigned using an arbitrary weighting function. Classification is performed using accumulated weight scores instead of the frame likelihood values. This frame level approach was adapted by Zheng et al. specifically to eliminate channel variability [58]. For the model likelihood scores of a test frame, a non-linear transformation was applied in order to accentuate the frame score between speakers and promote consistency between scores over all frames for the same speaker. Results of an experimental evaluation on the NIST 2000 database showed that combining frame level likelihood normalization with a GMM-UBM recognizer significantly improved identification error due to channel distortion, provided the score transformation parameters were adequately tuned.

D. Model Compensation Approaches

Model based compensation involves the modification of the trained GMM distributions in order to learn the noise characteristics, and thus increase the robustness of the recognition decisions (see Fig. 19). Such methods which address channel based distortions have been developed for speech recognition, however adaptation based approaches that adjust model parameters to represent test data are not readily applicable in speaker recognition due to the accompanying loss of speaker discriminative information. A simple approach to the model based compensation of telephone handset mismatch utilizes handset dependent background models to separate true and false speaker verification score distributions [59]. A handset classifier was applied during enrollment to ensure that normalization of scores on this model were performed with the matched handset type. Experimental evaluation showed that using handset dependent models to normalize the claimant scores reduced mismatch bias and improved false alarm rates. For speaker identification tasks it is only the mismatch between trained model conditions and testing conditions which must be compensated. An extension of handset dependent modeling to identification tasks would therefore require the construction of speaker specific GMMs in each channel condition.

Probabilistic factor analysis approaches utilize large amounts of multicondition enrollment data in order to jointly model inter-speaker and channel variability [61]. The motivation of this method is to allow adaptation based channel compensation for speaker recognition, while preventing reductions in inter-speaker variability (as is observed when applying traditional speech recognition adaptation techniques). In combining the speaker supervector prior formulation from the eigenvoice MAP, eigenchannel MAP and classical MAP techniques, speaker and channel factors are used to represent the inter-speaker and channel variation respectively between supervectors from the testing session and the enrolled speaker model [61]. For text-independent verification tasks on the NIST 2000 database, factor analysis produced a significant improvement over handset detection based approaches [62]. Despite recent research to improve its computational requirements [63], the factor analysis approach is limited by the computational complexity in both training (to estimate the speaker-independent hyper-parameters) and testing (to calculate the model likelihoods).

In practical applications it is unlikely that sufficient training data will be available for each speaker over all channel types which require compensation. As depicted in Fig. 19, the Speaker Model Synthesis (SMS) method addresses this issue by learning model parameter changes between different channels allowing synthesis of speaker models in unseen training conditions [60]. A channel independent multicondition GMM is firstly constructed using all data from a collection of channels. Using channel specific data, MAP adaptation is then applied to produce channel dependent GMMs, and transformations between different channel models are learned. For speaker training under a particular channel type, the effect of mismatch during testing may be reduced by synthesis of the training channel type to the

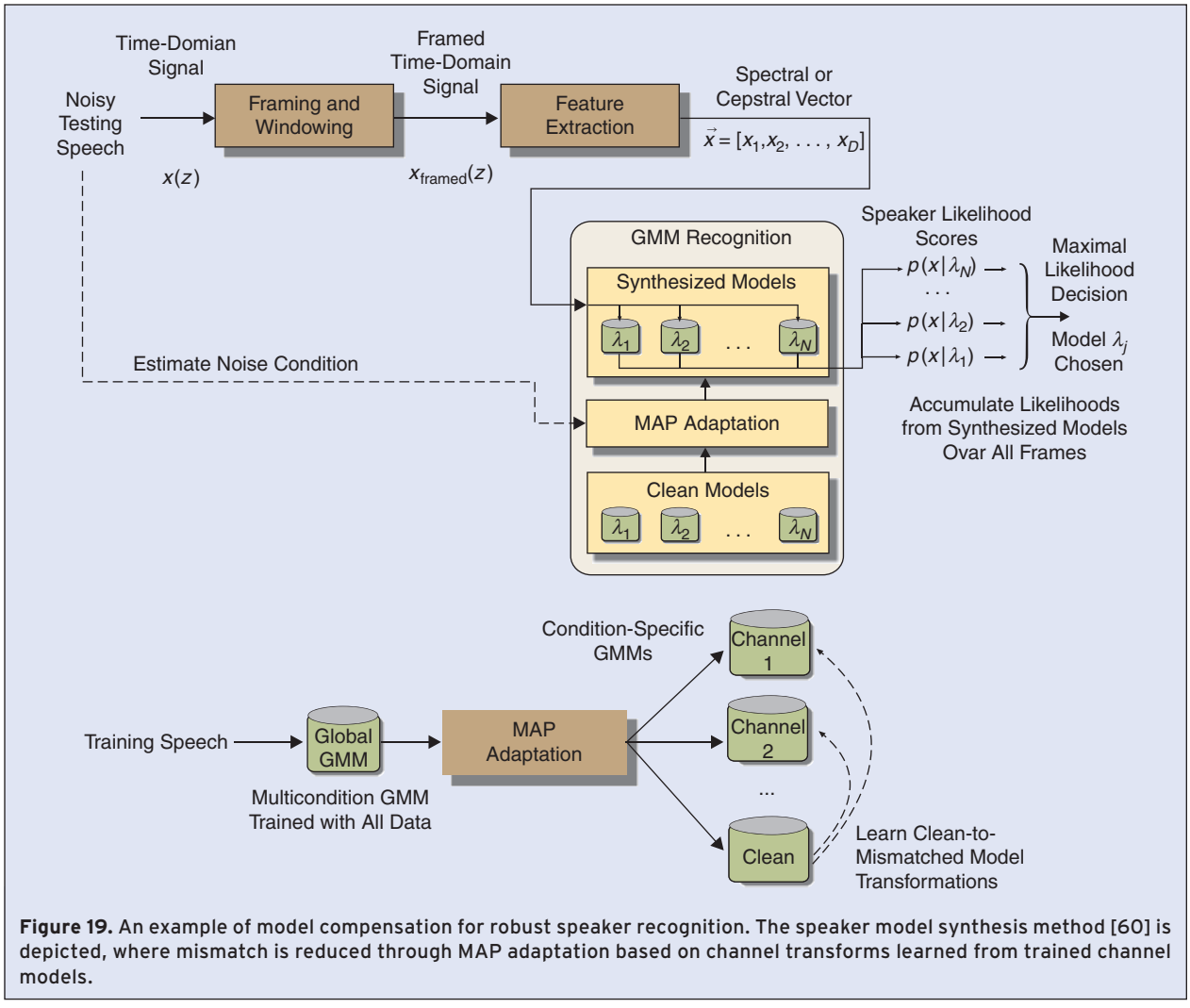

Figure 19. An example of model compensation for robust speaker recognition. The speaker model synthesis method [60] is depicted, where mismatch is reduced through MAP adaptation based on channel transforms learned from trained channel models.

testing channel type using the channel dependent transform [60].

The implementation of model synthesis for identification requires information about the mismatch to be extracted from the test speech waveform. Murthy and colleagues proposed a channel mismatch compensation method which learns a *synthetic variance distribution* (SVD) from stereo data recorded from multiple handsets [64]. Utilizing this distribution, a handset independent transformation is applied to the component variances within the trained speaker GMMs with the aim of increasing robustness to channel effects. Experimental evaluation on speaker identification tasks using the NIST1996databaseshowedthatfixedtargettransformation of the variances based on the SVD reduced the EER compared to the uncompensated case.

Model domain adaptation approaches may also provide robustness to environmental noise effects. Proposed for robust HMM-based speech recognition, parallel model combination (PMC) constructs noise corrupted speech models by combining the individual HMMs which model the speech and noise [16]. The technique is similar to HMM decomposition [65], although PMC may operate with cepstral based features by transforming the model parameters back into the log-spectral domain before combination via the mismatch function. This type of adaptation has been shown to be effective for compensating against additive noise for speech recognition [16], [96] and has been recently applied to the robust text-dependent verification task [67]. For an evaluation on the YOHO database [68], applying PMC to HMM-based text-dependent verification significantly improves the EER (relative to the MFCC + CMN baseline) in stationary and non-stationary noise cases. However, the application of PMC to GMM text-independent speaker identification tasks may be problematic due to the unconstrained nature of the speech and the subsequent lack of temporal information

captured by the speaker models (since phonemes cannot be easily modeled).

VII. Missing Data Robustness

Despite significant past research, the traditional approaches to providing robustness for speaker recognition have various limitations. Although feature domain compensation provides a generic and widely useful robustness framework which is model independent, it has limited effectiveness against non-linear channel effects and non-stationary additive distortion. For speaker identification, feature domain compensation methods can effectively reduce channel distortion effects, but cannot handle environmental distortion without the availability of matched models. Frame-based normalization may reduce the severity of these distortions when occurring sporadically over short intervals, but not when the majority of the test utterance is affected. Model-based approaches can generally provide strong robustness to channel effects and some additive distortions, however they have the disadvantage of requiring modifications to the speaker models and typically need noise knowledge to perform the adaptation. These methods are thus generally not suitable for providing robustness in rapidly changing environments. With the increasing demand for speaker recognition in modern practical applications, effective robustness strategies must provide resistance to transient and non-stationary environments of an unknown nature.

Missing data approaches provide compensation against arbitrary disturbances within a speech signal, and are thus capable of dealing with the problem of environmental noise. These methods are based on the observation that speech signals have a high degree of redundancy, where information about the underlying speech characteristics persists even in the presence of extreme distortion. Initially formulated as a technique for recovering partially occluded objects in image recognition [69], the missing data approach has a conceptual relationship with the human auditory system and its ability to process corrupted speech signals [70], [71]. The approach is based on a time-frequency analysis of the input speech signal, and the subsequent quantification of noise in each individual time-frequency point. Recognition occurs by utilizing the time-frequency regions labeled as speech dominant such that the effect of the noise is dramatically reduced. In this section missing data approaches for speaker identification are examined, including an introduction to the fundamentals of missing data processing, an outline of previous mask estimation methods for speaker identification, and a review of recently proposed model based constraints for missing data systems. Finally, we discuss the potential of approaches aimed at combining information from the signal and the trained models for increased speaker identification robustness.

A. Spectrographic Representations and Reliability Masking

In missing data approaches the time-frequency representation of the speech signal is used to model its noise corruption. Consider a clean speech sampled signal in the time domain represented by $s(z)$, where z is the sample index. Applying a windowing function, taking the magnitude of the FFT, and passing the output through a Mel-filterbank (as in (1)) produces the power spectral representation $S_{\mathrm{p}}(t, f)$, which represents the speech signal energy at time frame t and frequency channel f. The final time-frequency representation is obtained by applying a compressive non-linear function which, in this analysis, is assumed to be a logarithm: $S(t, f) = \log S_{\mathrm{p}}(t, f)$, where $S(t, f)$ is the log-spectral value at location (t, f) of the signal's spectrogram.

Corrupting the speech signal with an additive uncorrelated noise process with power spectrum $N_{\mathrm{p}}(t, f)$ produces a received signal with power spectra $X_{\mathrm{p}}(t, f)$, which is equivalent to the sum of the clean speech and noise spectra:

$$X_{\mathrm{p}}(t, f) = S_{\mathrm{p}}(t, f) + N_{\mathrm{p}}(t, f), \tag{19}$$

where the complete power spectral representations of the corrupted speech, clean speech and noise signals are given by X_{p}, S_{p} and N_{p} respectively. Let the corresponding log-spectral representations be denoted as X, S and N. The clean speech spectrogram S consists of a series of (log-)spectral vectors $\vec{s}_t = S(t)$ for time frame t, and each spectral vector has individual spectral components $s_{tf} = S(t, f) \in \vec{s}_t$ for each channel f of a D-dimensional filterbank. Applying equivalent notation to the corrupted speech and noise signals their spectral vectors at time t are denoted by $\vec{x}_t$ and $\vec{n}_t$ respectively, and the individual spectral components at time-frequency location (t, f) by x_{tf} and n_{tf} respectively.

Due to the time varying nature of the speech and noise spectra, the severity of the distortion varies over time and frequency. The strength of the distortion at a given (t, f) location is quantified by the signal-to-noise ratio (SNR): $SNR(t, f) = S_{\mathrm{p}}(t, f)/N_{\mathrm{p}}(t, f)$. The spectrogram of the corrupted speech signal typically contains regions of high SNR, where the speech contribution dominates, and regions of low SNR where the noise characteristics are dominant [72]. It is the occurrence of the low SNR regions in the spectrogram which decreases the performance of

speech processing systems, and the relative proportion of these low SNR regions increases with increasing noise strength (compare Fig. 20(a) and 20(b)).

Conceptually, missing data processing is based on the idea that by identifying the speech and noise dominant parts of the corrupted speech spectrogram the noise induced degradation can be minimized, and recognition performance may approximate that obtained by full knowledge of S. For a given corrupted speech signal vector $\vec{x}_t$, individual time-frequency components with a high SNR are similar to their corresponding clean speech spectral values and are labeled 'reliable'. Conversely, time-frequency components in $\vec{x}_t$ with a low SNR significantly differ from the clean speech value and are labeled 'unreliable'. The assignment of reliability decisions to components within all frames is represented by a time-frequency reliability mask M (see Fig. 20(c)). For a binary mask decision domain the components are considered as reliable ($m_{tf} = 1$) or unreliable ($m_{tf} = 0$) with absolute certainty, while soft masking allows for uncertainty in the decision by interpreting the mask value as a probability of reliability ($m_{tf} \in [0, 1]$) [73]. In this review we are concerned only with binary reliability decisions. Given the binary mask vector $\vec{m}_t$, the spectral vector $\vec{x}_t$ can be separated into reliable (r) and unreliable (u) component vectors according to $\vec{x}_{t_r} = \{x_{tf} | m_{tf} = 1, f = 1, 2, \ldots, D\}$ and $\vec{x}_{t_u} = \{x_{tf} | m_{tf} = 0, f = 1, 2, \ldots, D\}$. For similarly separated vectors of the clean speech signal $\vec{s}_{t_r}$ and $\vec{s}_{t_u}$ the following relations are assumed between clean and corrupted speech reliability sub vectors

$$\vec{x}_{t_r} = \vec{s}_{t_r} \tag{20}$$

$$\vec{x}_{t_u} \geq \vec{s}_{t_u}. \tag{21}$$

The first relation is a consequence of the compressive nature of the log operation on the spectral values: for a reliable component $S_p(t, f) > N_p(t, f)$ in the log domain we have $x_{tf} = X(t, f) = \log[S_p(t, f) + N_p(t, f)] \approx \log[S_p(t, f)] \approx s_{tf}$. Since it is assumed that the spectral values of $S_p(t, f)$ and $N_p(t, f)$ are non-negative, it follows that $N_p(t, f) \geq 0$ and $0 \leq S_p(t, f) \leq X_p(t, f)$. In the log domain $0 \leq s_{tf} \leq x_{tf}$, and this bounding constraint may also be utilized in the recognition process (see Section VII-B2).

The robustness provided by missing data compensation is dependent on the accuracy of the reliability labeling. When full a priori noise knowledge is available ideal spectrographic masks can be constructed according to an SNR-based criterion:

$$m_{tf}^{\text{oracle}} = \begin{cases} 1 & \text{if } 10 \log_{10}[S_p(t, f)/N_p(t, f)] > \theta_{\text{oracle}}, \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where θ_{oracle} is a threshold in dB. Construction of the oracle time-frequency mask is shown in Fig. 21. When $\theta_{\text{oracle}} = 0$ dB the reliability assignment is determined by whether the speech energy exceeds the noise energy, however in previous studies the value of θ_{oracle} has been increased to attempt to ensure that only speech dominant points are reliably labeled [74]. Although there is no single optimal value for the threshold, typically $\theta_{\text{oracle}} \in [-3 \text{ dB}, 3 \text{ dB}]$ for the calculation of these ideal ('oracle') masks [72], [75].

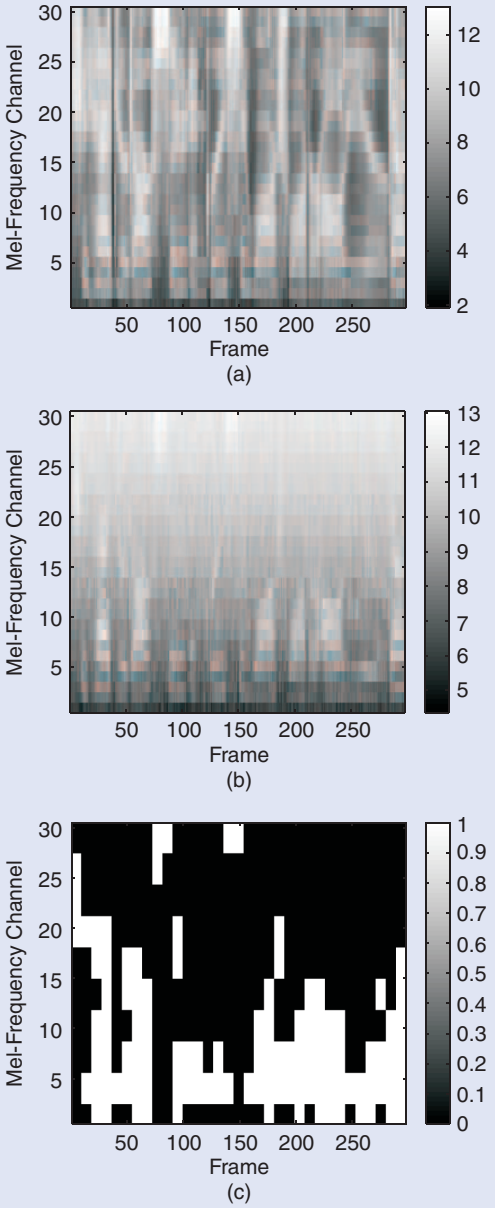

Figure 20. (a) Spectrograms for a speech utterance from the TIMIT database in clean conditions, and (b) in additive white noise at 5 dB SNR. (c) Also shown is the reliability mask containing (ideal) reliability labels for time-frequency regions in the utterance. (a) Clean spectrogram. (b) Speech and noise spectrogram. (c) Oracle reliability mask.

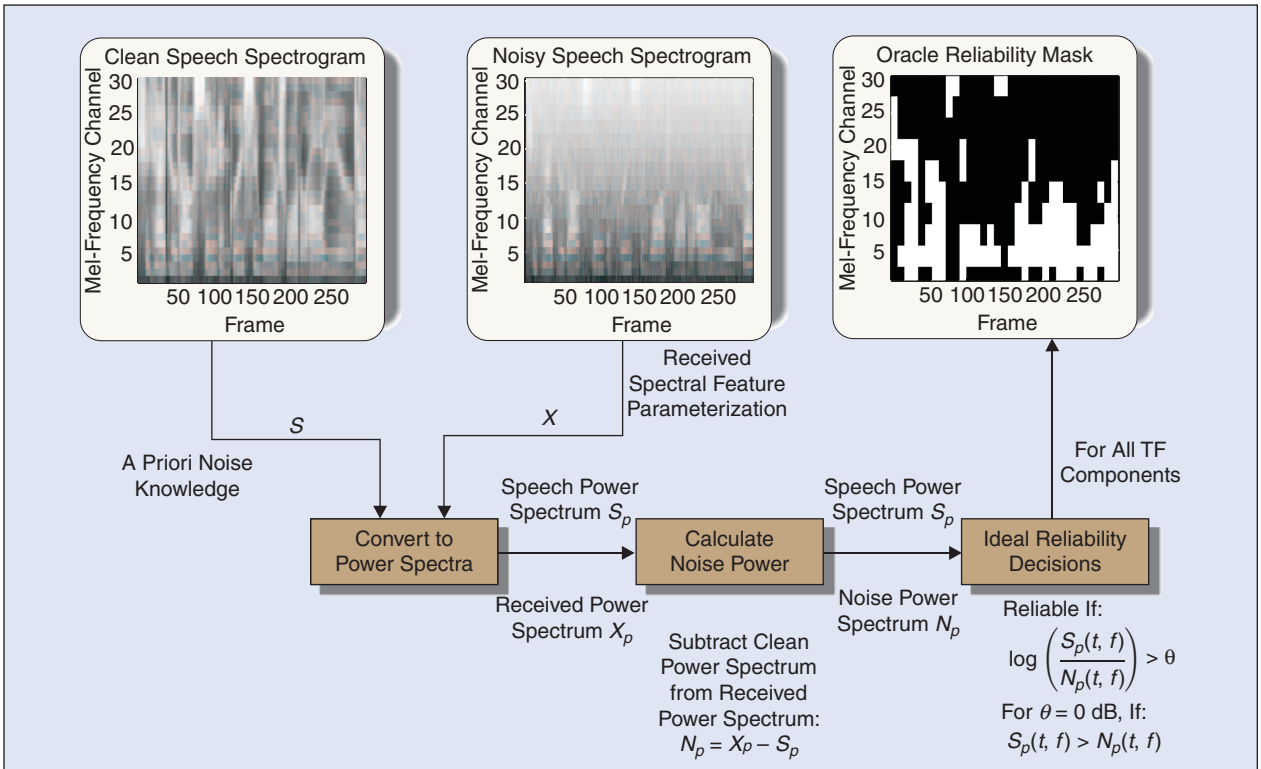

Figure 21. The calculation of the ideal time-frequency reliability mask given the clean (a priori) and the corrupted (received) test utterance spectrograms. By converting the log-spectrographic representations back into the power spectral domain, the noise power may be calculated and reliability labels assigned to each time-frequency (TF) point. A threshold of $\theta_{oracle} = 0$ dB is used.

B. GMM Recognition with Missing Data

Given the corrupted received speech signal feature representation X, and the binary time-frequency reliability mask which partitions each vector $\vec{x}_t \in X$ into reliable and unreliable subsets $\vec{x}_{t_r}$ and $\vec{x}_{t_u}$, missing data recognition methods must perform classification only with this partial knowledge. There are two distinct paradigms which may be used to achieve this [72]:

1) **Feature Vector Compensation**, where the log-spectrogram of the clean speech signal S is reconstructed by utilizing the reliable subsets $\vec{x}_r$ to estimate the true clean speech values of the unreliable subset (i.e to find $\vec{s}_u \leq \vec{x}_u$ for all frames).
2) **Classifier Compensation**, where the computation of the frame likelihood $p(\vec{x} \mid \lambda)$ within the GMM recognizer is modified to accommodate the presence of unreliable data.

1) Feature Vector Compensation

These approaches utilize the reliability information as input to a classification problem which estimates the true value of the unreliable spectral components in a vector based on a priori knowledge of the speech spectrographic structure (as in Fig. 22). Specifically, clean speech training produces a Gaussian estimate of the clean speech spectrogram values conditional on the observed reliable components. There are two main variants to feature compensation: *cluster-based reconstruction* and *covariance-based reconstruction*.

a) Cluster-Based Reconstruction

The cluster-based approach assumes that spectrographic component values are time invariant, and models the spectral vectors within a clean speech signal as independent and identically distributed random variables. The distribution of clean spectral vectors, as obtained from a training data set, is assumed to consist of clusters where all vectors belonging to the cluster share the cluster distribution [75]. Assuming the cluster distributions are Gaussian the distribution of an arbitrary spectral vector $\vec{x}$ is given:

$$P(\vec{x}) = \sum_{v=1}^{V} g_v \frac{1}{(2\pi \mid \Theta_v \mid)^{D/2}} \exp\left\{ -\frac{1}{2}(\vec{x} - \vec{\mu}_v)' \Theta_v^{-1} (\vec{x} - \vec{\mu}_v), \right. \tag{23}$$

where D is the vector dimension, g_v is the a priori probability that $\vec{x}$ belongs to cluster v, and μ_v, Θ_v are the mean

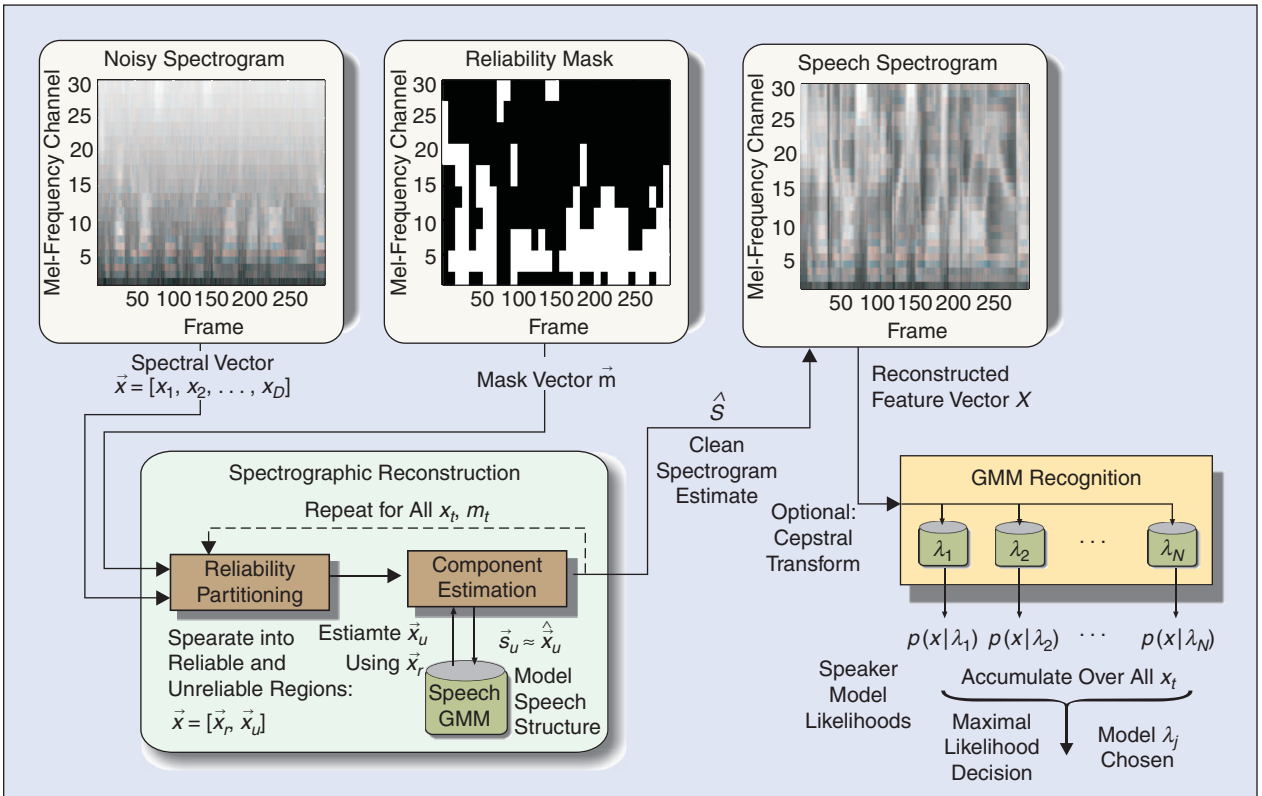

Figure 22. Missing data recognition via reconstruction of the speech spectrogram. For a corrupted spectrogram and corresponding reliability mask, knowledge of the structure of the speech is captured by a classifier which estimates the true value of unreliable components based on the reliably declared components.

and variance respectively of cluster v [75]. The first step in the re-estimation process is the determination of the cluster membership $v_{\vec{x}}$ for speech vector $\vec{x}$ with reliable and unreliable subsets $\vec{x}_r$ and $\vec{x}_u$:

$$\hat{v}_{\vec{x}} = \underset{v}{\operatorname{argmax}} \{p(\vec{x}_r | v) P(v)\}. \tag{24}$$

Since only the values of $\vec{x}_r$ approximate the clean speech signal, the cluster membership estimate $\hat{v}_{\vec{x}}$ is obtained via a ML decision by integrating over the unreliable components to obtain the marginal density $p(\vec{x}_r | v)$. Once the cluster membership has been identified MAP estimation is applied on the relevant cluster distribution to estimate the clean values $\hat{\vec{x}} \approx \vec{s}_u$ corresponding to the received subset $\vec{x}_u$, conditional on $\vec{s}_u \leq \vec{x}_u$ [72], [75], [76]:

$$\hat{\vec{x}}_u = \vec{\mu}_{v,u} + \Theta_{v,ur} \Theta_{v,uu}^{-1} (\vec{x}_r - \vec{\mu}_{v,r}), \tag{25}$$

where $\vec{s}_r \approx \vec{x}_r$ and $\mu_{v,r}$, $\mu_{v,u}$, $\Theta_{v,ru}$, $\Theta_{v,rr}$ are the mean and variance parameters partitioned according to subsets $\vec{x}_r$ and $\vec{x}_u$. Recognition (or cepstral transformation) may then proceed with the imputed spectral vector $\vec{x} = (\vec{x}_r, \hat{\vec{x}}_u)$.

b) Covariance-based Reconstruction

In correlation-based approaches the reconstruction of an individual component is dependent on its correlation with all other components, and spectral vectors are thus modeled as samples of a stationary Gaussian random process. Estimation of the Gaussian process parameters are performed using clean speech training, where the order of emission of the spectral vectors has no influence on their model means and inter-element covariances. That is, for a received spectral vector $\vec{x}_t$ with components $x_{t_1 f_1}, x_{t_2 f_2} \in \vec{x}_t$ the mean of component x_{tf} and covariance between components $x_{t_1 f_1}, x_{t_2 f_2}$ is given respectively as [75]:

$$\mu(t, f) = E[\vec{x}_{tf}] = \mu(f), \tag{26}$$

$$C(t_1, t_2, f_1, f_2) = E[(x_{t_1 f_1} - \mu_{f_1})(x_{t_2 f_2} - \mu_{f_2})]. \tag{27}$$

The estimated unreliable component values $\hat{\vec{x}}_{t_u}$ are produced using bounded MAP as in the cluster-based case. However for simplicity the procedure may be constrained such that only reliable components within a certain neighborhood of the given unreliable component contribute to the estimation. In work by Raj and Stern [72] this neighborhood $\vec{x}_{t_r}^{(n)}$ is defined as all reliable components in the spectrogram which have normalized

covariance of greater than or equal to 0.5 with at least one element in $\vec{x}_{t_u}$.

2) Classifier Compensation

In classifier compensation approaches the classifier itself is modified to perform recognition with the partitioned spectral vector. The two most prominent variants are *class-conditional imputation* and *marginalization*.

a) Class-Conditional Imputation

Class-conditional imputation involves a reconstruction of the unreliable spectral components $\vec{x}_u$ which is specific to the distribution of a particular HMM state [77]. For speaker recognition a GMM is assumed and so the state output likelihood becomes equivalent to the model likelihood $p(\vec{x}\,|\,\lambda)$. For a corrupt spectral vector partitioned as $\vec{x} = (\vec{x}_r, \vec{x}_u)$, the model specific unreliable spectral estimate is obtained via MAP using the distribution of model λ_k:

$$\hat{\vec{x}}_{u,k} = \underset{\vec{x}}{\operatorname{argmax}} \{p(\vec{x}\,|\,\vec{x}_r, \lambda_k)\}. \tag{28}$$

Classification is performed by producing class specific spectrogram estimates $X_k = (X_{r,k}, X_{u,k})$ allowing a ML decision to be made over all spectral vectors as in (9).

b) Marginalization

In marginalization the spectral components of the input observation are used directly to perform optimal recognition based on the reliable (r) and unreliable (u) subsets. As demonstrated in Fig. 23, this approach replaces the standard Gaussian output likelihood $p(\vec{x}\,|\,\lambda)$ with the marginal distribution of the reliable components $p(\vec{x}_r\,|\,\lambda)$, which is conditional on the unreliable components. For a given reliability partition $\vec{x} = (\vec{x}_r, \vec{x}_u)$, the GMM parameters for each speaker model may also be separated according to:

$$\vec{\mu}_i = (\vec{\mu}_{r_i}, \vec{\mu}_{u_i}) \text{ and } \Sigma_i = \begin{bmatrix} \Sigma_{rr_i} & \Sigma_{ru_i} \\ \Sigma_{ur_i} & \Sigma_{uu_i} \end{bmatrix}, \tag{29}$$

where i is the mixture index. The marginal likelihood is obtained by integrating over the distribution of the conditional unreliable components [74], [78], [79]:

$$p(\vec{x}_r\,|\,\lambda) = \sum_{i=1}^{M} g_i \mathcal{N}(\vec{x}_r; \vec{\mu}_{r_i}, \Sigma_{rr_i}) \int_{\vec{x}_L}^{\vec{x}_H} \mathcal{N}(\vec{x}_v; \vec{\mu}_{u|r_i}, \Sigma_{u|r_i}) d\vec{x}_v, \tag{30}$$

where g_i is the weight of the ith mixture and $\mathcal{N}(\cdot)$ is a multivariate Gaussian as given by (5). The conditional mean $\vec{\mu}_{u|r_i}$ and covariance $\Sigma_{u|r_i}$ parameters are calculated as:

$$\vec{\mu}_{u|r_i} = \vec{\mu}_{u_i} + \Sigma'_{ru_i} \Sigma_{rr_i}^{-1} (\vec{x}_{r_i} - \vec{\mu}_{r_i}), \tag{31}$$

$$\Sigma_{u|r_i} = \Sigma_{uu_i} - \Sigma'_{ru_i} \Sigma_{rr_i}^{-1} \Sigma_{ru_i}. \tag{32}$$

The contribution of the unreliable component distribution to the marginal density is dependent on the choice of integration bounds $\vec{x}_L$ and $\vec{x}_H$. In the absence of any information about the unreliable components the bounds extend to infinity ($[\vec{x}_L, \vec{x}_H] = [-\infty, \infty]$) resulting in the bounded integral contribution evaluating to 1.0, and the marginal density thus becomes dependent only on the reliable components. However, in the case of log-spectral features it is known that the true feature values $\vec{s}_u$ are constrained by zero and the observed values $\vec{x}_u$. When the upper and lower limits are set to match these constraints ($[\vec{x}_L, \vec{x}_H] = [\vec{0}\,, \vec{x}_u]$) bounded marginalization is performed [79]. Past research has shown that, although more computationally expensive, the use of bounded marginalization produces far superior recognition performance to full marginalization due to the ability to penalize models based on the unreliable components' known range of values [78]. If the covariance structure within the speaker models is diagonal, such that $\Sigma_i = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2)_i \in \mathbb{R}^D$, then the multivariate reliable and conditional unreliable distributions reduce to products over the univariate Gaussian distributions corresponding to each component. However, despite the dominance of diagonal covariance modeling in speech recognition, recent research has shown that for missing data speaker identification full covariance modeling significantly outperforms diagonal modeling due to the ability to capture the relations between the correlated spectral components [80].

3) Discussion

In comparing marginalization and imputation techniques, marginalization offers the advantage of performing optimal recognition directly based on the spectral values and the given reliability information. Previous work shows the high recognition rates obtainable using marginalization approaches for both speech and [72], [74], [81] and speaker [80] recognition tasks. When constrained to the log-spectral feature domain marginalization based recognition is superior to imputation approaches, particularly at low SNRs [82]. However, with marginalization approaches the recognizer is limited to utilizing the log-spectral feature domain, or other feature domains which localize the corruption in individual time-frequency points. Spectrographic reconstruction approaches allow the transformation of the final estimated clean spectral vectors to the cepstral domain, which is widely regarded as favorable to recognition. When performing this transformation following spectrogram reconstruction, the cluster-based imputation method was observed to outperform spectral based

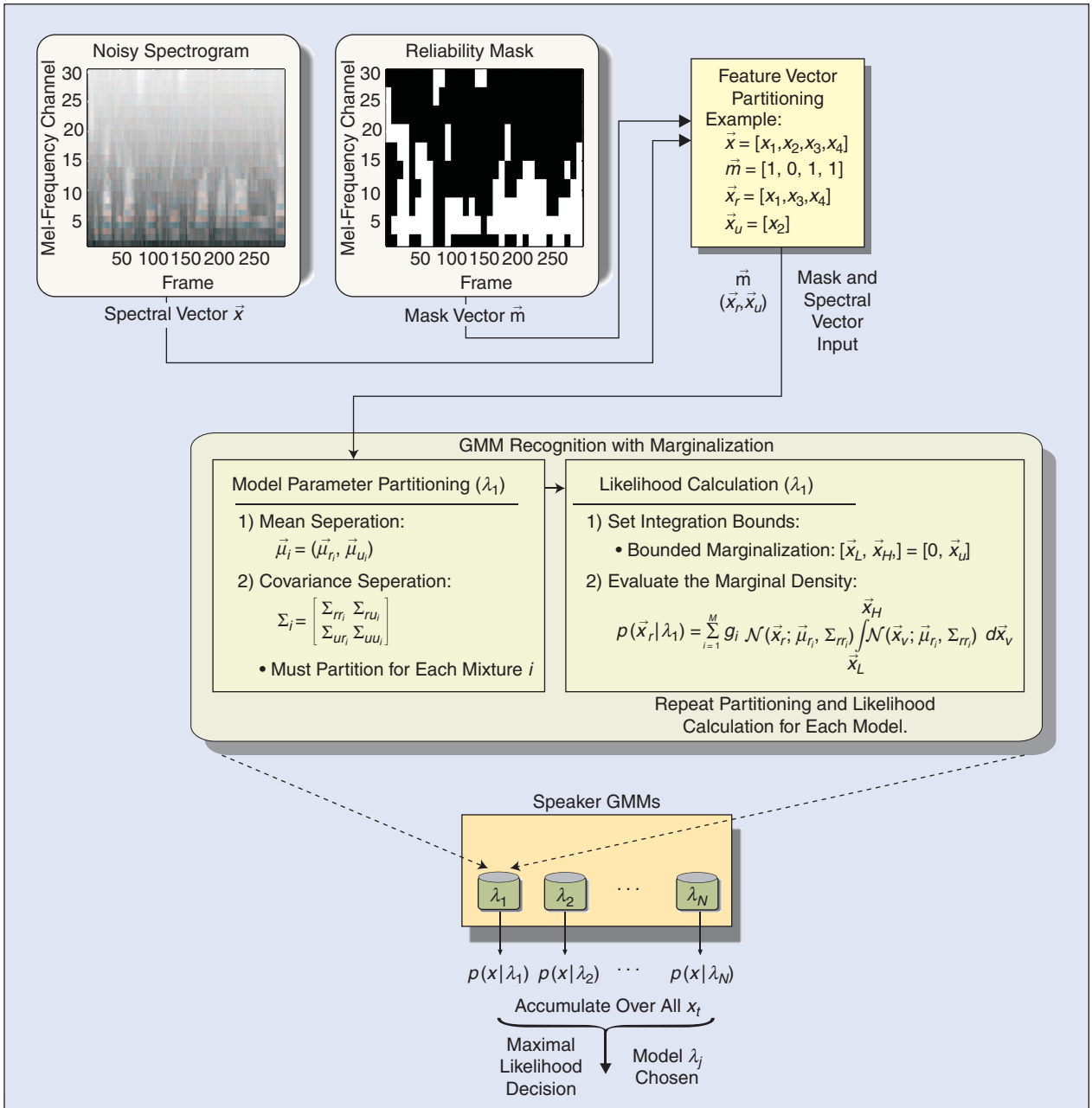

Figure 23. Missing data recognition via marginalization. For each spectral vector the reliable component distribution and conditional unreliable component distribution are utilized to produce a marginal likelihood score.

marginalization. Although this evaluation showed class-conditional and covariance based imputation as inferior to the cluster-based variant for general robust speech recognition, other work suggests these methods do have uses for specific types of spectrographic noise corruption [75], [77].

Other factors must also be considered when interpreting the results by Raj and colleagues [82], such as the use of a smaller dimensional log-spectral vector for marginalization in comparison to the cepstral vectors produced post-reconstruction. In addition, the use of diagonal covariance models distinctly biases the results towards the imputation methods, since the extracted cepstral features are largely uncorrelated while spectral features show significant correlation. Marginalization based systems should thus dramatically benefit from the use of full covariance models. The computational requirements for performing recognition with either approach should also be considered. Marginalization is disadvantaged here due to the need to calculate a computationally expensive multivariate integral, although recent research has introduced

techniques such as sub-band processing to reduce this overhead [83].

C. Mask Estimation for Speaker Identification

Despite the fundamental differences between the recognition paradigms, the robustness provided by any missing data compensation strategy is critically dependent on the accuracy of the time-frequency reliability decisions. In the case of complete a priori noise knowledge the true SNR is known for each time-frequency point allowing oracle masks to be constructed (see (22)). When these ideal reliability masks are available missing data recognition produces high robustness even under extreme noise distortion [81]. In practice the absence of a priori noise knowledge forces an estimation of the reliability decisions, where past research has largely concentrated on developing techniques which accurately reproduce the oracle mask. Traditional methods for missing data mask estimation utilize the properties of the speech signal to calculate the reliability decisions, and these *bottom-up* approaches can be categorized as SNR-based estimation techniques, auditory and perceptual estimation techniques, and classification-based techniques. Examples for each of the distinct estimation approaches are discussed below, with specific emphasis on their applications for robust speaker recognition.

1) SNR-Based Estimation

In SNR-based techniques the time-frequency reliability decisions are calculated by a direct estimation of the power spectrum of the corrupting noise signal. The spectral subtraction method [84] can be utilized to achieve this, where an estimate of the average noise power spectrum is obtained by assuming silence (no speech energy) within the first several frames of the utterance. For a received noise corrupted power spectrum $X_{\mathrm{p}}(t,f)$ the estimated clean speech power spectrum is produced by:

$$\hat{S}_{\mathrm{P}}(t,f) = \begin{cases} X_{\mathrm{p}}(t,f) - \hat{N}_{\mathrm{p}}(t,f) & \text{if } X_{\mathrm{p}}(t,f) - \hat{N}_{\mathrm{p}}(t,f) > \gamma X_{\mathrm{p}}(t,f), \\ \gamma X_{\mathrm{p}}(t,f) & \text{otherwise,} \end{cases} \tag{33}$$

where $\hat{N}_{\mathrm{p}}(t,f)$ is the estimated noise power spectrum calculated over the first T_{avg} frames as $\hat{N}_{\mathrm{p}}(t,f) = 1/T_{\mathrm{avg}} \sum_{\mathrm{T}=1}^{T_{\mathrm{avg}}} X_{\mathrm{p}}(\tau,f)$, and γ is a small spectral flooring factor ($\gamma \approx 0$) ensuring non-negativity for components in the estimated clean power spectrum [84]. An SNR criterion for binary reliability decisions based on this spectral subtraction is given by

$$m_{tf}^{\mathrm{ss}(\theta)} = \begin{cases} 1 & \text{if } 10\log_{10}\left(\dfrac{\hat{S}_{\mathrm{p}}(t,f)}{\hat{N}_{\mathrm{p}}(t,f)}\right) > \theta, \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

where θ is the subtraction energy threshold in dB (see Fig. 24). A 0 dB criterion ($\theta = 0$ dB) implies that time-frequency points are reliable if the estimated clean speech energy exceeds the estimated noise energy: $\hat{S}_{\mathrm{P}}(t,f)/\hat{N}_{\mathrm{P}}(t,f)$. A negative energy criterion approach was proposed by Drygajlo and El-Maliki [85], where a component was labeled as reliable if the received signal power spectral value exceeded the estimated noise power spectral value:

$$m_{tf}^{-\mathrm{Energy}} = \begin{cases} 1 & \text{if } |X_{\mathrm{p}}(t,f)| > |\hat{N}_{\mathrm{p}}(t,f)|, \\ 0 & \text{otherwise,} \end{cases} \tag{35}$$

where the noise estimate $\hat{N}_{\mathrm{p}}(t,f)$ is produced using a voice activity detector (VAD). By combining spectral subtraction type enhancement (as in (34)) with negative energy criterion based missing data, a significant EER reduction is obtained for speaker verification in the presence of stationary additive noises [85]. This approach has been extended through the use of soft spectral subtraction for the speech enhancement and reliability masking stages [86]. Here subtraction of the noise estimate from the squared DFT magnitudes, performed prior to their input into the Mel-filterbank, decreases the error in each feature component. The resulting masks give an EER improvement compared to hard spectral subtraction when applied to the speaker verification task.

For use as a pure missing feature estimator, previous work on missing data speech recognition reports that the 0 dB criterion is superior to the negative energy criterion [74], [87]. By increasing the criterion threshold the noise energy may be overestimated allowing noisy points which would otherwise be included to be removed, at the expense of also removing speech dominated regions. The SNR-based estimation methods which utilize classical spectral subtraction via the 0 dB SNR reliability criterion are limited by their first order (mean only) representation of the noise. To account for variations in the noise magnitude a statistical detector was proposed which models the distribution of the noise in each frequency band [88]. This method produced improved verification performance compared to the standard 0 dB criterion.

The drawback of spectral subtraction based SNR mask estimation is its lack of robustness to non-stationary and transient noises. For these types of disturbances using the generalized noise estimate within a small number of frames to produce the entire noise spectrogram often results in inaccurate estimation. In

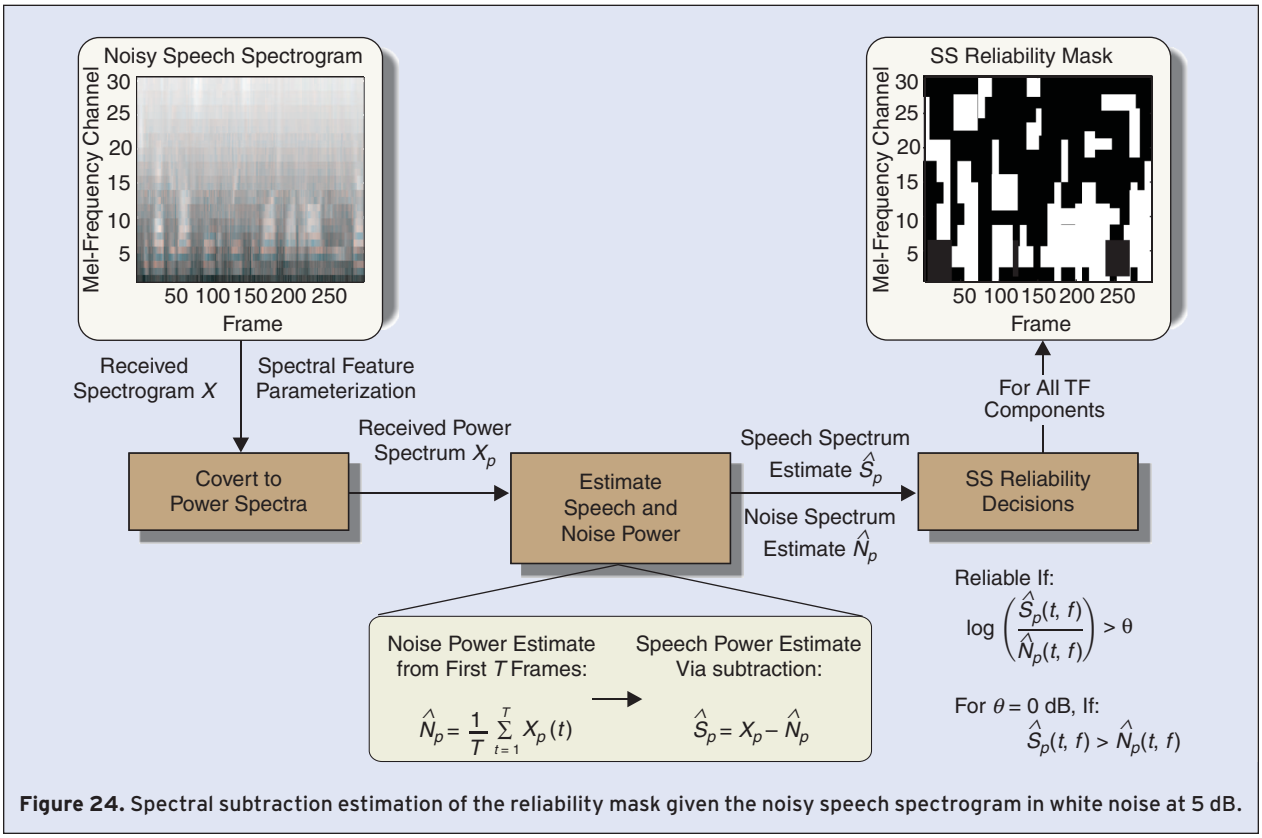

Figure 24. Spectral subtraction estimation of the reliability mask given the noisy speech spectrogram in white noise at 5 dB.

addition to spectral subtraction based approaches, other methods have been proposed to estimate the local SNR including the Vector Taylor Series method [75], and various spectral estimation techniques such as energy clustering, weighted averaging, envelope tracking and Hirsh histograms [89]. Although offering marginally improved performance compared to standard spectral subtraction estimation, their performance remains poor in non-stationary environments.

2) Auditory and Perceptual Estimation

In these types of mask estimation techniques reliability decisions for components of the received spectrogram are based on perceptually motivated criteria and the properties of the human auditory system. Computational auditory scene analysis (CASA) aims to model the human auditory ability to identify, segregate and process sounds from different sources, and is thus naturally suited to the problem of separating target speech signals from other noise sources [90], [91]. For missing data applications CASA approaches may be utilized to identify time-frequency regions which are dominated by a single source (the target speaker), and hence calculate reliability decisions without explicit local SNR estimation. A harmonicity grouping strategy was proposed by Barker et al. [92] based on the knowledge that energy within voiced speech is organized around harmonics of the fundamental frequency. In this approach harmonic groups are identified within voiced speech frames, and a decision process is consulted to determine whether the group belongs to the speech or noise source. A simplistic decision process which labels all identified groups as reliable is implemented and, when used in conjunction with a standard local SNR estimator, speech recognition performance which exceeds the multicondition baseline can be obtained. The disadvantage of the approach is the difficulty in identifying the source of the individual harmonicity groups in the case of distortions which have speech like properties.

Recently CASA approaches based on pitch information have been utilized to perform robust speaker identification [93]. Reliability decisions are achieved via monaural voiced speech segregation obtained from accurate pitch contour tracking [94]. For typically troublesome unresolved harmonics, the technique generates segments based on amplitude modulation (AM) in addition to temporal continuity and groups them according to AM rates. The voiced speech segregation technique was evaluated on a small 38 speaker set from the TIMIT database using marginalization to modify the diagonal covariance based GMM recognizer. Estimation of the voiced

speech segregation mask was found to considerably outperform estimation via spectral subtraction when tested in non-stationary cocktail party and rock music noises. However, the performance gap between the oracle and voiced speech segregation masks remains large, particularly at low SNRs.

Motivated by the importance of voiced speech regions for speaker discrimination [95], other CASA mask estimation methods have also focused on extracting voicing information from the corrupted speech signal. To identify the voicing character of speech spectra without explicit pitch estimation Jančovič and Köküer propose a measure based on the distance between the shape of the short-term signal and frame analysis window derived spectra [96]. The calculation of this voicing distance (VD) is summarized as follows: Firstly the magnitude spectra is produced via the Short-Time Fourier Transform (STFT) including zero-padding to ensure spectra smoothness. The voicing distance at DFT channel k is given by the Euclidean distance between magnitude spectra

$$\mathrm{VD}(k) = \left[\frac{1}{2M_c + 1} \sum_{\tau=-M_c}^{M_c} \left(|S_{\mathrm{STFT}}(k+\tau)| - |W(\tau)| \right)^2 \right]^{\frac{1}{2}}, \tag{36}$$

where M_c is the number of components in the speech spectra around k, and $S_{\mathrm{STFT}}(k)$ and $W(m)$ are values of the normalized STFT and windowed spectra respectively [96]. The VD for each filterbank channel f is produced by summing the individual distances from DFT channels which contribute to the filterbank output:

$$\mathrm{VD}^{\mathrm{FB}}(f) = \frac{1}{X_{\mathrm{FB}}(f)} \sum_{k=k_f}^{k_f+K_f-1} \mathrm{VD}(k) \cdot G_f(k) \cdot |S_{\mathrm{STFT}}(k)|^2, \tag{37}$$

where $G_f(k)$ is the frequency response of filterbank f, k_f and K_f are the lowest component and total number of components of $G_f(k)$ respectively, and $X_{\mathrm{FB}}(f)$ is the spectral output value of the filterbank. Finally the voicing distances are post-processed to help eliminate outliers. Experimental evaluation showed that the VD measure is related to the local SNR of a voiced filterbank channel, and this motivates the construction of a voicing mask which assigns time-frequency reliability decisions according to:

$$M^{\mathrm{voiced}}(t,f) = \begin{cases} 1 & \text{if } \mathrm{VD}_t^{\mathrm{FB}}(f) < \beta, \\ 0 & \text{otherwise,} \end{cases} \tag{38}$$

where β is an empirically determined threshold [97]. Jančovič and Köküer combine the estimated voicing decisions with a standard SNR mask where reliability decisions are formed via the SNR criterion using a stationary noise estimate. The combined masking approach was evaluated for marginalization based speaker identification on additive noise corrupted TIMIT data. The results confirmed a performance improvement for the combined strategy compared to the use of noise masking alone, particularly for the non-stationary noise case [97].

$$m_{tf}^{\mathrm{Bayes}} = \begin{cases} 1 \text{ if } p(\lambda_{m=1}^{\mathrm{Bayes}}(f)) p(\vec{s}_{tf} | \lambda_{m=1}^{\mathrm{Bayes}}(f)) > p(\lambda_{m=0}^{\mathrm{Bayes}}(f)) p(\vec{s}_{tf} | \lambda_{m=0}^{\mathrm{Bayes}}(f)), \\ 0 \text{ otherwise,} \end{cases} \tag{39}$$

CASA approaches to missing data mask estimation have also been developed to handle convolutional disturbances. Palomäki et al. proposed a modulation filtering scheme to produce reliability decisions for spectral features under reverberative distortion effects [98]. This method exploits the ability of modulation filtering to emphasize the strong speech regions (i.e the onsets) and remove regions contaminated by distortions. In an evaluation for robust speech recognition the modulation masking based missing data system showed large improvements over the uncompensated MFCC baseline. However, for low direct-to-reverberative ratios and large reverberation (T_{60}) times there is significant performance difference between utilizing the practical mask estimator and masks based on a priori knowledge. In further work for estimating reliability decisions in the presence of reverberation, Palomäki et al. have developed a binaural auditory model which uses spatial locality queues to identify time-frequency regions in the auditory scene that originate from a common azimuth [99]. When the resulting reliability masks are passed to a marginalization based missing data recognizer, this binaural model is shown to improve significantly over the MFCC baseline for reverberation times of up to $T_{60} = 2.7s$ and for spatial separations of 20° or larger.

3) Classifier-Based Estimation

In contrast to SNR-based and perceptual estimation techniques which formulate reliability decisions by directly utilizing properties of the speech signal, classifier based methods use supervised model training and a subsequent classification process to determine the labeling. In the Bayesian approach this classification is achieved via the modeling of distributions of a set of reliability features extracted from the noise distorted speech signal. If the features are designed such that they exploit the characteristics of the speech signal itself (rather than those of the corrupting noise), then probabilistic evaluation of the feature values will indicate whether the associated time-frequency point is reliable. Consider a spectral component x_{tf}, and a set of corresponding reliability features $\vec{s}_{tf}$. Under the Bayesian approach the binary reliability decision for x_{tf} is given by (39), where

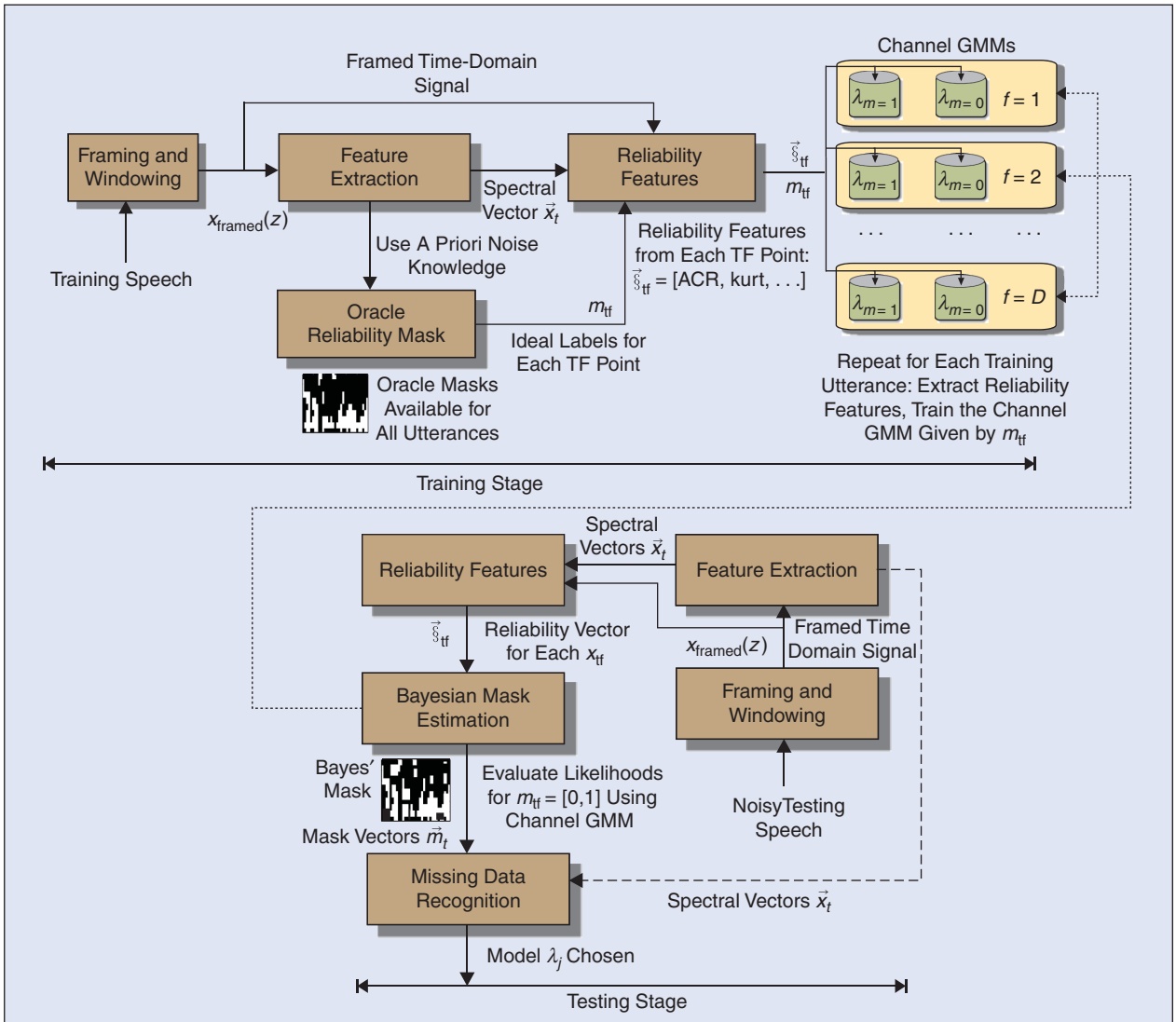

Figure 25. Classifier-based reliability mask estimation using the Bayesian approach proposed by Seltzer et al. Oracle mask training is used to construct statistical models for a set of reliability features in both unreliable and reliable speech regions. For a test speech utterance, a mask estimate is constructed by extracting the reliability features and performing classification on the trained distributions for reliable and unreliable decisions.

$\lambda_{m=1}^{\text{Bayes}}(f)$ and $\lambda_{m=0}^{\text{Bayes}}(f)$ are the trained distributions for a reliable and unreliable decision respectively in frequency channel f, and $p(\vec{\S}_{tf} | \lambda_{m=1}^{\text{Bayes}}(f))$ and $p(\vec{\S}_{tf} | \lambda_{m=0}^{\text{Bayes}}(f))$ are the corresponding probability densities of reliability and unreliability for channel f based on the reliability feature vector $\vec{\S}_{tf}$.

Using this framework Gaussian mixture classification was performed by Raj using band spectral energy and its derivatives as features [75]. Seltzer et al. further develop this idea by computing features designed to estimate reliability based on the characteristics of the speech signal [100]. These included: the *comb filter ratio* to compare energy in voiced regions to energy in non-harmonic regions, the *autocorrelation peak ratio* to measure signal periodicity, the *sub-band to full-band energy ratio* representing spectral shape, *sub-band energy to noise floor ratio* to estimate the noise floor energy, *kurtosis* used to measure signal Gaussianity, and *spectral valley flatness* to measure SNR. Using oracle masks, separate classifiers were trained for unvoiced and voiced speech types within each channel since the lack of harmonicity in unvoiced regions prevents the use of the pitch based features (see Fig. 25). In evaluations for robust speech recognition classifier estimation outperforms traditional spectral subtraction estimators in all noise conditions, but particularly for the non-stationary cases [100], [101].

An extension to this work proposes the incorporation of spectral variation across time and frequency

to improve upon simple white noise classifier training [102]. Since white noise only approximates the corruption induced by other noises if mask component estimates across sub-bands are independent, colored noise training is instead used by Kim et al. to give improved performance in non-stationary cases. In subsequent work, Kim and Stern attempt to alleviate the problem of data insufficiency for the frequency bands by the independent processing of each band, and also propose a more accurate voiced/unvoiced decision process [103].

Relevance vector machine (RVM) classification has been recently proposed to construct the reliability mask decisions through the direct modeling of the STFT coefficients [104]. Compared to the use of an SVM classifier, the RVM approach produces masks of similar accuracy with the advantage of a reduced computational complexity. Weiss and Ellis also compared the performance of the produced RVM masks to CASA pitch [105] based masking for speech recognition using reconstruction. They report that the RVM produces superior recognition rates due to its tendency to favor deletion errors over inclusion errors (Section IV-D further discusses mask error types).

4) Discussion

It is observed from past evaluations that both auditory and classifier-based techniques can produce superior recognition performance to simplistic SNR-based methods. The advantage of CASA type approaches to mask estimation is their ability to construct decisions based on the properties of the speech signal's spectra. This allows a more accurate identification of speech dominant spectrographic regions compared to SNR-based approaches which assume a generalization of noise characteristics observed in a small number of speech free frames. This ensures that CASA techniques will be preferable to SNR methods such as spectral subtraction in difficult environments.

In comparing classifier mask estimation methods to the SNR and perceptual techniques several comments can be made. Since classifier estimation attempts to model the properties of the speech, the quality of the reliability decisions produced do not depend on obtaining an accurate noise estimate. Similarly to CASA methods, classifier estimation are thus effective for compensating against stationary and non-stationary noises, where previous studies have shown that Bayesian approaches in particular can provide far higher recognition rates in music type disturbances compared to traditional SNR estimation [100].

A comparison between classifier techniques and the various CASA and perceptual approaches is difficult due to the diversity in the applications of the proposed algorithms, and the lack of a clear evaluation procedure within missing data based speech processing research [106]. However, with generality it can be stated that a disadvantage of the classifier based approach is its weakness to noises which share similar spectral characteristics to those of the speech signal. In this case the reliability features would be unable to distinguish speech and noise dominant time frequency components causing poor reliability mask accuracy. CASA approaches may be able to utilize other information such as spatial locality in order to assist in the reliability labeling, and should thus be more effective for speech shaped noise compensation or co-channel identification tasks.

D. Mask Errors and Model-Based Processing

The weakness of traditional approaches to missing data is their reliance on an accurate estimation of the reliability mask. In practice mask estimation methods which utilize the properties of the speech or noise signals (termed *bottom-up* methods) often contain errors. In the case of binary masking there are two distinct types of error: the inclusion of truly unreliable time-frequency points ('inclusion' errors), and the deletion of truly reliable time-frequency points ('deletion' errors). For standard 'bottom-up only' missing data the recognizer has no protection against these errors, particularly in the case of inclusion corruption which typically reduces the (log-) likelihood score of the true model. As a result recognition rates obtained from practically estimated masks are significantly lower than those obtained using ideal masks (compare the oracle and estimated masks in Fig. 26).

Recent research has attempted to increase the accuracy of the reliability decisions by utilizing information within the trained acoustic models. The active perception approach is an example of such a top-down robustness strategy. Based on the dominance of high intensity spectral components in human hearing, this method provides robustness for speech recognition using the feature energy of the trained clean models [107]. The approach exploits the observation that, for an additive environmental disturbance of a given energy, in the log-spectral domain time-frequency regions with high speech energy are less affected compared to regions of low speech energy. Consequently, for speech recognition, feature components with high model energy are assumed to correspond to a specific speech sound and are retained, while components with low model energy are assumed to be non-robust and are ignored. Utilizing missing data theory the observation is then divided into components 'mandatory' for recognition $\vec{x}_R$, and non-robust components $\vec{x}_N$, giving the posterior probability of state q_j as

$$P(q_j|\vec{x}) = \frac{p(\vec{x}_R(q_j)|q_j)p(q_j)}{p(\vec{x}_{R(q_j)})}, \tag{40}$$

where the partitioning of mandatory and non-robust features (the reliability partitioning) is state dependent. Cranen and de Veth evaluate this method on connected digit robust speech recognition, where the state dependent reliable subset is determined by choosing thresholds for which a fixed proportion of values observed in training fell below [107]. Although performance improved significantly in clean speech, in mismatched noise the performance benefit is substantially less. The major drawbacks of this approach are its assumption of high energy component reliability and the need to tune thresholds to form the state dependent reliable subset. In non-stationary mismatched conditions this means the reliable subset chosen may have a large number of errors compared to the ideal bottom-up decisions.

For speaker recognition, universal compensation [108] is a model based robustness technique which combines missing data theory and multicondition training to compensate against arbitrary noise types. The first step in the method is the construction of multicondition models for each speaker using clean training data and noisy training data obtained by corrupting the clean data with wide band flat spectrum noise at a range of SNRs. For an arbitrary unknown test utterance spectrum which experiences full-band corruption with respect to the clean model spectrum, only partial-band corruption may be experienced with respect to the multicondition model spectra. This full-band to partial-band conversion is achieved by performing a search over the feature space of each multicondition model to find components which best match the corresponding components in the test spectrum [108] (see Fig. 27). Using only the matched components a score is produced for each multicondition model, whose scores are then combined to give an overall recognition score.

Producing the matched feature subset for each speaker model is achieved using probabilistic union modeling [109], [110]. For a given test spectral vector $\vec{x} = (x_1, x_2, \ldots, x_D)$, let a matched component subset to speaker model λ at SNR level $l \in [1, 2, \ldots, L]$ be $\vec{x}^{(\lambda, l)}$. The probability of test observation $\vec{x}$ given model λ is

$$p(\vec{x}|\lambda) = \sum_{l=1}^{L} p(l|\lambda)p(\vec{x}^{(\lambda, l)}|\lambda, l), \tag{41}$$

where $p(l|\lambda)$ is the SNR level l prior for model λ and $p(\vec{x}^{(\lambda, l)}|\lambda, l)$ is the likelihood of the matched subset on the multicondition model (λ, l). This is a form of missing data since only the components of the matched subset $\vec{x}^{(\lambda, l)} \in \vec{x}$ are used to evaluate the likelihood for each model spectrum, where it is assumed that $\vec{x}^{(\lambda, l)}$ can be defined as the subset $\vec{x}_{\text{sub}}$ which maximizes $p(\vec{x}_{\text{sub}}|\lambda, l)$. Normalization is needed to compensate for different matched subset sizes and so the conditional probability must be replaced with the posterior defined as:

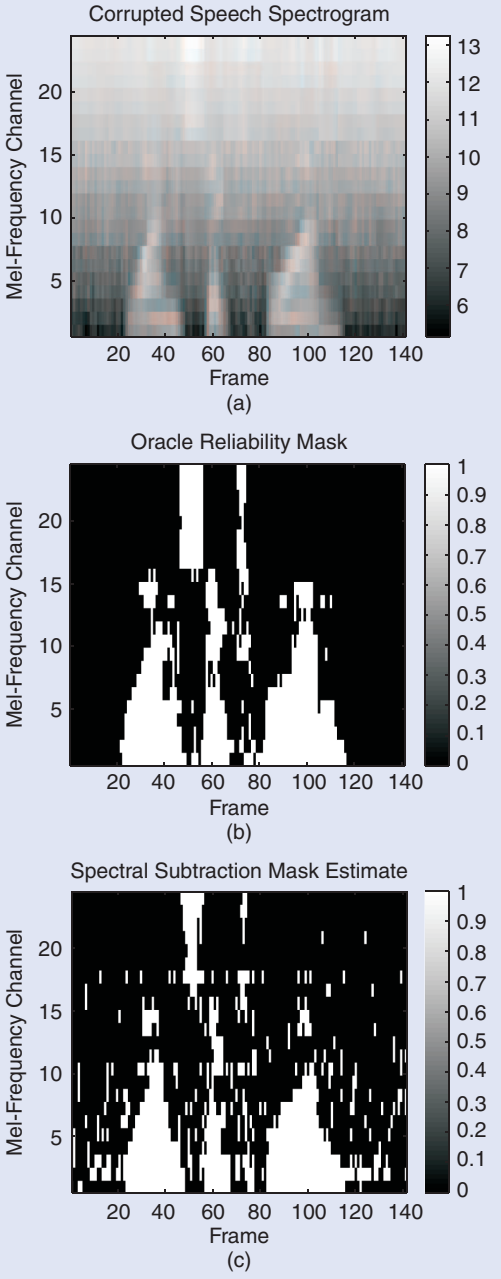

Figure 26. Binary time-frequency mask errors: a comparison between the oracle reliability mask (b) and a spectral subtraction (SS) estimated mask (c) for a TiDigits utterance corrupted with 10 dB white noise. Regions of inclusion and deletion error can be clearly observed within the SS mask, where a threshold of $\theta = 0$ dB is used to produce the estimate. (a) corrupted speach spectrogram; (b) oracle reliability mask; (c) spectral subtraction mask estimation.

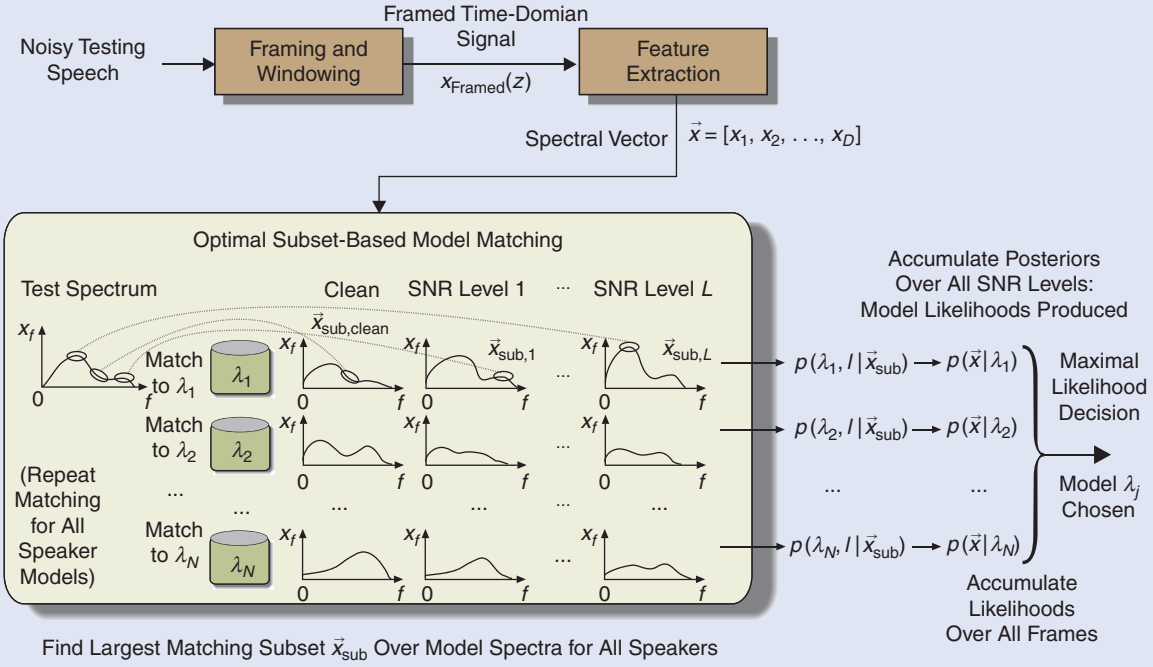

Figure 27. Universal Compensation: spectral feature component matching is performed on multicondition models of each speaker to produce an optimal reduced feature set which maximizes recognition for the given model.

$$p(\lambda, l|\vec{x}_{\text{sub}}) = \frac{p(\vec{x}_{\text{sub}}|\lambda, l)p(\lambda, l)}{\sum_{\lambda', l'} p(\vec{x}_{\text{sub}}|\lambda', l')p(\lambda', l')}. \quad (42)$$

It can be shown that the probability of the test observation given the speaker model $p(\vec{x}|\lambda)$ is proportional to the optimized posterior $p(\lambda, l|\vec{x}_{\text{sub}})$ over all SNR levels l. Since $p(\lambda, l|\vec{x}_{\text{sub}})$ favors large subset sizes, by finding the largest matched subset $\vec{x}_{\text{sub}}$ for each model (λ, l), the probability of $p(\vec{x}|\lambda)$ is also optimized [108]. Universal compensation has been extensively evaluated in the context of speaker identification [108], [111], and the results show that the method approximates multicondition model performance in known noise conditions, and surpasses the multicondition model in unknown conditions. Despite its ability to provide high robustness, the disadvantage of the approach is the required exhaustive search over the feature space of both the clean and multicondition models for each speaker.

E. Combining Information Sources

Although the use of purely top-down approaches provides an alternative to traditional bottom-up methods (which are vulnerable to estimation errors), constructing accurate reliability decisions using model information alone is typically computationally intensive, especially for high feature dimensionality. A solution is to combine bottom-up and top-down sources of information, with the goal of utilizing knowledge from the trained models to prevent recognizer exposure to errors within the bottom-up mask estimates. Two combined bottom-up top-down approaches have demonstrated success for missing data speech recognition: the mask model based approach, and the multisource decoder.

1) Mask Modeling for Missing Data Speech Recognition

The most intuitive and efficient way to perform top-down processing of estimated reliability masks is to directly utilize the parameters of the trained models. For speech recognition a mask modeling approach has been proposed, where a mask model is estimated for each HMM state and mixture in order to predict the expected reliability decisions in a given noise condition [112]. The method introduces a mask probability term $P(M|Q, W)$ to express the likelihood of a given mask M, corresponding to observation spectrogram X, given state sequence Q, and hypothesized word sequence W. By expressing the estimated word sequence as

$$\hat{W} = \underset{W}{\text{argmax}}\, P(X|M, Q, W)P(M|Q, W)P(Q|W)P(W), \quad (43)$$

where $P(Q|W)$ and $P(W)$ are the state transition and language model priors respectively, the ML decision is made by combining the observation likelihood $P(X|M, Q, W)$

with the mask probability $P(M|Q, W)$. The observation likelihood is computed via marginalization using the estimated bottom-up SNR reliability mask, while the mask probability is obtained by training mask models using oracle masks in multiple noise conditions. For mask vector $\vec{m}_t$, $P(\vec{m}_t|q, l)$ is estimated for HMM state q and mixture l by firstly calculating the posterior probability $P(l|\vec{x}_t, q)$ from the speech HMM parameters as in [112]. The mask model parameter for feature channel f in state q and mixture l is given by

$$\mu_{f,l,q} = \frac{\sum_{t:\vec{x}_t \in q} P(l|\vec{x}_t,q) \cdot m_{tf}}{\sum_{t:\vec{x}_t \in q} P(l|\vec{x}_t,q)}, \tag{44}$$

and the mask probability is the product of the individual feature parameter values within the model:

$$P(\vec{m}_t|l, q) = \prod_{f=1}^{D} \mu_{f,l,q}^{m_{tf}} \ (1 - \mu_{f,l,q})^{1-m_{tf}}. \tag{45}$$

Essentially the mask probability term uses top-down knowledge to provide protection against errors in the bottom-up mask which may otherwise produce a high observation likelihood on an incorrect word sequence. Experimental evaluations for connected digit recognition have shown that the use of mask probabilities provides reasonable improvements over standard bottom-up only missing data for a variety of noise conditions [113].

2) Multisource Decoding

While the mask modeling approach separates the top-down and bottom-up decision processes by utilizing 'static' information from the trained HMMs, multisource decoding [114] is a combined strategy which integrates the reliability mask estimation and missing data recognition stages. Specifically, the multisource decoder combines CASA principles and top-down use of the recognizer likelihoods in order to find accurate reliability labels in non-stationary noise conditions. Bottom-up processes are firstly employed to identify spectro-temporal regions which are likely to belong to the same source. Efficiency is ensured by dividing the time-frequency representation into fragments, each of which represents a region dominated by single source as determined by the bottom-up stage. For representations consisting of N fragments, a top-down search is performed to find the most likely speech model sequence for each of the possible 2^N labels. By performing a simultaneous search over the fragment labeling space using bottom-up processes and the word string space using the recognizer likelihoods, the multisource decoder outperforms standard bottom-up only missing data for speech recognition in non-stationary environments [114].

3) Discussion

Although developed for robust speech recognition, the mask modeling and multisource decoding techniques illustrate the potential of combining bottom-up and top-down sources of information to enhance the robustness of missing data systems while maintaining computational efficiency. These two methods considered exhibit two distinct combination strategies: the combination of the independently produced bottom-up estimates and static model parameter based top-down decisions within the classification scores, and the integration of the reliability labeling and missing data recognition stages via a top-down search which operates on a collection of time-frequency fragments proposed by a bottom-up stage. The advantage of the mask modeling approach is computational efficiency. By training mask models for each frequency channel using multicondition oracle reliability decisions, fast correction of errors within the bottom-up mask can be achieved. This is primarily due to the ability to perform all significant computation associated with the top-down processing offline during or following model training. The top-down search performed by the multisource decoder results in a potentially prohibitive amount of computation depending on the choice of fragment size. However, this may be reduced by through the use of dynamic programming to combine equivalent hypotheses at the conclusion processing for each fragment [115]. The search complexity thus becomes a function of the maximum number of simultaneous fragments, which remains constant over utterance length unlike the fragment count. It is expected that the multisource decoder is capable of producing more accurate results compared to approaches such as mask modeling, since the former uses top-down information specific to the utterance being processed while the latter relies on generic top-down information from a training or validation set. An additional advantage of the multisource approach is its conceptual extendibility to search across multiple simultaneous models, and this may permit the recognition of both voices in simultaneous speech [114].

Despite the robustness shown by both the mask modeling and multisource decoding methods for their intended speech recognition applications, the direct adaptation of these approaches for missing data speaker recognition is problematic. In its current form the mask modeling approach is not applicable to speaker recognition since the trained models capture only speaker discriminative information, as opposed to the speech information. The absence of temporal evolution in the information captured by the speaker GMMs would also have the effect of constraining any robustness information available from the models to

the spectral frequency domain. The absence of speech modeling also prevents the use of multisource decoding as a combined estimation and recognition strategy for speaker recognition, since the word sequence search to verify the bottom-up labeling cannot be performed. Designing a combined bottom-up top-down approach for missing data speaker recognition essentially requires the discriminative information within the speaker models to be related to the quality or 'correctness' of the speech reliability decisions. The lack of research focused on developing combined missing data approaches for speaker recognition is a result of the difficulty in achieving this requirement, and also perhaps of the lesser emphasis which speaker recognition specific applications have traditionally received in missing data research activities. However, based on the success of combined approaches for missing data speech recognition, the implementation of similar methods for speaker recognition should provide a significant improvement in robustness compared to existing bottom-up missing data techniques.

VIII. Summary

In the first part of this article a tutorial on closed-set, text-independent, speaker identification systems was presented. Speaker identification is basically a pattern recognition problem with the added complexity of dealing with time-series, non-stationary data. As a consequence vectors of features need to be derived from the acoustic data, with the MFCC concatenated with delta and acceleration temporal derivatives and subject to cepstral mean normalization (CMN) being the most popular. For identification of speakers, individual models of speakers need to be defined based on the feature vectors from utterances of that speaker. In speaker recognition the GMM has enjoyed wide success, especially when augmented by adaptation strategies such as in the GMM-UBM approach. With a GMM standard Maximum-Likelihood (ML) scoring can be used to identify the speaker of an unknown utterance. However the recent interest in discriminative based Support Vector Machines (SVM) approaches relying on optimal separation of classes has yielded successful identification systems especially in cases of data sparsity (limited amount of data). To date GMM and SVM based systems, their derivatives and hybrids represent the state of the art in speaker recognition technologies.

This part of the tutorial concluded with experimental evaluations in mismatched, limited training data and additive noise environments. In clean conditions high recognition rates, in excess of 95% on 64 speakers, were achieved. The GMM-UBM and GMM-SVM systems were found to be especially robust when confronted with limited training data highlighting the importance of utilizing some form of background model (the UBM). With mismatched channels the importance of CMN was underscored by an 18% reduction in performance when CMN was not adopted. Although mismatched conditions can be handled by some form of normalization (e.g. CMN) the performance in the presence of additive noise produced significant degradation in recognition, dropping by as much as 20% even when only relatively mild noise (30 dB SNR for the white noise case) impacted on the environment. Evidently robustness, especially against additive noise, is an important issue for speaker recognition systems.

In the latter part of this article we have provided a review of robustness techniques for speaker recognition. An overview of previously proposed methods for noise compensation in speech processing was firstly presented. These techniques are primarily suited to compensating against channel effects, and rely on assumptions such as stationarity or knowledge of the noise characteristics in order to effectively compensate against environmental disturbances. To perform speaker recognition in commercial applications improved robustness against environmental distortions, and in particular against non-stationary and transient effects, is required. Missing data recognition is naturally suited to this problem due to its ability to compensate for arbitrary unknown environmental effects within speech signals. These approaches are critically dependent on the accuracy with which individual time-frequency regions within a speech signal can be identified as speech or noise dominant. In practical situations the absence of a priori noise knowledge requires an estimation of these reliability decisions in the form of a reliability mask. Past approaches to reliability mask estimation for speaker recognition were then reviewed including SNR-based methods, auditory and perceptual criteria and classification-based techniques. While SNR-based methods offer simplicity by attempting a direct estimation of the noise spectra from speech free regions, their performance is generally inferior to auditory and classifier techniques, which identify regions of speech dominance by utilizing perceptual cues (such as locality or pitch information) and specifically designed features respectively.

Regardless of the technique used to estimate the reliability decisions, in difficult noise conditions these bottom-up estimation methods will produce reliability mask errors which adversely affect recognition performance. Top-down approaches are thus examined as a solution to address the vulnerability of traditional missing data systems to mask errors. Firstly, pure top-down methods are reviewed which utilize model information to

construct the reliability mask. Finally, motivated by the need for a more efficient recognition process than top-down estimation alone may provide, recent missing data approaches which combine bottom-up and top-down sources of information are reviewed. Following a discussion of these combined methods and their significance in the context of speaker recognition, we conclude that combining information sources provides the next logical step in improving the robustness of GMM based missing data speaker recognition systems.

Roberto Togneri (M'89-SM'04) received the B.E. degree in 1985, and the Ph.D degree in 1989 both from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now currently an Associate Professor. Dr. Togneri is a member of the Signals and Systems Engineering Research Group and heads the Signal and Information Processing Lab. His research activities include signal processing and robust feature extraction of speech signals, statistical and neural network models for speech and speaker recognition, and related aspects of communications, information retrieval, and pattern recognition. He has published over 80 refereed journal and conference papers in the areas of spoken language and information systems, was co-author of the book *Fundamentals of Information Theory and Coding Design*, Chapman & Hall/CRC, 2003 and is the chief investigator on two Australian Research Council Discovery Project research grants from 2010 to 2013.

Daniel Pullella received the B.E. and B.Comp.Sci degrees from the University of Western Australia, Crawley, in 2006.

From 2005 to 2006, he was a Research Intern at the Signal and Information Processing Lab, University of Western Australia. In 2007, he joined the Signals and Systems Engineering Research Group (SSERG), University of Western Australia, where he has just completed his Ph.D. on robust speaker identification using time-frequency masking. His research interests are in the areas of pattern recognition and signal processing, with specific applications to robust speaker recognition, noise compensation for speech processing, and robust speech feature extraction.

References

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Systems Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.

[2] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2002, vol. 4, pp. 4072–4075.

[3] J. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE, vol.* 85, no. 9, pp. 1437–1462, 1997.

[4] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Mag.*, vol. 11, no. 4, pp. 18–32, 1994.

[5] S. Furui, "Recent advances in speaker recognition," *Pattern Recognit. Lett.*, vol. 18, no. 9, pp. 859–872, 1997.

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

[7] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 4, pp. 430–451, 2004.

[8] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition: A need for caution," *IEEE Signal Processing Mag.*, vol. 26, no. 2, pp. 95–103, 2009.

[9] "Special issue on speaker recognition," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 1–266, 2000.

[10] "Special section on speaker and language recognition," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 7, pp. 1951–2115, 2007.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 29, no. 2, pp. 254–272, 1981.

[12] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-plp speech analysis technique," *in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1992, vol. 1, pp. 121–124.

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop Speaker Recognition*, 2001, pp. 213–218.

[14] C. L. Nikias and J. M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Processing Mag.*, vol. 10, no. 3, pp. 10–37, 1993.

[15] S. Wenndt and S. Shamsunder, "Bispectrum features for robust speaker identification," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1997, vol. 2, pp. 1095–1098.

[16] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1992, vol. 1, pp. 233–236.

[17] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Mag.*, vol. 13, no. 5, 1996, pp. 58–71.

[18] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoustic. Soc. Amer.*, vol. 55, p. 1304, 1974.

[19] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, 1994.

[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[21] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 133–147, 1998.

[22] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[23] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "Hidden Markov model toolkit (htk) version 3.4 user's guide," 2002.

[24] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[25] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, 1995.

[26] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process.* Lett., vol. 13, no. 5, pp. 308–311, 2006.

[27] J. Bilmes. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden

Markov models," *Int. Comput. Sci. Inst.* [Online]. 4. Available: http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf
[28] O. Thyes, R. Kuhn, P. Nguyen, and J. Junqua, "Speaker identification and verification using eigenvoices," in *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, 2000, vol. 2, pp. 242–245.
[29] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, 2000.
[30] R. Zilca, "Text-independent speaker verification using utterance level scoring and covariance modeling," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 363–370, 2002.
[31] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.
[32] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in *Proc. IEEE Neural Networks Signal Process. Workshop*, 2000, vol. 2, pp. 775–784.
[33] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 210–229, 2006.
[34] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *Darpa Timit: Acoustic-Phonetic Continuous Speech Corps CD-ROM*, 1993.
[35] J. Campbell and D. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1999, vol. 2, pp. 829–832.
[36] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. (2005). SVM and kernel methods Matlab toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France [Online]. Available: http://asi.insa-rouen.fr/enseignants/arakotom/toolbox/
[37] H. Hirsch. (2005). FaNT—Filtering and noise adding tool [Online]. Available: http://dnt.kr.hsnr.de/download.html
[38] I. R. G.712. (1996). Transmission performance characteristics of pulse code modulation channels [Online]. Available: http://www.itu.int/rec/T-REC-G.712-199611-S/en
[39] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1995, vol. 1, pp. 329–332.
[40] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, 1976.
[41] A. Acero and X. Huang, "Augmented cepstral normalization for robust speech recognition," in *Proc. IEEE Automatic Speech Recognition Workshop,* Snow Bird, *UT*, 1995, pp. 146–147.
[42] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proc. ARPA Human Language Technology Workshop*, 1993, pp. 69–74.
[43] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, and G. Tong, "Integrating RASTA-plp into speech recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1994, vol. 1, pp. 421–424.
[44] J. Markel and A. Gray, *Linear Prediction of Speech. Secaucus,* NJ: Springer-Verlag, New York, 1982.
[45] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 2, pp. 117–125, 1995.
[46] K. T. Assaleh and R. J. Mammone, "New lp-derived features for speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 630–638, 1994.
[47] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, 1998, pp. 260–267.
[48] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP),* 2003, vol. 2, pp. 53–56.
[49] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Magnitude-only estimation of handset nonlinearity with application to speaker recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1998, vol. 2, pp. 745–748.
[50] D. A. Reynolds, "Htimit and llhdb: Speech corpora for the study of handset transducer effects," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, 1997, vol. 2, pp. 1535–1538.
[51] L. P. Heck, Y. Konig, M. K. Snmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Commun.*, vol. 31, no. 2–3, pp. 181–192, 2000.
[52] V. Chandran, D. Ning, and S. Sridharan, "Speaker identification using higher order spectral phase features and their effectiveness vis-a-vis mel-cepstral features," in *Biometric Authentication.* Berlin: Springer-Verlag, 2004, vol. 3072, pp. 1–20.
[53] D. Ning and V. Chandran, "The effectiveness of higher order spectral phase features in speaker identification," in *Proc. ODYSSEY 2004—The Speaker and Language Recognition Workshop,* Toledo, Spain, pp. 245–250.
[54] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. European Conf. Speech Communication Technology (Eurospeech),* Rhodes, Greece, 1997, pp. 963–966.
[55] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 42–54, 2000.
[56] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2003, vol. 2, pp. 49–52.
[57] K. P. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Philadelphia, PA, 1996, pp. 1764–1767.
[58] Z. Rong, Z. Shuwu, and X. Bo, "Text-independent speaker identification using gmm-ubm and frame level likelihood normalization," in *Proc. Int. Symp. Chinese Spoken Language Process.*, 2004, pp. 289–292.
[59] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP),* 1997, vol. 2, pp. 1071–1074.
[60] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP),* Beijing, China, 2000, pp. 495–498.
[61] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2004, vol. 1, pp. 37–40.
[62] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. ODYSSEY 2004—The Speaker and Language Recognition Workshop,* Toledo, Spain, pp. 219–226.
[63] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2005, vol. 1, pp. 637–640.
[64] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 554–568, 1999.
[65] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1990, pp. 845–848.
[66] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, 1996.
[67] W. Lit Ping and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2001, pp. 457–460.
[68] J. P. Campbell, "Testing with the yoho CD-ROM voice verification corpus," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1995, vol. 1, pp. 341–344.
[69] S. Ahmed and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann, 1993, pp. 393–400.
[70] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.
[71] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Commun.*, vol. 35, no. 3–4, pp. 141–177, 2001.
[72] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 101–116, 2005.

[73] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: Soft data modelling for noise robust asr," Stratford upon-Avon, U.K., Tech. Rep., 2001.
[74] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2001.
[75] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Pittsburgh, PA, Carnegie Mellon Univ., 2000.
[76] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of damaged spectrographic features for robust speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2000, vol. 1, pp. 357–360.
[77] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. European Conf. Speech Communication Technology (Eurospeech)*, 1999, pp. 2837–2840.
[78] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1997, vol. 2, pp. 863–866.
[79] A. C. Morris, M. P. Cooke, and P. D. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust asr," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1998, vol. 2, pp. 737–740.
[80] M. Kühne, D. Pullella, R. Togneri, and S. Nordholm, "Towards the use of full covariance models for missing data speaker recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, Las Vegas, Navada, 2008, pp. 4537–4540.
[81] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1994, pp. 1555–1558.
[82] B. Raj, M. L. Seltzer, and R. M. Stern, "Robust speech recognition: The case for restoring missing features," in *Proc. European Conf. Speech Communication Technology (Eurospeech)*, 2001.
[83] D. Pullella, M. Kuhne, and R. Togneri, "Sub-band partitioning for full covariance based missing data speaker recognition," *Int. J. Inform. Syst. Sci.*, (Advances in Information and Systems Sciences Series), vol. 3, no. 3–4, pp. 641–648, 2009.
[84] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.
[85] M. El-Maliki and A. Drygajlo, "Missing features detection and handling for robust speaker verification," in *Proc. European Conf. Speech Communication Technology (Eurospeech)*, Budapest, Hungary, 1999, pp. 975–978.
[86] M. T. Padilla, T. F. Quatieri, and D. A. Reynolds, "Missing feature theory with soft spectral subtraction for speaker verification," in *Proc. European Conf. Speech Communication Technology (Interspeech)*, 2006, vol. 1, pp. 913–916.
[87] R. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," in *Proc. European Conf. Speech Communication Technology (Eurospeech)*, Rhodes, Greece, 1997, pp. 37–40.
[88] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. Consistent and Reliable Acoustic Cues (CRAC) Workshop*, Aalborg, Denmark, 2001.
[89] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study," in *Proc. European Conf. Speech Communication Technology (Eurospeech)*, Budapest, Hungary, 1999, pp. 2407–2410.
[90] G. J. Brown and D. Wang, *Separation of Speech by Computational Auditory Scene Analysis* (ser. Speech Enhancement). New York: Springer-Verlag, 2005.
[91] P. D. Green, M. P. Cooke, and M. D. Crawford, "Auditory scene analysis and hidden Markov model recognition of speech in noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1995, vol. 1, pp. 401–404.
[92] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, 2001, pp. 213–216.
[93] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2006, vol. 1, pp. 645–648.
[94] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.
[95] J. P. Eatock and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 1994, vol. 1, pp. 133–136.
[96] P. Jančovič and M. Köküer, "Estimation of voicing-character of speech spectra based on spectral shape," *IEEE Signal Process. Lett.*, vol. 14, no. 1, pp. 66–69, 2007.
[97] P. Jančovič and M. Köküer, "Employment of voicing information of speech spectra for noise-robust speaker identification," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Poznan, Poland, 2007.
[98] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1–2, pp. 123–142, 2004.
[99] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, 2004.
[100] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, 2004.
[101] M. L. Seltzer, B. Raj, and R. M. Stern, "Classifier-based mask estimation for missing feature methods of robust speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Beijing, China, 2000, vol. 3, pp. 538–541.
[102] W. Kim, R. M. Stern, and H. Ko, "Environment-independent mask estimation for missing-feature reconstruction," in *Proc. European Conf. Speech Communication Technology (Interspeech)*, 2005, pp. 2637–2640.
[103] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, 2006, vol. 1, pp. 305–308.
[104] R. Weiss and D. Ellis, "Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking," in *Proc. Workshop Statistical Perceptual Audition (SAPA)*, 2006, pp. 31–36.
[105] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, 2003.
[106] C. Cerisara, S. Demange, and J. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, 2007.
[107] B. Cranen and J. de Veth, "Active perception: Using a priori knowledge from clean speech models to ignore non-target features," in *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, 2004.
[108] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions—A universal compensation approach," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, 2005, vol. 1, pp. 617–620.
[109] J. Ming, P. Jančovič, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 403–414, 2002.
[110] J. Ming, D. Stewart, P. Hanna, P. Corr, J. Smith, and S. Vaseghi, "Robust speaker identification using posterior union models," in *Proc. European Conf. Speech Communication Technology (Interspeech)*, Geneva, Switzerland, 2003, pp. 2645–2648.
[111] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 5, pp. 1711–1723, 2007.
[112] P. Jančovič and M. Köküer, "On the mask modeling and feature representation in the missing-feature asr: Evaluation on the consonant challenge," in *Proc. European Conf. Speech Communication Technology (Interspeech)*, Brisbane, Australia, 2008, pp. 1777–1780.
[113] M. Köküer and P. Jančovič, "Incorporating mask modelling for noise-robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 3929–3932.
[114] J. Barker, M. Cooke, and D. Ellis, "Integrating bottom-up and top-down constraints to achieve robust asr: The multisource decoder," in *Proc. Consistent and Reliable Acoustic Cues (CRAC) Workshop*, Aalborg, Denmark, 2001, pp. 63–66.
[115] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sound sources," in *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, 2000, pp. 270–273.