Generalized Mutual Interdependence Analysis of Noisy Channels

Heiko Claussen, Justinian Rosca, Viswanathan Ramasubramanian, and Subramani Thiyagarajan

Abstract The main motivation for our present work is to reliably perform voice (or signal) detection for a source of interest from a single microphone recording. We rely on the assumption that the input signal contains invariant information about the channel, or transfer function from each source to the microphone, which could be reliably exploited for signal detection and classification. In this chapter we employ a nonconventional method called generalized mutual interdependence analysis (GMIA) that proposes a model for the computation of this hidden invariant information present across multiple measurements. Such information turns out to be a good characteristic feature of a signal source, transformation, or composition that fits the model. This chapter introduces a unitary and succinct description of the underlying model of GMIA, and the formulation and solution of the corresponding optimization problem. We apply GMIA for feature extraction in the problem of own-voice activity detection, which aims at classification of a near-field channel based on access to prior information about GMIA features of the channel. It is extremely challenging to recognize the presence of voice in noisy scenarios with interference from music, car noise, or street noise. We compare GMIA with MFCC and cepstral-mean features. For example, GMIA performs with equal error rates below 10 % for music interference of SNRs down to -20 dB.

Keywords Acoustic signal classification • Voice detection • Feature extraction • Near-field channels • Speaker verification • Head related transfer functions • Mutual interdependence analysis • Hearing aids

H. Claussen • J. Rosca (✉)
Siemens Corporation, Corporate Research, 755 College Road East, Princeton, NJ 08540, USA
e-mail: heiko.claussen@siemens.com; justinian.rosca@siemens.com

V. Ramasubramanian • S. Thiyagarajan
Siemens Corporate Research and Technologies-India, Bangalore, India
e-mail: V.Ramasubramanian@siemens.com; thiyagarajan.s@siemens.com

T.D. Andrews et al. (eds.), *Excursions in Harmonic Analysis, Volume 1,*
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-0-8176-8376-4_18,

1 Introduction

Our goal is to compute a simplified statistical data representation that retains invariant information that is necessary for subsequent tasks such as classification or prediction. Methods such as Fisher's linear discriminant analysis (FLDA) [10], canonical correlation analysis (CCA) [14], or ridge regression [23] extract "optimal" representations of a dataset. For instance, FLDA defines a projection space that maximizes the ratio of the between- and within-class scatter of the training data to reduce the dimensionality of the input. CCA assumes one common source in two datasets. The dimensionality of the data is reduced by retaining the space that is spanned by pairs of projecting directions in which the datasets are maximally correlated. In contrast, ridge regression finds a linear combination of the inputs that best fits a desired response. In this chapter, we review an alternative second-order statistical criterion to find an "optimal" dataset representation, called GMIA. We aim to define an invariant computation or feature of high dimensional instances of a single class, which does not change within its class, where the number of input instances N is smaller than their dimensionality D.

We further consider the application of GMIA to the system identification problem of an acoustical channel, as follows. Multiple people (representing the multiple inputs of a linear acoustic system) could be engaged in conversational speech. Audio could be captured using multiple microphones, which are the system outputs available for identification of the linear time invariant system representing the channels. Each transfer function input to output can be modeled as an FIR filter, and the system can be modeled as a MIMO FIR acoustic system. Such a scenario, encountered not just in acoustics but also in communications and other areas, is conventionally addressed by blind source separation (for source estimation) and blind channel identification techniques (for channel identification).

In this section we are interested in one sensor only, and we aim to exploit partial additional information about the channel or source in order to recognize if a particular channel, and consequently its source, is active. For example, practical problems abstracted by this scenario are the own-voice activity detection (OVAD) for hearing aids and headsets. The channel of interest corresponds to the invariant channel of the owner's voice to a single microphone. Detecting when the owner's voice is active, in contrast to external active speakers or noises, is of importance for automatic processing (e.g., in the hearing aid). We are interested in a semi-blind solution to OVAD, which exploits training information about the owner's channel (and possibly the owner's voice) to assess if the currently identified active channel fits the owner in contrast to external sources of sound.

Methods to blindly or semi-blindly identify the channel include second order and higher-order statistical approaches. The latter require large amounts of data to achieve good recognition performance, while second-order methods promise speed and efficiency. We will apply GMIA, a second-order method, to effectively capture the invariant own-voice channel information in noisy scenarios. Other applications, in addition to OVAD for hearing aids and headsets, are the detection of the owner's

voice in videoconferencing, the detection and tracking of slowly varying dynamic speech channels in interactive speech gaming, or the detection of active speech channels in hands free communication. All could exploit a GMIA-based approach to the corresponding single-input single-output (SISO) problem to address more complex MIMO channel detection solutions.

The outline of this chapter is as follows. In Sect. 2 we discuss the importance of voice detection applications and present related work. Section 3 revisits the generalized mutual interdependence analysis (GMIA) method [4–7]. In Sect. 4 we bring in a generative model for GMIA(λ) parameterized by λ and demonstrate the effect of noise on the extracted features. Section 5 analyzes the applicability of GMIA for channel extraction and classification from monaural speech. In Sect. 6 we evaluate the performance of GMIA for OVAD and compare these results with mel-frequency cepstral coefficients (MFCC) and cepstral-mean (CM)-based approaches. We draw conclusions in Sect. 7.

2 Motivation and Related Work

Signal detection in continuous or discrete time is a cornerstone problem in signal processing. One particularly well-studied instance in speech and acoustic processing is voice detection, which subsumes a solution to the problem of distinguishing the most likely hypothesis between one assuming speech presence and a second assuming the presence of noise. Furthermore, when multiple people are speaking, it is difficult to determine if the captured audio signal is from a speaker of interest or from other people. Speech coding, speech/signal processing in noisy conditions, and speech recognition are important applications where a good voice/signal detection algorithm can substantially increase the performance of the respective system.

Traditionally, voice detection approaches used energy criteria such as short-time SNR estimation based on long-term noise estimation [22], likelihood ratio test of the signal and exploiting a statistical model of the signal [3], or attempted to extract robust features (e.g., the presence of a pitch [9], the formant shape [15], or the cepstrum [13]) and compare them to a speech model. Diffuse, nonstationary noise, with a time-varying spectral coherence, plus the presence of a superposition of spatially localized but simultaneous sources make this problem extremely challenging when using a single sensor (microphone).

Not surprisingly, during the last decade, researchers have focused on multimodality sensing to make this problem tractable. Multiple channel voice detection algorithms take advantage of the extra information provided by additional sensors. For example, [21] blindly identify the mixing model and estimates a signal with maximal signal-to-interference-ratio (SIR) obtainable through linear filtering. Although the filtered signal contains large artifacts and is unsuitable for signal estimation it was proven ideal for signal detection. Another example, is the WITTY (Who is Talking to You) project from Microsoft [24], which deals with the voice detection problem by means of integrated heterogeneous sensors

(e.g., a combination of a close-talk microphone and a bone-conductive microphone). Even further, multimodal systems using both microphones and cameras have been studied [17].

The main motivation for our present work is to perform voice (or signal) detection for the source of interest with the reliability of multimodal approaches such as WITTY but in the absence of additional sensors such as a bone-conducting microphone. We will demonstrate that a single microphone signal contains invariant information about what may be the channel, or transfer function from each source to the microphone, which could be reliably exploited for signal detection and classification (e.g., OVAD). We use GMIA [6] to extract this invariant information for both reference (training) and testing, and further to compare classification performance on the OVAD problem to MFCC and CM-based approaches.

Mutual interdependence analysis (MIA) was first introduced by Claussen et al. [4] to extract a representation, also called common or mutual component, which is equally correlated with all the inputs. After successfully applying MIA to text-independent speaker verification and illumination-robust face recognition [5], the method was generalized to GMIA [6] to account for different noise levels and to relax the requirement for equal correlation of the common component with each input. A conclusive up-to-date statement of GMIA is presented in [7]. In the next section we review GMIA and some of its properties.

3 Generalized Mutual Interdependence Analysis

In the following let $\mathbf{x}_i \in \mathbb{R}^D$ denote the ith input vector $i = 1 \ldots N$ and a column of the input matrix $\mathbf{X}$. Moreover, $\mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$, $\underline{\mathbf{1}}$ is a vector of ones and $\mathbf{I}$ represents the identity matrix.

Extracting a common component $\mathbf{s} \in \mathbb{R}^D$ in the inputs $\mathbf{X}$ can be defined as finding a direction in $\mathbb{R}^D$ that is equally correlated with the inputs. That is:

$$\zeta \underline{\mathbf{1}} = \mathbf{X}^T \cdot \mathbf{s} \quad \text{where } \zeta \text{is a constant.} \tag{1}$$

This is an underdetermined problem if $D \geq N$. MIA finds an estimate of $\mathbf{s}$, i.e., a direction denoted by $\mathbf{w}_{\text{MIA}} \in \mathbb{R}^D$ that minimizes the projection scatter of the inputs $\mathbf{x}_i$, under the linearity constraint to be in the span of $\mathbf{X}$. That is, $\mathbf{w} = \mathbf{X} \cdot \mathbf{c}$. Generally, MIA is used to extract a common component from high-dimensional data $D \geq N$. Its cost function is given as:

$$\mathbf{w}_{\text{MIA}} = \underset{\mathbf{w}, \mathbf{w}=\mathbf{X}\cdot\mathbf{c}}{\arg\min} \left(\mathbf{w}^T \cdot \left(\mathbf{X} - \mu \cdot \underline{\mathbf{1}}^T\right) \cdot \left(\mathbf{X} - \mu \cdot \underline{\mathbf{1}}^T\right)^T \cdot \mathbf{w} \right). \tag{2}$$

By solving Eq. (2) in the span of the original inputs rather than mean subtracted inputs, a closed-form solution can be found [4]:

$$\mathbf{w}_{\text{MIA}} = \zeta \mathbf{X} \cdot \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \underline{\mathbf{1}}. \tag{3}$$

The properties of MIA are captured in the following theorems:

Theorem 1. *The minimum of the criterion in Eq. (2) is zero if the inputs* $\mathbf{x}_i$ *are linearly independent.*

If inputs are linearly independent and span a space of dimensionality $N \leq D$, then the subspace of the mean subtracted inputs in Eq. (2) has dimensionality $N-1$. There exists an additional dimension in $\mathbb{R}^N$, orthogonal to this subspace. Thus, the scatter of the mean subtracted inputs can be made zero. The existence of a solution where the criterion in Eq. (2) becomes zero is indicative of an invariance property of the data.

Theorem 2. *The solution of Eq. (2) is unique (up to scaling) if the inputs* $\mathbf{x}_i$ *are linearly independent.*

This is shown by the existence of the closed-form solution in Eq. (3). However, it is important to note that, if $\mathbf{w}$ is not constrained to the span of the inputs, any combination $\hat{\mathbf{w}}_{\text{MIA}} + \mathbf{b}$ with $\mathbf{b}$ in the nullspace of $\mathbf{X}$ is also a solution. Also, the MIA problem has no defined solution if the inputs are zero mean, that is, if $\mathbf{X} \cdot \underline{\mathbf{1}} = \underline{\mathbf{0}}$. The reason is that there exists $\mathbf{w} = \underline{\mathbf{0}}$ in the span of the inputs as a trivial solution to Eq. (2).

The MIA data model in Eq. (1) is extended in [6] to incorporate measurement noise $\mathbf{n} \sim \mathcal{N}(\underline{\mathbf{0}}, \mathbf{C_n})$ and to relax the equal correlation constraint from $\zeta\underline{\mathbf{1}}$ to $\mathbf{r}$:

$$\mathbf{r} = \mathbf{X}^T \cdot \mathbf{w} + \mathbf{n}. \tag{4}$$

We assume $\mathbf{w}$ to be a random variable. Our goal is to estimate $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{w}, \mathbf{C_w})$ assuming that $\mathbf{w}$ and $\mathbf{n}$ are statistically independent. Given the model in Eq. (4), the generalized MIA criterion (GMIA) is defined as:

$$\mathbf{w}_{\text{GMIA}} = \mu_\mathbf{w} + \mathbf{C_w} \cdot \mathbf{X} \cdot \left(\mathbf{X}^T \cdot \mathbf{C_w} \cdot \mathbf{X} + \mathbf{C_n}\right)^{-1} \cdot \left(\mathbf{r} - \mathbf{X}^T \cdot \mu_\mathbf{w}\right) \tag{5}$$

$$= \mu_\mathbf{w} + \left(\mathbf{X} \cdot \mathbf{C_n^{-1}} \cdot \mathbf{X}^T + \mathbf{C_w^{-1}}\right)^{-1} \cdot \mathbf{X} \cdot \mathbf{C_n^{-1}} \cdot \left(\mathbf{r} - \mathbf{X}^T \cdot \mu_\mathbf{w}\right). \tag{6}$$

Throughout the remainder of the document, the GMIA parameters are $\mathbf{C_w} = \mathbf{I}$, $\mathbf{C_n} = \lambda\,\mathbf{I}$, $\mathbf{r} = \zeta\underline{\mathbf{1}}$ and $\mu_\mathbf{w} = \underline{\mathbf{0}}$. We refer to this parameterization by

$$\text{GMIA}(\lambda) = \zeta\mathbf{X} \cdot \left(\mathbf{X}^T \cdot \mathbf{X} + \lambda\,\mathbf{I}\right)^{-1} \cdot \underline{\mathbf{1}}. \tag{7}$$

When $\lambda \to \infty$, the GMIA solution represents the mean of the inputs. Indeed, the inverse $\left(\mathbf{X}^T \cdot \mathbf{X} + \lambda\,\mathbf{I}\right)^{-1} \to \frac{1}{\lambda}\mathbf{I}$ simplifying the solution to $\mathbf{w}_{\text{GMIA}} \to \frac{\zeta}{\lambda}\mathbf{X} \cdot \underline{\mathbf{1}}$. Furthermore, MIA [solution to Eq. (3)] is equivalent to GMIA(λ) when $\lambda = 0$. In the rest of this chapter, we denote MIA by GMIA(0) to emphasize their common theoretical foundation.

4 Generative Signal Model for GMIA

This section evaluates the behavior of GMIA(λ) for different types and intensities of additive distortions. In particular, we evaluate the effect of noise components that are either recurring uncorrelated components or Gaussian noise. We use the generative signal model in [7] to generate synthetic data with various properties. In contrast to published work we show a gradual change in the intensities of the different noise types and compare the feature extraction result to the true feature desired. This allows an interpretation of GMIA(λ) and analysis of its performance on data with unknown noise conditions from the field.

Assume the following generative model for input data $\mathbf{x}$:

$$\begin{aligned} \mathbf{x}_1 &= \alpha_1 \mathbf{s} + \mathbf{f}_1 + \mathbf{n}_1 \\ \mathbf{x}_2 &= \alpha_2 \mathbf{s} + \mathbf{f}_2 + \mathbf{n}_2 \\ &\vdots \\ \mathbf{x}_N &= \alpha_N \mathbf{s} + \mathbf{f}_N + \mathbf{n}_N, \end{aligned} \tag{8}$$

where $\mathbf{s}$ is a common, invariant component or feature we aim to extract from the inputs, α_i, $i = 1,\ldots,N$ are scalars (typically all close to 1), $\mathbf{f}_i$, $i = 1,\ldots,N$ are combinations of basis functions from a given orthogonal dictionary such that any two are orthogonal, and $\mathbf{n}_i$, $i = 1,\ldots,N$ are Gaussian noises. We will show that GMIA estimates the invariant component $\mathbf{s}$, inherent in the inputs $\mathbf{x}$.

Let us make this model precise. As before, D and N denote the dimensionality and the number of observations. Additionally, K is the size of a dictionary $\mathbf{B}$ of orthogonal basis functions. Let $\mathbf{B} = [\mathbf{b}_1,\ldots,\mathbf{b}_K]$ with $\mathbf{b}_k \in \mathbb{R}^D$. Each basis vector $\mathbf{b}_k$ is generated as a weighted mixture of maximally J elements of the Fourier basis which are not reused ensuring orthogonality of $\mathbf{B}$. The actual number of mixed elements is chosen uniformly at random, $J_k \in \mathbb{N}$ and $J_k \sim \mathscr{U}(1,J)$. For $\mathbf{b}_k$, the weights of each Fourier basis element i are given by $w_{jk} \sim \mathscr{N}(0,1)$, $j = 1,\ldots,J_k$. For $i = 1,\ldots,D$ (analogous to a time dimension) the basis functions are generated as:

$$b_k(i) = \frac{\sum_{j=1}^{J_k} w_{jk} \sin\left(\frac{2\pi i\, \alpha_{jk}}{D} + \beta_{jk}\frac{\pi}{2}\right)}{\sqrt{\frac{D}{2}\sum_{j=1}^{J_k} w_{jk}^2}}$$

with

$$\alpha_{jk} \in \left[1,\ldots,\frac{D}{2}\right]; \beta_{jk} \in [0,1]; \left[\alpha_{jk},\beta_{jk}\right] \neq \left[\alpha_{lp},\beta_{lp}\right] \forall j \neq l \text{ or } k \neq p.$$

In the following, one of the basis functions $\mathbf{b}_k$ is randomly selected to be the common component $\mathbf{s} \in [\mathbf{b}_1,\ldots,\mathbf{b}_K]$. The common component is excluded from the basis used to generate uncorrelated additive functions $\mathbf{f}_n$, $n = 1,\ldots,N$. Thus only $K-1$ basis functions can be combined to generate the additive functions $\mathbf{f}_n \in \mathbb{R}^D$. The actual number of basis functions J_n is randomly chosen, similarly to J_k, with $J = K-1$. The randomly correlated additive components are given by:

$$f_n(i) = \frac{\sum_{j=1}^{J_n} w_{jn} c_{jn}(i)}{\sqrt{\sum_{j=1}^{J_n} w_{jn}^2}}$$

with

$$\mathbf{c}_{jn} \in [\mathbf{b}_1, \ldots, \mathbf{b}_K]\,; \mathbf{c}_{jn} \neq \mathbf{s}, \forall j, n; \mathbf{c}_{jn} \neq \mathbf{c}_{lp}, \forall j \neq l \text{ and } n = p.$$

Note that $\|\mathbf{s}\| = \|\mathbf{f}_n\| = \|\mathbf{n}_n\| = 1, \forall n = 1, \ldots, N$. To control the mean and variance of the norms of common, additive, and noise components in the inputs, each component is multiplied by the random variable $a_1 \sim \mathcal{N}(m_1, \sigma_1^2)$, $a_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ and $a_3 \sim \mathcal{N}(m_3, \sigma_3^2)$, respectively. Finally, the synthetic inputs are generated as:

$$\mathbf{x}_n = a_1 \mathbf{s} + a_2 \mathbf{f}_n + a_3 \mathbf{n}_n \tag{9}$$

with $\sum_{i=1}^{D} x_n(i) \approx 0$. The parameters of the artificial data generation model are chosen as $D = 1{,}000$, $K = 10$, $J = 10$, and $N = 20$.

Throughout the experiments we keep the parameters $m_1 = 1$, $\sigma_1 = 0.1$, $\sigma_2 = 0.1$ and $m_3 = 0$ of the distributions for a_1, a_2, and a_3 constant. We vary the mean amplitude m_2 of the recurring uncorrelated components and the variance σ_3 of the Gaussian noise and illustrate its effect on GMIA(0), GMIA(λ), and the sample mean in Fig. 1. The figure shows a matrix of 3D histograms for different parameters m_2 and σ_3. Each point in a histogram represents an experiment for a given value of λ (x-axis). The y-axis indicates the correlation of the GMIA solution with $\mathbf{s}$, the true common component. The intensity (z-axis) of the point represents the number of experiments, in a series of random experiments, where we obtain this specific correlation value for the given λ. Overall, we performed $1{,}000$ random experiments with randomly generated inputs using various values of λ per histogram.

Results show that a change in the mean amplitude m_2 of the recurring uncorrelated components $\mathbf{f}_i$ has a minimal effect on GMIA(0) but greatly affects the correlation coefficient of $\mathbf{s}$ with the sample mean. That is, the sample mean results is a good representation of $\mathbf{s}$ only if m_2 is low and the common component $\mathbf{s}$ is dominant in the data. Moreover, this indicates that GMIA(0) succeeds in finding a good representation of $\mathbf{s}$.

The second row of Fig. 1 shows that an increased variance σ_3 of the noise can improve the GMIA(0) result. The increased noise level appears to acts as a regularization in the matrix inversion when computing GMIA. This has the same effect as an increased value of the regularization parameter λ.

Moreover, the experiments show that the results for all λ suffer for high noise variances σ_3, but that the spectral mean is affected the most. In all experiments, GMIA(λ) performs equally or outperforms GMIA(0) and the spectral mean. This demonstrates that GMIA is more versatile than the spectral mean in extracting a common component from data with an unknown and possibly varying distortion. In the following section we evaluate how the extraction results are affected for nonstationary, real-world data such as speech.

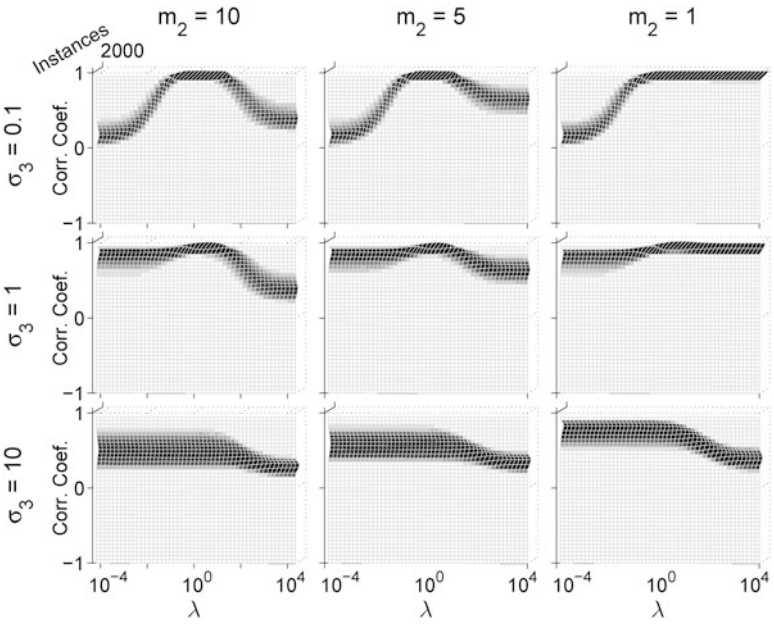

Fig. 1 Histograms of GMIA(λ) extraction performance for different levels of additive Gaussian noise and uncorrelated additive components $\mathbf{f}_i$. The mean of the inputs extracts the common component $\mathbf{s}$ well for low energy contributions of $\mathbf{f}_i$. Small levels of Gaussian noise result in a drop of the GMIA(0) performance. Larger amount of Gaussian noise results first in an improved GMIA(0) performance and later in a reduced extraction result overall λ. High levels of noise are better addressed by GMIA(0) then the mean

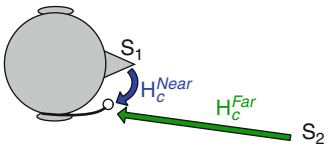

Fig. 2 Own-voice activity detection (OVAD) scenario: person is wearing a headset; speech from near-field (own), S_1, or far-field (external sources), S_2, is recorded on a single nearby microphone. The signal incorporates channel information, e.g., $\mathbf{H}_c^{\text{Near}}$ or $\mathbf{H}_c^{\text{Far}}$, respectively

5 Channel Extraction from Mono Speech Recordings

Lets us consider a single microphone recording of near-field and far-field nonoverlapping conversational speech as in Fig. 2. As noted in Sect. 2, a potential application of GMIA is to extract channel features in the context of owner speech detection for hearing aids. This problem has been referred to as OVAD, to imply the recognition of when the wearer of the hearing-aid (owner) is talking and when an external speaker is talking in a conversation (between the owner and the external speaker).

Such a detection facilitates, e.g., the hearing aid signal processing to be adapted dynamically to own-voice (OV) or external-speaker (EXT) characteristics.

We aim for an understanding of the domain and timescales where real-world acoustic data (e.g., conversational speech) fits the generative model studied in Eq. (9). As a first step, in this section, we review the model for the recorded signal and its dependence on speaker and channel characteristics. We use data from one or more speakers for fixed positions (i.e., exhibiting common channels), as in Fig. 2, to extract channel information using GMIA. Later, in Sect. 6, we address the OVAD problem.

5.1 Speech and Channel Models

A speech signal can be modeled as an excitation that is convolved with a linear dynamic filter, which represents the channel including the microphone characteristic, the channel impulse response of the environment, and the vocal tract. The excitation signal can be modeled for voiced speech as a periodic signal and for unvoiced speech as random noise [8, p. 50]. Let $\mathbf{E}^{(p)}$, $\mathbf{H}_v^{(p)}$, $\mathbf{H}_c$, and $\mathbf{S}^{(p)}$ be the spectral representations of the excitation or pitch signal (covering the lungs and vocal chords), the vocal tract filter (covering the mouth, tongue, teeth, lips, and nasal cavity), the external channel impulse response, and the speech signal parts of person p, respectively. Note that the channel impulse response implicitly depends on the spatial location of the receiver. This can vary substantially from near-field to far-field, or even over different far-field only or near-field only locations. If the environment of the speaker is invariant (e.g., the speaker does not move significantly) and we make simplifying assumptions to idealize the spectrum and capture important features at the timescale of interest, assume the data can be modeled as: $\mathbf{S}^{(p)} = \mathbf{E}^{(p)} \cdot \mathbf{H}_v^{(p)} \cdot \mathbf{H}_c$. For person p and instance[1] i, we obtain:

$$\log \mathbf{S}_i^{(p)} = \log \mathbf{E}_i^{(p)} + \log \mathbf{H}_v^{(p)} + \log \mathbf{H}_c. \tag{10}$$

$\mathbf{E}_i^{(p)}$ is nonstationary in general for timescales larger than the pitch period.[2] $\mathbf{H}_v^{(p)}$ may capture invariant characteristics of the speaker's vocal tract as well as phoneme-specific characteristics (and underlying speech neural control) that can be considered stationary and hence invariant within phonetic timescales, in keeping with the quasistationary assumptions of the speech process.[3] This fundamental

[1]The instance i implicitly represents the timescale of interest, e.g., a timescale of the order of the pitch period (10–20 ms) or of the order of the average word period (500 ms).

[2]The spectrum of the excitation changes slowly for voiced sounds and appears unchanged although radically different over the duration of a consonant, at the phonetic timescale.

[3]A detailed analysis of these components of the speech production model is beyond present scope.

model of speech production extended with the external channel transfer function is the basis for defining inputs $\mathbf{x}_i$ and the corresponding timescales where various components play the role of $\mathbf{s}$ and $\mathbf{f}_n$ from Eq. (9).

For example, [7] use training data from different nonlinearly distorted channels for each person from various portions of the NTIMIT database [11]. The intuition was that the channel variation results in a low contribution of the channel in the GMIA extract while the vocal tract characteristic $\log \mathbf{H}_v^{(p)}$ is retained. In contrast, in this chapter, we considered training instances $\mathbf{x}_i$ from multiple people exploring an identical external channel $\mathbf{H}_c$ (e.g., from the same external position and using the same microphone, which is the case for own-voice recordings in OVR). In this case the $\log \mathbf{E}_i^{(p)}$ and $\log \mathbf{H}_v^{(p)}$ components in Eq. (10) play the role of the orthogonal components $\mathbf{f}_n$ in our synthetic model (Eq. (9)), while $\log \mathbf{H}_c$ is the invariant. In such a setup, GMIA can be used to identify invariant characteristics of the channel (e.g., near-field channel for OVR).

We use various portions of the TIMIT database [12] for our experiments in this section. TIMIT contains speech from 630 speakers that is recorded with a high quality microphone in a recording studio like environment. Each speaker is represented by 10 utterances. We convolve the TIMIT speech with a head-related transfer function (HRTF) to simulate various invariant channels. The output of an algorithm for channel identification can thus be compared directly with the true HRTF used to generate the data.

We chose a HRTF from a position on the right side of a dummy head with a source distance of 20 cm, azimuth of 0° and at an elevation of -30° as invariant channel, and a HRTF for the right side of the dummy head with a source distance of 160 cm, azimuth of 0° and at an elevation of 0° as external channel. The HRTF data has been obtained from [18]. Thereafter, the data is windowed with half overlapping Hann windows of 0.2 s length and transferred into the power spectral domain.

Our goal is to apply GMIA to extract channel information and evaluate if GMIA representations can be used to distinguish different channels. Person-dependent information is minimized by introducing variation in the excitation $\mathbf{E}_i^{(p)}$ using speech from both voiced and unvoiced signals. Note that speech signals contain silence periods where no channel information is present. Furthermore, voiced speech is sparse in the spectral domain. Therefore, not all parts of the channel characteristic are fully represented at all times. Clearly, the channel does not equally correlate with the spectral information of the speech from different time windows. A GMIA representation will be computed separately from speech of the same or multiple speakers.

5.2 Speaker Model

For one person p_0, consider the vector $\mathbf{x}_i$ obtained from a speech clip i:

$$\mathbf{x}_i = \log \mathbf{S}_i^{(p_0)} = \left(\log \mathbf{H}_c + \log \mathbf{H}_v^{(p_0)}\right) + \left(\log \mathbf{E}_i^{(p_0)}\right) \approx \mathbf{s} + \mathbf{f}_i. \tag{11}$$

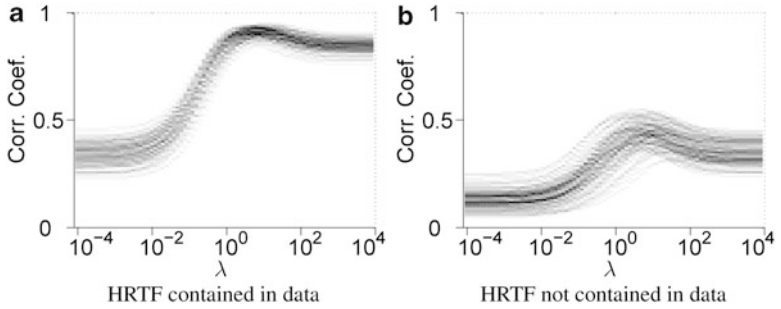

Fig. 3 Histograms as vertical slices of the plot of the correlation coefficients between GMIA(λ), for a fixed value of λ, from single-speaker data and two different HRTF's. *Dark gray bins* represent a large number and *light gray bins* a low number of instances with a particular correlation coefficient. (**a**) The HRTF used to generate the speech data is well represented by the GMIA(λ) result for $\lambda = 10^1$, resulting in a mean correlation coefficient of 0.9. (**b**) An HRTF that is not contained in the speech data minimally correlates with the GMIA extract

We use data as above for one single person and with channels for near- and far-field given by the HRTFs to the right side of the dummy head. According to the data model in Eq. (11) we expect that GMIA computes a common component capturing information about both the channel and the speaker characteristics. Indeed, $\log \mathbf{H}_c + \log \mathbf{H}_v^{(p_0)}$ is invariant to the actual clip i used as input. Next we compute GMIA and correlate the result with known channel information (HRTF) to verify our hypothesis.

All experiments are repeated for 100 speakers and various values of λ. Figure 3a illustrates the histogram of the correlation coefficients of the GMIA extract from the near-field speech with the ground truth near-field HRTF for a 20 cm source/receiver distance. Note that both $\mathbf{w}_{\mathrm{GMIA}(10^{-4})} \approx \mathbf{w}_{\mathrm{MIA}}$ and $\mathbf{w}_{\mathrm{GMIA}(10^{4})} \approx \mu$ (the mean of the inputs) do not compute maximal correlation coefficients. The median correlation value at $\lambda = 10^1$ is 0.9, demonstrating that GMIA can extract good representations of the original HRTF. In contrast, Fig. 3b shows histograms of the correlation coefficients with the HRTF from a far-field position (160 cm source/receiver distance) that was not used in the data generation. The low correlation coefficients indicate that channel characteristics are well separable with the extracted GMIA representations.

Note that Fig. 3a is similar to Fig. 1 for $\sigma_3 = 0.1$ and $m_2 = 5$, which represents the case where the common component intensity varies over different training instances. This confirms that for speech the channel is not equally correlated with the spectral information from different time windows.

5.3 Channel Model

The previous subsection shows that the GMIA projection correlates well with the channel and that it can be used as feature for channel detection or as classifier of the

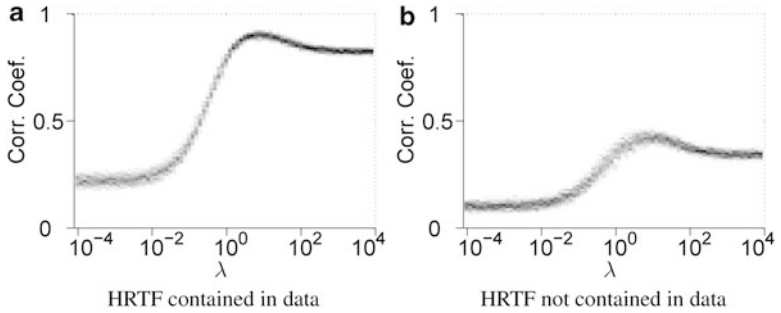

Fig. 4 Histograms (vertical slices of the plot) of the correlation coefficients between GMIA(λ) from multiple-speaker data and two different HRTF's. *Dark gray bins* represent a large number and *light gray bins* a low number of instances with a particular correlation coefficient. (**a**) The HRTF that is convolved with the data is well extracted for GMIA(λ) with $\lambda = 10^1$ resulting in a mean correlation coefficient of 0.9. The variance of the result is lower than for GMIA(λ) from single-speaker data (see Fig. 3) (**b**) The HRTF that is not contained in the data only minimally correlates with the GMIA extract

channel. We would like to make the model in Eq. (11) more precise and eliminate the speaker dependence as much as possible. For this we use data from multiple speakers p_i with $i = 1 \ldots N$ as follows:

$$\mathbf{x}_i = \log \mathbf{S}_i^{(p_i)} = (\log \mathbf{H}_c) + \left(\log \mathbf{E}_i^{(p_i)} + \log \mathbf{H}_v^{(p_i)} \right) \approx \mathbf{s} + \mathbf{f}_i. \tag{12}$$

We expect to compute a common component that better captures the channel. The experiment is performed as follows. First, a number of speakers, corresponding to the number of training instances N in Sect. 5.2, are selected randomly from the TIMIT database. One of their 10 utterances is randomly selected and convoluted with the previously chosen near-field HRTF. Thereafter, one speech segment (e.g., 0.2 s long) is randomly selected from each speaker. These segments are thereafter used to extract a GMIA representation in the log-spectral domain. The experiment is repeated for 100 randomly selected sets of speakers and various values of λ. Figure 4a shows a histogram of the correlation coefficients of the GMIA result and the ground truth for the channel, the near-field HRTF. Figure 4b illustrates the correlation coefficients between the GMIA extract and the HRTF from the external channel (160 cm source/receiver distance) that was not used in the data generation.

Indeed, Fig. 4 shows a reduced variance of the correlation coefficients for different speakers compared to Fig. 3 and thus a more reliable extraction of the channel. GMIA will be further used for channel estimation in the OVAD problem.

6 Own Voice Activity Detection

Section 5 demonstrated the efficacy of GMIA to extract channel features using a known HRTF as the channel convolved with the TIMIT speaker data under both speaker model and channel model formulations. In this section, we extend this

further to a realistic scenario of OVAD using the same large speaker database convolved with near-field and far-field HRTFs to closely approximate own-voice speakers and external speakers.

In the experimental scenarios used here with such data for OVAD, though the underlying HRTF information is available (as was used in Sect. 5 for measuring the correlation coefficients between extracted MIA features and the reference HRTF), we assume the underlying HRTF information to be unknown and unavailable, thereby treating them as implicit in the speech data (as is the case with real recorded OV and EXT speaker data at an hearing aid); for this purpose, the underlying OV and EXT channel information are equivalently considered only in the form as available by means of estimates of channel information from held-out reference data, such as by the GMIA extraction proposed here. Thus, in this scenario, how well the GMIA-based features offer a good own-voice recognition performance when set in a channel detection framework will serve to demonstrate the effectiveness of GMIA to extract the underlying channel information from the actual OV and EXT speech data.

Toward this, we demonstrate in this section the use of GMIA-based channel features for OVAD in a verification framework posed as an hypotheses testing problem. Further, in order to provide a comparative reference for the GMIA-based approach, we consider two alternate approaches: one using cepstral mean as an alternate channel feature and set in the same verification framework, and the other using the conventional speech feature, namely, MFCC, set in a speaker verification framework. We work with a 100-speaker database convolved with near-field and far-field HRTFs to closely represent own-voice and external speakers. The performance of these three verification systems are given and compared in terms of the equal-error-rate (EER) measure. Additionally, given that GMIA is specifically formulated to handle real-world data with additive noise, we also demonstrate the effectiveness of GMIA for noisy data by considering three noise conditions, namely, street, car, and music noises, at different SNRs (clean, 20 dB, 10 dB, 0 dB, -10 dB and -20 dB) and show how its parameterization (in terms of λ—the assumed noise variance) allows a superior performance at a range of optimal λ, in comparison to the other two approaches (cepstral-mean- and MFCC-based speaker-verification).

6.1 GMIA Based Verification Framework for OVAD

Given the conversational speech signal, the OVAD problem can be reduced to that of detecting the underlying channel. This in turn involves extracting the channel feature from the speech signal and classifying it as own-voice or external-speaker channel, thereby comprising a 2-class problem. Alternately, this can also be viewed as a "own-voice verification" problem (e.g., as in speaker-verification), set in a hypothesis testing framework of deciding between the two hypotheses:

H_0: Input speech segment is own voice.
H_1: Input speech segment is not own voice (i.e., external speaker).

The verification framework adopted here is essentially as in speaker-verification, which is well established [2, 19]. We outline this here as adopted to the channel verification task: Given a set of OV and EXT speakers, a test OV speaker is verified as OV speaker with respect to a set of OV channel features extracted from another set of OV speakers. The latter is referred to as "reference" OV channel features, and serves to represent the underlying unknown OV channel, as extracted by GMIA; such a channel information, by virtue of being speech- and speaker independent, can be expected to be invariant across a set of OV speakers and to generalize to an unseen test OV speaker. Likewise, a test external (EXT) speaker can be verified as "not OV" speaker against the same set of reference OV channel features. In general, a set of test OV/EXT speakers represented in terms of their channel features are verified in this manner with respect to another set of reference OV channel features, thus constituting a statistically robust channel verification task.

Ideally, the OV test GMIA features ought to yield high correlation scores (or alternately, low distance scores) with OV reference channel features, while the EXT test GMIA features yield low correlation scores with the OV reference channel features. If the features represent the OV and EXT channels well and offer good separability in the GMIA feature space, the corresponding OV and EXT score distributions are also well separated. An optimal threshold is determined on the OV and EXT score distributions which minimizes false rejections (fr, which is the number of true OV features rejected as "not OV") and false acceptances (fa, which is the number of true EXT features accepted as "OV"). The corresponding EER of (Prob(fr), Prob(fa)) is reported as the OVR system performance, with lower EER implying a better performance.

6.2 Alternate Approaches

In order to provide a baseline reference to the OVAD by GMIA-based channel features as discussed above, we also consider two other alternatives to OVAD: one using an alternate channel feature extraction, namely, the "cepstral mean," and another using a speaker-verification approach wherein OVR is carried out in terms of verifying whether the input speaker is the wearer or not.

6.2.1 Cepstral-Mean-Based OVAD

The mean vector obtained from GMIA for large λ ($\lambda \to \infty$) corresponds to the mean of the log-spectral vectors in a clip (analysis window for extracting a GMIA vector). Alternately, one can consider the mean of the cepstral vectors derived by an inverse FFT or DCT of the log-spectral vectors, as is done for deriving cepstral coefficients or MFCCs in speech recognition [16]. This mean vector, referred to as "cepstral-mean" (CM) in speech recognition, is popularly used in the context

of cepstral mean normalization (CMN) for channel compensation [1, 16]. Here, it is already a well established concept that the cepstral mean of the log spectra of long speech intervals approximates the channel cepstra and that subtraction of this long-term averaged cepstral-mean from the individual frames of cepstral features removes the channel effect, thereby rendering the resultant cepstral vectors robust to channel variability (such as arising from channel differences in telephony speech recognition due to differences in handset, physical channel media, wireless network channels, etc., particularly between training and test conditions).

6.2.2 Speaker-Verification-Based OVAD

In a OVAD task, the OV speaker is fixed and given and can be made to provide training data to define OV models that characterize the OV speaker. By this, the OVAD task can be alternately defined as a conventional speaker-verification task of treating the OV speaker as the target speaker and EXT speakers as the impostor speakers. For this, it becomes necessary to use conventional "speaker" feature representations, such as MFCC [2, 19]. In this case, the OV speaker is represented by a statistical model (GMM) or a nonparametric model (VQ) in the MFCC feature space.

The distribution of the MFCC vectors (and the GMM- or VQ-based representation of this distribution) of a speaker characterizes the unique acoustic signature or footprint of that speaker in the MFCC feature space as manifesting in the unique spectral characteristics of his voice, manner of articulation of the different sounds of the language (phonemes), and spectral dynamics (which can be potentially captured in the delta and delta-delta MFCCs). The OV and EXT speaker data occupy different regions in the feature space, by virtue of the fact that the spectral characteristics of each of these speech is a result of convolution with different channels (here, HRTF). An OV speaker model thereby offers a better match with OV test speaker data than with EXT test speaker data, which then becomes the primary basis of OVAD by MFCC-based speaker verification. The verification task is thus essentially as described in Sect. 6.1, but constituting a "speaker" verification (as against "channel" verification, since the MFCC features here serve as "speaker" features) in this case taking the form of computing OV scores between OV test MFCC vectors and the OV models and EXT scores between EXT test MFCC vectors and the OV models, subsequently forming the OV and EXT score distributions and then determining the EER.

6.3 Experimental Setup

Here, we present the experimental details of the three OVAD tasks, namely, GMIA-based channel verification, cepstral-mean (CM)-based channel verification, and MFCC-based speaker-verification. These three frameworks are as described

generically earlier in Sects. 6.1 and 6.2. While the three tasks have specific differences due to their underlying idiosyncratic frameworks, they share an overall experimental scenario, comprising the following common aspects.

All the OVAD experiments use a randomly selected (but fixed) subset of 100 speakers from the TIMIT database (of 630 speakers) as the test set of OV and EXT speakers, with each speaker having 10 sentences, each 3 to 4 s duration. The fixed subset of 100 test speakers is convolved with single fixed near-field and far-field HRTFs to generate the own voice and external type of speakers, respectively (referred to as OV and EXT henceforth); the HRTFs used here are as described in Sect. 5. In order to examine the noise robustness of GMIA and the two alternate approaches, we consider three different noise conditions, namely, street, car, and music, and five SNRs for each of these noise conditions (20 dB, 10 dB, 0 dB, −10 dB, and 20 dB), in addition to the clean case. The specific noise data is added to the original clean TIMIT sentences at the desired SNR subsequent to the HRTF convolutions, i.e., to the OV and EXT data.

We now describe the specific variations in the experiments for each of the three OVAD tasks.

6.3.1 GMIA-Based OVAD

While the 100 speakers as defined above constitutes the test data, GMIA experiments use a set of 300 speakers (different from the 100 test speakers) to define the "reference" OV channel feature. This is motivated by the channel model formulation in Sect. 5.3, where a GMIA vector is extracted in a speaker-independent manner. Here, a single GMIA reference vector is extracted from the 300-speaker clean data, i.e., with $N = 300$, as defined in Sect. 5.3.

For the noise-added experiments, only the test data is made noisy, while the above reference GMIA vector is extracted and kept fixed from clean 300 speaker data. For the purposes of examining and establishing the noise-robust advantage intrinsic to GMIA through its parameter λ, the GMIA-based channel verification experiments are conducted for λ varying over the range of $[10^{-4}$ to $10^{4}]$. One such experiment (for a given λ) consists of using 100 test OV and EXT speaker data and computing 1 GMIA vector for each speaker (from the entire duration of 30 to 40 s of that speaker, corresponding to $N = 300$–400 in $\mathbf{X}$ of Eq. (7)). The test database of 100 speakers thus yields 100 OV and EXT scores, from which the EER corresponding to the given λ is obtained. For a given noise-type and SNR, EER is obtained as a function of λ over the range $[10^{-4}$ to $10^{4}]$. Such an EER-vs-λ curve is obtained for all the 6 SNRs (clean, 20 dB, 10 dB, 0 dB, −10 dB, and 20 dB), for each noise type (street, car, and music).

6.3.2 Cepstral-Mean-Based OVAD

The experimental framework for this task uses the identical test set of 100 speakers as above, while differing only in the way the reference cepstral-mean channel

feature vector is derived and in how the test set scores are computed in a leave-one-out framework, in order to offer a statistically robust verification task; this is outlined below.

For a given speaker (OV or EXT), a cepstral-mean vector is computed from the entire duration of that speaker (30 to 40 s, yielding 300–400 cepstral vectors, each obtained using framesize of 200 ms and overlap of 100 ms). The cepstral vector for each frame is obtained by a DCT of the log-spectral vector.

For a given test OV speaker (among 100 test speakers), the remaining 99 OV speakers are defined as the reference channel speakers. 1 cepstral-mean vector is computed for each of these 99 speakers (from clean data), thereby providing 99 clean reference channel vectors (for that test OV speaker). One score is computed between the test cepstral-mean vector (from the entire duration of that test speaker) and the reference cepstral-mean vector (from among the 99 reference vectors) which has the highest correlation with the test cepstral-mean vector. For the given test OV speaker, the corresponding EXT speaker (the same speaker in the 100 speaker database, but now from the EXT set) is used to compute the EXT score with respect to the same OV reference channel vectors.

The above is repeated for each of the 100 test OV speakers as the test speaker (with the remaining 99 speakers forming the reference channel set), thereby yielding 100 OV and EXT scores, from which the score distribution is formed and EER determined; this corresponds to a specific noise type and SNR. EERs are obtained for all 5 SNRs and clean cases for the 3 noise types (street, car, and music).

6.3.3 MFCC-Based OVAD

This OVAD task differs in several respects from the above two channel verification tasks, in that it is essentially a speaker verification task and therefore has a fairly different experimental setup, though sharing the broad parameters with the above tasks to allow for a fair comparison.

The primary feature for this task is the MFCC vector computed with a framesize of 20 ms and overlap of 10 ms, constituting quasistationary timescales as required to derive spectral information of speech data. This yields 100 MFCC vectors per second of speech data, and each TIMIT speaker (of duration 30–40 s) has about 3000–4000 vectors. The MFCC feature vector used here is derived with a set of 40 triangular filters applied on the log spectra of a frame followed by DCT on the filter energy outputs to yield the cepstral coefficients; the MFCC vector used is of dimension 36, consisting of 12 cepstral coefficients (coefficients 2 to 13, with the first energy co-efficient not used, thereby making the feature insensitive to signal energy), 12 delta and 12 delta-delta coefficients.

The verification task here is set in the leave-one-out framework (as defined for the cepstral-mean task). For a given test speaker, the remaining 99 speakers are used to define the reference OV speakers against which the test speaker MFCCs are scored. Each of these 99 speakers is represented by a VQ codebook of size 64, considered adequate from established speaker-identification tasks [20].

A scoring window is defined for the test data for deriving a score with respect to the reference VQ codebooks. The scoring windows used here are 1, 2, 4, 8, 16 and 30 s. For a specific scoring window duration, an accumulated dissimilarity (distance) score is computed for the window with respect to each of the 99 VQ codebooks. The accumulated score for a VQ codebook is the sum of the individual scores of the MFCC vectors in the window, the individual score of a vector being the distance between the vector and the nearest codevector in the VQ codebook. The final score of the test window is determined as the minimum across the 99 VQ codebooks, i.e., a window of test vectors has a single score with respect to the best scoring reference VQ codebook.

For a given test window duration, OV and EXT scores are computed over the test data duration of a speaker and score distributions formed from such scores from all test speakers in the above leave-one-out framework; an EER is obtained for each test window duration for a given noise type and SNR. For the different noise types and SNRs, only the test data is subjected to noise, while the reference VQ codebooks are maintained as derived from clean data.

6.4 OVAD Results Analysis

In this section, we present results of the above three OVAD tasks (GMIA based channel verification, CM-based channel verification, and MFCC-based speaker-verification) for different noise types and SNRs. The performance of the three verification approaches are given in terms of EER, as defined earlier in Sect. 6.1, in street, car, and music noises, respectively, for different SNRs.

6.4.1 OVAD for GMIA, CM, and MFCC in Noisy Conditions

Figure 5a–c show EER as a function of λ for GMIA. As expected, the EER shows a pronounced dependence on λ, consequently offering the best performance at $\lambda = 10^0$ consistently for both clean and noisy cases. This is in agreement with the similar dependence and optimality shown by the correlation coefficients for the experiments reported in Sect. 5.3. Optimal results are obtained similarly for values of $\lambda = 0.1$–10. This validates the importance of the parameterization of GMIA in terms of λ to handle real-world noisy data.

Fig. 5 Own-voice activity detection with GMIA(λ), MFCC, and CM for various noise types and levels. (**a**) Street noise above 0, dB SNRs enables GMIA-based own-voice activity detection with EERs below 10 %. GMIA(10^0) achieves best results. (**b**) Car noise above -10 dB SNRs enables GMIA-based own-voice activity detection with EERs below 5 %. There is a clear improvement for $\lambda = 10^0$ over the spectral mean. (**c**) Music noise is least affecting the GMIA-based own-voice activity detection. (**d**) CM performs mostly below the spectral mean and by a large margin below GMIA(10^0). MFCC performs below GMIA(10^0) for high SNRs and at level for low SNRs

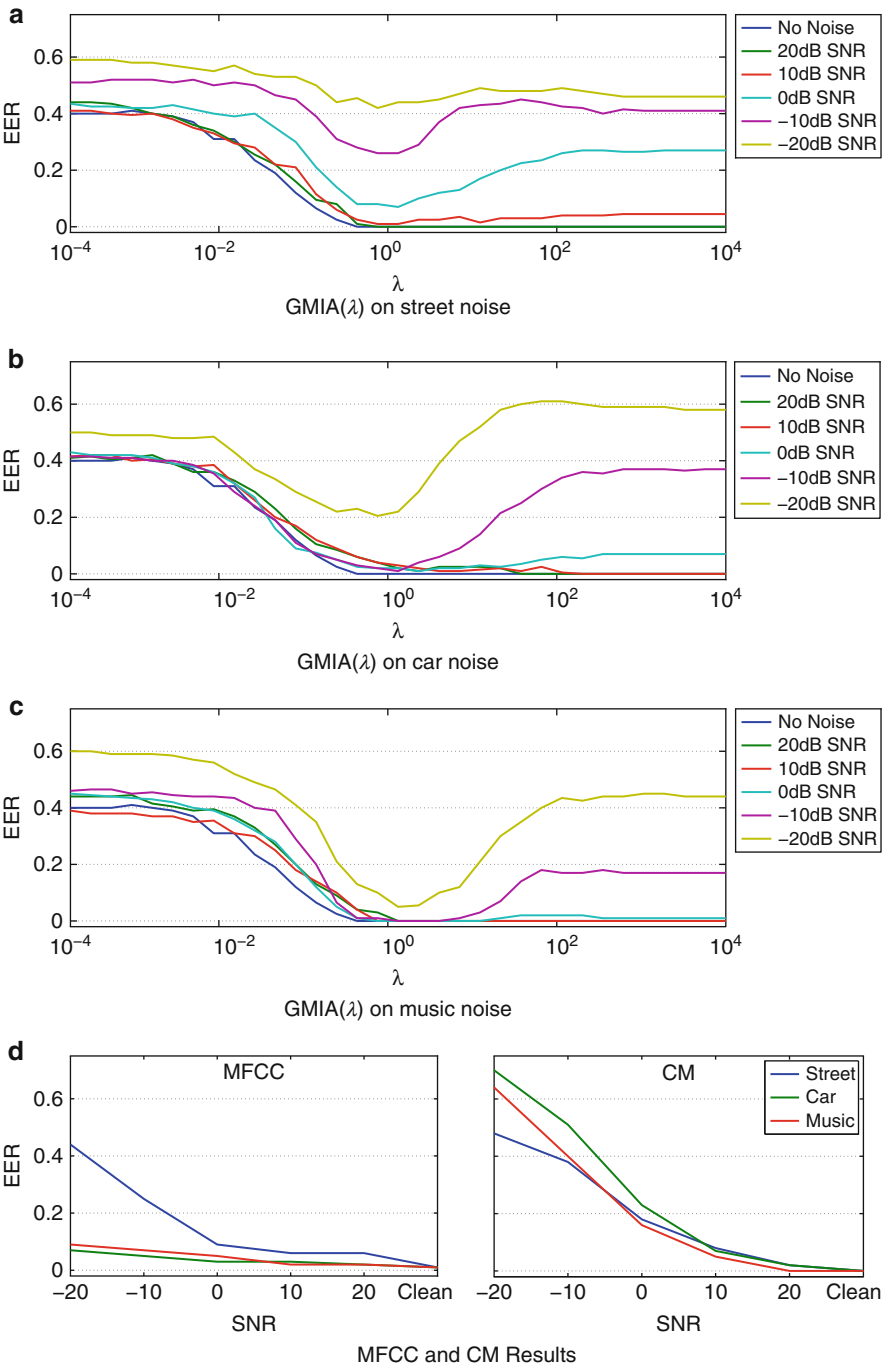

Fig. 5 (continued)

More specifically, it can be noted that for low SNRs and for all noise types, the optimal λ results in a prominent dip in EER, when compared to MIA (for $\lambda = 10^{-4}$) and the spectral mean (for $\lambda = 10^4$). This is in line with the basis of the channel model formulation in Sect. 5.3, indicating the ability of GMIA (at optimal λ) to extract an invariant component in the presence of a higher degree of uncorrelated additive components $\mathbf{f}_i$ [in Eq. (12)], in this case corresponding to large variability in log-spectral components corrupted with higher levels of noise (lower SNRs).

With regard to MFCCs, Fig. 5d shows that MFCC offers competitive performance to GMIA (comparable or even lower EERs at times, such as for street noise at -20 dB and -10 dB and car noise at -20 dB) for lower SNRs, while the optimal GMIA performances are better than MFCC for high SNRs. The better performance of GMIA over MFCCs (particularly for high SNR cases) is accounted for as follows. MFCC-based speaker-verification approach attempts to model the OV (or EXT) space as the feature space spanned by the speech of the owner (or external) speaker (i.e., spanned by all the phonetic realizations as is unique to a speaker) and hence implicitly captures both the channel and speaker information. This in turn makes the feature space occupied by the OV and EXT speaker data to be large and diffuse, leading to potentially higher overlap of their feature spaces and a consequent higher overlap of the OV and EXT score distributions with associated higher EERs. In contrast, the GMIA features represent the channel information directly with minimal associated speaker information (as was evident from the results in Fig. 4, where the channel model, being extracted in a speaker-independent manner, offers lower variance of the correlation coefficients) and consequently better separability between the OV and EXT spaces and associated lower EERs.

Within the channel modeling framework, the alternative cepstral-mean features (Fig. 5d) have higher EERs than the "spectral mean" of GMIA at $\lambda = 10^4$ (i.e., the asymptotic performance for GMIA for $\lambda \to \infty$), particularly for lower SNRs. Moreover, the EERs for cepstral mean are significantly higher than the best GMIA EERs for all noise types and SNRs. In general, while CM offers reasonably comparable performance at clean conditions, it degrades severely with increase in noise levels and has poor noise robustness. When compared to MFCC, MFCC clearly outperforms CM for all cases.

6.4.2 OVAD for GMIA, CM, and MFCC for Varying Test Durations

Figure 6 shows an important computational aspect of GMIA—the duration over which a single GMIA vector is computed. In this figure, EER-vs-λ is shown for varying durations (1, 2, 4, 8, and 16 s) over which the GMIA vector is computed in the test data. GMIA exhibits no particular sensitivity to this duration (at the optimal λ) for clean case (Fig. 6a). Even 1 s of data is sufficient to realize a 0 % EER for the clean case at the optimal λ.

However, for the noisy case (car noise at 0 dB) in Fig. 6b, the EER curve worsens with decrease in the duration (from 16 s to 1 s). For 1 s data, even the EER at optimal λ is as high as $\sim$30 % and it needs 4 s of data to enable EERs $\sim$8 %.

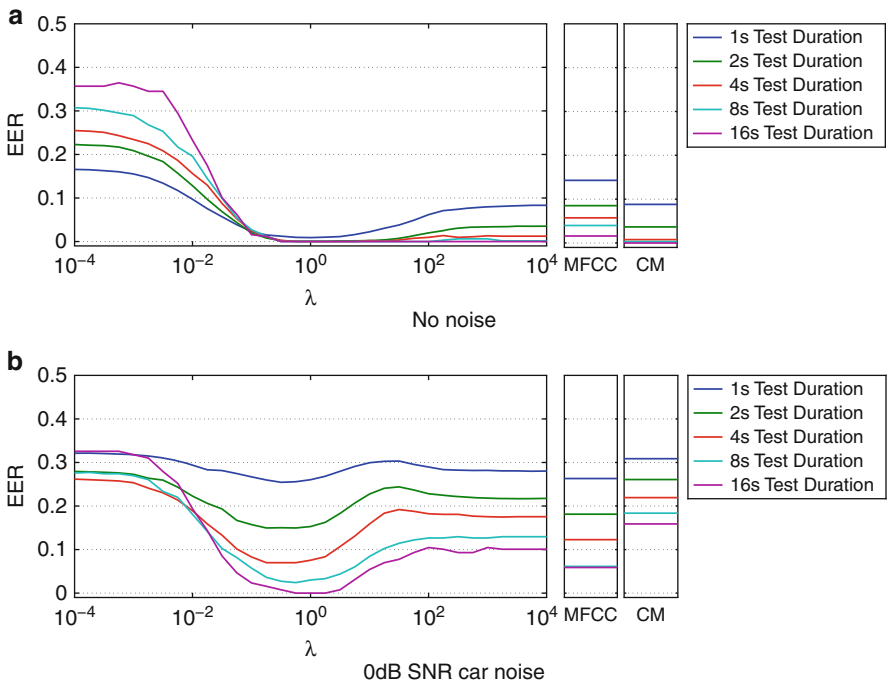

Fig. 6 Own-voice activity detection with GMIA(λ) and MFCC for various test durations. (**a**) GMIA(10^0) and the spectral mean both outperform the MFCC in case of no noise. (**b**) GMIA(10^0) outperforms MFCC for car noise with long test durations and achieves similar results for short test durations. MFCC performs better then the spectral mean

This shows that channel extraction with GMIA requires large amounts of data to enable noise-robust extraction, i.e., larger data implying sufficient uncorrelated components [**f** in Eq. (12)] to enable their cancellation and reliable extraction of the common channel component. This will impact online applications, where shorter durations (over which an OVAD decision is reported) will be clearly preferred.

Considering MFCC, GMIA(10^0) offers better performance than MFCC for the clean case. For the noisy case (Fig. 6b), GMIA(10^0) is again better than MFCC for longer durations, but comparable for shorter durations. The dependence of MFCC on longer durations is consistent with previously reported results on MFCC-based speaker verification where it is known that test durations of the order of 5–10 s are necessary to achieve optimal performance [20]; this is primarily due to the fact that such speaker verification relies on having long acoustic signature of the speaker to yield a sufficiently discriminating accumulated score.

Considering CM, for clean cases, CM has comparable performance to the spectral mean (GMIA(10^4)); however, for the noisy case, CM is worse than MFCC and also the spectral mean (GMIA(10^4)), indicating that CM is more sensitive to noise than GMIA, though it can offer comparable performance to the spectral mean for clean conditions.

7 Conclusion

GMIA is a low-complexity second-order statistical method for projecting data in a subspace that captures invariant properties of the data. This chapter summarizes the theory behind GMIA in a unitary presentation and most importantly carries the reader through a succession of increasingly difficult application examples. The examples come from a conspicuous albeit well-studied signal processing problem: voice (signal) activity detection and classification. We show how real-world conversational speech data should be modeled to fit the GMIA assumptions. From there, low-complexity GMIA computations can induce reliable features that are used for classification under noisy conditions and operate with small amounts of data. Furthermore, our results push the state of the art and are intriguing. For example, GMIA features perform better than cepstral power and mel-frequency cepstral coefficient features, particularly in noisy conditions, and are amenable to online (real-time) detection algorithms. More significantly, the approach opens the door for a large number of possible applications where a signal source (e.g., a speaker), characterized by a slow varying or invariant channel that is learned can be tracked from single channel data. The GMIA approach derived and applied in this chapter resonates with the principle of doing more with less, which will certainly find new applications in discrete time signal processing in the near future.

References

1. Benesty, J., Sondhi, M.M., Huang, Y.: Handbook of Speech Processing. Springer, Berlin (2008)
2. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A.: A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. **4**, 430–451 (2004)
3. Cho, Y., Al-Naimi, K., Kondoz, A.: Improved voice activity detection based on a smoothed statistical likelihood ratio. In: International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 737–740. IEEE, Los Alamitos, CA (2001)
4. Claussen, H., Rosca, J., Damper, R.: Mutual interdependence analysis. In: Independent Component Analysis and Blind Signal Separation, pp. 446–453. Springer, Heidelberg (2007)
5. Claussen, H., Rosca, J., Damper, R.: Mutual features for robust identification and verification. In: International Conference on Acoustics, Speech and Signal Processing, pp. 1849–1852. Las Vegas, NV (2008)
6. Claussen, H., Rosca, J., Damper, R.: Generalized mutual interdependence analysis. In: International Conference on Acoustics, Speech and Signal Processing, pp. 3317–3320. Taipei, Taiwan (2009)
7. Claussen, H., Rosca, J., Damper, R.I.: Signature extraction using mutual interdependencies. Pattern Recognit. **44**, 650–661 (2011)
8. Deng, L., O'Shaughnessy, D.: Speech Processing: A Dynamic and Optimization-Oriented Approach. Signal Process. Commun. Dekker, New York (2003)
9. ETSI: Digital cellular telecommunication system (phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech traffic channels; general description. Technical Report V.7.0.0, ETSI (1999)

10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**, 179–188 (1936)
11. Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J.: NTIMIT. Linguistic Data Consortium, Philadelphia CDROM (1993). http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S2
12. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia CDROM (1993). http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1
13. Haigh, J., Mason, J.: Robust voice activity detection using cepstral features. In: IEEE Region 10 Conference TENCON, vol. 3, pp. 321–324. IEEE (1993)
14. Hotelling, H.: Relation between two sets of variates. Biometrika **28**, 322–377 (1936)
15. Hoyt, J.D., Wechsler, H.: Detection of human speech in structured noise. In: International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 237–240. IEEE (1994)
16. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing: A guide to Theory, Algorithm, and System Development. Prentice Hall, New York (2001)
17. Liu, P., Wang, Z.: Voice activity detection using visual information. In: International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 609–612. Montreal, Canada (2004)
18. Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., Wu, X.: Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. IEEE Trans. Audio, Speech Lang. Process. **17**(6), 1124–1132 (2009)
19. Reynolds, D.A., Campbell, W.M.: Text-independent speaker recognition. In: Benesty, J., Sondhi, M., Huang, Y. (eds.) Handbook of Speech Processing and Communication, pp. 763–781. Springer GMBH, New York (2007)
20. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech, Audio Process. **3**(1), 72–83 (1995)
21. Rosca, J., Balan, R., Fan, N., Beaugeant, C., Gilg, V.: Multichannel voice detection in adverse environments. In: European Signal Processing Conference (2002)
22. Srinivasan, K., Gersho, A.: Voice activity detection for cellular networks. In: IEEE Speech Coding Workshop, pp. 85–86 (1993)
23. Tikhonov, A.: On the stability of inverse problems. Doklady Akademii Nauk SSSR **39**(5), 195–198 (1943)
24. Zhang, Z., Liu, Z., Sinclair, M., Acero, A., Deng, L., Huang, X., Zheng, Y.: Multi-sensory microphones for robust speech detection, enhancement and recognition. In: International Conference on Acoustics, Speech and Signal Processing, pp. 781–784. IEEE (2004)