Blind Channel Magnitude Response Estimation in Speech Using Spectrum Classification

Nikolay D. Gaubitch, *Member, IEEE*, Mike Brookes, *Member, IEEE*, and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—We present an algorithm for blind estimation of the magnitude response of an acoustic channel from single microphone observations of a speech signal. The algorithm employs channel robust RASTA filtered Mel-frequency cepstral coefficients as features to train a Gaussian mixture model based classifier and average clean speech spectra are associated with each mixture; these are then used to blindly estimate the acoustic channel magnitude response from speech that has undergone spectral modification due to the channel. Experimental results using a variety of simulated and measured acoustic channels and additive babble noise, car noise and white Gaussian noise are presented. The results demonstrate that the proposed method is able to estimate a variety of channel magnitude responses to within an Itakura distance of $d_I \leq 0.5$ for SNR ≥ 10 dB.**

***Index Terms*—Blind channel estimation, GMM.**

I. Introduction

When speech is captured by a microphone positioned at some distance away from the talker, the spectrum of the observed speech will be modified due to propagation through the acoustic channel between talker and microphone. We define the channel as encompassing the combined effects of the acoustic environment, the positioning of the microphone and the characteristics of the microphone and associated sound capturing equipment. In all practical cases the captured signal will contain ambient background noise in addition. Thus, the observed signal at the microphone can be expressed as

$$x(n) = s(n) * h(n) + \nu(n), \tag{1}$$

where $s(n)$ is the desired speech signal, $h(n)$ is the channel impulse response, $\nu(n)$ is additive observation noise and $*$ denotes convolution.

The channel can degrade perceived quality and reduce intelligibility of the captured speech signal; the reduction in intelligibility due to the channel effects becomes more severe in the presence of noise [1]. Knowledge of the channel characteristics offers the potential to design an equalization filter which enhances the speech. It can also provide information about the capturing environment and has been used in, for example, codec identification [2]. Consequently, channel estimation has developed into an important topic in acoustic signal processing [3].

There exist many robust methods for the identification of frequency response functions when both the input signal, $s(n)$, and the observed signal, $x(n)$, are known [4]–[7]. However, in practical scenarios only the observed signal, $x(n)$, is available which leads to the need for unsupervised, blind channel estimation procedures as proposed by Stockham *et al.* in [8]. In the case of multiple microphones, spatial information is available; the channels between the source and each microphone are different while the source is the same. This fact is exploited by several channel estimation algorithms [3], [9], [10]. In contrast, single microphone blind channel estimation is inherently a more challenging problem since the spatial information is not available and it is therefore necessary to exploit alternative information such as a speech production model or other characteristics of the speech signals. For example, Stockham *et al.* [8] derived a method based on homomorphic deconvolution [11] and showed that the channel can be found as the difference between the average log-spectrum of the clean signal and that of the observation. The method, however, requires that a representative version of the clean signal is available. Alternative methods based on [8] use the Long Term Average Speech Spectrum (LTASS) of clean speech to obtain the channel magnitude spectrum from the observed speech [12], [13]. A different approach is taken by Hopgood and Rayner [14] who proposed using a time-invariant Autoregressive (AR) model for the channel and a time-varying AR model for the speech signal; this was later extended to the case of moving talkers [15]. Although good results were demonstrated for low order channels, the method performs less well with more complex channels such as acoustic impulse responses in rooms. Moreover, the method requires knowledge of the channel order—a common restriction in blind channel identification.

In this paper, we consider single channel blind estimation of the magnitude spectrum of an unknown channel. The log-spectrum estimate of the underlying clean speech in each frame is found using RASTA filtered Mel-frequency Cepstral Coefficients (MFCC) [16] and a Gaussian Mixture Model (GMM) based classifier [17] and it is subtracted from the observed speech. This method is an extension and generalization of the of the method by Stockham *et al.* [8] and the LTASS based blind channel estimation methods in [13], [12] as will be discussed in Section III. A preliminary version of the algorithm was presented in [18].

The remainder of the paper is organized as follows. In Section II the classification-based algorithm is derived and its relation to the previous methods is shown in Section III. In Section IV the effects of noise are discussed and robustness to

Manuscript received September 11, 2012; revised February 23, 2013, May 21, 2013; accepted June 05, 2013. Date of publication June 20, 2013; date of current version July 22, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rongshan Yu.

The authors are with the Centre for Law Enforcement Audio Research (CLEAR), Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: ndg@imperial.ac.uk; mike.brookes@imperial.ac.uk; p.naylor@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2270406

additive noise is introduced to our approach through the use of an additional processing step. Simulation results demonstrating the effects of different parameters of the algorithm and its performance with a variety of noise types and channels are given in Section V. Finally, conclusions from this work are drawn in Section VI.

II. Channel Magnitude Spectrum Estimation

A. Preliminaries

Expressing (1) in the frequency domain and using the Short Time Fourier Transform (STFT) we can write

$$X(k,l) \approx S(k,l)H(k) + V(k,l), \tag{2}$$

for frequency bin k and time frame l. This approximation relies on the frame length of the STFT being large compared to the impulse response $h(n)$; a detailed examination of its validity is given in [19]. The effects of the frame length on the channel estimation are discussed in Section V. We assume that the channel varies much more slowly than the speech and, therefore, $H(k)$ does not vary significantly with l. In the noiseless case, $V(k,l) \equiv 0$, we can write

$$|X(k,l)|^2 = |S(k,l)|^2 |H(k)|^2. \tag{3}$$

It was shown in [8] that given prior knowledge of the magnitude spectrum of the speech signal, $|S(k,l)|$, we can estimate the log-magnitude spectrum of the channel as

$$\underline{\hat{H}}(k) = \frac{1}{L}\sum_{l=1}^{L}(\underline{X}(k,l) - \underline{S}(k,l)), \tag{4}$$

where $\underline{A}(k,l) = \log(|A(k,l)|)$, $\hat{A}$ denotes an estimate of A and L is the total number of speech frames. The effects of noise on the estimation will be discussed in Section IV.

In practice, $\underline{S}(k,l)$ is not known but can be estimated as $\underline{S}(k,l) \approx \underline{\hat{S}}(k,l)$. In the remainder of this section we will discuss a method for finding such estimates based on a clean speech model. The accuracy of the channel estimation will depend on the match between the estimated speech log-spectrum $\underline{\hat{S}}(k,l)$ and the true speech spectrum $\underline{S}(k,l)$. Since the absolute level of the speech is unknown, the channel can be estimated only to within an unknown scale factor.

The channel estimate $\underline{\hat{H}}(k)$ in (4) includes the true channel response convolved with the Fourier transform of the STFT window. The effect of varying the STFT frame length is examined in Section V-C.

B. Clean Speech Model

The channel estimation procedure relies on a trained GMM [17] clean speech model with M mixtures and a known average log-spectrum associated with each mixture. The procedure for obtaining this from clean speech training examples, $s(n)$, is illustrated in the upper panel of Fig. 1. This is an offline process that only needs to be performed once.

Given a training data-set of clean speech, the speech signal, $s(n)$, is divided into overlapping windowed frames

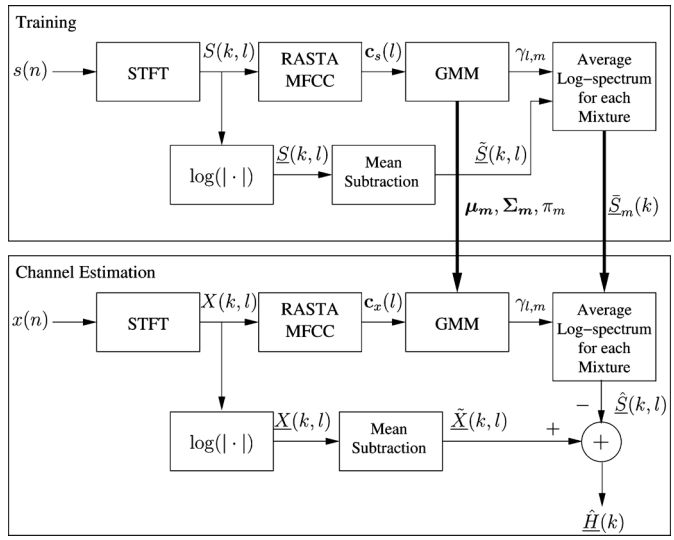


Fig. 1. System diagram of the channel estimation algorithm.

and the STFT applied to give $S(k,l)$. A feature vector, $\mathbf{c}_s(l) = [c_s(1,l)\ c_s(2,l)\ \ldots\ c_s(N,l)]$, of N RASTA-MFCCs is calculated for the lth frame and the corresponding log-spectrum for that frame, $\underline{S}(k,l)$, is obtained. The mean of the log-spectrum is subtracted

$$\tilde{\underline{S}}(k,l) = \underline{S}(k,l) - \frac{1}{K}\sum_{k=0}^{K-1}\underline{S}(k,l), \tag{5}$$

where K defines the number of frequency points in the STFT. RASTA filtered MFCC coefficients [16] are employed since they will be used to distinguish between different speech spectra in filtered speech and RASTA-MFCCs are designed to be robust to channel effects.

The feature vectors, $\mathbf{c}_s(l)$, are used to train an M-mixture GMM defined by the means, $\boldsymbol{\mu}_m$, diagonal covariances, $\boldsymbol{\Sigma}_m$, and weights, π_m, of each mixture. The mixture probabilities, $\gamma_{l,m}$, are calculated as [17]

$$\gamma_{l,m} = \frac{\pi_m \mathcal{N}(\mathbf{c}_s(l) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=1}^{M}\pi_j \mathcal{N}(\mathbf{c}_s(l) \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{6}$$

where $\mathcal{N}(\mathbf{c}_s(l) \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ denotes a multivariate Gaussian distribution.

We combine $\gamma_{l,m}$ and $\tilde{\underline{S}}(k,l)$ to obtain a weighted average of the short-term log-spectra over all available frames of training data and thus, obtain the set of M average clean speech log-spectra:

$$\bar{\underline{S}}_m(k) = \frac{\sum_{l=1}^{L}\gamma_{l,m}\tilde{\underline{S}}(k,l)}{\sum_{l=1}^{L}\gamma_{l,m}}, \quad \forall k. \tag{7}$$

In this way, each mixture is associated with a clean speech spectrum, which is formed from the weighted average of the speech spectra assigned to a particular mixture.

C. Classification-Based Channel Estimation

The clean speech model from Section II-B is now used to estimate the unknown channel as shown in the lower

panel of Fig. 1. The STFT is applied to the observed speech signal, $x(n)$ and an N-dimensional feature vector, $\mathbf{c}_x(l) = [c_x(1,l)\ c_x(2,l)\ \ldots\ c_x(N,l)]$, with the corresponding log-spectrum for that frame, $\underline{X}(k,l)$, is obtained; the log-spectrum mean is subtracted as in (5).

The estimate of the clean speech log-spectrum, $\underline{\hat{S}}(k,l)$, is obtained using the feature vectors, $\mathbf{c}_x(l)$, and the GMM parameters $(\boldsymbol{\mu}_m,\ \boldsymbol{\Sigma}_m,\ \pi_m)$. The probability of feature vector $\mathbf{c}_x(l)$ arising from each of the M mixtures is calculated as in (6). Thus, for each mixture $m = 1,\ 2,\ \ldots,\ M$, we obtain the probability, $0 \leq \gamma_{l,m} \leq 1$, of $\mathbf{c}_x(l)$ arising from that mixture and $\sum_m \gamma_{l,m} = 1$. Using these probabilities, we calculate the estimate of the clean speech spectrum for the lth frame as a weighted average of the average clean-speech spectra, $\underline{\bar{S}}_m(k)$, associated with each mixture:

$$\underline{\hat{S}}(k,l) = \sum_{m=1}^{M} \gamma_{l,m} \underline{\bar{S}}_m(k), \quad \forall k. \tag{8}$$

Using this weighted approach rather than a Maximum Likelihood (ML) selection can be advantageous when there is an uncertainty in the classification.

The channel magnitude response can now be estimated according to (4) with $\underline{S}(k,l) \approx \underline{\hat{S}}(k,l)$ calculated as either the ML or the Minimum Mean Squared Error (MMSE) estimate of the clean speech models obtained from the GMM. Having a rich model of clean speech spectra, $\underline{\bar{S}}_m(k)$, facilitates more rapid and more accurate channel estimation, as will be discussed in Sections III and IV.

III. Relationship to Previous Methods

In previous work the clean speech spectrum was approximated using the average log-spectrum from a clean version of the audio signal [8] or the more generic LTASS [12], [13] such that in (4) $\underline{\hat{S}}(k,l) = \underline{S}_{\mathrm{LTASS}}(k),\ \forall l$. This LTASS-based approach is a special case of the GMM-based classification method, in the case of only one mixture. It can be seen from the derivation of the average speech spectra in Section II-B that, when $M = 1$, the likelihood of the frame belonging to that mixture is $\gamma_{l,m} = 1,\ \forall l$. Consequently, from (7) we can write

$$\begin{aligned} \frac{\sum_{l=1}^{L} \gamma_{l,m} \underline{\tilde{S}}(k,l)}{\sum_{l=1}^{L} \gamma_{l,m}} &= \frac{1}{L} \sum_{l=1}^{L} \underline{\tilde{S}}(k,l) \\ &\approx \underline{S}_{\mathrm{LTASS}}(k). \end{aligned} \tag{9}$$

Increasing the number of mixtures, so as to classify different sounds in the speech separately, decreases the estimation time and increases the accuracy of the channel estimation because a more accurate estimate of the clean speech average spectrum is obtained at each and any time instant; there is less reliance on the match between a talker's speech and the generic spectral representation of the LTASS.

IV. Channel Magnitude Estimation in Noise

In most scenarios, noise is present in the observed speech signal, which may limit the channel estimation accuracy. The adverse effects of additive noise on the channel estimation procedure are potentially twofold. First, noise will introduce errors in the classification since the GMM parameters are derived from clean speech, which may result in wrong selection of the clean speech average spectrum. Second, it will introduce a bias in the channel estimation in (4) [8]. It was shown in [8] that the bias can be compensated for where the compensating factor is found based on the assumptions that the audio signal is a stationary random process and that noise and speech are independent. In the following, we do not make an assumption on the speech statistics and derive an expression that is based on the Signal-to-Noise Ratio (SNR).

Consider the power spectrum of the noisy observation in (2)

$$|X|^2 = |S|^2|H|^2 + |V|^2 + 2|S||H||V|\cos\theta, \tag{10}$$

where $\theta = \angle S + \angle H - \angle V$ and we have omitted the frame and the frequency indices for reasons of clarity. Dividing by $|S|^2|H|^2$ on both sides of (10) and taking the log gives

$$\underline{X} - \underline{S} = \underline{H} + \epsilon \tag{11}$$

with

$$\epsilon = 0.5\log(\zeta^{-1} + 2\zeta^{-0.5}\cos\theta + 1) \tag{12}$$

and

$$\zeta = \frac{|S|^2|H|^2}{|V|^2} \tag{13}$$

is the SNR. Equation (12) shows that the error in the channel estimate is related to the SNR. In the remainder of this section, we show how this can be used in order to add noise robustness.

A. Noise Bias Correction

It is evident from (11) that the bias in the channel magnitude estimate due to noise can be corrected given an estimate, $\hat{\epsilon}$, of the error term, ϵ, in (12). In practice, this requires knowledge of the SNR, ζ, and of the phase term θ. While there are several existing methods for noise spectral estimation [20], [21], estimators for the phases are not readily available. However, it will be shown in the following that (12) can be approximated closely in terms of the SNR only.

In order to calculate $E\{\epsilon\}$, the expected value of (12) over θ we consider three separate cases: $\zeta = 1, \zeta > 1$ and $\zeta < 1$. The phase term θ is the difference between the speech and noise phases and can be assumed uniform if the two signals are independent. For positive SNR $\zeta > 1$ we can write (12) as

$$\begin{aligned} \epsilon &= 0.5\log\left(|1 + \zeta^{-0.5}e^{j\theta}|^2\right) \\ &= \Re(\log(1 + \zeta^{-0.5}e^{j\theta})), \end{aligned} \tag{14}$$

where $\Re(a)$ denotes the real component of a. Next, taking the expectation over θ and substituting $z = e^{j\theta} \Rightarrow \frac{\mathrm{d}z}{\mathrm{d}\theta} = jz$, we have that

$$\begin{aligned} E\{\epsilon\} &= \Re\left(\frac{1}{2\pi}\int_{\theta=0}^{2\pi} \log(1 + \zeta^{-0.5}e^{j\theta})\mathrm{d}\theta\right) \\ &= \Re\left(\frac{1}{2j\pi}\oint_{|z|=1} \log(1 + \zeta^{-0.5}z)z^{-1}\mathrm{d}z\right) = 0, \end{aligned} \tag{15}$$

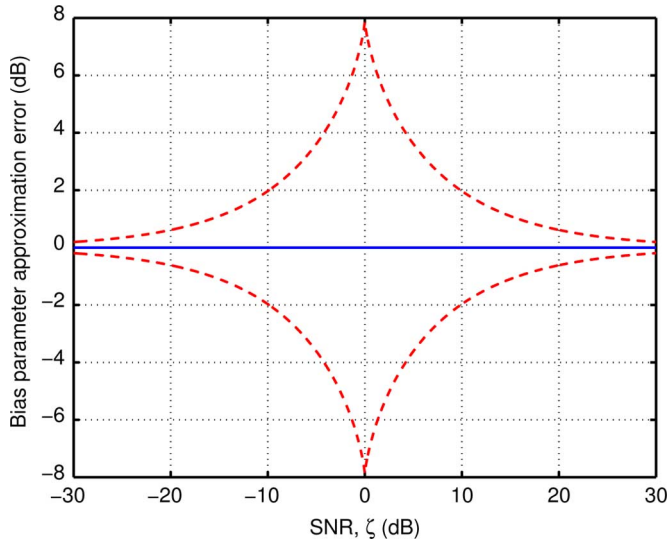


Fig. 2. Mean approximation error (solid line) ±1 standard deviation (dashed line) for the approximate estimation of (12) by (18).

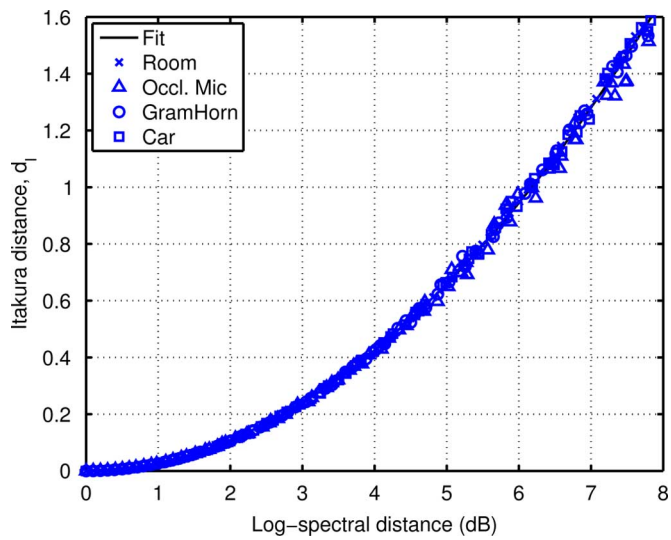


Fig. 3. Scatter plot and a quadratic polynomial fit for Log-spectral distance versus Itakura distance for four different channels.

where the integrand in the contour integral is analytic within the unit circle and hence the integral equals zero. For negative SNR $\zeta < 1$ (12) can be written as

$$\epsilon = -0.5\log\zeta + 0.5\log(1 + 2\zeta^{0.5}\cos\theta + \zeta), \tag{16}$$

which, following the same procedure as for (15) results in

$$\begin{aligned} E\{\epsilon\} &= -0.5\log\zeta + \Re\left(\frac{1}{2\pi}\int_{\theta=0}^{2\pi}\log(1+\zeta^{0.5}e^{j\theta})\mathrm{d}\theta\right) \\ &= -0.5\log\zeta + \Re\left(\frac{1}{2j\pi}\oint_{|z|=1}\log(1+\zeta^{0.5}z)z^{-1}\mathrm{d}z\right) \\ &= -0.5\log\zeta. \end{aligned} \tag{17}$$

For the special case of $\zeta = 1$, continuity implies that $E\{\epsilon\} = 0$. Combining the results from (15) and (17) the noise error term is written as

$$\hat{\epsilon} = \frac{1}{2L}\sum_{l=1}^{L}\log\left(\max\left(\frac{1}{\hat{\zeta}}, 1\right)\right). \tag{18}$$

The accuracy of this expression is illustrated in Fig. 2 where it can be seen that the mean error for a frame of estimation is zero and that the standard deviation of the error has a peak of 8 dB at 0 dB SNR but decreases rapidly at higher or lower SNRs; as the estimates are averaged over many frames, the error will generally be close to zero. The SNR is estimated as

$$\hat{\zeta} = \frac{\max(|X|^2 - |\hat{V}|^2, \beta|X|^2)}{|\hat{V}|^2}, \tag{19}$$

where, $0 < \beta \ll 1$ is the spectral floor parameter and $|\hat{V}|^2$ is the noise spectrum estimate, which we obtain using the MMSE method [22], [21]. This term $\hat{\epsilon}$ can now be subtracted from the channel magnitude estimate obtained with the algorithm described in Section II. The benefits of the bias correction will be demonstrated by the simulation results in Section V.

V. Simulation Results

Simulation results are now presented to demonstrate the performance of the channel estimation algorithm and to investigate the effects of different parameters.

A. Experimental Setup

The speech data of the TIMIT corpus was used in the evaluation. TIMIT contains 6300 sentences; ten sentences spoken by each of the 438 male and 192 female talkers. Two of the ten sentences are dialect diagnostics and use the same text for all talkers while the remaining sentences represent a total of 2342 distinct texts. The corpus is divided into a training set (462 talkers) and a test set (168 talkers) with entirely distinct sentence texts apart from the dialect diagnostics. The duration of each utterance is approximately three seconds and the sampling frequency is $f_s = 16$ kHz.

Following the procedure described in Section II-B, the full training set was used to train the GMM and to calculate the average log-spectra of clean speech. Processing was performed using Hanning windowed frames overlapping by 50%. From each frame, $N = 12$ RASTA-MFCCs were calculated and used to train the GMM. The choices of frame length and number of mixtures are discussed in Section V-C. The complete TIMIT test set was then used for the channel estimation experiments. The testing and training data sets are exclusive. The test sentences for each of the 168 talkers were concatenated to form one utterance per talker with an approximate duration of 30 s.

In these experiments three sets of acoustic channel responses were considered: (i) The identity channel, $|H(k)| = 1,\ \forall k$, (ii) synthetic room responses with a range of reverberation times, (iii) four measured real impulse responses. The synthetic room responses were generated using the source-image method [23] for a rectangular room with dimensions $6 \times 5 \times 4$ m. The source and microphone were positioned in the room separated by 1.5 m; reverberation times, T_{60}, included 0.05 s and a range between

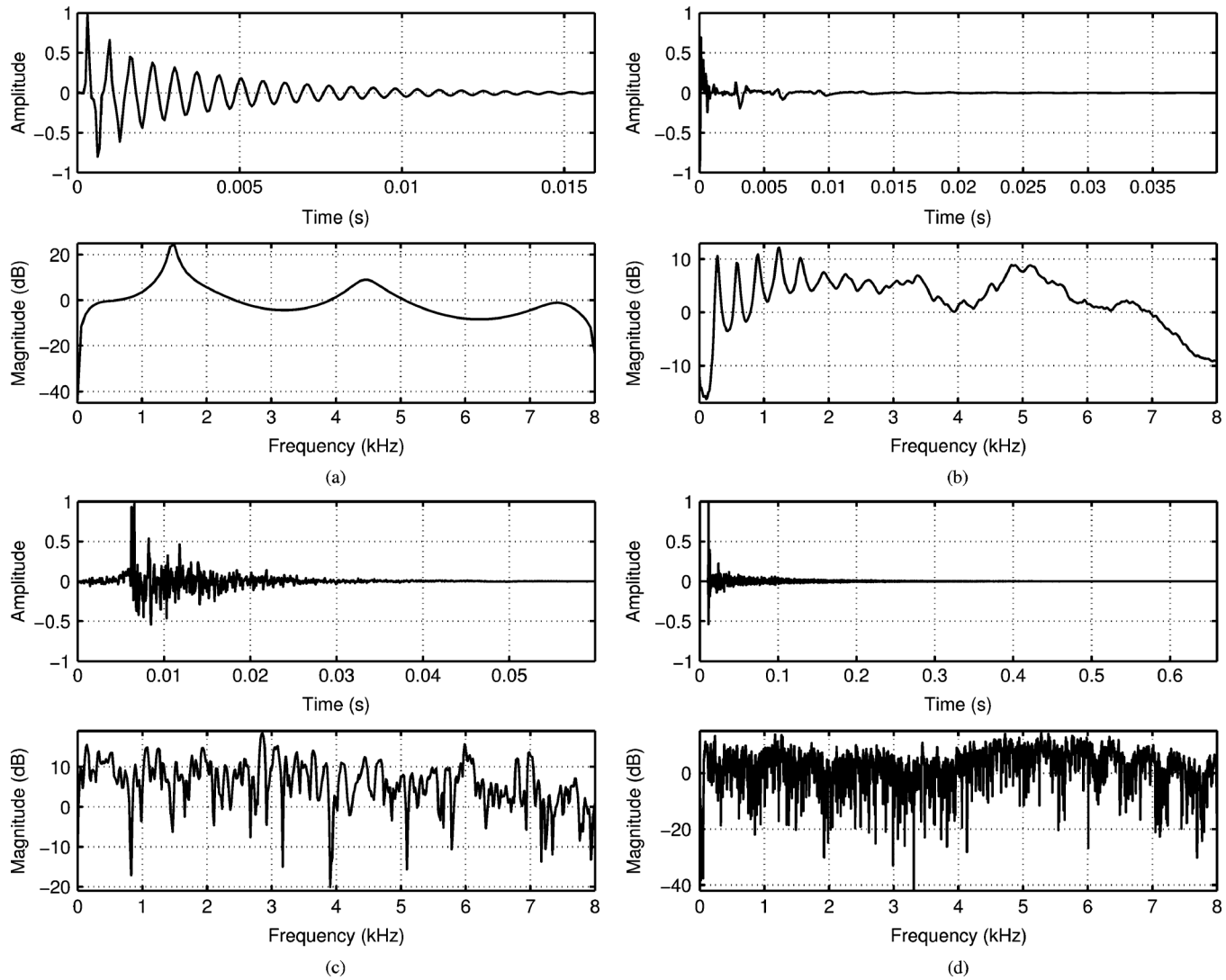


Fig. 4. Acoustic impulse responses and corresponding magnitude spectra. All the impulse responses are truncated at their estimated reverberation time. (a) Occluded microphone. (b) Gramophone horn. (c) Car cabin. (d) Reverberant room.

0.1 and 1 s in 0.1 s increments. Source and microphone positions were changed for each reverberation time keeping the separation distance constant in order to introduce variation in the magnitude responses between the different conditions. The third set includes four measured acoustic channel responses, which were chosen to represent different spectral characteristics: an occluded microphone channel, a gramophone horn, a car cabin and a reverberant room; the impulse responses and their corresponding magnitude spectra are shown in Fig. 4. Additionally, three types of noise were considered: babble noise, car noise and white Gaussian noise at SNRs in the range of -10 dB to 60 dB in steps of 10 dB. The samples containing noise and channel were generated as follows: (i) clean speech was convolved with the impulse response under investigation; (ii) the active speech level was calculated from the reverberant speech according to ITU-T P.56 [22], [24]; (iii) the noise was added with its level adjusted to achieve the desired SNR.

B. Spectral Distance Metric

The proposed channel estimation method was employed to estimate a variety of channel magnitude spectra which were evaluated by comparison with the magnitude spectra of the true channels. Since the estimation is only defined up to a scale factor, the scale-independent Itakura spectral distance defined by [25] was used

$$d_I = \log\left(\frac{1}{K}\sum_{k=0}^{K-1}\frac{|H(k)|^2}{|\hat{H}(k)|^2}\right) - \frac{1}{K}\sum_{k=0}^{K-1}\log\left(\frac{|H(k)|^2}{|\hat{H}(k)|^2}\right). \tag{20}$$

It is interesting to see the relationship of d_I to the widely used log-spectral distance. The four measured channel responses defined in Section V-A were perturbed by adding normally distributed random errors to their magnitude responses to create different levels of log-spectral distance and the resulting Itakura distance was computed in each case. The resulting data points are shown in Fig. 3 together with a quadratic polynomial fit of the points. The log-spectral distance is thus approximately given by $\sqrt{40d_I}$. Following the discussion in Section (II-A), the true channel magnitude response, $|H(k)|$, was convolved with the frequency response of the STFT window before the evaluation. It can be seen that Itakura distances of 0.1–0.5 correspond to

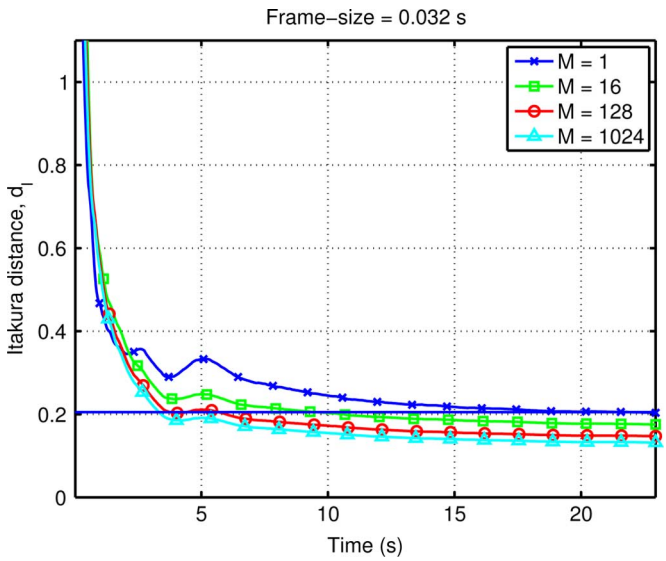


Fig. 5. Effect of increasing the duration of the utterance on channel estimation accuracy for different number of mixtures, M. The horizontal line at $d_I = 0.2$ represents the convergence threshold.

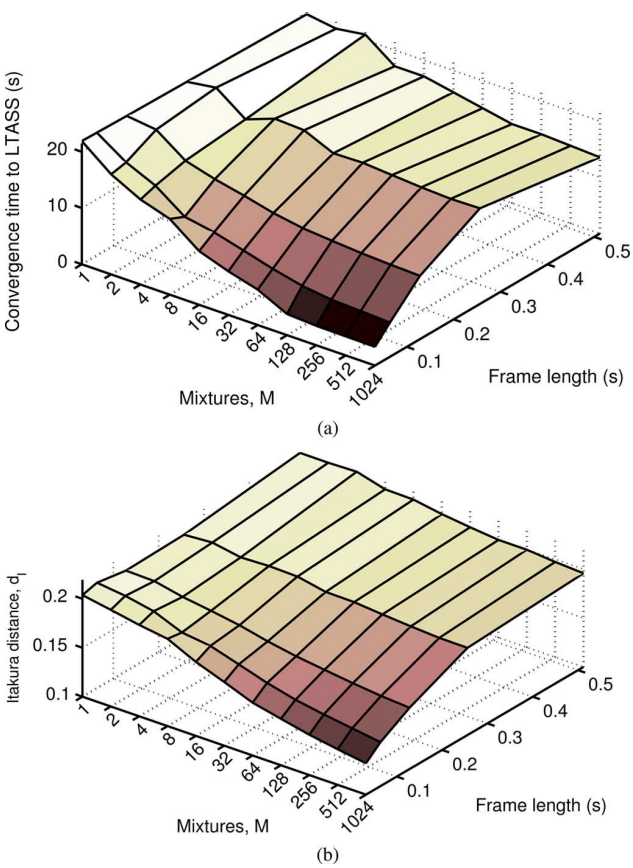


Fig. 6. Effects of frame length and number of mixtures on the convergence time and convergence accuracy. (a) Convergence time vs. number of mixtures and frame length. (b) Convergence accuracy vs. number of mixtures and frame length.

log-spectral distances of 1.5–4.5 dB; this, as will be seen in the following experimental results, is the general operating range of the proposed method.

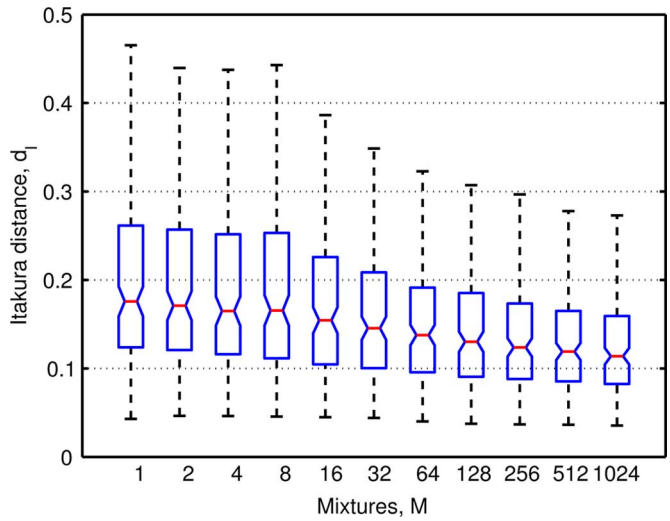


Fig. 7. Variance of estimation accuracy across talkers as a function of number of mixtures. The line in the box is the median, the box edges are the quartiles and the whiskers correspond to approximately ± 2.7 standard deviations, covering 99.3% of the data.

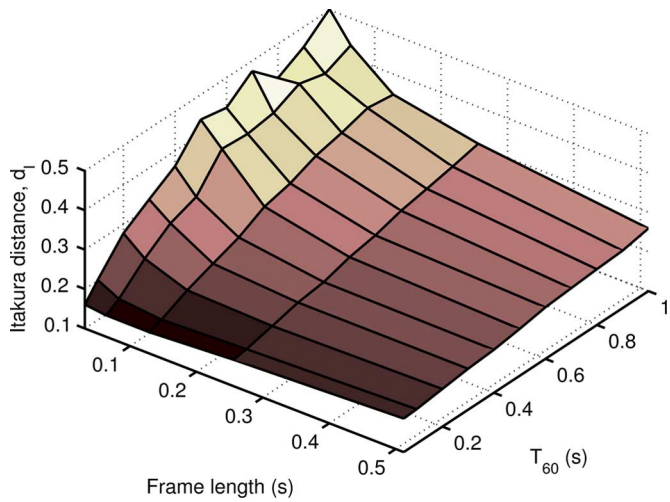


Fig. 8. Estimation accuracy in terms of Itakura distance as a function of reverberation time and frame length.

C. Algorithm Parameters

The key parameters of the algorithm are now investigated including the number of mixtures and the frame length. Frame lengths of 32, 64, 128, 256, 512 ms and $M = 2^m$ mixtures for $m = 0, 1, \ldots, 10$ were considered. As discussed in Section III, the method is equivalent to that using LTASS [13], [12] when $M = 1$, which served as the baseline method for comparison.

The objective was to investigate these parameters in relation to the time required for the estimation and accuracy of the estimates. Using the identity channel, $|H(k)| = 1,\ \forall k$, the channel estimation algorithm was applied directly on clean speech from the TIMIT test set, which should result in an estimate of a flat channel response over all frequencies, $|\hat{H}(k)| = \eta, \forall k$. All combinations of frame length and number of mixtures were considered. In each case, the available speech data for the estimation was increased in increments of half a frame length until the complete length of the utterance was reached. A random offset was applied so that the estimation always initiated with speech present but not necessarily at the beginning of a word;

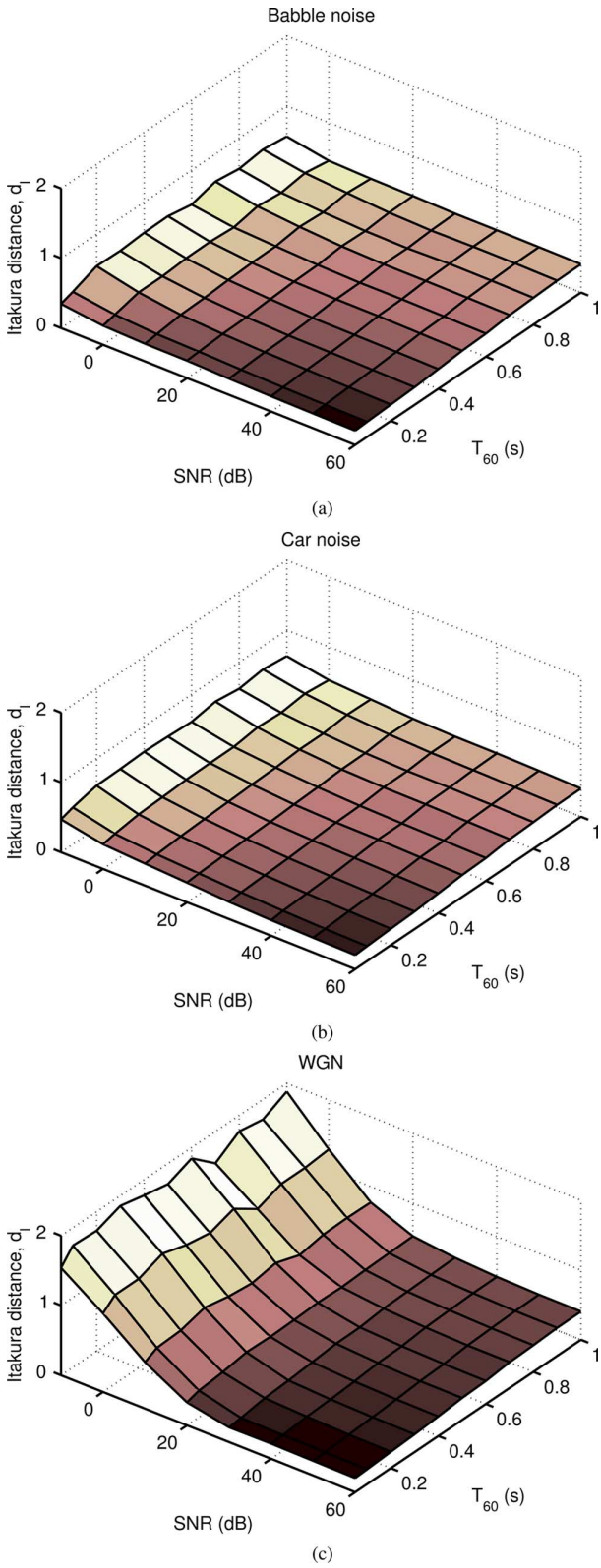


Fig. 9. Channel magnitude estimation in noise—simulated room impulse responses. (a) Babble noise. (b) Car noise. (c) White Gaussian noise.

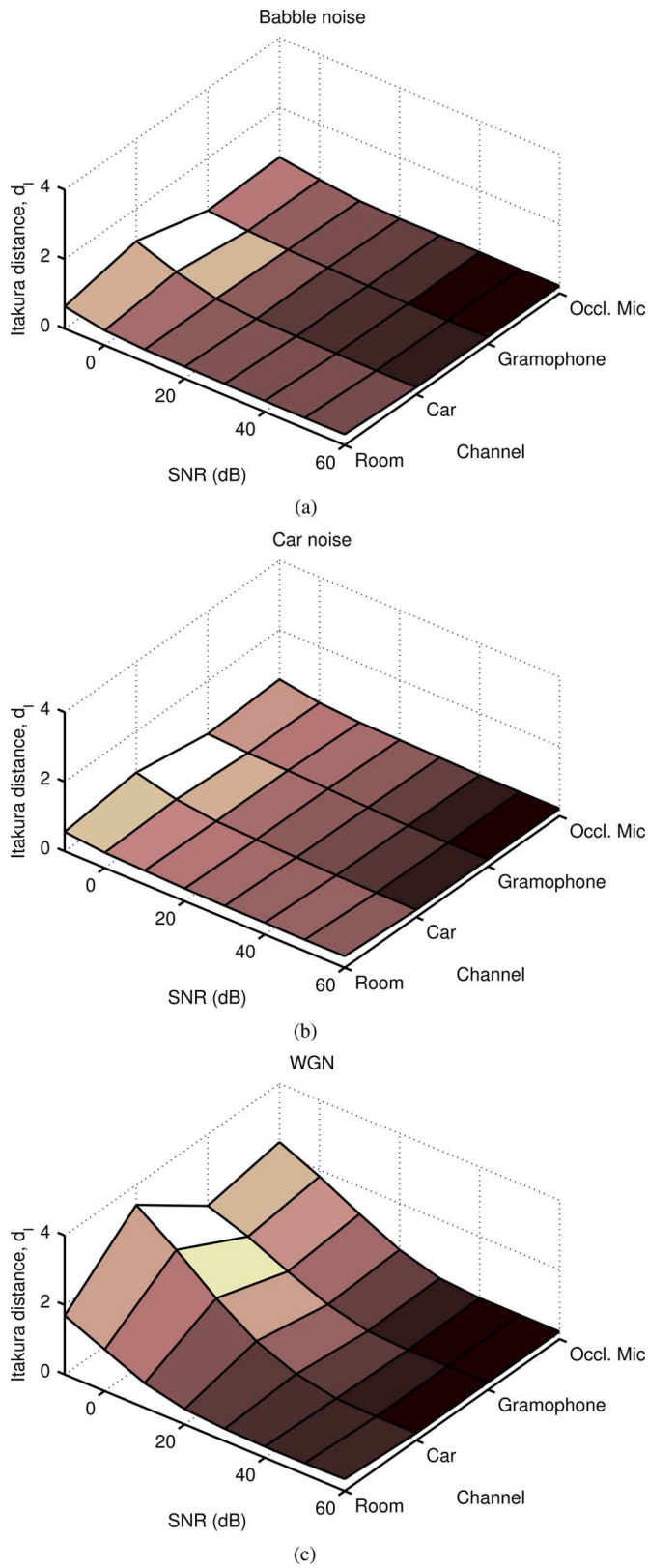


Fig. 10. Channel magnitude estimation in noise—measured responses. (a) Babble noise. (b) Car noise. (c) White Gaussian noise.

this was done in order to provide realistic and fair conditions for the timing evaluation. The Itakura distance was measured at each increment. An example result with a frame length of 32 ms and four different number of mixtures averaged over all talkers is shown in Fig. 5. It can be seen that in all cases the estimation error reduces rapidly during the first 1–4 s before converging to a limit of $d_I \leq 0.2$; the time it takes to reach this limit decreases

with increasing number of mixtures and also the final estimation accuracy at the end of the utterance is improved.

Based on these observations, a robust estimate of the convergence threshold was defined as the 10th percentile of d_I over all speakers and frame lengths; this is indicated in Fig. 5 for the case of one mixture, $M = 1$, by the horizontal line at $d_I \approx 0.2$.

Additionally, the time to reach the convergence threshold for $M = 1$ and the final convergence accuracy for each M and for the different frame lengths were measured. The results are shown in Fig. 6. The following observations can be made:

1) the convergence time decreases with an increased number of mixtures and reaches a lower limit of approximately 2.5 s at $M \geq 256$,
2) there is an improvement in the accuracy of up to $d_I \approx 0.06$ at $M \geq 256$,
3) the improvement in speed and accuracy is smaller as the frame length increases.

These results are intuitively consistent. The improvement in speed and accuracy with increased number of mixtures comes from the fact that a more accurate model of the underlying speech spectrum is used in the estimation at each frame. As the frame length increases the average spectrum for each mixture becomes less distinct and more closely resembles that of the LTASS. Thus, the error increases because of the larger inter-talker variability of the model.

We also investigated the spread of the estimation accuracy across talkers. The results for a frame length of 32 ms are shown in Fig. 7. For each value of M, the plot shows the median, the quartiles and the range of Itakura distances. The spread of the errors becomes significantly smaller for $M \geq 256$ compared to the LTASS based approach, $M = 1$.

In the next experiment, the simulated acoustic impulse responses were used to study the effects of estimation accuracy as a function of the reverberation time, T_{60}, and the frame length. The results are shown in Fig. 8 where we observe that the estimation accuracy decreases significantly if the frame length is less than about half the channel length. For example, at a frame length of 32 ms and $T_{60} = 1$ s the error is $d_I = 0.5$; the Itakura distance is halved as the frame length is increased to 0.5 s and the majority of the error is due to mismatch between the true speech spectra and the average models. On the other hand, a frame length that is longer than the impulse response has little effect on the estimation accuracy.

From the results this far, it appears that $M \geq 256$ is a reasonable choice for the number of mixtures and that the frame length should be as short as possible, both in terms of convergence speed and accuracy. However, as noted above, spectral smoothing will occur if the frame length is substantially shorter than the channel response. Consequently, there is a trade off between the convergence time and the accuracy, which could be controlled depending on the channel order. It was found empirically that a good choice of frame length is 128 ms for many typical practical scenarios such as those used in the following experiments.

D. Blind Channel Estimation in Noise

Based on the results in Section V-C, a GMM with 256 mixtures and frame length of 128 ms was used. The first experiment was conducted with the simulated acoustic impulse responses and the babble, car and white Gaussian noises, as described in Section V-A. The results, averaged over the 168 talkers, are shown in Fig. 9. It can be seen that, an estimation with accuracy of $d_I \leq 0.5$ is obtained for all noise types for $\mathrm{SNR} \geq 10$ dB. The greatest degradation due to noise is observed in the case of white Gaussian noise. This is to be expected, since both car noise and babble noise have similar shape to the long-term average speech spectrum and thus largely cancel out in the estimation procedure. This is due to the relationship between bias and SNR shown in (18): if the noise spectrum is of a similar shape as the long term speech spectrum, $\hat{\epsilon}$ will be close to a constant over all frequencies and will result in a scalar shift of the estimated channel.

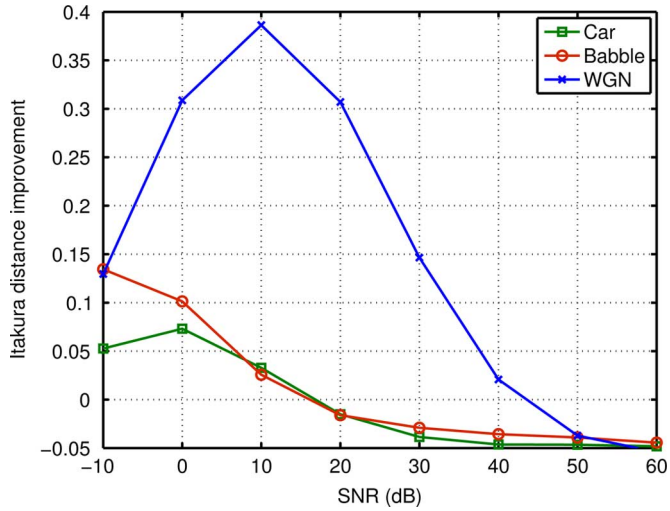


Fig. 11. Estimation improvement in terms of Itakura distance using bias correction compared to direct estimates.

The same experiment was performed using the measured channels in Fig. 4. The results are shown in Fig. 10, and show the same trends as for the simulated acoustic impulse responses.

The results in Figs. 9 and 10 are the outcome of the channel estimation with noise bias reduction from Section IV-A. Fig. 11 shows the gain in Itakura distance for the noisy estimates when using the bias correction compared with estimates without the correction. The results are averaged over all simulated and measured impulse responses. The greatest benefit is seen for white Gaussian noise, which is not surprising since it covers the entire spectrum. For babble and car noise, the procedure is beneficial at $\mathrm{SNR} < 20$ dB. At high SNRs, the procedure slightly reduces the accuracy of channel response estimate because of errors in the SNR estimate.

Finally, Fig. 12 shows an example estimate for each of the four channels estimated from one talker in car noise at $\mathrm{SNR} = 60$ dB and $\mathrm{SNR} = 10$ dB. It can be seen that the estimates at $\mathrm{SNR} = 60$ dB for the occluded microphone, the gramophone horn and the car cabin response are very close to the true channel response $(d_I \leq 0.15)$, while the error is larger $(d_I = 0.28)$ for the room response. This is due to the use of a window that covers only about 1/5 of the impulse response length. Nevertheless, large scale features of the magnitude response are captured even in this case. When the noise level is increased such that the $\mathrm{SNR} = 10$ dB, the estimation becomes worse; this is mainly in

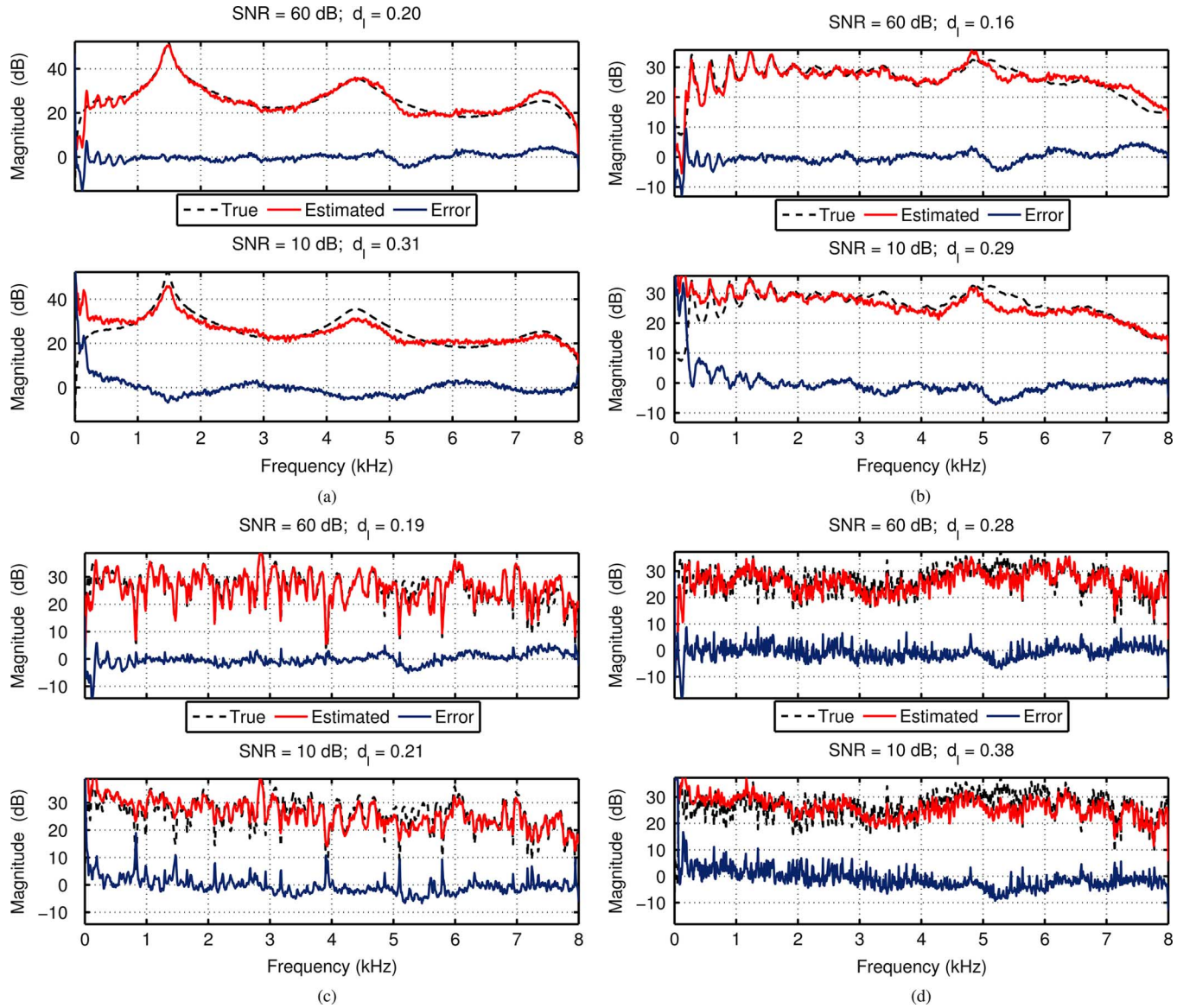


Fig. 12. Example of true and estimated magnitude responses estimated from one talker in car noise at 60 dB SNR (upper plots) and 10 dB SNR (lower plots) for each of four measured impulse responses; the difference between the true and the estimated responses, $\text{Error} = 10\log_{10}\hat{H} - 10\log_{10}H$, is also shown as the lowest trace on each plot. (a) Occluded microphone. (b) Gramophone horn. (c) Car cabin. (d) Reverberant room.

the frequency region below 300 Hz where the noise spectrum is most significant and also where there are deep nulls in the channel responses, where there is little signal above the noise. However, the main features of the channels are identified correctly as desired.

VI. Conclusion

A new classification-based algorithm has been proposed for the identification of the magnitude response of an unknown channel from an observed reverberant single-channel speech signal including additive noise. The algorithm is based on spectrum classification using channel robust RASTA-MFCC features and a GMM-based classifier. There is a trained average spectrum associated with each class that is used for the identification of the channel. A previously published method based on LTASS has been shown to be a special case of our algorithm. In comparison to the baseline of the LTASS-based method, the new algorithm is more accurate and requires up to ten times less data to estimate the channel response. Experimental results using a wide range of simulated and measured acoustic impulse responses showed that channel magnitude responses can be identified with the new method to an accuracy of $d_I \leq 0.5$ for all noise types for $\text{SNR} \geq 10$ dB. Even where there are errors, the large-scale features, including overall spectral shape, spectral peaks and valleys, of the channel responses are identified correctly.

References

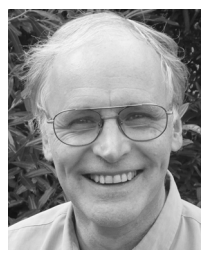
[1] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3937–3946, Dec. 2008.

[2] D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive CODEC identification algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4477–4480.

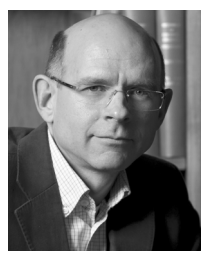
[3] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. New York, NY, USA: Springer, 2010.

[4] J. S. Bendat and A. G. Persol, *Random Data Analysis and Measurement Procedures*. New York, NY, USA: Wiley, 1986.
[5] P. R. White, M. H. Tan, and J. K. Hammond, "Analysis of the maximum likelihood, total least squares and principal component approaches for frequency response function estimation," *J. Sound Vibr.*, vol. 290, pp. 676–689, 2006.
[6] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NY, USA: Prentice-Hall, 2002.
[7] P. Loganathan, A. Khong, and P. Naylor, "A class of sparseness-controlled algorithms for echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1591–1601, Nov. 2009.
[8] T. G. Stockham, T. M. Cannon, and R. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, no. 4, pp. 678–692, Apr. 1975.
[9] Y. Huang, J. Benesty, and J. Chen, "Identification of acoustic MIMO systems: Challenges and opportunities," *Signal Process.*, vol. 6, no. 86, pp. 1278–1295, 2006.
[10] L. Tong and S. Perreau, "Multichannel blind identification: from subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, Oct. 1998.
[11] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 3, pp. 437–466, Sep. 1968.
[12] S. J. Wenndt and A. J. Noga, "Blind channel estimation for audio signals," in *Proc. IEEE Aerospace Conf.*, 2004, vol. 5, pp. 3144–3150.
[13] N. D. Gaubitch, M. Brookes, and P. A. Naylor, "Blind channel identification in speech using the long-term average speech spectrum," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Glasgow, U.K., Aug. 2009.
[14] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 476–488, Sep. 2003.
[15] J. R. Hopgood, C. Evers, and S. Fortune, "Bayesian single channel blind dereverberation of speech form a moving talker," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. New York, NY, USA: Springer, 2010, ch. 8, pp. 219–268.
[16] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[18] N. D. Gaubitch, M. Brookes, P. A. Naylor, and D. Sharma, "Single-microphone blind channel identification in speech using spectrum classification," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Barcelona, Spain, Aug. 2011.
[19] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[20] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.
[21] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[22] D. M. Brookes, VOICEBOX: A speech processing toolbox for MATLAB, [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997
[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[24] ITU-T, "Objective measurement of active speech level," Int. Telecomm. Union (ITU-T) Rec. P.56, Mar. 1993.
[25] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition, ser. Signal Processing Series*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

Nikolay D. Gaubitch obtained the M.Eng. degree in computer engineering from Queen Mary, University of London in 2002 and the Ph.D. degree in acoustic signal processing from Imperial College London in 2007. Between 2007 and 2012 he was a research associate with the Centre for Law Enforcement Audio Research (CLEAR). His research interests span various topics within the field of enhancement of reverberant and noisy speech signals using one ore more microphones. Specifically, he has worked on blind room transfer function estimation and equalization, speech quality and intelligibility estimation and audio processing using ad-hoc microphone arrays. He is currently a postdoctoral researcher at Delft University of Technology.



Mike Brookes is a Reader (Associate Professor) in Signal Processing in the Department of Electrical and Electronic Engineering at Imperial College London. After graduating in Mathematics from Cambridge University in 1972, he spent four years at the Massachusetts Institute of Technology before returning to the UK and joining Imperial College. Within the area of speech processing, he has concentrated on the modeling and analysis of speech signals, the extraction of features for speech and speaker recognition and on the enhancement of poor quality speech signals. Since 2007 he has been the Director of the Home Office sponsored Centre for Law Enforcement Audio Research (CLEAR) which investigated techniques for processing heavily corrupted speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB.



Patrick A. Naylor (M'89–SM'07) received his B.Eng. degree in electronic and electrical engineering from the University of Sheffield, U.K., in 1986 and the Ph.D. degree from Imperial College, London, U.K., in 1990. Since 1990 he has been a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. His research interests are in the areas of speech, audio and acoustic signal processing. He has worked in particular on adaptive signal processing for dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multi-channel speech enhancement and speech production modelling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the UK, USA and in mainland Europe. He is an associate editor of IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING and an associate member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.