SINGLE-MICROPHONE BLIND CHANNEL IDENTIFICATION IN SPEECH USING SPECTRUM CLASSIFICATION

Nikolay D. Gaubitch, Mike Brookes, Patrick A. Naylor, and Dushyant Sharma

Centre for Law Enforcement Audio Research (CLEAR)
Imperial College London, UK

ABSTRACT

We propose an algorithm for blind estimation of the magnitude response of a channel using the observations of a single microphone. The algorithm employs channel robust RASTA filtered Mel-frequency cepstral coefficients as features and a Gaussian mixture model based classifier to generate a dictionary of average speech spectra. These are then used to infer the channel response from speech that has undergone spectral modification in the capturing process. Simulation results using babble noise, car noise and white Gaussian noise are presented, which demonstrate that the proposed method is able to estimate a variety of channel responses to within $3-4$ dB in terms of weighted spectral distance; and it is more accurate than a previously published method.

1. INTRODUCTION

When a speech signal is captured by a microphone positioned at some distance away from the talker, the spectrum of the observed speech will be modified by the transmission channel between talker and microphone. The channel combines the effect of the acoustic environment, the positioning of the the microphone and the characteristics of the sound capturing equipment. In addition, there may be further degradations by the sound capturing equipment and ambient background noise. The overall process can be expressed as

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

where $s(n)$ is the desired speech signal, $h(n)$ are the total channel effects, $\nu(n)$ is additive observation noise and $*$ denotes convolution. Our objective is to identify the magnitude spectrum of the channel $h(n)$ using only the microphone observation $x(n)$.

Channel effects and noise can have detrimental effects on captured speech [1]. These most often reduce the perceptual experience of a listener and may, when the channel and the noise are predominant, result in loss of intelligibility. Furthermore, channels and noise can deteriorate the performance of speech recognition, speaker identification applications. Knowledge of the channel response can be used to enhance observed speech but could be used to obtain information about the recording equipment and the acoustic environment.

Much previous work in blind channel estimation has considered the case of multiple microphones [1]. The single-microphone case is inherently more difficult as no spatial information is available compared with the multi-microphone case. Previous work on single channel identification has considered using a Bayesian framework for parametric single channel estimation [2] and the use of the Long-Term Average Speech Spectrum (LTASS) [3, 4, 5].

The method proposed in this paper can be seen as a generalization of the LTASS method in [3]. Whereas [3] modelled speech using a single long-term average spectrum, our approach divides this into multiple classes of average speech spectra. The log-spectral magnitude of the channel is inferred by subtracting the spectrum in a frame of observed speech from the closest matching template of clean speech average spectrum. The advantage of this approach is that a more accurate estimate of the channel can be obtained at each estimation instant and so, less data is required for the estimation. Furthermore, there will be less variation in the estimation accuracy between speakers and utterances.

The remainder of this paper is organized as follows. The channel estimation algorithm is described in Section 2. Experimental results demonstrating the Gaussian Mixture Model (GMM) based algorithm are given in Section 3 and conclusions are drawn in Section 4.

2. THE BLIND CHANNEL ESTIMATION ALGORITHM

The algorithm consists of two stages:

1. training of the GMM to derive K classes of average speech spectra and
2. inference of an unknown channel.

First, the general principle behind the algorithm is introduced and then each of the two stages is described in detail.

2.1 Preliminaries

It is customary to process speech in the frequency domain using short overlapping frames and, accordingly, from (1) we can write

$$X_l(k) = S_l(k)H_l(k) + V_l(k), \tag{2}$$

where $A_l(k)$ denotes the Short-Time Fourier Transforms (STFTs) of the lth frame of $a(n)$ and k is the frequency bin index. Furthermore, under the assumption that the signal and the noise magnitudes and phases are independent, and that the channel is stationary or varies much slower than the speech, we can write

$$E\{|X_l(k)|^2\} = E\{|S_l(k)|^2\}|H(k)|^2 + E\{|V_l(k)|^2\}, \tag{3}$$

where $E\{\cdot\}$ denotes expectation. In the noise-free case, $V_l(k) = 0, \forall l$, and if we have some prior knowledge of the spectrum of the speech signal, $|S_l(k)|$, we can obtain an estimate of the magnitude spectrum of the channel channel with

$$|\hat{H}_l(k)| = \exp\left(Z_{X,l}(k) - \hat{Z}_{S,l}(k)\right), \tag{4}$$

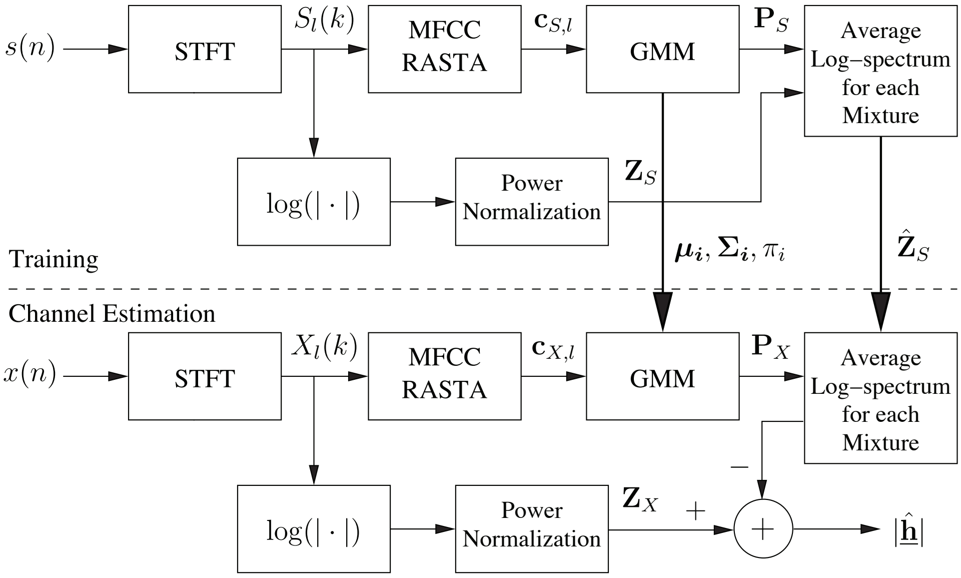

Figure 1: System diagram of the proposed algorithm.

where $Z_{A,l} = \log(|A_l(k)|)$, $\hat{A}_l(k)$ denotes an estimate of $A_l(k)$, and we have considered the instantaneous estimates of the expectations in (3).

The problem is then how to find $\hat{Z}_{S,l}(k)$. A solution that has been considered is to use the LTASS [3, 4]. This has shown to give approximate channel estimations but requires long and phonetically balanced observed utterances for the LTASS approximation to become valid. Instead, we propose using a classifier in order to find a finer grid of average speech spectra that are closely related to the different sounds in speech. In this case, more precise estimates are obtained for each frame of speech, resulting in more rapid and more accurate channel estimates.

2.2 Classes of average speech spectra

The first stage in the algorithm is to derive the K classes of average spectra following the procedure shown in the upper half of Fig 1. In order to do this, we need a suitable set of features to represent a frame of the speech spectrum and a classification algorithm.

Ideally, the features should not be affected by the channel; the reason for this will be clarified further in Section 2.3. A suitable candidate are the RASTA filtered Mel-Frequency Cepstral Coefficients (MFCC-RASTA), which were developed as robust features for speech recognition. RASTA processing performs bandpass filtering in the cepstral domain in order to reduce channel effects [6]. Thus, for the lth frame of clean speech we obtain a feature vector $\mathbf{c}_{S,l}$ of $N_\mathbf{c}$ MFCC-RASTA coefficients.

We use the feature vectors, $\mathbf{c}_{S,l}$, to train a K-mixture GMM [7] to obtain the means, μ_i, diagonal covariances, Σ_i, and weights, π_i, of each mixture. An important component that we will use are the relative mixture probabilities – the probability that feature vector $\mathbf{c}_{S,l}$ belongs to the ith mixture– defined as [7]

$$p(z_i = 1 \mid \mathbf{c}_{S,l}) = \frac{\pi_l \mathcal{N}(\mathbf{c}_{S,l} \mid \mu_i, \Sigma_i)}{\sum_{j=1}^{K} \pi_l \mathcal{N}(\mathbf{c}_{S,l} \mid \mu_j, \Sigma_j)}, \qquad (5)$$

where $z_i \in \{0,1\}$.

Subsequently, we form a $K \times L_S$ matrix, $\mathbf{P}_S$, comprising the K mixture probabilities for the L_S available speech frames and an $L_S \times N_{FT}$ matrix, $\mathbf{Z}_S$, with the short-term log-spectra of the speech for each frame, where N_{FT} defines the number of points in the short-term Fourier transform. In order to avoid issues with signal level differences that may arise in the identification process, the log-spectra are all normalized by subtracting their mean according to

$$\bar{Z}_{S,l}(k) = Z_{S,l} - \frac{1}{N_{FT}} \sum_{k=0}^{N_{FT}-1} Z_{S,l}(k), \quad \forall l. \qquad (6)$$

This process is not to be confused with cepstral mean subtraction which aims to neutralize the channel; this normalization only affects the log-spectral magnitude.

Finally, we combine $\mathbf{P}_S$ and $\bar{\mathbf{Z}}_S$ to perform a weighted average of the short-term log spectra and to obtain a set of K average speech spectra

$$\hat{\mathbf{Z}}_S = \mathbf{P}_S \bar{\mathbf{Z}}_S, \qquad (7)$$

where $\hat{\mathbf{Z}}_S$ is a $K \times N_{FT}$ matrix whose ith row represents the average log-spectrum corresponding to the ith mixture.

2.3 Channel estimation

The channel estimation stage is shown in the lower half of Fig. 1. We now make use of the GMM parameters, together with $\hat{\mathbf{Z}}_S$ from (7) to estimate the unknown channel. In a similar fashion to the procedure of Section 2.2, we begin by forming a set of feature vectors, $\mathbf{c}_{X,l}$, from the observed speech signals, $x(n)$. Note, that this is different from Section 2.2 where the features were extracted from clean speech. Next, the features and the GMM parameters are used with (5) to find the relative probabilities for each feature vector from the observed speech and a relative probability matrix, $\mathbf{P}_X$, is formed. We create an $L_X \times N_{FT}$ matrix $\bar{\mathbf{Z}}_X$ with normalized short-term log-spectra of the observed speech for each frame.

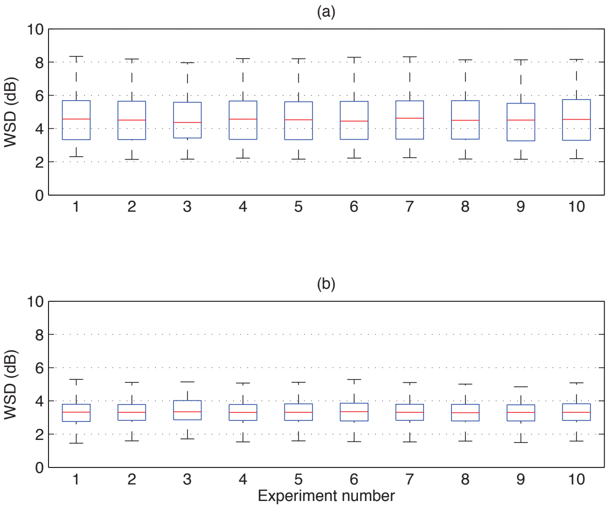

Figure 2: Ten different channels comprising one conjugate pair of poles and one conjugate pair of zeros. Channel estimates for (a) LTASS based estimation and (b) GMM based estimation using 80 utterances. The line in the box is the median, the box edges are the 25th and the 75th percentiles and the whiskers correspond to approximately ±2.7 standard deviations.

The magnitude spectrum of the channel can be estimated up to an unknown scale factor using

$$|\underline{\hat{\mathbf{h}}}| = \exp\left(\frac{\left[\bar{\mathbf{Z}}_X - \mathbf{P}_X^T\hat{\mathbf{Z}}_S\right]^T \mathbf{1}}{L_X}\right), \tag{8}$$

where $|\underline{\hat{\mathbf{h}}}| = [|\hat{H}(0)|\ |\hat{H}(1)|\ \ldots\ |\hat{H}(N_{FT})|]$, superscript ^{T} denotes matrix transpose and $\mathbf{1}$ is a $L_x \times 1$ vector of with all elements equal to one. Here, $\mathbf{P}_X$ acts as a selection matrix and generates a weighted average spectrum based on the class probabilities and the average spectrum templates of clean speech.

3. EXPERIMENTAL RESULTS

In this section, we present a set of experimental results to demonstrate some of the abilities of the channel estimation algorithm. The performance is also compared to the LTASS based algorithm from [3].

The weighted root mean squared log-spectral distance described in [3] is used as a quantitative measure to evaluate the estimated channels. The metric compares two power spectra $P_1(k)$ and $P_2(k)$ according to

$$\text{WSD} = \left[\frac{\sum_{k=0}^{N-1} W(k)|e(k)|^2}{\sum_{k=0}^{N-1} W(k)}\right]^{\frac{1}{2}} \text{dB}, \tag{9}$$

where

$$e(k) = 10\log_{10}\left(\frac{P_1(k)}{P_2(k)}\right), \tag{10}$$

and $W(k)$ is a frequency dependent weight function, which combines A-weighting and LTASS. Thus, the weighting utilizes properties of speech signals and of human hearing.

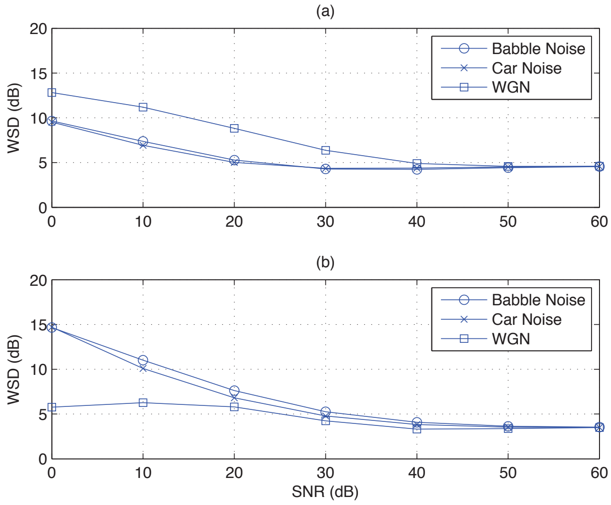

Figure 3: Estimation of a single pole-pair channel in noise using (a) LTASS based method. and (b) GMM based method.

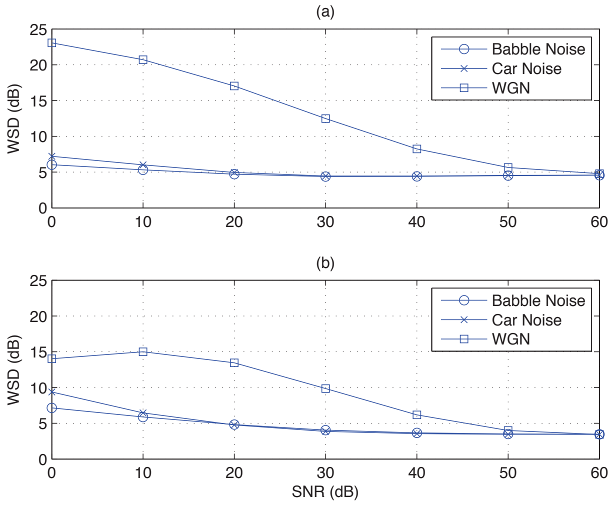

Figure 4: Estimation of a single zero-pair channel in noise using (a) LTASS based method. and (b) GMM based method.

For all experiments, data is drawn from the TIMIT database. The TIMIT training set constitutes anechoic, noise-free recordings from 422 male and 184 female talkers, with ten sentences from each talker; the duration of each sentence is approximately three seconds and the sampling frequency is $f_s = 16$ kHz. Following the procedure described in Section 2.2, we used the full training set of TIMIT to train the GMM and to define the classes of average log-spectra. The same set was also used to calculate the LTASS as in [3]. We used 32 ms frames overlapping by 50% and multiplied by a Hanning window. From each frame we calculated $N_\mathbf{c} = 12$ MFCC-RASTA coefficients and used these to train a GMM of $K = 1024$ mixtures.

The core TIMIT test set was used for the channel estimation examples; it consists of 240 sentences (including the dialect sentences), ten sentences each from 16 male and 8 female talkers. In this way, we use different data from that used for training the GMM and for estimation of the LTASS. The sentences were concatenated using three sentences to create one utterance, which results in a total of 80 utterances of male and female talkers.

In the first experiment, we used ten randomly generated channels comprising one conjugate pole pair and one conjugate zero pair without additive noise. The results are shown in the box plot in Fig. 2 (a) for the LTASS based approach and (b) for the GMM based method. It can be seen from this plot that the variation in estimation accuracy for the different utterances is greatly reduced by the GMM method and also the overall accuracy is improved by about 2 dB. This is to be expected since the GMM based approach defines a a much finer grid of average speech spectra. There is little dependence on the channel for both algorithms.

Next, we generated two fixed channels, one with a conjugate pair of poles and one with a conjugate pair of zeros. We then added noise to the filtered signals, varying the SNR between 0 and 60 dB. Three different types of noise were considered: babble noise, car noise and White Gaussian Noise (WGN). Figure 3 shows the channel identification results in terms of weighted spectral distance for the single-pole channel and Fig. 4 shows the results obtained from the single-zero channel. It can be seen from these results that the GMM based method is more robust to noise compared to the LTASS based method, for SNR > 10 dB. The most notable improvement covering the full range of SNRs is for WGN. WGN flattens the average spectrum of the observed speech and violates the LTASS assumption – the channel estimate will tend to an inverted LTASS as SNR $\rightarrow -\infty$. On the other hand the GMM based method is able to select more appropriate spectra for this and, thus, reduces the error. However, a more thorough study of the classification errors caused by the noise and the effects this has on the channel estimation errors are required and are left as future work.

Finally, we show an illustrative example with a real measured channel. The objective is to demonstrate the performance of the algorithm with realistic data and to relate the numbers of the weighted spectral distance to a typical estimation example. The measured microphone response was convolved with the clean speech samples. The true and the estimated channels are shown in Fig. 5. The weighted spectral distance in this example is 5.5 dB for the estimation using the LTASS based approach and 3.1 dB for the estimation with the GMM based algorithm. We see that the important large scale components (the position of the three poles in this case) have been identified correctly.

4. CONCLUSIONS

We have developed an algorithm for blind identification of the magnitude spectrum of a stationary or slowly time-varying channel. The key principle of the algorithm is the classification of average log-spectra of clean speech with a Gaussian Mixture Model (GMM) and MFCC-RASTA features. The GMM is then used with frames of speech that has undergone a channel to find the best matching template of clean speech from which the former is subtracted. In this way, the remainder after the subtraction is the unknown channel.

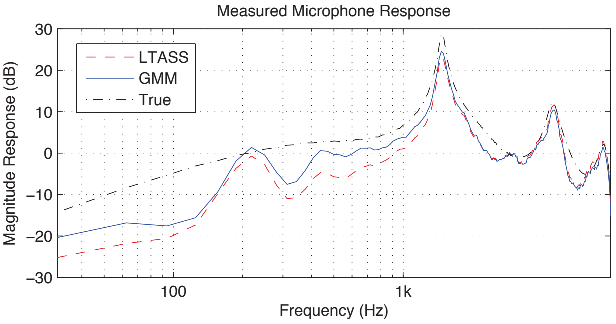

Figure 5: Estimates of a measured microphone response in the noise-free case. The WSD is 5.5 dB with LTASS based estimation and 2.9 dB with GMM based estimation.

The operational properties of the algorithm were demonstrated using a variety of simulated channels and a measured channel comprising a microphone response. Based on the current results, the proposed GMM based approach showed lower estimation variance from different utterances compared to an earlier LTASS based algorithm. The simulations also considered the performance in various types of additive noise, where the results indicated that the performance is much dependent on the spectral properties of the noise and also of the channel. However, the GMM based approach was generally more robust to noise compared to the LTASS method and significantly so in the case of white Gaussian noise. Overall, the estimation accuracy of the method in the noise-free case is in the range of $3-4$ dB in terms of a weighted log-spectral distance.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[2] 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.

[3] N. D. Gaubitch, M. Brookes, and P. A. Naylor, “Blind channel identification in speech using the long-term average speech spectrum,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Aug. 2009.

[4] S. J. Wenndt and A. J. Noga, “Blind channel estimation for audio signals,” in *IEEE Aerospace Conf.*, vol. 5, 2004, pp. 3144–3150.

[5] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. E. Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, , T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, “An international comparison of long-term average speech spectra,” *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

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

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.