# Model based Estimation of STP parameters for Binaural Speech Enhancement

Kavalekalam, Mathew Shaji; Nielsen, Jesper Kjær; Christensen, Mads Græsbøll; Boldt, Jesper Bünsow

*Published in:*
2018 26th European Signal Processing Conference (EUSIPCO)

*DOI (link to publication from Publisher):*
10.23919/EUSIPCO.2018.8553145

*Publication date:*
2018

*Document Version*
Accepted author manuscript, peer reviewed version

# Model based Estimation of STP parameters for Binaural Speech Enhancement

Mathew Shaji Kavalekalam, Jesper Kjær Nielsen,
Mads Græsbøll Christensen
*Audio Analysis Lab, CREATE*
*Aalborg University*
Aalborg, Denmark
{msk, jkn, mgc}@create.aau.dk

Jesper Boldt
*GN Hearing*
Ballerup, Denmark
jboldt@gnresound.com

***Abstract*—This paper deals with the estimation of the short-term predictor (STP) parameters of speech and noise in a binaural framework. A binaural model based approach is proposed for estimating the power spectral density (PSD) of speech and noise at the individual ears for an arbitrary position of the speech source. The estimated PSDs can be subsequently used for enhancement in a binaural framework. The experimental results show that taking into account the position of the speech source using the proposed method leads to improved modelling and enhancement of the noisy speech.**

***Index Terms*—autoregressive modelling, binaural speech enhancement**

## I. Introduction

Understanding of speech in difficult listening situations like cocktail party scenarios is a major issue for the hearing impaired. Speech enhancement capabilities of a hearing aid (HA) in such scenarios have been observed to be limited. Generally, a hearing impaired person is fitted with HAs at both ears. With the recent developments in HA technology, HAs are able to communicate with each other through a wireless link and share information. This enables binaural processing of signals. Binaural processing of noisy signals has shown to be more effective than processing the noisy signal independently at each ear [1]. Some binaural speech enhancement algorithms with multiple microphones present in each hearing aid have been previously proposed in [2], [3].

However, in this work we are concerned with binaural speech enhancement algorithms with access to only one microphone per HA. This is obseved in in-the-ear (ITE) HAs, where the space constraints limit the number of microphones per HA. Some of the existing algorithms with a single microphone present in each hearing aid are [4]–[6]. These algorithms perform the enhancement in the frequency domain by assuming that the speech and noise components are uncorrelated, and do not take into consideration the dynamics of speech production process. It was recently proposed in [7], [8] to perform binaural enhancement of speech while taking into account the speech production model. The filter parameters here consists of the STP parameters of speech and noise. STP parameters constitute of the autoregressive (AR) parameters representing the spectral envelope and the excitation variance corresponding to the gain of the envelope. These parameters can be used to parametrically model the speech and noise PSDs at the individual ears. The estimation of these filter parameters in [7], [8] assumed that the speaker source is in the nose direction of the listener. Due to this assumption, the speech PSDs at the two ears were modelled in [7], [8] using the same set of STP parameters. This type of modelling might not be appropriate if the speaker is not in the nose direction. This scenario is of interest, as it has been observed in [9], [10], that the Speech Reception Threshold (SRT) is not always the minimum when the speaker is in the nose direction. It was noticed that the listeners often tend to orient their head away from the speech source for an improvement in the SRT. Thus, in this paper, we propose a method to take the position of the speaker into account while estimating the speech and noise PSDs at the two ears. This leads to the estimation of individual speech PSDs for the two ears. A codebook based approach, which takes into account the a priori information regarding the speech and noise AR spectral envelopes is proposed to estimate the STP parameters. The method proposed in this paper uses a multiplicative update method [11] commonly used in non-negative matrix factorisation (NMF) applications [12] to estimate the gain parameters corresponding to the speech and noise AR processes.

The remainder of the paper is structured as follows. Section II motivates the problem and also introduces the signal model used in the paper. Section III explains the proposed method of estimating the speech and noise STP parameters in detail. Experiments and results are presented in Section IV followed by conclusion in Section V.

## II. Motivation

In this section we introduce the signal model and motivate this work. The binaural noisy signals at the left/right ear, denoted by $z_{l/r}(n)$ is written as

$$z_{l/r}(n) = s_{l/r}(n) + w_{l/r}(n) \qquad \forall n = 0, 1, 2 \ldots, \quad (1)$$

where $s_{l/r}(n)$ is the clean speech component and $w_{l/r}(n)$ is the noise component. A very popular way to represent the

This work was supported by Innovations fund Denmark (Grant no: 99-2014-1).

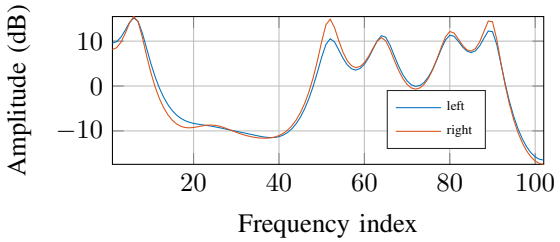


Fig. 1: Gain normalised spectral envelopes for the left and right channel

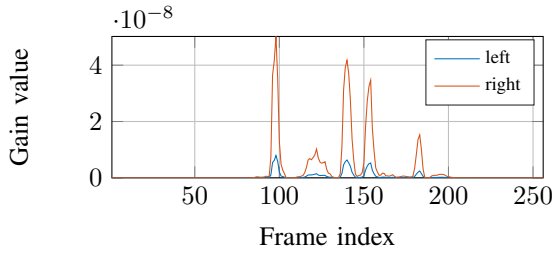


Fig. 2: Plot of the excitation variances for the left and right channel

clean speech component is in the form of an AR process. In [7], [8], it was assumed that the target speaker is located in the nose direction of the listener. Due to this assumption, the clean speech component at both ears were represented using AR processes having the same set of STP parameters. This modelling is reasonable as long as the speaker is in the nose direction of the listener. However, it might not be an appropriate model for the case when speaker is not present in the nose direction. Here, we have conducted a few simulations to show the properties of the parameters corresponding to the speech component present at the left and right microphones. The speaker position is set to be 40 degree right of the listener at a distance of 80 cm. Fig. 1 shows a snapshot of the gain normalised spectral envelopes for the left and right channel. It can be seen that the gain normalised spectral envelopes at the left and right channels have approximately the same content. In comparison to the AR spectral envelopes, it can be seen from Fig. 2, that there is considerable difference in the excitation variances between the left and right channels. This can be explained due to the head shadowing effect, which leads to an attenuation of the intensity at the ear on the far side (left ear in this case). Motivated by these observations in figures 1 and 2, we model the speech component at the left and right ears using the same spectral envelope but different excitation variances as

$$s_{l/r}(n) = \Big(\sum_{i=1}^{P} a_i s_{l/r}(n-i)\Big) + u_{l/r}(n), \tag{2}$$

where $\{a_i\}_{i=1}^{P}$ is the set of speech AR parameters and $u_{l/r}(n)$ is white Gaussian noise (WGN) with zero mean and excitation variance $\sigma^2_{u_{l/r}}(n)$. It is also assumed that the noise component at both ears have similar spectral shape. This is due to the diffuse noise field assumption. The noise components can be similarly expressed as an AR process of order $Q$ as follows,

$$w_{l/r}(n) = \Big(\sum_{i=1}^{Q} b_i w_{l/r}(n-i)\Big) + v(n). \tag{3}$$

where $\{b_i\}_{i=1}^{Q}$ is the set of noise AR parameters and $v(n)$ is white Gaussian noise (WGN) with zero mean and excitation variance $\sigma^2_v(n)$. STP parameters corresponding to speech and noise are considered to be constant over a duration of 25ms.

## III. Model Based Estimation of STP Parameters

The speech and noise STP parameters required for the enhancement are estimated frame-wise using a codebook based approach [7], [13]. The estimation of these parameters uses a priori information about the speech and noise spectral envelopes present in trained codebooks in the form of Linear Prediction Coefficients (LPC). These trained parameters offers us an elegant way to take into account prior information regarding the noise type and speaker of interest. Here, we use a Bayesian framework for estimating the STP parameters. The random variables (r.v) corresponding to the parameters to be estimated are represented as $\boldsymbol{\theta} = [\boldsymbol{\theta}_s\ \boldsymbol{\theta}_w] = [\mathbf{a}; \sigma^2_u; \mathbf{b}; \sigma^2_v; c]$, where $\mathbf{a}, \mathbf{b}$ corresponds to r.v representing the speech and noise AR parameters, $\sigma^2_u, \sigma^2_v$ representing the speech and noise excitation variances and $c$ corresponds to the scale parameter that relates to the excitation variance between the left and right ear *i.e.* $\sigma^2_{u_l} = \sigma^2_u$ and $\sigma^2_{u_r} = c \times \sigma^2_u$. In this work, scale parameter is considered time varying, to take into account the changes in speaker position. Fig. 3 shows a basic block diagram of the enhancement framework, where it can be seen that the STP parameters are estimated jointly using the information at the left and right channels. Thus, the MMSE estimate of the parameter vector

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{z}_l, \mathbf{z}_r) = \int_{\Theta} \boldsymbol{\theta} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{z}_l, \mathbf{z}_r)} d\boldsymbol{\theta}, \tag{4}$$

where $\mathbf{z}_l$ and $\mathbf{z}_r$ is a frame of length $N$ of noisy speech at the left and right ears respectively. Let us define $\boldsymbol{\theta}^{\mathrm{ML}}_{ij} = [\mathbf{a}_i; \sigma^{2,\mathrm{ML}}_{u,ij}; \mathbf{b}_j; \sigma^{2,\mathrm{ML}}_{v,ij}; c^{\mathrm{ML}}_{ij}]$ where $\mathbf{a}_i$ is the $i^{th}$ entry of speech codebook (of size $N_s$), $\mathbf{b}_j$ is the $j^{th}$ entry of the noise codebook (of size $N_w$) and $\sigma^{2,\mathrm{ML}}_{u,ij}, \sigma^{2,\mathrm{ML}}_{v,ij}$ and $c^{\mathrm{ML}}_{ij}$ represents the maximum likelihood (ML) estimates of the excitation variances and the scale parameter respectively for the $ij$ th

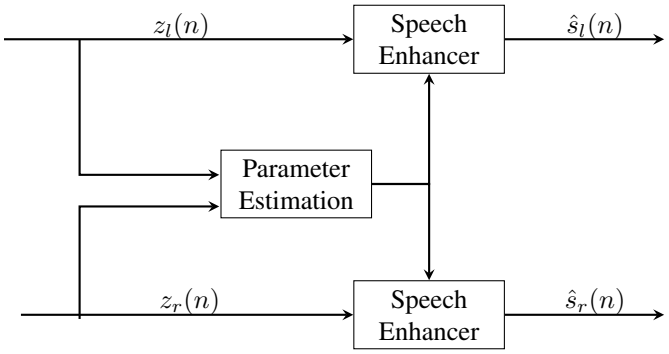


Fig. 3: Basic block diagram of the binaural enhancement framework

combination of the codebook entries. Using the above definition, (4) is approximated as [13]

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{N_s}\sum_{j=1}^{N_w} \boldsymbol{\theta}_{ij}^{\mathrm{ML}} \frac{p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})p(\boldsymbol{\theta}_{ij}^{\mathrm{ML}})}{p(\mathbf{z}_l, \mathbf{z}_r)}, \tag{5}$$

where the MMSE estimate is expressed as a weighted linear combination of $\boldsymbol{\theta}_{ij}^{\mathrm{ML}}$ with weights proportional to $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$. It is assumed that the left and right noisy signal are conditionally independent given $\boldsymbol{\theta}_{ij}^{\mathrm{ML}}$, which leads to $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$ being written as,

$$p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}}) = p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}}). \tag{6}$$

As the scale term is not used for modelling the spectrum at the left ear the likelihood $p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$ is expressed as

$$p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\mathrm{ML}}) = p(\mathbf{z}_l|[\mathbf{a}_i; \sigma_{u,ij}^{2,\mathrm{ML}}; \mathbf{b}_j; \sigma_{v,ij}^{2,\mathrm{ML}}]). \tag{7}$$

Similarly $p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$ is expressed as

$$p(\mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}}) = p(\mathbf{z}_r|[\mathbf{a}_i; c_{ij}^{\mathrm{ML}} \times \sigma_{u,ij}^{2,\mathrm{ML}}; \mathbf{b}_j; \sigma_{v,ij}^{2,\mathrm{ML}}]) \tag{8}$$

Logarithm of the likelihood $p(\mathbf{z}_l|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$ can be written as being proportional to the negative of Itakura-Saito (IS) divergence between the noisy periodogram at the left ear $P_{z_l}(k)$ and the modelled noisy spectral envelope $\hat{P}_{z_{l,ij}}^{\mathrm{ML}}(k)$, where $k$ corresponds to the frequency index [13]. Using the same result for the right ear, $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij}^{\mathrm{ML}})$ can be written as

$$p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij}^{\mathrm{ML}}) = K \exp\Bigg( -\frac{N}{2}\Big( d_{\mathrm{IS}}\big[P_{z_l}(k), \hat{P}_{z_{l,ij}}^{\mathrm{ML}}(k)\big] + d_{\mathrm{IS}}\big[P_{z_r}(k), \hat{P}_{z_{r,ij}}^{\mathrm{ML}}(k)\big]\Big)\Bigg) \tag{9}$$

where $\hat{P}_{z_{l,ij}}^{\mathrm{ML}}(k)$ and $\hat{P}_{z_{r,ij}}^{\mathrm{ML}}(k)$ are denoted as

$$\hat{P}_{z_{l,ij}}^{\mathrm{ML}}(k) = \frac{\sigma_{u,ij}^{2,\mathrm{ML}}}{|A_s^i(k)|^2} + \frac{\sigma_{v,ij}^{2,\mathrm{ML}}}{|A_w^j(k)|^2}, \tag{10}$$

$$\hat{P}_{z_{r,ij}}^{\mathrm{ML}}(k) = \frac{c_{ij}^{\mathrm{ML}}\sigma_{u,ij}^{2,\mathrm{ML}}}{|A_s^i(k)|^2} + \frac{\sigma_{v,ij}^{2,\mathrm{ML}}}{|A_w^j(k)|^2} \tag{11}$$

and $1/|A_s^i(k)|^2$ is the spectral envelope corresponding to the $i^{th}$ entry of the speech codebook, $1/|A_w^j(k)|^2$ is the spectral envelope corresponding to the $j^{th}$ entry of the noise codebook. For a particular combination of the speech and noise codebook entries, the ML estimates of the excitation variances are estimated by maximising $p(\mathbf{z}_l, \mathbf{z}_r|\boldsymbol{\theta}_{ij})$. This is equivalent to minimising the total IS distortion as seen in (9) given by

$$T_{\mathrm{IS}} = d_{\mathrm{IS}}[P_{z_l}(k), \hat{P}_{z_{l,ij}}(k)] + d_{\mathrm{IS}}[P_{z_r}(k), \hat{P}_{z_{r,ij}}(k)], \tag{12}$$

where $\hat{P}_{z_{l,ij}}$ and $\hat{P}_{z_{r,ij}}$ has the same form as in (10) and (11). Here, we use a multiplicative method to estimate the excitation variances and scale term that leads to a minimisation of the cost function in (12). In the multiplicative update method, the value of the variable at $(l+1)^{\text{th}}$ iteration is computed by multiplying the value of the variable at $l^{th}$ iteration with the ratio between the negative component of the the gradient and the positive component of the gradient, which is mathematically written as [12], $\phi^{l+1} \leftarrow \phi^l \frac{\nabla f(\phi^l)_-}{\nabla f(\phi^l)_+}$, where $\phi$ is the variable of interest. Taking the derivative of (12) with respect to speech and noise excitation variances, and the scaling term $c$, we get

$$\frac{\partial T_{\mathrm{IS}}}{\partial \sigma_{u,ij}^2} = \frac{1}{N}\sum_{k=1}^{N} \frac{\frac{1}{|A_s^i(k)|^2}}{\hat{P}_{z_{l,ij}}(k)} - \frac{\frac{P_{z_l}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{l,ij}}(k)^2} + \frac{\frac{c_{ij}}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{c_{ij}P_{z_r}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \tag{13}$$

$$\frac{\partial T_{\mathrm{IS}}}{\partial \sigma_{v,ij}^2} = \frac{1}{N}\sum_{k=1}^{N} \frac{\frac{1}{|A_w^j(k)|^2}}{\hat{P}_{z_{l,ij}}(k)} - \frac{\frac{P_{z_l}(k)}{|A_w^j(k)|^2}}{\hat{P}_{z_{l,ij}}(k)^2} + \frac{\frac{1}{|A_w^j(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{P_{z_r}(k)}{|A_w^j(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \tag{14}$$

$$\frac{\partial T_{\mathrm{IS}}}{\partial c_{ij}} = \frac{1}{N}\sum_{k=1}^{N} \frac{\frac{\sigma_{u,ij}^2}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)} - \frac{\frac{\sigma_{u,ij}^2 P_{z_r}(k)}{|A_s^i(k)|^2}}{\hat{P}_{z_{r,ij}}(k)^2} \tag{15}$$

Using the multiplicative update rule, the values for the excitation noise variances are computed iteratively as shown below

$$\sigma_{u,ij}^{2(l+1)} \leftarrow \sigma_{u,ij}^{2(l)} \frac{\sum_{k=1}^{N} \frac{P_{z_l}(k)}{|A_s^i(k)|^2\hat{P}_{z_{l,ij}}(k)^2} + \frac{c_{ij}^{(l)}P_{zr}(k)}{|A_s^i(k)|^2\hat{P}_{z_{r,ij}}(k)^2}}{\sum_{k=1}^{N} \frac{1}{|A_s^i(k)|^2\hat{P}_{z_{l,ij}}(k)} + \frac{c_{ij}^{(l)}}{|A_s^i(k)|^2\hat{P}_{z_{r,ij}}(k)}} \tag{16}$$

$$\sigma_{v,ij}^{2(l+1)} \leftarrow \sigma_{v,ij}^{2(l)} \frac{\sum_{k=1}^{N} \frac{P_{z_l}(k)}{|A_w^j(k)|^2\hat{P}_{z_{l,ij}}(k)^2} + \frac{P_{zr}(k)}{|A_w^j(k)|^2\hat{P}_{z_{r,ij}}(k)^2}}{\sum_{k=1}^{N} \frac{1}{|A_w^j(k)|^2\hat{P}_{z_{l,ij}}(k)} + \frac{1}{|A_w^j(k)|^2\hat{P}_{z_{r,ij}}(k)}} \tag{17}$$

$$c_{ij}^{(l+1)} \leftarrow c_{ij}^{(l)} \frac{\sum_{k=1}^{N} \frac{\sigma_{u,ij}^{2(l)}P_{zr}(k)}{|A_s^i(k)|^2\hat{P}_{z_{r,ij}}(k)^2}}{\sum_{k=1}^{N} \frac{\sigma_{u,ij}^{2(l)}}{|A_s^i(k)|^2\hat{P}_{z_{r,ij}}(k)}}. \tag{18}$$

It should be noted that $\hat{P}_{z_{l,ij}}(k)$ and $\hat{P}_{z_{r,ij}}(k)$ used in (16), (17) and (18) in the $l^{\text{th}}$ iteration is computed using excitation variances and the scale parameter from the $(l-1)^{\text{th}}$ iteration. We have summarised the proposed algorithm for estimating the speech and noise STP parameters in Algorithm 1.

**Algorithm 1** Summary of the estimation framework

```
while new time-frames are available do
    for ∀i ∈ N_s do
        for ∀j ∈ N_w do
            compute the ML estimates of excitation noise variances and the scale term (σ^{2,ML}_{u,ij}, σ^{2,ML}_{v,ij}, c^{ML}_{ij}) using (16), (17) and (18)
            compute the modelled spectrum for left channel P̂^{ML}_{z_{l,ij}} and right channel P̂^{ML}_{z_{r,ij}} using (10) and (11) respectively
            compute the likelihood values p(z_l, z_r|θ^{ML}_{ij}) using (9)
        end for
    end for
    Get the estimates of STP parameters (σ̂²_u, {â_i}^P_{i=1}, σ̂²_v, {b̂_i}^Q_{i=1}, ĉ) using (5)
end while
```

## IV. Experiments

This section will elaborate on the experiments used to evaluate the proposed algorithm. The test audio files used for the experiments consisted of speech from the GRID database [14] re-sampled to 8 kHz. The noise signal used is a binaural babble recording from the ETSI database [15], which was recorded with two microphones placed on a dummy head. Binaural noisy signals were generated by convolving the clean speech signal with binaural anechoic head related impulse responses (HRIR) corresponding to ITE HAs obtained from [16] and adding the binaural noise signals to the convolved signals. The experiments were performed for different positions of the speakers (the position of the speaker is defined as in Fig. 4). The speech and noise STP parameters required for the enhancement process are estimated every 25 ms, as explained in Section III. For our experiments, we have used a speech codebook of 64 entries, which was generated using the generalised Lloyd algorithm [17] on a training sample of 2-4 minutes of HRIR convolved speech from the specific speaker of interest. Using a speaker specific codebook instead of a general speech codebook leads to improvement in performance, and a comparison between the two is studied in [18]. The HRIR used for convolving the training signal corresponded to zero degrees, whereas the test signals consisted of speech coming form different directions. It should be noted that the sentences used for training the codebook was not included in the test sequence. The noise codebook consisting of only 8 entries, is generated using thirty seconds of noise signal. The audio samples used for training the noise signal was different from audio samples used for testing. The AR order for the speech and noise signal is chosen to be 14. The codebooks as well as MATLAB code for generating the codebooks will be available at `https://tinyurl.com/mskcreatevbn`. We have evaluated the proposed method in terms of the accuracy in the estimation of STP parameters as well as the enhancement performance.

### A. Accuracy in the estimation of STP parameters

This section evaluates the proposed algorithm in terms of the accuracy in the estimation of STP parameters. Fig. 5 shows the plots of the true and estimated speech excitation variances (for the left and right channels) for speaker position at 30 degree to the left of the listener at a distance of 80 cm, for a particular test signal. It can be seen that the proposed method captures the difference in speech excitation variances between the two channels. We now evaluate the ability of the proposed algorithm to deal with changes in the speaker position. For the experiments, the position of the speaker has been varied from $-15$ degree to 0 degree at frame index 149 and from 0 degree to 10 degree at frame index 285 at a distance of 80 cm. Fig. 6 shows the estimated value of the scale parameter along the frame index for different speaker positions. It can be seen from Fig. 6 that the $\hat{c}$ has a value of approximately 0.2 until frame index 141 and then changes to approximately 1 from frame index 149 until 282, and finally changes to 2 from then onwards. The $\hat{c}$ for the first one third portion has a value less than 1 as the speaker is located to the left of the listener. In this case, the level of the signal at the right ear is attenuated in comparison to the level at the left ear, due to the head shadowing effect. For the second portion $\hat{c}$ is approximately 1 as the speaker is located in front of the listener. As the speaker position is changed to 10 degree right of listener, $\hat{c}$ has a value of around 2, as the speaker is closer to the right ear. The position of the speaker can be easily tracked without any delay using the proposed method, as the scale parameter is estimated for every frame index. Moreover, the proposed method does not require the knowledge of the speaker position at any stage to initialise the value of the scale parameter. It should be noted that the scale parameter is relevant only in the speech active regions. Thus, the aberrations present in Fig. 6 can be explained by the speech being absent in certain time frames.

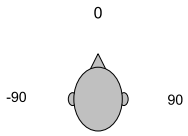


Fig. 4: Figure showing the top view of the listener. Position of the speaker has been varied for the experiments

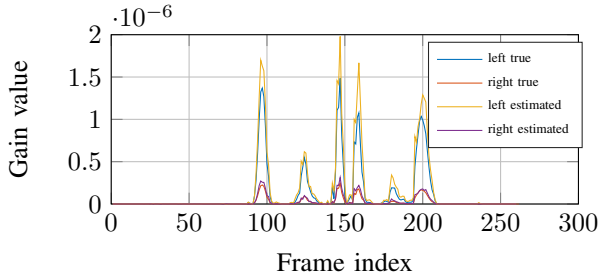


Fig. 5: Plot of the true and estimated speech excitation variances

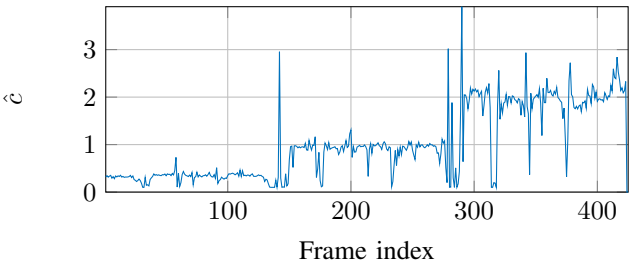


Fig. 6: Plot of the estimated scale parameter ($\hat{c}$)

Next, we compute the total IS divergence between the observed noisy periodograms and the modelled spectrums for test signals taken form the GRID database. This measure shows the ability of the estimated parameters to fit the observed noisy spectrum. For this experiment, the position of the speaker is varied around the listener for two different distances at SNR = 5 dB. Table I shows the computed IS divergences for different speaker positions for the proposed method and the method in [7] which we denote as BSTP. It should be noted that

the excitation gains in [7] were calculated by minimising an approximate cost function as opposed to here. Thus, to make a fair comparison, we have used the multiplicative update method [11] for computing the excitation variances as used here for [7]. It can be seen that the estimation of the STP parameters using the proposed method leads to a reduced IS divergence between the modelled and the observed spectrums.

TABLE I: Table showing total IS divergence between the modelled noisy spectrum and the observed noisy periodograms (left + right channels) for different speaker positions

| | | Angle of the speaker | | | |
|---|---|---|---|---|---|
| | Distance (cm) | $-85$ | $-75$ | $-65$ | $-55$ |
| Proposed | 80 | **3.61** | **3.75** | **3.73** | **3.65** |
| | 300 | **3.62** | **3.73** | **3.72** | **3.62** |
| BSTP [7] | 80 | 3.98 | 4.30 | 4.35 | 4.20 |
| | 300 | 3.85 | 4.16 | 4.25 | 4.08 |

### B. Enhancement performance

We now evaluate the benefit of incorporating the speaker position for enhancement. The framework that we have used for the experiments is similar to [7] where a fixed lag Kalman smoother is used for enhancement on each channel. Fig. 7 shows the short-term objective intelligibility (STOI) [19] scores obtained for the two methods when the speaker is at a position of $-50$ degree at 300 cm. The STOI score shown in the Fig. 7 corresponds to the score obtained for the better ear. We have compared the propsed method to BSTP and dual channel speech enhancement method proposed in [4] which we denote here as TwoChSS. It can be seen that taking into account the position of the speaker using the proposed method leads to improvement in the STOI scores especially in low SNR region. It can be seen that TwoChSS degrades the performance of the signal in terms of STOI. This is mainly due to the assumption in TwoChSS that the speaker is in the nose direction of the listener. It should also be noted that the performance of the proposed method and BSTP is similar when the speaker is in the nose direction as $\hat{c} \approx 1$.

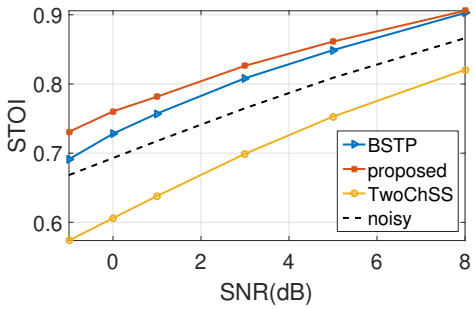


Fig. 7: Comparison of the STOI scores when the speaker is 50 degrees to the left of the speaker

## V. Conclusion

This paper proposed a model based approach for estimating the STP parameters of speech and noise in a binaural framework. The proposed method is able to take into account the position of the speaker while estimating the parameters which leads to an improved modelling of the observed spectrum in comparison to a previous method proposed in [7]. The estimated parameters are subsequently used for enhancement of speech in a binaural framework.

## References


[1] T. V. D. Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[3] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 342–355, 2010.

[4] M. Dorbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," in *European Signal Processing Conference, 1996. EUSIPCO 1996. 8th*. IEEE, 1996, pp. 1–4.

[5] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[6] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–14, 2006.

[7] M. S. Kavalekalam, M. G. Christensen, and J. B. Boldt, "Binaural speech enhancement using a codebook based approach," *Proc. Int. Workshop on Acoustic Signal Enhancement*, 2016.

[8] ——, "Model based binaural enhancement of voiced and unvoiced speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing. Proceedings*, 2017.

[9] J. A. Grange and J. F. Culling, "The benefit of head orientation to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 703–712, 2016.

[10] ——, "Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4061–4072, 2016.

[11] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 3, pp. 457–468, 2017.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[13] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] ETSI202396-1, "Speech and multimedia transmission quality; part 1: Background noise simulation technique and background noise database." 2009.

[16] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–10, 2009.

[17] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[18] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook based approach," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 2016.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.