MODEL BASED BINAURAL ENHANCEMENT OF VOICED AND UNVOICED SPEECH

Mathew Shaji Kavalekalam[1], Mads Græsbøll Christensen[1] and Jesper B. Boldt[2]

[1]Audio Analysis Lab, AD:MT, Aalborg University, Denmark {msk,mgc}@create.aau.dk
[2]GN Resound A/S, DK 2750, Ballerup, Denmark {jboldt}@gnresound.com

ABSTRACT

This paper deals with the enhancement of speech in presence of non-stationary babble noise. A binaural speech enhancement framework is proposed which takes into account both the voiced and unvoiced speech production model. The usage of this model in enhancement requires the Short term predictor (STP) parameters and the pitch information to be estimated. This paper uses a codebook based approach for estimating the STP parameters and a parametric binaural method is proposed for estimating the pitch parameters. Improvements in objective score are shown when using the voiced-unvoiced speech model in comparison to the conventional unvoiced speech model.

***Index Terms*—** speech enhancement, Kalman filter, autoregressive models

1. INTRODUCTION

Understanding of speech in difficult listening situations like cocktail party scenarios is a major issue for hearing impaired. Thus, the primary objectives of a speech enhancement system present in a hearing aid are to improve the quality and intelligibility of the degraded speech. Generally, a hearing impaired person is fitted with hearing aids at both ears. This enables the hearing aids to communicate with each other and share information between the hearing aids. Binaural processing of noisy signals has shown to be more effective than processing the noisy signal independently at each ear [1]. Apart from a better noise reduction performance, binaural algorithms make it possible to preserve the binaural cues such as inter-aural time difference and inter-aural level difference which contribute to spatial release from masking [2]. Some binaural speech enhancement algorithms with multiple microphones present in each hearing aid are [3, 4] and with a single microphone present in each hearing aid are [5, 6, 7]. The above mentioned algorithms have shown improvements in speech intelligibility in comparison to bilateral methods of enhancement. Most of the algorithms stated above perform the enhancement in the frequency domain by assuming that the speech and noise components are uncorrelated, and do not take into account the speech production model.

This work was supported by Innovations fund Denmark

In this paper we propose a binaural speech enhancement algorithm, which takes into account the voiced and unvoiced speech production model. A very conventional method to model clean speech for enhancement purposes is using the autoregressive (AR) model with white Gaussian noise as the excitation signal [8]. However, this model is not very suitable for representing voiced speech. Thus it was proposed in [9] to use a modified model for representing both kinds of speech. This model takes into account both the voiced and unvoiced speech by modifying the excitation signal that is used in the AR model. In this paper, we propose to use this voiced-unvoiced speech model in a binaural speech enhancement framework. This framework requires the speech and noise STP parameters (which consists of the AR parameter vector and excitation variance) and the pitch parameters to be estimated. Speech and noise STP parameters are estimated using a codebook based approach, and a parametric binaural method is proposed to estimate the pitch parameters.

The remainder of the paper is structured as follows. Section 2 explains the signal model and the assumptions that will be used in the paper. Section 3 explains the speech enhancement framework in detail. Experiments and results are presented in Section 4 followed by conclusion in Section 5.

2. SIGNAL MODEL

The binaural noisy signals at the left/right ear, denoted by $z_{l/r}(n)$ is expressed as shown in (1), where $s_{l/r}(n)$ is the clean speech component and $w_{l/r}(n)$ is the noise component which are assumed to be uncorrelated.

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

It is assumed that the target speaker is located in front of the user. Due to this assumption, the clean speech component at both the ears are considered to be approximately equal. A very common way to represent the clean speech component is in the form of an AR process (of order P) which is represented as

$$s(n) = \Big(\sum_{i=1}^{P} a_i(n) s(n-i) \Big) + u(n) \tag{2}$$

where $u(n)$ is white noise signal with variance $\sigma_u^2(n)$. Although this model is suitable for representing unvoiced

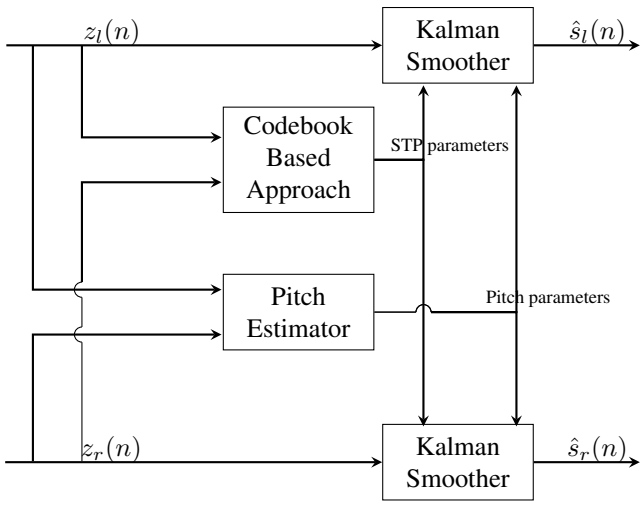

Fig. 1: Basic block diagram of the enhancement framework

speech, it is not appropriate for modelling voiced speech. The enhancement framework used here models $u(n)$ as

$$u(n) = b(n, p_n)u(n - p_n) + d(n), \tag{3}$$

where $d(n)$ is white Gaussian noise with variance $\sigma_d^2(n)$, p_n is the instantaneous pitch period and $b(n, p_n)$ is the degree of voicing. In portions of voiced speech, $b(n, p_n)$ is assumed to be close to 1 and the variance of $d(n)$ is assumed to be small, whereas in portions of unvoiced speech, $b(n, p_n)$ is assumed to be close to zero which then simplifies into the conventional unvoiced AR model. It is also assumed that listener is present in a diffuse noise field such that the noise component at both ears have similar spectral shape.

3. METHOD

Figure 1 shows the basic block diagram of the proposed enhancement framework. The noisy signals at the left and right ears are enhanced using a fixed lag Kalman smoother (FLKS) which requires the STP parameters and pitch parameters. The usage of identical filter parameters on both the ears leads to the preservation of binaural cues. STP parameters are estimated using a codebook based approach which is explained in 3.1 and the pitch parameters are estimated using a parametric method explained in section 3.2.

3.1. Binaural Codebook based estimation of STP parameters

The speech and noise STP parameters required for the enhancement are estimated using a codebook based approach [10, 11], which assumes that the clean speech component and noise component are AR processes. The estimation of these parameters uses the a priori information about speech and noise spectral shapes stored in trained codebooks in the form of Linear Prediction Coefficients (LPC). The random variables (r.v) corresponding to the parameters to be estimated are concatenated to form a single vector $\theta = [\theta_s\ \theta_w] = [\mathbf{a}; \sigma_u^2; \mathbf{c}; \sigma_v^2]$, where $\mathbf{a}, \mathbf{c}$ corresponds to r.v representing the speech and noise LPC, and σ_u^2, σ_v^2 representing the speech and noise excitation variances. The MMSE estimate of the parameter vector is written as

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

where $\mathbf{z}_l$ and $\mathbf{z}_r$ is frame of noisy speech at the left and right ears respectively. Let us define $\theta_{ij} = [\mathbf{a}_i; \sigma_{u,ij}^{2,ML}; \mathbf{c}_j; \sigma_{v,ij}^{2,ML}]$ where $\mathbf{a}_i$ is the ith entry of speech codebook (of size N_s), $\mathbf{c}_j$ is the jth entry of the noise codebook (of size N_w) and $\sigma_{u,ij}^{2,ML}, \sigma_{v,ij}^{2,ML}$ represents the maximum likelihood (ML) estimates of the excitation variances. The discrete counterpart of (4) can be written as

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

where the MMSE estimate is expressed as a weighted linear combination of θ_{ij} with weights proportional to $p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij})$. Assuming that the left and right noisy signal are conditionally independent given θ_{ij}, $p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij})$ can be written as

$$p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij}) = p(\mathbf{z}_l|\theta_{ij})p(\mathbf{z}_r|\theta_{ij}) \tag{6}$$

Logarithm of the likelihood $p(\mathbf{z}_l|\theta_{ij})$ can be written as the negative of Itakura-Saito distortion between noisy spectral envelope at the left ear $P_{z_l}(\omega)$ and modelled noisy spectral envelope $\hat{P}_z^{ij}(\omega)$ [10]. Using the same result for the right ear, $p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij})$ can be written as

$$p(\mathbf{z}_l, \mathbf{z}_r|\theta_{ij}) = \exp\left(-\left(d_{\mathrm{IS}}(P_{z_l}(\omega), \hat{P}_z^{ij}(\omega)) + d_{\mathrm{IS}}(P_{z_r}(\omega), \hat{P}_z^{ij}(\omega))\right)\right) \tag{7}$$

where

$$\hat{P}_z^{ij}(\omega) = \frac{\sigma_{u,ij}^{2,ML}}{|A_s^i(\omega)|^2} + \frac{\sigma_{v,ij}^{2,ML}}{|A_w^j(\omega)|^2} \tag{8}$$

and $1/|A_s^i(\omega)|^2$ is the spectral envelope corresponding to the ith entry of the speech codebook, $1/|A_w^j(\omega)|^2$ is the spectral envelope corresponding to the jth entry of the noise codebook. More details regarding this method is available in [11] and the references therein.

3.2. Binaural pitch estimation

In this paper, we propose a parametric binaural method to estimate the pitch parameters. This method uses the harmonic model to represent the clean speech as a sum of L harmonically related complex sinusoids. Using the harmonic model, noisy signal at the left ear can be represented as

$$\mathbf{z}_l = \mathbf{V}\mathbf{D}_l\mathbf{q} + \mathbf{w}_l \tag{9}$$

where $\mathbf{z}_l$ and $\mathbf{w}_l$ is a length N frame of noisy and noise samples at the left ear, $\mathbf{q}$ is a vector of complex amplitudes, $\mathbf{V}$ is

the Vandermonde matrix which is defined as $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_L]$, where $\mathbf{v}_l = [1\ e^{j\omega_0 l} \dots e^{j\omega_0 l(N-1)}]^T$ and $\mathbf{D}_l$ is the directivity matrix from the source to the left ear. This matrix contains an angle dependent delay and magnitude term along the diagonal, which can be designed using the values found in [12]. Similarly, the noisy signal at the right ear is written as

$$\mathbf{z}_r = \mathbf{V}\mathbf{D}_r\mathbf{q} + \mathbf{w}_r. \tag{10}$$

Defining $\mathbf{y} = [\mathbf{z}_l^T \mathbf{z}_r^T]^T$, the two equations are combined and rewritten as follows,

$$\mathbf{y} = \begin{bmatrix}\mathbf{V}\mathbf{D}_l \\ \mathbf{V}\mathbf{D}_r\end{bmatrix}\mathbf{q} + \begin{bmatrix}\mathbf{w}_l \\ \mathbf{w}_r\end{bmatrix} = \mathbf{H}\mathbf{q} + \mathbf{w}. \tag{11}$$

First the ML estimate of amplitude vector $\mathbf{q}$ is obtained as

$$\hat{\mathbf{q}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} \tag{12}$$

Noise variance estimate for the left and right channel is obtained as

$$\hat{\sigma}_l^2 = 1/N||\mathbf{z}_l - \mathbf{V}\mathbf{D}_l\hat{\mathbf{q}}||^2 \tag{13}$$

$$\hat{\sigma}_r^2 = 1/N||\mathbf{z}_r - \mathbf{V}\mathbf{D}_r\hat{\mathbf{q}}||^2. \tag{14}$$

Subsequently, the ML estimate of pitch and model order is estimated jointly using the MAP model selection [13, 14] as

$$\{\hat{\omega}_0, \hat{L}\} = \underset{\{L\in\mathcal{L},\omega_0\in\Omega_0\}}{\arg\min} N(\ln\hat{\sigma}_l^2 + \ln\hat{\sigma}_r^2) + \frac{3}{2}\ln N + L\ln N + 0.5(2L+1)\ln N_{ch} \tag{15}$$

where Ω_0 is the set of candidate fundamental frequencies, $\mathcal{L}$ is the set of candidate model orders, N_{ch} is the number of channels which is 2 in this case. It should also be noted that the usage of a complex signal model in (9) and (10) requires the real noisy signals to be converted into the complex domain by the means of Hilbert transform. The degree of voicing ($b(n, p_n)$) is calculated by taking the ratio between the energy present at the harmonics and the total energy present in the signal.

3.3. Enhancement by FLKS for voiced speech

The estimated STP and pitch parameters are subsequently used for enhancement by FLKS as explained below. The usage of FLKS (with a smoother delay of $d_s \geq P$) from a speech enhancement perspective requires the AR signal model in (2) to be written as a state space form as shown below

$$\mathbf{s}_{l/r}(n) = \mathbf{A}(n)\mathbf{s}_{l/r}(n-1) + \mathbf{\Gamma}_1 u(n) \tag{16}$$

where $\mathbf{s}_{l/r}(n) = [s_{l/r}(n)s_{l/r}(n-1)\dots s_{l/r}(n-d_s)]^T$ is a $(d_s+1)\times 1$ vector containing the d_s+1 recent speech samples, $\mathbf{\Gamma}_1 = [1, 0\dots 0]^T$ is a $(d_s+1)\times 1$ vector and $\mathbf{A}(n)$ is the $(d_s+1)\times(d_s+1)$ speech state evolution matrix written as

$$\mathbf{A}(n) = \begin{bmatrix} a_1(n) & a_2(n) & \dots & a_P(n) & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 1 & 0 & \vdots & \dots & 0 \\ 0 & \dots & \dots & 1 & 0 & \dots & 0 \\ \vdots & \dots & \dots & 0 & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 0 & 1 & 0 \end{bmatrix} \tag{17}$$

Using (3), the state space equation for the excitation signal is written as

$$\mathbf{u}(n) = \mathbf{B}(n)\mathbf{u}(n-1) + \mathbf{\Gamma}_2 d(n) \tag{18}$$

where $\mathbf{u}(n) = [u(n)u(n-1)\dots u(n-p_{max}+1)]^T$, p_{max} is the maximum pitch period, $\mathbf{\Gamma}_2 = [1, 0\dots 0]^T$ is a $(p_{max})\times 1$ vector and

$$\mathbf{B}(n) = \begin{bmatrix} b(n,1) & b(n,2) & \dots & b(n,p_{max}) \\ & \mathbf{I} & & \mathbf{0} \end{bmatrix} \tag{19}$$

is a $p_{max}\times p_{max}$ matrix where $b(n,i) = 0\ \forall i \neq p_n$. The state space equation for the noise signal is written as

$$\mathbf{w}_{l/r}(n) = \mathbf{C}(n)\mathbf{w}_{l/r}(n-1) + \mathbf{\Gamma}_3 v(n), \tag{20}$$

where $\mathbf{w}_{l/r}(n) = [w_{l/r}(n)w_{l/r}(n-1)\dots w_{l/r}(n-Q+1)]^T$, $\mathbf{\Gamma}_3 = [1, 0\dots 0]^T$ is a $Q\times 1$ vector and

$$\mathbf{C}(n) = \begin{bmatrix} c_1(n) & c_2(n) & \dots & c_Q(n) \\ & \mathbf{I} & & \mathbf{0} \end{bmatrix} \tag{21}$$

is $Q\times Q$ matrix. The concatenated state space equation is

$$\mathbf{x}_{l/r}(n+1) = \mathbf{F}(n)\mathbf{x}_{l/r}(n) + \mathbf{\Gamma}_4 g(n+1), \tag{22}$$

where $\mathbf{x}_{l/r}(n+1) = \begin{bmatrix}\mathbf{s}_{l/r}(n) \\ \mathbf{u}(n+1) \\ \mathbf{w}_{l/r}(n)\end{bmatrix}$, $\mathbf{\Gamma}_4 = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{\Gamma}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma}_3\end{bmatrix}$, $g(n+1) = \begin{bmatrix}d(n+1) \\ v(n)\end{bmatrix}$, and $\mathbf{F}(n) = \begin{bmatrix}\mathbf{A}(n) & \mathbf{\Gamma}_1\mathbf{\Gamma}_2^T & \mathbf{0} \\ \mathbf{0} & \mathbf{B}(n+1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}(n)\end{bmatrix}$. The measurement equation is given by

$$z_{l/r}(n) = \mathbf{\Gamma}^T\mathbf{x}_{l/r}(n), \tag{23}$$

where $\mathbf{\Gamma} = [\Gamma_1^T\mathbf{0}^T\Gamma_3^T]^T$. The final state space equation and measurement equation denoted by (22) and (23) respectively, is subsequently used for the formulation of the FLKS equations (24 - 28). The prediction stage of the FLKS, which computes the a priori estimates of the state vector ($\hat{\mathbf{x}}_{l/r}(n|n-1)$) and error covariance matrix ($\mathbf{M}(n|n-1)$) is written as

$$\hat{\mathbf{x}}_{l/r}(n|n-1) = \mathbf{F}(n-1)\hat{\mathbf{x}}_{l/r}(n-1|n-1) \tag{24}$$

$$\mathbf{M}(n|n-1) = \mathbf{F}(n-1)\mathbf{M}(n-1|n-1)\mathbf{F}(n-1)^T + \mathbf{\Gamma}_4\begin{bmatrix}\sigma_d^2(n) & 0 \\ 0 & \sigma_v^2(n-1)\end{bmatrix}\mathbf{\Gamma}_4^T. \tag{25}$$

σ_d^2 used in (25) is substituted with the value obtained for σ_u^2 obtained from (5). Kalman gain is computed as shown in (26)

$$\mathbf{K}(n) = \mathbf{M}(n|n-1)\mathbf{\Gamma}[\mathbf{\Gamma}^T\mathbf{M}(n|n-1)\mathbf{\Gamma}]^{-1}. \tag{26}$$

Correction stage of the FLKS, which computes the a posteriori estimates of the state vector and error covariance matrix is given by

$$\hat{\mathbf{x}}_{l/r}(n|n) = \hat{\mathbf{x}}_{l/r}(n|n-1) + \mathbf{K}(n)[z_{l/r}(n) - \mathbf{\Gamma}^T\hat{\mathbf{x}}_{l/r}(n|n-1)] \tag{27}$$

$$\mathbf{M}(n|n) = (\mathbf{I} - \mathbf{K}(n)\mathbf{\Gamma}^T)\mathbf{M}(n|n-1). \tag{28}$$

Finally, the enhanced signal at time index $n - d_s$ is obtained by taking the $d_s + 1^{th}$ entry of the a posteriori estimate of the state vector as $\hat{s}_{l/r}(n-d_s) = \hat{\mathbf{x}}_{l/r_{(d_s+1)}}(n|n)$.

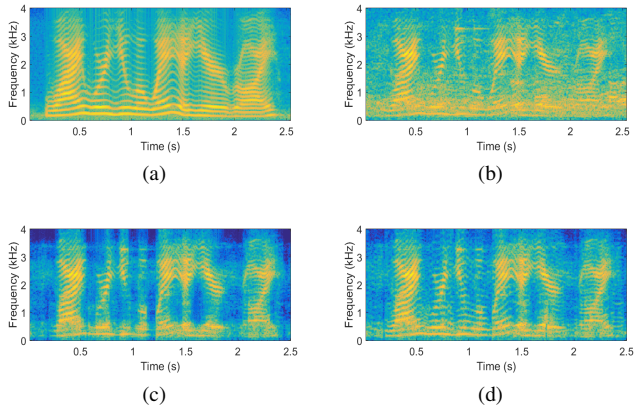

Fig. 2: Spectrograms of (a) clean signal, (b) noisy signal (SNR = 3 dB) at the left microphone and enhanced signals: (c) UV, (d) V-UV

		SNR (dB)			
		-5	-2	1	4
Noisy	male	0.6486	0.7178	0.7827	0.8387
	female	0.6305	0.7003	0.7668	0.8259
UV	male	0.6882	0.7655	0.8334	0.8841
	female	0.6568	0.7332	0.8035	0.8603
V-UV	male	0.7036	0.7803	0.8436	0.8893
	female	0.6857	0.7635	0.8277	0.8762

Table 1: comparison of STOI scores

		SNR (dB)			
		-5	-2	1	4
Noisy	male	1.6313	1.7754	1.8434	1.9577
	female	1.2924	1.4868	1.6331	1.7941
UV	male	1.8666	2.0467	2.2978	2.5326
	female	1.5066	1.7273	1.9487	2.1744
V-UV	male	1.8720	2.1006	2.3396	2.5489
	female	1.6088	1.8429	2.0625	2.2626

Table 2: comparison of PESQ scores

4. EXPERIMENTS

This section contains the experiments used to evaluate the proposed algorithm. Objective measures that have been used for the evaluation are Short time objective intelligibility (STOI) measure [15] and Perceptual evaluation of subjective quality (PESQ). The test audio files used for the experiments consisted of speech from the CHiME [16] and Eurom database[17] re-sampled to 8 KHz. The noise signal used is binaural babble recordings from the ETSI background noise database [18], which was recorded with two microphones placed on a dummy head. Binaural noisy signals were generated by convolving the clean speech signal with anechoic binaural head related impulse responses (HRIR) corresponding to in the ear hearing aids obtained from [19] and adding the binaural noise signals to the convolved signals. The speech and noise STP parameters required for the enhancement process is estimated every 25 ms using the codebook based approach, as explained in section 3.1. Speech codebook of 64 entries is generated using the Generalised Lloyd algorithm (GLA) [20] on a training sample of 2-4 minutes of HRIR convolved speech from the specific speaker of interest. Using a speaker specific codebook (which requires speaker identification) instead of a generalised speech codebook leads to improvement in performance, and a comparison between the two is studied in [21]. 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 two minutes of noise signal. The AR model order for both the speech and noise signal is chosen to be 14. The pitch period and degree of voicing is estimated using the parametric binaural method explained in 3.2. The cost function in (15) was evaluated on a 0.5 Hz grid for fundamental frequencies in the range 80-400 Hz. For each fundamental frequency candidate ω_0, the model orders considered were $\mathcal{L} = \{1, \ldots, \lfloor 2\pi/\omega_0 \rfloor\}$. The estimated pitch and STP parameters are subsequently used for enhancement as explained in section 3.3. Enhancement framework that uses the unvoiced model for enhancement, which does not use the pitch information is denoted as UV [11]. Enhancement framework that uses the voiced-unvoiced model for enhancement along with the pitch parameters estimated from the binaural noisy signal is denoted as V-UV. Table 1 shows the comparison of the STOI scores averaged for the left and right channels for the noisy, UV and V-UV. Table 2 shows the PESQ scores. It can be seen that using a voiced-unvoiced speech model is beneficial in comparison to using the conventional unvoiced model, for both the female and male speakers. It should be noted that the improvement of using the voiced-unvoiced model instead of conventional unvoiced model is more pronounced amongst the female speakers than the male speakers. The spectrograms of the different signals are shown in figure 2. It can be seen from figure 2c that using an unvoiced model for the enhancement results in removal of weak harmonics present in the clean signal whereas using the voiced-unvoiced model for enhancement preserves the harmonics as can be seen from figure 2d.

5. CONCLUSION

This paper proposed a binaural speech enhancement framework that takes into account the speech production process. The proposed method requires the pitch parameters and the STP parameters to be estimated. A parametric binaural method is proposed to estimate the pitch parameters and the STP parameters were estimated using a codebook based method. Using the modified voiced-unvoiced model in the place of conventional unvoiced AR model for enhancement shows considerable improvement in STOI and PESQ scores.

6. 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] A. Bronkhorst and R Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, 1991.

[9] Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 510–524, 1999.

[10] 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.

[11] 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.

[12] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE transactions on speech and audio processing*, vol. 6, no. 5, pp. 476–488, 1998.

[13] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[14] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. IEEE, 2012, pp. 409–412.

[15] 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.

[16] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*, 2015.

[17] D. Chan, A. Fourcin, D. Gibbon, B Granstrom, et al., "Eurom-a spoken language resource for the eu," in *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech'95*, 1995, pp. 867–880.

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

[19] 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.

[20] 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.

[21] 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.