Modeling Temporal Dependency for Robust Estimation of LP Model Parameters in Speech Enhancement

Chun Hoy Wong, Tan Lee, Yu Ting Yeung , P.C. Ching*

Department of Electronic Engineering
The Chinese University of Hong Kong
Hong Kong SAR, China
{chwong, tanlee, ytyeung, pcching}@ee.cuhk.edu.hk

10.21437/Interspeech.2015-403

Abstract

This paper presents a novel approach to robust estimation of linear prediction (LP) model parameters in the application of speech enhancement. The robustness stems from the use of prior knowledge on the clean speech and the interfering noise, which are represented by two separate codebooks of LP model parameters. We propose to model the temporal dependency between short-time model parameters with a composite hidden Markov model (HMM) that is constructed by combining the speech and the noise codebooks. Optimal speech model parameters are estimated from the HMM state sequence that best matches the input observation. To further improve the estimation accuracy, we propose to perform interpolation of multiple HMM state sequences such that the estimated speech parameters would not be limited by the codebook coverage. Experimental results demonstrate the benefits and effectiveness of temporal dependency modeling and states interpolation in improving the segmental signal-to-noise ratio, PESQ and spectral distortion of enhanced speech.

Index Terms: speech enhancement, linear predictive (LP) model parameters, hidden Markov model (HMM)

1. Introduction

Speech enables effective communication between human beings. Messages are delivered and received by speaking and listening to speech. Speech signals are often contaminated by different types of noise and interference in real-world situations [1]. The goal of speech enhancement is to suppress the noise and recover the original speech with good quality and intelligibility. Conventional approaches include spectral subtraction [2], Kalman filter [3] and sub-spaced methods [4]. More recent works exploit prior knowledge about speech parameters [5], [6] and [7]. In [5], a method of speech enhancement based on prior knowledge of short-term predictor parameters was proposed. These parameters refer to the linear prediction (LP) coefficients and the excitation variance. They represent the spectral envelope of speech [1]. Using a codebook for LP coefficients of speech and another codebook for noise, the LP coefficients of clean speech are estimated by searching for the pair of codewords that give the best match between the modeled noisy spectrum and the observed one. The estimated LP coefficients and the corresponding excitation variance are then used to construct a Wiener filter for speech enhancement.

* Yu Ting Yeung is currently affiliated with Stanley Ho Big Data Decision Analytics Research Centre, The Chinese University of Hong Kong.

In this paper, we propose to incorporate the temporal dependency in LP coefficients into the process of codebook search, so as to capture the dynamic characteristics of speech. Hidden Markov models (HMM) have been widely used to model the temporal dynamics of speech. Factorial HMM (FHMM) was developed to deal with multiple sound sources [9] [10]. In our application, speech and noise are the two sound sources. Each of them can be modeled by an HMM, in which each codeword in the respective codebook represents a state. The noise-corrupted speech is then modeled by a single HMM that is obtained by combining states from the speech HMM and the noise HMM. The estimated LP coefficients can be obtained from the optimal state sequence determined by the Viterbi algorithm. To alleviate the problem of limited codebook coverage, we propose an interpolation approach that involves multiple state sequences inferred from the HMM of noisy speech. Experimental results show the proposed methods outperform the conventional codebook based approach and the inclusion of temporal dependency leads to significant performance gain.

2. HMM for LP parameters estimation

In this section, we describe how the codebook approach to LP parameters estimation is extended to an HMM based approach. Let $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{w}$ denote the noisy speech, the clean speech and the noise, respectively, in a specific short-time frame. The additive signal model is assumed, i.e.

$$\mathbf{y} = \mathbf{s} + \mathbf{w}. \tag{1}$$

In the codebook approach [5], a codebook of size M is built for the clean speech and a codebook of size N is built for the noise. The codebook entries are the LP coefficients computed from the respective signals. Let $\mathbf{a_s^m} = [a_{s_1}^m a_{s_2}^m ....a_{s_p}^m]^T$ and $\mathbf{a_w^n} = [a_{w_1}^n a_{w_2}^n ....a_{w_q}^n]^T$ be the m^{th} and n^{th} codewords in the speech and noise codebooks respectively, where $m \in \{1, 2, ..., M\}$ and $n \in \{1, 2, ..., N\}$.

In continuous speech, LP coefficients of neighboring frames are related. To capture such temporal dependency, a hidden Markov model (HMM) is formulated to model the noisy observation. There are $M \times N$ states in this HMM, each being associated with a pair of speech codeword and noise codeword, i.e.,

$$\mathbf{a_y^i} = \mathbf{a_y^{mn}} = (\mathbf{a_s^m}, \mathbf{a_w^n}) \tag{2}$$

Assume that speech and noise are independent. Following [9] and [10], the prior probability and the transition probability in the HMM are given by,

$$p_{\mathbf{y}}(\mathbf{a_y^{mn}}) = p_{\mathbf{s}}(\mathbf{a_s^m}) p_{\mathbf{w}}(\mathbf{a_w^n}) \tag{3}$$

$$p_{\mathbf{y}}(\mathbf{a}_{\mathbf{y}}^{\mathbf{m_2 n_2}}|\mathbf{a}_{\mathbf{y}}^{\mathbf{m_1 n_1}}) = p_{\mathbf{s}}(\mathbf{a}_{\mathbf{s}}^{\mathbf{m_2}}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m_1}}) p_{\mathbf{w}}(\mathbf{a}_{\mathbf{w}}^{\mathbf{n_2}}|\mathbf{a}_{\mathbf{w}}^{\mathbf{n_1}}) \quad (4)$$

The state-level observation probability is assumed to be Gaussian [5], i.e.,

$$p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m}}, \mathbf{a}_{\mathbf{w}}^{\mathbf{n}}; \sigma_s^2, \sigma_w^2) = \frac{1}{(2\pi)^{\frac{L}{2}} |\mathbf{R_y}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{y}^T (\mathbf{R_y})^{-1} \mathbf{y} \right) \quad (5)$$

where $\mathbf{a}_{\mathbf{s}}^{\mathbf{m}}$ and $\mathbf{a}_{\mathbf{w}}^{\mathbf{n}}$ are the speech and noise codewords associated with the respective HMM state, and σ_s^2 and σ_w^2 denote the excitation variances of speech and noise respectively. L is the number of samples within the short-time frame.

The covariance matrix is given by $\mathbf{R_y} = \mathbf{R_s} + \mathbf{R_w}$, where

$$\mathbf{R_s} = \sigma_s^2 (\mathbf{V_s^T V_s})^{-1}, \quad \mathbf{R_w} = \sigma_w^2 (\mathbf{V_w^T V_w})^{-1}. \quad (6)$$

$\mathbf{V_s}$ and $\mathbf{V_w}$ are both $L \times L$ lower triangular Toeplitz matrices. Their first columns are $[1 a_{s1}^m a_{s2}^m ... a_{sp}^m 0...0]^T$ and $[1 a_{w1}^n a_{w2}^n ... a_{wq}^n 0...0]^T$ respectively.

The observation probability is estimated with the optimal excitation variances, i.e., $p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m}}, \mathbf{a}_{\mathbf{w}}^{\mathbf{n}}; \sigma_s^{2*}, \sigma_w^{2*})$[5], where σ_s^{2*} and σ_w^{2*} are determined by,

$$\{\sigma_s^{2*}, \sigma_w^{2*}\} = \arg \max_{\sigma_s^2, \sigma_w^2} p_{\mathbf{y}}(\mathbf{y}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m}}, \mathbf{a}_{\mathbf{w}}^{\mathbf{n}}; \sigma_s^2, \sigma_w^2). \quad (7)$$

Assuming small modeling error between the observed LP spectrum $P_y(\omega)$ and the modeled LP spectrum $\hat{P}_y(\omega) = \frac{\sigma_s^{2*}}{|A_s^m(\omega)|^2} + \frac{\sigma_w^{2*}}{|A_w^n(\omega)|^2}$, the optimal excitation variances are given as [5],

$$C \begin{bmatrix} \sigma_s^{2*} \\ \sigma_w^{2*} \end{bmatrix} = D \quad (8)$$

where

$$C = \begin{bmatrix} \left|\left| \frac{1}{P_y^2(\omega)|A_s^m(\omega)|^4} \right|\right| & \left|\left| \frac{1}{P_y^2(\omega)|A_s^m(\omega)|^2|A_w^n(\omega)|^2} \right|\right| \\ \left|\left| \frac{1}{P_y^2(\omega)|A_s^m(\omega)|^2|A_w^n(\omega)|^2} \right|\right| & \left|\left| \frac{1}{P_y^2(\omega)|A_w^n(\omega)|^4} \right|\right| \end{bmatrix} \quad (9)$$

$$D = \begin{bmatrix} \left|\left| \frac{1}{P_y(\omega)|A_s^m(\omega)|^2} \right|\right| \\ \left|\left| \frac{1}{P_y(\omega)|A_w^n(\omega)|^2} \right|\right| \end{bmatrix} \quad (10)$$

$$||f(\omega)|| = \int |f(\omega)| d\omega \quad (11)$$

$$A_s^m(\omega) = 1 - \sum_{i=1}^p a_{s_i}^m e^{-ji\omega}, \; A_w^n(\omega) = 1 - \sum_{i=1}^q a_{w_i}^n e^{-ji\omega} \quad (12)$$

Given a noisy input signal, the optimal HMM state sequence can be inferred by the Viterbi algorithm using the transition probability and the observation probability. The estimated LP coefficients are obtained by finding the speech codewords that are associated with the optimal states. Subsequently a Wiener filter is constructed at each time frame for speech enhancement,

$$H(\omega) = \frac{\frac{\sigma_s^{2*}}{|A_s^m(\omega)|^2}}{\frac{\sigma_s^{2*}}{|A_s^m(\omega)|^2} + \frac{\sigma_w^{2*}}{|A_w^n(\omega)|^2}} \quad (13)$$

3. Inference of multiple HMM state sequences

In the HMM based approach described in Section 2, a single optimal HMM state sequence is used. The estimated LP coefficients are directly given by the codebook entries. This means that the result of estimation is limited by the coverage of the codebook. To alleviate this problem and improve the estimation accuracy, we propose to perform interpolation on multiple HMM state sequences.

Consider K different HMM state sequences (paths) for an input signal of T frames. The following simplified notations are used in this section:

- $\pi(i)$: prior probability of state $\mathbf{a}_{\mathbf{y}}^{\mathbf{i}}$, $i = 1, 2, ..., MN$
- $a(i, j)$: transition probability from state $\mathbf{a}_{\mathbf{y}}^{\mathbf{i}}$ to $\mathbf{a}_{\mathbf{y}}^{\mathbf{j}}$
- $b_t(i)$: observation probability at state i for time frame t
- $i_t^k = (m_t^k, n_t^k)$: the state inferred for the k^{th} path at frame t, with m_t^k and n_t^k being the indices of the corresponding speech and noise codewords

Depending on the applications, there are different ways of inferring multiple state sequences from an HMM. In this application, we have the following considerations:

- The inferred paths should have the highest probabilities among all possible paths.
- The inferred states of different paths should be non-overlapping, i.e., $i_t^{k_1} \neq i_t^{k_2}$. This is to increase the diversity and avoid having the same codewords to dominate the result of interpolation.

We propose an algorithm that is able to efficiently produce sub-optimal paths for the above conditions. Table 1 explains the details of this algorithm. For $k = 1$, the path is inferred by the standard Viterbi algorithm (line 1 to 15). For the k^{th} path, we first consider the state at the last time frame T. Similar to the Viterbi algorithm, state j that gives the k^{th} maximum value of $\alpha(j, T)$ is chosen as i_T^k (line 17). Backtracking is performed at i_k^T (line 18-21) in approximating the path of the k^{th} highest likelihood. The constraint of $j \neq i_t^r, r = 1, 2, ..., k-1$ (line 19) is imposed for the requirement of non-overlapping paths. In case that the backtracking is infeasible, i.e., no state is eligible, the constraint would be removed (line 20-21).

4. Interpolation of multiple state sequences

Let $\mathbf{a}_{\mathbf{y}}^{\mathbf{i_t^k}} = (\mathbf{a}_{\mathbf{s}}^{\mathbf{m_t^k}}, \mathbf{a}_{\mathbf{w}}^{\mathbf{n_t^k}})$ be the LP coefficients inferred at time frame t in the k^{th} path, where $k = 1, 2, ..., K$ and $t = 1, 2, ..., T$. The LP coefficients are transformed into line spectral frequencies (LSF) (i.e., $\mathbf{a}_{\mathbf{y}}^{\mathbf{i_t^k}} \rightarrow \mathbf{l}_{\mathbf{y}}^{\mathbf{i_t^k}}$) for interpolation [11]. This is to ensure the stability of the interpolated LP spectrum.

The interpolated parameters $\hat{\theta}^t = \{\hat{\mathbf{l}}_{\mathbf{s}}^{\mathbf{m_t}}, \hat{\mathbf{l}}_{\mathbf{w}}^{\mathbf{n_t}}, \hat{\sigma}_s^{2t*}, \hat{\sigma}_w^{2t*}\}$ are obtained by

$$\hat{\theta}^t = \sum_{k=1}^K \theta^{kt} w(k, t) \quad (14)$$

where $\theta^{kt} = \{\mathbf{l}_{\mathbf{s}}^{\mathbf{m_t^k}}, \mathbf{l}_{\mathbf{w}}^{\mathbf{n_t^k}}, \sigma_{s^{t}}^{2^k*}, \sigma_{w^{t}}^{2^k*}\}$, $\mathbf{l}_{\mathbf{s}}^{\mathbf{m_t^k}}$ and $\mathbf{l}_{\mathbf{w}}^{\mathbf{n_t^k}}$ denote the inferred LSF parameters, and $\sigma_{s^{t}}^{2^k*}$, $\sigma_{w^{t}}^{2^k*}$ are the respective excitation variances as derived in (8).

The weight $w(k, t)$ is given by a heuristic scheme as follows,

$$w(k, t) = \frac{p_{\mathbf{s}}(\mathbf{a}_{\mathbf{s}}^{\mathbf{m_t^k}}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m_{t-1}^k}}) p_{\mathbf{w}}(\mathbf{a}_{\mathbf{w}}^{\mathbf{n_t^k}}|\mathbf{a}_{\mathbf{w}}^{\mathbf{n_{t-1}^k}})}{\sum_{i=1}^K p_{\mathbf{s}}(\mathbf{a}_{\mathbf{s}}^{\mathbf{m_t^i}}|\mathbf{a}_{\mathbf{s}}^{\mathbf{m_{t-1}^i}}) p_{\mathbf{w}}(\mathbf{a}_{\mathbf{w}}^{\mathbf{n_t^i}}|\mathbf{a}_{\mathbf{w}}^{\mathbf{n_{t-1}^i}})} \quad (15)$$

Table 1: *Multiple-path inference algorithm*

1: $k = 1$,$t = 1$
2: **for** $i = 1$ **to** MN **do**
3: $\alpha(i,t) = \pi(i)b_t(i)$
4: $\delta(i,t) = 0$
5: **end for**
6: **for** $t = 2$ **to** T **do**
7: **for** i **to** MN **do**
8: $\alpha(i,t) = b_t(i)\max_j\{\alpha(j,t-1)a(j,i)\}$
9: $\delta(i,t) = \operatorname{argmax}_j\{\alpha(j,t-1)a(j,i)\}$
10: **end for**
11: **end for**
12: $i_T^k = \operatorname{argmax}_j\{\alpha(j,T)\}$
13: **for** $t = T-1$ **to** 1 **do**
14: $i_t^k = \delta(i_{t+1}^k, t+1)$
15: **end for**
16: **for** $k = 2$ **to** K **do**
17: $i_T^k = \operatorname{argmax}_{\substack{j:j\neq i_T^r \\ r=1,2,\ldots,k-1}} \{\alpha(j,T)\}$
18: **for** $t = T-1$ **to** 1 **do**
19: $i_t^k = \operatorname{argmax}_{\substack{j:j\neq i_t^r \\ r=1,2,\ldots,k-1}} \{\alpha(j,t)a(j,i_{t+1}^k)\}$
20: **if** $\alpha(i_t^k,t)a(i_t^k,i_{t+1}^k) == 0$ **then**
21: $i_t^k = \operatorname{argmax}_j\{\alpha(j,t)a(j,i_{t+1}^k)\}$
22: **end if**
23: **end for**
24: **end for**

The weighting scheme reflects that codewords with higher transition probability should carry heavier weight in the interpolated parameters.

The interpolated parameters $\hat{\theta}^t$ are then converted back to LP coefficients for constructing the Wiener filter as in (13).

5. Experiments

In this section, we compare the performances of the HMM based approach versus the baseline codebook method (CB) as proposed in [5]. The proposed HMM approach takes temporal dependency of LP coefficients into consideration while the CB method performs LP parameters estimation for individual frames without considering the transitional relation among them.

5.1. Experiment set-up

All speech data were obtained from the TIMIT database [12]. For training of the speech codebook, 200 utterances were randomly selected from the database. Half of the training utterances were from male speakers and half from female speakers. The total duration of training speech was about 10 minutes. LP analysis of order 12 was applied to the training utterances with a Hanning window of 30 msec and 75% frame overlap. The generalized Lloyd's algorithm (GLA) was used to build the speech codebook using the Itakura-Saito distance. The codebook size was set to be 1024.

For training of the noise codebook, we used noise signals from NOISE-ROM 0 [13]. Two different types of noise, namely white noise and destroyer noise, were selected as representatives of wide-band and narrow-band noises respectively. The duration of training data was 2 minutes for each type of noise. The LP analysis order was 6 and the codebook size was 8. Other settings were the same as for the speech codebook.

A total of 20 test utterances were randomly selected from TIMIT. Half of the training utterances were from male speakers and half from female speakers. They were different from the training speech. Ten test utterances were used to determine the optimal number of HMM paths for interpolation. The other ten were used to evaluate and compare the performance of different speech enhancement methods. Noisy speech utterances were generated by adding white noise or destroyer noise to the clean speech utterances at various signal-to-noise ratios (SNR). The noise samples were also taken from NOISE-ROM 0. They were different from those used in codebook training.

We used the segmental signal-to-noise ratio (SSNR), perceptual evaluation of speech quality (PESQ)[14] and the spectral distortion (SD) to evaluate the performance of speech enhancement. SD is defined based on the LP spectrum as,

$$SD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}(10\log_{10}\frac{P_s(\omega)}{\hat{P}_s(\omega)})^2 d\omega} \qquad (16)$$

where $P_s(\omega)$ and $\hat{P}_s(\omega)$ are the LP spectra of clean speech and enhanced speech respectively.

5.2. Determining the number of HMM paths

In the HMM approach, LP parameters estimation can be done either from a single path or by interpolation of multiple paths. The following experiment aims to investigate how the number of paths affects the estimation accuracy and hence the performance in speech enhancement. The HMM approach was experimented with different number of HMM paths and the SSNR of enhanced speech was measured. Figures 1 and 2 show the results on white noise and destroyer noise respectively.

The results clearly show the benefit of using multiple HMM paths and the effectiveness of the proposed interpolation method. For both types of noise, the SSNR improves significantly when the number of paths is increased from one (single path) to a few hundreds. However, when the path number exceeds a certain value, the performance of speech enhancement shows a declining trend. This is related to the failure in fulfilling the backtracking condition. For the inference of a path with large index, there is a high chance that the condition of non-overlapping states can not be satisfied because many states have been assigned to other paths with lower indices. As a result, the inferred paths would have overlapping states, meaning that repeated speech codewords would be used for interpolation. The reduced diversity makes the interpolation less effective.

By examining the curves in the figures, the optimal number of HMM paths (k^*) for white noise and destroyer noise were found to be 600 and 400 respectively, with which the highest SSNR was achieved. These settings were used in the subsequent experiments.

5.3. Performance comparison of different methods

In this section, the performances of three different LP parameters estimation methods are compared:

- CB: The baseline codebook approach [5].
- HMM ($K = 1$): The HMM approach without interpolation (single path).

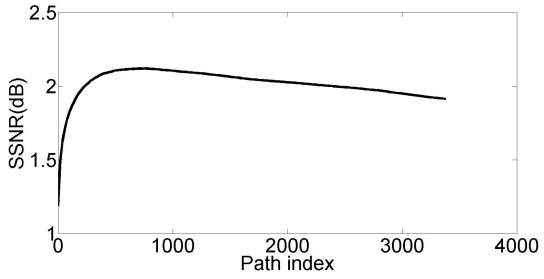

Figure 1: *SSNR of enhanced speech vs path number: white noise, input SNR = 5dB*

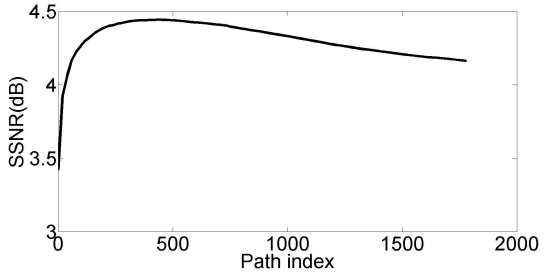

Figure 2: *SSNR of enhanced speech vs paths number: destroyer noise, input SNR = 5dB*

- HMM ($K = k^*$): The HMM approach with interpolation of k^* paths. The values of k^* are given as in Section 5.2.

The average values of SSNR, PESQ and SD obtained by the three methods are shown as in Tables 2 and 3. The SSNR, PESQ and SD for noisy speech are also given for reference. With the baseline CB approach, the SSNR of noisy speech is increased by about 3 dB for white noise and by about 5 dB for destroyer noise. The PESQ and SD are also improved significantly. Comparing the performances of CB and HMM ($K = 1$), the effectiveness of modeling temporal dependency is clearly seen. For white noise and destroyer noise at 0 dB SNR, the SSNR is increased by 0.4 dB and 0.2 dB respectively. The improvement on PESQ and SD measures is not significant. With the interpolation of multiple HMM paths, all performance measures are further improved to a noticeable extent for both types of noise and different SNRs. The overall SSNR improvement achieved by the proposed algorithm is about 5-6 dB and the SD is reduced by 40-50%. Sample files of enhanced speech can be downloaded at "`http://www.ee.cuhk.edu.hk/~chwong/download/Samples.zip`".

6. Conclusion

This study has confirmed the importance of temporal information in speech and demonstrated the benefit of modeling temporal dependency in LP model parameters estimation. HMM provides a straightforward way of temporal modeling and an established framework of estimating temporal sequences. The proposed interpolation method effectively exploits multiple HMM state sequences and leads to a robust estimation that extends beyond the coverage of the pre-trained codebook. As shown in [15], robust estimation of LP parameters can be integrated with speech periodicity enhancement to significantly improve the quality and intelligibility of enhanced speech.

Table 2: *Speech enhancement performance on white noise*

Input SNR(dB)	Methods	SSNR	PESQ	SD
5dB	Noisy speech	-2.3	1.9	25.3
	CB	0.9	2.3	17.5
	HMM ($K = 1$)	1.2	2.3	16.7
	HMM ($K = k^*$)	2.2	2.7	12.1
0dB	Noisy speech	-5.0	1.6	29.6
	CB	-1.7	1.9	20.9
	HMM ($K = 1$)	-1.3	1.9	20.0
	HMM ($K = k^*$)	-0.3	2.4	14.3

Table 3: *Speech enhancement performance on destroyer noise*

Input SNR(dB)	Methods	SSNR	PESQ	SD
5dB	Noisy speech	-2.0	2.4	16.6
	CB	3.1	2.6	11.1
	HMM ($K = 1$)	3.2	2.6	11.1
	HMM ($K = k^*$)	4.2	2.8	10.0
0dB	Noisy speech	-4.7	2.0	20.5
	CB	0.5	2.2	13.8
	HMM ($K = 1$)	0.7	2.2	13.7
	HMM ($K = k^*$)	1.7	2.4	11.8

7. References

[1] John R. Deller, Jr. , John H. L. Hansen, John G. Proakis, *Discrete Time Processing of Speech Signals,* IEEE Press, 2000.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[3] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding" *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 39, no. 9, pp. 1732-1742, Aug.1991.

[4] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement" *IEEE Trans. Speech Audio Process.*, vol. 3, no 4, pp. 251-266, Jul. 1995.

[5] S. Srinivasan, J. Samuelsson, and W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio. Speech, Language Process.*, vol. 14, no.1, pp. 163-176, 2006.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language. Process.*, vol. 21, no. 10, pp. 2140-2151, 2013.

[7] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, Language. Process.,* vol. 20, no. 6, pp. 1698-1712, 2012.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989

[9] A. N. Deoras and A. H. Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spo-

ken by multiple talkers on one audio channel," *IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2004

[10] A. Betkowska, K. Shinoda, S. Furui, "Speech Recognition using FHMMS Robust Against Non-stationary Noise," *IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 1029 - 1032, April. 2007

[11] K. K. Paliwal and W. B. Kleijn, *Speech Coding and Synthesis*, Elsevier Science Publication, Amsterdam, ch. 12, pp. 433-468, 1995

[12] The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT), *NIST Speech Disc. CD1-1.1.*, 1990

[13] NOISE-ROM 0, *Institute for Perception-TNO, PO Box 23, 3769 ZG Soesterberg, The Netherlands*.

[14] P.C., Loizou, *Speech enhancement: theory and practice*, CRC Press, 2007

[15] Feng Huang, Tan Lee, W. B. Kleijn and Ying-Yee Kong, "A method of speech periodicity enhancement using transform-domain signal decomposition," *Speech Communication*, vol. 67, pp. 102 - 112, 2015