INCORPORATING DYNAMIC TRACK INFORMATION FOR ALL-POLE PARAMETER ESTIMATION IN NOISE

Ruofei Chen and Cheung-Fat Chan

Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
ruofechen2@student.cityu.edu.hk, itcfchan@cityu.edu.hk

ABSTRACT

In this paper, we present a novel post-processing scheme to improve autoregressive (AR) speech parameter estimation in noise. The proposed technique exploits temporal correlation between dynamic all-pole parameters to capture natural speech evolution. To achieve this, a Kalman tracking scheme is proposed to track line spectrum frequency (LSF) trajectories with system parameter learned directly from processed online data. To facilitate the online system identification, a heuristic approach is initially proposed to preliminarily remove "musical tones" caused by conventional frame-based methods in LSF domain. Through performance evaluation based on a study of spectrogram and objective measures, it is demonstrated that the proposed post-processing scheme successfully restores natural and smooth evolution of speech dynamics, and in the meantime, effectively removes processing artifacts caused by conventional methods in various conditions.

Index Terms— temporal correlation, autoregressive(AR) model, line spectrum frequencies(LSF), musical tones, Kalman filter

1. INTRODUCTION

In classical speech enhancement, the reduction of background noise is often realized by a frame-based suppression gain function that is optimally derived from estimated clean and noise power spectrum coefficients. However, by taking advantage of the autoregressive (AR) modeling of speech, the gain function can be effectively approximated with AR spectrum samples, and hence the speech enhancement problem is transferred into an all-pole parameter estimation problem. There are several distinguishable advantages of using AR modeling in this context. First, by operating in reduced dimension, it is computational cost-efficient to perform online iterative parameter learning (e.g. iterative Wiener filtering (IWF) using maximum $a\ posteriori$ (MAP) techniques [1] and Kalman filtering using expectation-maximization (EM) algorithm [2]). Second, it is possible to impose constraints directly on all-pole parameters to limit the pole movement, in order to avoid the unrealistic spectral changes [3]. Furthermore, it facilitates the use of pre-cleaned AR codebooks with affordable size for data-driven methods [4].

However, there are certain trade-offs for these conventional AR-based methods. In non-data-driven approaches, AR parameter are directly estimated from noisy observations on a frame basis with certain optimal filtering rules. However, due to the fast varying nature of both speech and noise, it is extremely difficult to obtain low-variance AR estimate in low signal-to-noise ratio (SNR) regions. As a consequence, disturbing artifacts (commonly referred as "musical tones") are randomly created in time-frequency domain. Although various temporal constraints can be imposed to improve the estimate [3][5], due to gain fluctuation in frame-based processing, these approaches generate more or less "musical tones", depending on the operating conditions. On the other hand, in data-driven approaches [6][4], "musical tones" are effectively alleviated as AR estimates are replaced with pre-trained clean parameters. Nevertheless, several additional issues such as computational complexity, training model mismatch, and identification robustness are raised to limit their practical uses.

In this paper, we attempt to balance the above-mentioned trade-offs and we propose an online two-stage post-processing technique to re-estimate AR parameters by exploiting speech dynamic features. More specifically, a self-adaptive Kalman tracking technique is proposed to capture the long-term speech evolution and system parameters are learned directly from processed online data. To facilitate the online system parameter learning, a heuristic approach is initially proposed to preliminarily remove "musical tones" caused by conventional frame-based methods. The proposed method has several distinguishable differences as compared to conventional approaches for AR parameter estimation. First, it truly takes into account the speech dynamic track information rather than simply performing weighted averaging. In doing so, the spectral envelope evolution is well captured and hence the AR estimate react fast to the change of speech and noise signals. Second, the proposed enhancement process is training-free, and is performed in line spectral frequencies (LSF) domain with reduced dimension. Therefore, the computation load is significantly reduced and the improved AR estimates can be universally employed in related suppression gain functions. The effectiveness of the proposed method is evaluated based on based on a study of spectrogram and objective measures including log likelihood ratio (LLR) log spectral distance (LSD), and perceptual evaluation of speech quality (PESQ).

The remainder of this paper is organized as follows. In Section 2, the proposed two-stage post-processing scheme to improve AR estimate is presented. In Section 3, the performance of the proposed method is evaluated and compared with conventional methods. Finally, discussion and conclusion are addressed in Section 4.

2. PROPOSED POST-PROCESSING SCHEME

The main idea to obtain smooth and accurate AR estimates is to restore the original naturally evolving speech tracks, given preliminarily filtered observation which might contain "musical tones". Previous investigation [6] suggests that decent results can be obtained via Kalman filtering with system parameters learned from a parallel training corpus. Nevertheless, to facilitate the use of Kalman filter framework in training-free situations, a two-stage post-processing scheme is developed in this contribution. In the first stage, a heuristic musical-tone-removal step is designed to

combat "musical tones" and hence to improve the subsequent online Kalman parameter learning. In the second stage, an adaptive Kalman smoother is performed with system parameters such as the state transition matrix $\mathbf{F}$ (with temporal track information) and observation-state-mapping matrix $\mathbf{H}$ (with corruption type information) learned from musical-tone-removed LSFs.

2.1. Heuristic "Musical Tones" Removal

It is already verified in [3][5] that incorporating additional temporal information is extremely beneficial to combat "musical tones". In order to look for long-term speech evolution, we propose to exploit the speech dynamic feature in LSF domain. The major reason is twofold. First, due to the fact that "musical tones" often manifest as isolated peaks or "short ridges" [7] in time-frequency domain, it is possible to distinguish natural formants and "musical tones" with a series of LSF observation by inspecting the difference between adjacent odd-indexed and even-indexed LSFs as they are indicative of the formant bandwidth [3]. Second, it is shown in [6] that, with robust system identification, speech temporal trajectories can be effectively captured by tracking LSF parameters.

2.1.1. Classification

Initially, a series of N preliminarily filtered power spectra $|\hat{X}_\ell(\omega)|^2$ with $\ell = 1, 2, \ldots, N$ are converted to autocorrelation sequence according to Wiener-Khinchin relationship

$$r_{\ell,k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{X}_\ell(\omega)|^2 \cos(k\omega) d\omega \tag{1}$$

An order P linear predictive coding (LPC) analysis is then performed to compute LSFs for each time frame. The basic processing unit in the block-based analysis is a $P \times N$ dynamic feature matrix $\mathbf{U}$ with each column being a set of LSF coefficients $\theta_{\ell,k}$ with $k = 1, 2, \ldots, P$. In addition, a $P/2 \times N$ position matrix $\mathbf{S}$ is defined by the odd-indexed LSFs as

$$s_{\ell,k} = \theta_{\ell,2k-1}, \quad k = 1, 2, \ldots, P/2 \tag{2}$$

and a $P/2 \times N$ difference matrix $\mathbf{D}$ are defined by the difference between adjacent odd-indexed and even-indexed LSFs as

$$d_{\ell,k} = \min_{j=-1,1} |\theta_{\ell,2k+j} - \theta_{\ell,2k}|, \quad k = 1, 2, \ldots, P/2 \tag{3}$$

By comparing each element in $\mathbf{D}$ with a pre-defined bandwidth threshold δ, spectral peaks, either natural formants or artifacts, are identified if $d_{\ell,k} < \delta$. The next step is to distinguish "musical tones" from natural formants by evaluating the temporal track information. In each located peak with index $[\ell^*, k^*]$, a $(2\Delta_t + 1) \times (2\Delta_f + 1)$ stripe region ($\Delta_t > \Delta_f$) centered at $[\ell^*, k^*]$ is defined, where Δ_f and Δ_t are the frequency deviation tolerance and the formant bandwidth threshold, respectively. The current peak is classified as a component in a formant track if it extends at least one direction (backward or forward) along the time track. By defining the backward track as R_b, the forward track as R_f, it is interpreted mathematically as

$$\min\left(\frac{\sum_{i,j\in R_b} d_{i,j}}{L_b}, \frac{\sum_{i,j\in R_f} d_{i,j}}{L_f}\right) < \rho \tag{4}$$

where the L_b and L_f are the total number of points belong to R_b and R_f, respectively. ρ is the average formant bandwidth threshold to discriminate track and randomness. The current point (i, j) belongs to a backward/forward track if it satisfies the following

$$(i,j) \in R_b, \quad \text{if } i \in [\ell^* - \Delta_t, \ell^*], j \in [k^* - \Delta_f, k^* + \Delta_f], \text{and} |s_{\ell^*,k^*} - s_{i,j}| < \beta \tag{5}$$

$$(i,j) \in R_f, \quad \text{if } i \in [\ell^*, \ell^* + \Delta_t], j \in [k^* - \Delta_f, k^* + \Delta_f], \text{and} |s_{\ell^*,k^*} - s_{i,j}| < \beta \tag{6}$$

where $|s_{\ell^*,k^*} - s_{i,j}| < \beta$ explains the constraint that the current point (i, j) is connected to the center point (ℓ^*, k^*) only if it appears in adjacent frequency locations (in terms of formant position parameters), and the pre-defined threshold β defines the term "adjacent" in this context.

2.1.2. Processing Treatment

In order to remove an isolated peak that is identified with index (ℓ^*, k^*), a spectral smoothing procedure is developed to average out LSFs over a "safety" region without affecting other formants. More specifically, an additional backward and forward search (along the frequency track, started from k^*) is performed on the ℓ^{*th} column of the difference matrix $\mathbf{D}$ to find the last peak ($d_{\ell^*,k_{\text{lp}}} < \delta$) and the next peak ($d_{\ell^*,k_{\text{np}}} < \delta$) for this particular frame. The "safety" region is defined as $[2k_{\text{lp}} + 2, 2k_{\text{np}} - 2]$ (note that index k in matrix $\mathbf{D}$ may affect $[2k - 1, 2k + 1]$ region in θ). Consequently, the target peak is smoothed out by averaging LSFs over the region with the increment τ being

$$\tau = \frac{\theta_{\ell^*,2k_{\text{np}}-2} - \theta_{\ell^*,2k_{\text{lp}}+2}}{2(k_{\text{np}} - k_{\text{lp}} - 2)} \tag{7}$$

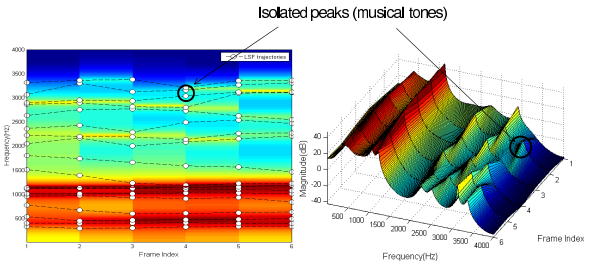
The complete algorithm for heuristic "musical tones" removal is summarized in Table 1. The effectiveness of the proposed approach is illustrated in Fig.1. By contrasting the proposed post-processing technique in Fig.1(a) with a conventional decision-directed Wiener filter (DD_WF) [5] in Fig.1(b), it is noted that the proposed approach effectively adjusts the irregular points in LSF trajectories and also removes the isolated peaks in the spectrograms.

Table 1: Proposed algorithm for heuristic "musical tones" removal

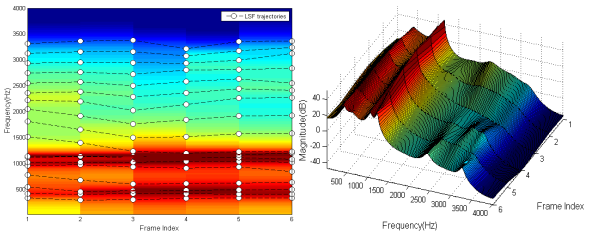
Initial condition: Position matrix $\mathbf{S}$, Difference matrix $\mathbf{D}$
// *Search the difference matrix* $\mathbf{D}$
For each $\ell \in [1, N]$ $\quad k \in [1, P/2]$
  // *locate a peak*
  If $d_{\ell,k} < \delta$
    // *Search the stripe centered at the peak* (ℓ, k)
    For each $i \in [\ell - \Delta_t, \ell + \Delta_t]$, $\quad j \in [k - \Delta_f, k + \Delta_f]$
      Classify the peak using (4)(5)(6)
      If it is an isolated peak
        Perform the smoothing using (7)
      End
    End
  End
End

2.2. Speech Dynamics Tracking

The heuristic approach developed in last section can effectively remove "musical tones". However, it is noted that there are still several issues to tackle. First, the formant is not substantially improved and it is often sharper than the original since conventional methods often emphasize high SNR regions. Second, the spectral envelopes are



(a) Processed by decision-directed Wiener filtering (musical tone retained)



(b) Processed by proposed heuristic filtering (musical tone removed)

Fig. 1: Comparison of LSF trajectories and 3D surface spectrograms

less ordered compared to the original since the temporal constraint is only imposed in classification, the smoothing is still performed on a frame-by-frame basis. In order to incorporate the true trajectory information, we propose to apply the dynamic tracking scheme on the heuristically enhanced LSFs to further improve the AR parameter estimation.

A Kalman tracking framework is initially proposed in [6] to model the noisy and clean LSF blocks as the observation and state sequences using a linear dynamical system (LDS). Given Kalman system parameters $\mathbf{\Theta} = \{\mathbf{F}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \hat{\mathbf{x}}_1, \mathbf{\Sigma}_1\}$ (where $\mathbf{F}$ is the state transition matrix, $\mathbf{H}$ is a state-observation mapping error, $\mathbf{Q}$ and $\mathbf{R}$ are state and observation error covariances, $\hat{\mathbf{x}}_1$ and $\mathbf{\Sigma}_1$ are initial state mean and error covariance, respectively), clean all-pole parameters can be readily derived by applying Kalman tracking over an analysis block. However, the major difference between the system identification in [6] and that in this contribution is that a training corpus is no longer required here, and Kalman system parameters are learned directly in the enhancement stage. Instead of using a parallel corpus, it is proposed to estimate the model parameters $\mathbf{\Theta}$ from the heuristically enhanced feature block $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N\}$ and the noisy observation block $\mathbf{Y}$, with diagonal constraints imposed on $\mathbf{F}$ and $\mathbf{H}$. The diagonal constraints are crucial in this context for two major reasons. First, it constrains the evolution of each LSF coefficient to be linear along the time track (cross-correlation between LSFs with different frequency locations are minimized). In doing so, $\mathbf{F}$ and $\mathbf{H}$ are forced to be smooth in nature and therefore tolerate no irregularity. Second, as only a single block is adopted in this learning, ill-condition might be posed in calculating the matrix inversion in ML estimates due to insufficient data. The diagonal constraints offer great flexibility in the track length requirement. To achieve this, the ML learning of state transition matrix $\mathbf{F}$ with diagonal constraints can be effectively defined by minimizing the following objective function

$$\mathcal{J} = \sum_{\ell=2}^{N} (\tilde{\mathbf{x}}_\ell - diag(\mathbf{f})\tilde{\mathbf{x}}_{\ell-1})^T (\tilde{\mathbf{x}}_\ell - diag(\mathbf{f})\tilde{\mathbf{x}}_{\ell-1}) \qquad (8)$$

with $\mathbf{f}$ being the vector containing diagonal elements of $\mathbf{F}$. The standard least square (LS) solution gives $\mathbf{f}$ as

$$\mathbf{f} = [\sum_{\ell=2}^{N} diag(\tilde{\mathbf{x}}_{\ell-1})]^{-1} \sum_{\ell=2}^{N} \tilde{\mathbf{x}}_\ell \qquad (9)$$

Similarly, the objective function to obtain a vector $\mathbf{h}$ that containing diagonal elements of $\mathbf{H}$ is

$$\mathcal{J} = \sum_{\ell=1}^{N} (\mathbf{y}_\ell - diag(\mathbf{h})\tilde{\mathbf{x}}_\ell)^T (\mathbf{y}_\ell - diag(\mathbf{h})\tilde{\mathbf{x}}_\ell) \qquad (10)$$

which gives

$$\mathbf{h} = [\sum_{\ell=1}^{N} diag(\tilde{\mathbf{x}}_\ell)]^{-1} \sum_{\ell=1}^{N} \mathbf{y}_\ell \qquad (11)$$

The state transition matrix is computed as $\mathbf{F} = diag(\mathbf{f})$ and corruption mapping matrix is computed as $\mathbf{H} = diag(\mathbf{h})$, respectively. The rest of model parameters are computed with re-estimated $\mathbf{F}$ and $\mathbf{H}$ using the standard ML estimator (in this case, a single training block) described in [6]. In the final stage, for each analysis block, the suppression gain function is recomputed using with AR spectrum samples constructed from tracked LSFs.

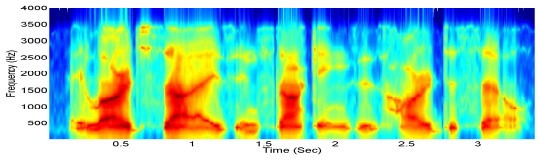
3. PERFORMANCE EVALUATION

The effectiveness of the proposed two-stage post-processing scheme is evaluated in this section. Fig.2 shows the envelope spectrograms (comprise of consecutive energy-normalized spectral envelopes) of a sentence "a large size in stockings is hard to sell" spoken by a male speaker in various conditions. By comparing Fig.2(a)-(c) it is observed that the classical WF with DD *a priori* SNR estimator (similar for other conventional spectral weighting techniques) effectively suppress the average noise floor while retaining considerable speech information. However, it is also noted in Fig.2(c) that plenty of "smeared spots" randomly reside on the processed spectrogram and the formant bandwidth is smaller as compared to the original in Fig.2(a). Nevertheless, as demonstrated in Fig.2(d), the above two common deficiencies of conventional methods are significantly improved by applying the proposed post-processing scheme. It is noticed that the "smeared spots" are removed and the entire evolution is natural and smooth with speech information well-retained.

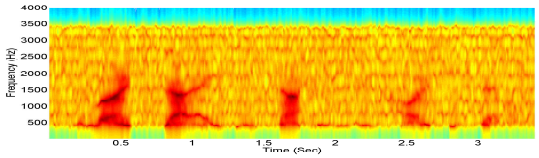
The objective measures consist of a log likelihood ratio (LLR) measure which assesses the dissimilarity between energy-normalized all-pole spectra, a log spectral distance (LSD) measure which taking into account the additional excitation variance information, and a standard perceptual evaluation of speech quality (PESQ) measure to evaluate the overall speech quality improvement. The proposed approach is evaluated with following experimental settings. Clean speech and noise are taken from IEEE sentence database and NOISEX-92 database, respectively. Clean speech is manually corrupted by additive noise at SNR level from 0dB to 10dB, with a step size of 5dB. Two types of noise, namely, car interior noise and babble noise are adopted. The sampling frequency is 8KHz. The block and frame duration are 112ms and 32ms, respectively. The frame shift is 8ms. The size of short-time Fourier transform (STFT)

Table 2: Objective evaluation results

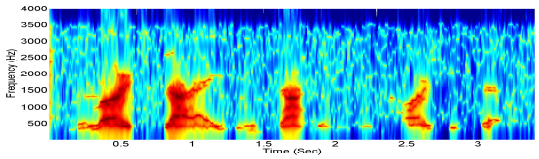
		Objective Evaluation Results								
		LLR Distance			LSD (in dB)			PESQ (out of 4.5)		
Noise		Input SNR			Input SNR			Input SNR		
Type	Method	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Car	NOISY	1.126	0.976	0.803	17.71	15.47	12.03	1.86	2.18	2.35
Interior	DD_WF	0.991	0.784	0.607	12.39	10.84	9.185	2.04	2.34	2.54
Noise	OT_WF	0.875	0.644	0.512	9.912	8.857	8.074	2.18	2.46	2.62
	CB_WF	0.623	0.434	0.334	6.541	5.642	4.779	2.35	2.51	2.73
	DD_HNM	–	–	–	–	–	–	2.21	2.39	2.62
	OT_HNM	–	–	–	–	–	–	2.32	2.50	2.71
	CB_HNM	–	–	–	–	–	–	2.42	2.62	2.80
Babble	NOISY	1.222	1.053	0.874	17.95	14.38	10.29	1.91	2.17	2.28
Noise	DD_WF	1.094	0.844	0.638	13.70	10.84	7.816	2.06	2.24	2.38
	OT_WF	0.884	0.751	0.598	9.446	8.612	6.426	2.18	2.34	2.46
	CB_WF	0.712	0.434	0.334	6.541	5.642	4.779	2.32	2.48	2.65
	DD_HNM	–	–	–	–	–	–	2.28	2.39	2.62
	OT_HNM	–	–	–	–	–	–	2.39	2.47	2.70
	CB_HNM	–	–	–	–	–	–	2.51	2.67	2.88



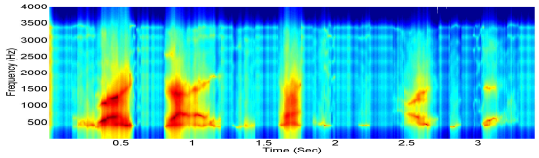
(a) clean speech



(b) noisy (corrupted by white noise at SNR = 5dB)



(c) Wiener filtering with DD *a priori* SNR estimator



(d) Proposed post-processing scheme

Fig. 2: Comparison of spectrograms in various conditions

is 256 and the order of LPC analysis is 18. Other parameters are empirically set as follows. The bandwidth threshold is $\delta = 0.1$, the frequency deviation threshold is $\beta = 0.1$, and average bandwidth threshold is $\rho = 0.12$, all in radius. Moreover, $\Delta_t = 4$ and $\Delta_f = 1$ are defined for stripe offsets.

In this evaluation, the online tracked (denoted with prefix OT_) AR parameters are fed into both classical Wiener filter (WF) gain function [1] and harmonic noise model (HNM) based analysis-synthesis framework [8], to compare with its non-data-driven (AR parameter estimated with classical DD estimator [5], denoted with prefix DD_) and data-driven (AR parameter estimated with pre-trained LDS codebook [6], denoted with prefix CB_) counterparts. The results of above three measures with various noise and SNR settings are shown in Table.2 (the LLR and LSD measures for HNM-based approaches is omitted as they use the identical AR estimates adopted in WF-based approaches). It is observed that the proposed method balances the trade-off between conventional decision-directed method and codebook driven method in obtaining natural and smooth AR estimates. It indicates that the proposed method can be employed as a post-processing tool to improve over conventional spectral weighting techniques, without imposing additional computational burden brought by codebook design. Besides, informal subjective evaluation also suggests that the proposed method effectively alleviates perceptual annoying "musical tones".

4. DISCUSSION AND CONCLUSION

This work is developed based on the fundamental LSF modeling of speech evolution proposed in [6]. The major difference is that a training corpus is no longer required by imposing diagonal constraints on state transition and state-observation mapping matrices. To achieve this, a novel heuristic musical-tone-removal approach is proposed to facilitate the online Kalman system parameter learning. The tracked LSF estimates can be connected with classical spectral weighting rules (e.g. [5]) or analysis-synthesis modules (e.g. [8]) for speech enhancement. Objective evaluation results demonstrate the effectiveness of the proposed technique.

5. REFERENCES

[1] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197–210, June 1978.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[3] J. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795 – 805, Apr. 1991.

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

[5] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 629–632, May 1996.

[6] R. Chen, C.-F. Chan, and H. C. So, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1324 –1336, May 2012.

[7] G. Zenton, K.-C. Tan, and T.G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287 –292, may 1998.

[8] R. F. Chen, C. F. Chan, H. C. So, J. Lee, and C. Y. Leung, "Speech enhancement in car noise environment based on an analysis-synthesis approach using harmonic noise model," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4413–4416, Apr. 2009.