A Correlation-Maximization Denoising Filter Used as An Enhancement Frontend for Noise Robust Bird Call Classification

Wei Chu, Abeer Alwan

Department of Electrical Engineering
University of California, Los Angeles
{weichu, alwan}@ee.ucla.edu

Abstract

In this paper, we propose a Correlation-Maximization denoising filter which utilizes periodicity information to remove additive noise in bird calls. We also developed a statistically-based noise robust bird-call classification system which uses the denoising filter as a frontend. Enhanced bird calls which are the output of the denoising filter are used for feature extraction. Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) are used for classification. Experiments on a large noisy corpus containing bird calls from 5 species have shown that the Correlation-Maximization filter is more effective than the Wiener filter in improving the classification error rate of bird calls which have a quasi-periodic structure. This improvement results in a 4.1% classification error rate which is better than the system without a denoising frontend and a system with a Wiener filter denoising frontend.

Index Terms: Correlation-Maximization filter, speech enhancement, bird call classification

1. Introduction

Bird songs are important in the communication between birds of specific species. A bird can listen to other birds and classify them as conspecific or heterospecific, neighbor or stranger, mate or non-mate, kin or non-kin [1]. It can also sing to other birds for mate attraction, danger alert, or territory defense [2]. Behavioral and ecological studies could benefit from automatically detecting and identifying species from acoustic recordings.

Researchers from the Ecology and Evolutionary Biology department at UCLA recorded calls from 5 species of Antbirds (Barred Antshrike, Dusky Antbird, Great Antshrike, Mexican Antthrush, Dot-winged Antwren) in a Mexican rainforest [3]. Different kinds of background noises are observed in the recordings, such as other bird chirps, insect sounds, and instrument noises. Pre-processing is needed to suppress the background noise and enhance the target bird call before feature extraction.

A prevailing denoising filter is the Wiener filter which estimates the additive noise spectrum and adaptively updates the frequency response of the filter [4].

Suppose that the clean signal $x[n]$ and the noisy signal $y[n]$ are wide sense stationary, and $x[n]$ and the additive noise $v[n]$ are uncorrelated. After minimizing the mean square error, we have the relationship between the spectrum of estimated signal $\hat{x}[n]$ and noisy signal $y[n]$ denoted by $\hat{X}(f)$ and $Y(f)$:

$$|\hat{X}(f)|^2 = H(f)|Y(f)|^2 \tag{1}$$

Supported in part by NSF

where $H(f)$ denotes the frequency response of the filter $h(n)$. The estimate of the Signal-to-Noise Ratio (SNR) at frequency f denoted by $\widehat{\mathrm{SNR}}(f)$ can be expressed as:

$$\widehat{\mathrm{SNR}}(f) = \frac{|\hat{X}(f)|^2}{|\hat{V}(f)|^2} \tag{2}$$

where $\hat{V}(f)$ is the estimated spectrum of the noise signal $v[n]$. Note that $|Y(f)|^2 = |\hat{X}(f)|^2 + |\hat{V}(f)|^2$ assuming $\hat{X}(f)$ and $\hat{V}(f)$ are orthogonal, The estimated clean spectrum can be expressed as:

$$|\hat{X}(f)|^2 = \frac{\widehat{\mathrm{SNR}}(f)}{1 + \widehat{\mathrm{SNR}}(f)}|Y(f)|^2 \tag{3}$$

Therefore, the noncausal Wiener filter converts the denoising problem into an SNR estimation problem [5].

According to our observation, the Wiener filter sometimes fails to identify background chirps as noise and enhances both the target and non-target chirps.

In order to suppress background chirps, we utilize the periodicity of the chirps in the bird call to develop a denoising filter which enhances the periodic structure of the target call. The coefficients of the filter are obtained through a gradient search approach which maximizes the value of a correlation based function. Therefore, we call it a Correlation-Maximization denoising filter.

In the following sections, we analyze the characteristics of some bird calls, design the Correlation-Maximization filter, and develop a statistically-based bird call classification system. We also discuss the advantage of the Correlation-Maximization filter over Wiener filtering in the bird call classification problem.

2. A Correlation-Maximization Filter

According to our observation, every Antbird call is quasi-periodic in terms of the interval between chirps, and the intervals slowly decrease with time. An example is shown in Figure 1.

In the following, we will discuss how to search an optimal correlation based denoising filter which can enhance this periodic structure.

2.1. Search Chirp Interval Using a Correlation Function

Suppose a bird call $x[n]$ is corrupted by an additive noise $v[n]$. The noisy acoustic signal $y[n]$ is the input to an FIR filter $h[n]$ of L taps. The output of the filter is the estimation of $x[n]$:

$$\hat{x}[n] = \sum_{k=1}^{L} h[k]y[n-k] \tag{4}$$

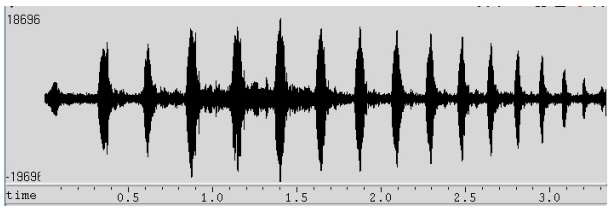

Figure 1: The waveform of a Great Antshrike (GAS) call

$y[n]$ is decomposed into M frames with a frame step size of Δ and a frame length of N, and we assume that $y[n]$ and $x[n]$ are wide sense stationary in each frame. Since the spectral distributions of different frames in a bird call are similar, a single $\mathbf{h}$ is assumed for each bird call. Therefore at frame m, the cross correlation function of $\hat{x}[n]$ at lag k denoted by $\phi_x^m[0,k]$ can be expressed as:

$$\begin{aligned}\phi_{\hat{x}}^m[0,k] &= \sum_{n=m\Delta}^{m\Delta+N-1-k} \hat{x}[n]\hat{x}[n+k] \qquad (5)\\ &= \sum_{p=1}^{L} h[p] \sum_{q=1}^{L} h[q] \sum_{n=m\Delta}^{m\Delta+N-1-k} y[n-p]y[n+k-q].\end{aligned}$$

Note that the lag k has K possible values, $k = k_0, k_1, ...k_{K-1}$. We can define an $L \times L$ cross correlation function matrix $\mathbf{\Phi}_y^m[0,k]$ for frame m at lag k. The element of $\mathbf{\Phi}_y^m[0,k]$ in row p and column q is expressed as:

$$\mathbf{\Phi}_y^m[0,k]_{pq} = \sum_{n=m\Delta}^{m\Delta+N-1-k} y[n-p]y[n+k-q]. \quad (6)$$

Therefore we have

$$\phi_{\hat{x}}^m[0,k] = \mathbf{h}^T \mathbf{\Phi}_y^m[0,k]\mathbf{h} \quad (7)$$

where $\mathbf{h} = [h[0], h[1], \cdots, h[L]]^T$ denotes the coefficients of the FIR filter. To confine the dynamic range of $\phi_{\hat{x}}^m[0,k]$, the normalized cross correlation function $\bar{\phi}_{\hat{x}}^m[0,k]$ is used as follows:

$$\bar{\phi}_{\hat{x}}^m[0,k] = \frac{\phi_x^m[0,k]}{\sqrt{\phi_x^m[0,0]\phi_x^m[k,k]}} \quad (8)$$

Note that $\bar{\phi}_{\hat{x}}^m[0,k] \in [-1,1]$.

It is possible to find the chirp interval in each frame over the denoised signal $\hat{x}[n]$. Dynamic programming can be used to minimize the distortion induced by background noise in the chirp interval search [6]. Because the objective of the dynamic programming is to search the path which has a minimum accumulative cost. The local cost of frame m at lag k is defined as $-\bar{\phi}_{\hat{x}}^m[0,k]$. Since the chirp interval is gradually decreasing over time, the cost of transitioning from lag k_i to k_j denoted by $d(k_i,k_j)$is defined as follows:

$$d(k_i,k_j) = e^{\alpha|k_i-\delta-k_j|} - 1 \qquad i,j = 0,\cdots,K-1 \quad (9)$$

where α and δ are pre-set empirically. α is a scaling factor, and δ is an estimate of how fast the chirp interval changes per second. This exponential function can impose more penalty than its linear counterpart on the transition cost in order to prevent chirp intervals from greatly varying between two consecutive frames. Note that when $k_j = k_i - \delta$, $d(k_i,k_j) = 0$.

Then, a trellis structure of $K \times M$ for dynamic programming is built, where M is the number of total frames, K is the number of the possible candidates at each frame. $\mathbf{s} = [s_1, s_2, \cdots, s_M]$ is used to denote an arbitrary valid path in the trellis.

2.2. Search The Optimal Denoising Filter

Usually in matched filtering [7], the optimal linear filter is obtained by maximizing the SNR in the presence of additive noise. We search the filter coefficients in a grid by minimizing a correlation-based cost function.

It can be assumed that an optimal filter $\mathbf{h}$ can enhance the periodic structure of the target bird call and remove the additive noise so that the minimum accumulative cost is achieved in the chirp interval search over the denoised signal. This assumption can be expressed as:

$$\mathbf{h}^* = \arg\min_{\mathbf{h}} \mathcal{F}(\mathbf{h},\mathbf{s}) \quad (10)$$

where $\mathbf{h}^*$ denotes the optimal denoising filter, the accumulative cost $\mathcal{F}(\mathbf{h},\mathbf{s})$ which is summation the accumulative local and transition costs is expressed as:

$$\mathcal{F}(\mathbf{h},\mathbf{s}) = \sum_{m=1}^{M} -\bar{\phi}_{\hat{x}}^m[0,s_m] + \sum_{m=1}^{M-1} d(s_m,s_{m+1}). \quad (11)$$

The gradients of $\mathcal{F}(\mathbf{h},\mathbf{s})$ w.r.t. $\mathbf{h}$ can be expressed as:

$$\begin{aligned}&\nabla_{\mathbf{h}}\mathcal{F}(\mathbf{h},\mathbf{s}) \qquad (12)\\ =\ &\nabla_{\mathbf{h}}\{-\sum_{m=1}^{M} \frac{\mathbf{h}^T\mathbf{\Phi}_y^m[0,s_m]\mathbf{h}}{\sqrt{\mathbf{h}^T\mathbf{\Phi}_y^m[0,0]\mathbf{h}}\sqrt{\mathbf{h}^T\mathbf{\Phi}_y^m[s_m,s_m]\mathbf{h}}}\} + \mathbf{0}\\ =\ &-\sum_{m=1}^{M} \frac{1}{\sqrt{\mathbf{h}^T\mathbf{\Phi}_y^m[0,0]\mathbf{h}\cdot\mathbf{h}^T\mathbf{\Phi}_y^m[s_m,s_m]\mathbf{h}}}\cdot\\ &[\frac{\mathbf{\Phi}_y^m[0,s_m]+\mathbf{\Phi}_y^m[0,s_m]^T}{\mathbf{h}^T\mathbf{\Phi}_y^m[0,s_m]\mathbf{h}} - \frac{1}{2}\frac{\mathbf{\Phi}_y^m[0,0]+\mathbf{\Phi}_y^m[0,0]^T}{\mathbf{h}^T\mathbf{\Phi}_y^m[0,0]\mathbf{h}}\\ &-\frac{1}{2}\frac{\mathbf{\Phi}_y^m[s_m,s_m]+\mathbf{\Phi}_y^m[s_m,s_m]^T}{\mathbf{h}^T\mathbf{\Phi}_y^m[s_m,s_m]\mathbf{h}}]\mathbf{h}\end{aligned}$$

Therefore, the gradient descent method can be used to search the optimal filter $\mathbf{h}^*(\mathbf{s})$ for a path $\mathbf{s}$. The minimum cost is achieved when $\nabla_{\mathbf{h}}\mathcal{F}(\mathbf{h},\mathbf{s}) = \mathbf{0}$. Note that $\mathbf{s}$ is independent of $\mathbf{h}$. The final optimal filter $\mathbf{h}^*$ can be searched using a brute-force method:

Algorithm 2.1: BRUTE-FORCE FILTER SEARCH **(h)**

Set the iteration time = I, the iteration stopping threshold = ϵ.
for all valid $\mathbf{s}$
do {
  Initialize $\mathbf{h}_0(\mathbf{s}) = [1,0,\cdots,0]^T$.
  for $i = 0$ **to** I
  do {
    $\mathbf{h}_{i+1}(\mathbf{s}) = \mathbf{h}_i(\mathbf{s}) - t_i\nabla_{\mathbf{h}}\mathcal{F}(\mathbf{h}_i(\mathbf{s}),\mathbf{s})$,
    where t_i is the step size;
    if $||\mathbf{h}_{i+1}(\mathbf{s}) - \mathbf{h}_i(\mathbf{s})||/||\mathbf{h}_i(\mathbf{s})|| < \epsilon$
      then $\mathbf{h}^*(\mathbf{s}) = \mathbf{h}_i(\mathbf{s})$, **break** .
  if $i == I$
    then $\mathbf{h}^*(\mathbf{s}) = \mathbf{h}_I(\mathbf{s})$.
The optimal path $\mathbf{s}^* = \arg\min_{\mathbf{s}} \mathcal{F}(\mathbf{h}^*(\mathbf{s}),\mathbf{s})$
The optimal filter $\mathbf{h}^* = \mathbf{h}^*(\mathbf{s}^*)$, **exit** .

2.3. Speed Up The Search: N-best Search

Instead of a grid search, we propose to search through a trellis similar to the N-best search in ASR.

There are K^M possible paths in a $K \times M$ trellis. Suppose the average iteration time of the gradient search is $\bar{I}$, this brute-force approach needs $K^M \times \bar{I}$ iterations which is computationally unacceptable.

If the gradient search stopped at iteration i, the optimal path among all the valid paths denoted by $\mathbf{s}_i^*$ can be expressed as:

$$\mathbf{s}_i^* = \arg\min_{\mathbf{s}} \mathcal{F}(\mathbf{h}_i(\mathbf{s}), \mathbf{s}). \quad (13)$$

Since $\mathbf{s}_i^*$ may not be equal to $\mathbf{s}^*$, we need to search through all possible paths; however, we can assume that $\mathbf{s}^*$ is within a path subset during each iteration. The subset is composed of the top N-best paths which are the output of the dynamic programming on the trellis. That means the gradient descent search only needs to be applied to the N-best paths, not all the paths at each iteration. Then the brute-force search approach can be improved into an N-best search:

Algorithm 2.2: N-BEST FILTER SEARCH($\mathbf{h}$)

Set the iteration time = I, the iteration stopping threshold = ϵ;
Set the N-best path number = J.
for $j = 0$ **to** J
  do Initialize an N-best (J) filter list $\mathbf{h}_0^j = [1, 0, \cdots, 0]^T$.
for $i = 0$ **to** I
  do
    for $j = 1$ **to** J
      do
        Use $\mathbf{h}_i^j$ to build j_{th} trellis by calculating $\bar{\phi}_{\tilde{x}}^m[k]$.
        Search N-best (J) paths in j_{th} trellis.
        for $k = 1$ **to** J
          do
            Select k_{th} best path denoted by $\mathbf{s}_i^{(j,k)}$,
            $\mathbf{h}_{i+1}^{(j,k)} = \mathbf{h}_i^j - t_i \nabla_{\mathbf{h}} \mathcal{F}(\mathbf{h}_i^j, \mathbf{s}_i^{(j,k)})$,
            where t_i is the step size.
    Sort $\mathbf{h}_{i+1}^{(j,k)}$, $j, k = 1, \cdots, J$, according to values of
    $\mathcal{F}(\mathbf{h}_{i+1}^{(j,k)}, \mathbf{s}_i^{(j,k)})$ in ascending order to obtain
    a sorted filter list denoted by $\tilde{\mathbf{h}}_{i+1}^l$, $l = 1, \cdots, J^2$
    for $j = 1$ **to** J
      do $\mathbf{h}_{i+1}^j = \tilde{\mathbf{h}}_{i+1}^j$.
    if $\max_{j=1\ldots J} \dfrac{||\mathbf{h}_{i+1}^j - \mathbf{h}_i^j||}{||\mathbf{h}_i^j||} < \epsilon$
    then $\mathbf{h}^* = \mathbf{h}_i^1$, **exit** .
if $i == I$
  then $\mathbf{h}^* = \mathbf{h}_I^1$, **exit** .

Although $J \times \bar{I}$ trellis building and dynamic programming operations are newly introduced in this N-best search approach, the total average gradient search iterations is reduced to $J^2 \times \bar{I}$ compared to the $K^M \times \bar{I}$ iterations in the brute-force search approach when the M is large. Typically, for Antbird calls, K = 49, $1 \leq M \leq 50$, J = 20.

3. Experiments

The Antbird call corpus contains 3366 bird calls from 5 species. We split the corpus into a training and testing set with a ratio of 2:1 as shown in Table 1. The training set is 85 minutes long and the testing set is 42 minutes long. The calls are 0.5 - 5.0 seconds long.

Table 1: Number of bird calls in the training and test sets. BAS: Barred Antshrike; DAB: Dusky Antbird; GAS: Great Antshrike; MAT: Mexican Antthrush; DWA: Dot-winged Antwren.

	BAS	DAB	GAS	MAT	DWA	Total
Training	240	888	350	609	159	2246
Testing	120	444	175	304	77	1120

Table 2: Classification error rate (%) on the test set. **W+**/**CM+**: feature extraction using the output of the Wiener/Correlation-Maximization based denoising filter

	GMM	HMM
MFCC	8.7	5.4
W+MFCC	5.9	4.9
CM+MFCC	5.3	4.6
CM+W+MFCC	**4.7**	**4.1**

The original single-channel acoustic signal is collected at a sampling rate of 44.1 kHz. The frequency range of the bird calls is from 500 to 6000 Hz. Thus, we use a band-pass filter with cutoff frequencies between 360 Hz and 6500 Hz to remove irrelevant frequency components. The signal is then downsampled to 16 kHz.

In the Correlation-Maximization denoising filter, the number of the filter taps (L) is 20. Since an analysis frame should contain at least two chirps to extract the chirp interval, and the bird chirp length ranges from 60 to 300 ms, the frame length is 600 ms, i.e. 9600 samples. The frame step size is 100 ms, and the correlation lag step is 5 ms. The number of lags is (300 - 60)/5 + 1 = 49. The maximum number of iterations (I) in the gradient search, and the number of N-best paths (J) are both 20. According to the experimental results, increasing L, I, or J does not boost the classification accuracy but does increase the computational cost.

A 39-dimension feature composed of the first 13 MFCCs and first and second derivatives is computed every frame for model training and testing.

In the GMM classifier, each species' model is set to have 256 Gaussians. In the HMM-based classifier, each species' model also has 256 Gaussians per state. The recognition network is the same as the one used in isolated word recognition, in which each species corresponds to a word node. Choosing the correct state number may enable finer modeling of a bird call. Since the number of chirps in a bird call varies and the state numbers are the same for all 5 species, state number 6, which is the minimum number of chirps in all the bird calls, is used for each species model and it also results in the lowest classification error rate among state numbers 1 to 9.

Classification results are shown in Table 2. The HMM-based classifier results in better performance than the GMM classifier when using the same features. After applying the Correlation-Maximization denoising filter, classification error rates of both GMM and HMM-based classifiers are lower than their counterparts using the Wiener filer. Since the Correlation-Maximization filter uses a long frame length to estimate the slow-varying noise and to capture interval periodicity, and the Wiener filter employs a relatively short frame length to track the fast changing noise, it is possible that cascading the two filters can further reduce error rates.

As shown in Table 3, the confusion matrix is used to analyze the classification errors. The calls of BAS, MAT, and DAB are less likely to be misclassified as other species compared to

Table 3: The confusion matrix of using CM+W+MFCC feature and HMM based classifier on the test set; **RE**: the number of errors divided by the total number of calls in the row; **PE**: the number of errors in the row divided by the total number of calls.

	BAS	DAB	GAS	MAT	DWA	RE(%)	PE (%)
BAS	120	0	0	0	0	0.0	0.0
DAB	1	430	7	5	1	3.2	1.2
GAS	6	3	149	17	0	14.9	2.3
MAT	0	0	0	304	0	0.0	0.0
DWA	0	4	2	0	71	7.8	0.5

those of GAS and DWA. The GAS→MAT errors (1.5%) accounted for more than 35% in the total errors (4.1%).

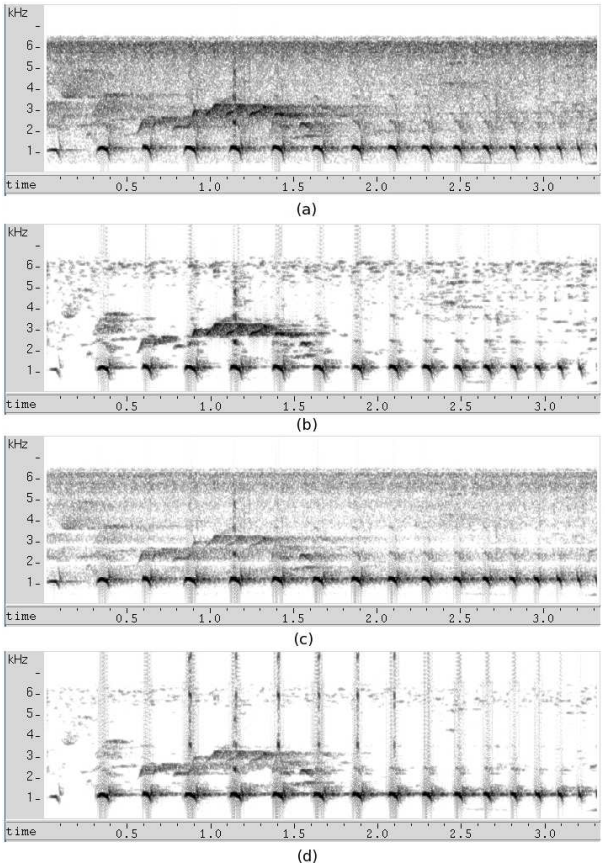

Figure 2: A Great Antshrike (GAS) call: (a) original spectrogram; (b) spectrogram after Wiener filtering; (c) spectrogram after Correlation-Maximization filtering; (d) spectrogram after Wiener and Correlation-Maximization filtering.

A GAS bird call is used to illustrate the difference between the Wiener and Correlation-Maximization filter. From Figure 2 (a), other bird chirps are observed from 0.6 to 1.6 seconds, and background noise can act as adverse factors to the classification task. As shown in Figure 2 (b), both target and non-target bird chirps are enhanced after Wiener filtering. That is because Wiener filter can not denoise discriminately. It can be seen from Figure 2 (c) that the Correlation-Maximization filter can suppress the non-target chirps while enhancing the target chirps. That is because the Correlation-Maximization filter is supposed

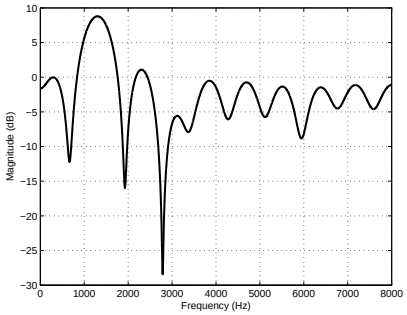

Figure 3: The frequency response of the Correlation-Maximization filter for a GAS call.

to only enhance the periodic structure of the target bird call. It is also shown in Figure 2 (d) that both non-target bird call and background noise are suppressed when cascading the Wiener filter and the Correlation-Maximization filter.

The frequency response of the optimal Correlation-Maximization denoising filter for this bird call is shown in Figure 3. The filter has a pass-band from 800 to 1750 Hz, which can enhance the frequency components of the target bird call, a stop-band from 2600 to 8000 Hz, and a dip around 2800 Hz can minimize the interference introduced by background noise and other bird chirps. Other filters were developed for other bird calls.

4. Conclusions

For bird calls which have a quasi-periodic structure in the time domain and a relatively invariant power spectral density across frames, a Correlation-Maximization based denoising filter is effective in enhancing the target bird calls which results in a reduction in classification error rate.

The advantage of the Correlation-Maximization based denoising filter over the Wiener filter is that it avoids estimating the SNR. Instead, it uses the periodicity of the bird call to instruct the denoising.

5. References

[1] P. Marler, "A comparative approach to vocal learning: song development in white-crowned sparrows," *J Comp Physiol Psychol*, vol. 71, pp. 1–25, 1970.

[2] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*, Cambridge University Press, New York, 1995.

[3] V. Trifa, A. Kirschel, and C. E. Taylor, "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2424–2431, 2008.

[4] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley Press, New York, 1949.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] D. Talkin, "Robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 497–518, 1995.

[7] G. Turin, "An introduction to matched filters," *IEEE Trans. on Information Theory*, vol. 6, no. 3, pp. 311–329, 1960.