PITCH EXTRACTION OF SPEECH SIGNALS USING AN EIGEN-BASED SUBSPACE METHOD

Takahiro Murakami\textsuperscript{a}, Munehiro Namba\textsuperscript{b}, Tetsuya Hoya\textsuperscript{c} and Yoshihisa Ishida\textsuperscript{d}

\textsuperscript{a,d}Department of Electronics and Communications, Meiji University
1-1-1, Higashi-Mita, Tama-Ku, Kawasaki, 214-8571, Japan
E-mail: \textsuperscript{a}murakami@bach.mind.meiji.ac.jp
\textsuperscript{d}ishida@isc.meiji.ac.jp
\textsuperscript{b}Department of Mathematics and Informatics, Tokyo Gakugei University
4-1-1, Nukuikitamachi, Koganei-City, Tokyo, 184-8501, Japan
E-mail: namba@u-gakugei.ac.jp
\textsuperscript{c}Laboratory for Advanced Brain Signal Processing, BSI-RIKEN
2-1, Hirosawa, Wakoh-City, Saitama, 351-0198, Japan
E-mail: hoya@bsp.brain.riken.go.jp

ABSTRACT

In this paper, we propose a novel method for detecting the fundamental frequencies of speech signals contaminated by noise. The proposed method exploits an eigen-based subspace principle to estimate unknown parameters of the noisy speech signal. In the proposed method, the estimated parameters are used for recovering the spectrum of the signal buried in noise, and then the restored spectrum is used for pitch extraction. Moreover, the proposed method reduces the computational complexity within the subspace method. In the simulation results, it is shown that the proposed method estimates more accurate fundamental frequencies than the conventional pitch estimation methods with saving the computational complexity.

1 INTRODUCTION

The fundamental frequency of speech signals is one of the most essential features in the human voice. Its estimation plays very important role in various speech applications, such as the automatic speaker recognizers, very low rate speech codings, and speech instruction systems for impaired people [1-4].

The cepstral method and the autocorrelation function method are well known methods for determining the pitch of the speech signals [1]. In the cepstral method, the influence of the vocal tract is removed by applying the logarithm operation. However, since the cepstral method is very sensitive to noise, the cepstral method is not able to estimate accurate fundamental frequencies in lower SNR environments. Therefore, in general, the cepstral method is used to the noise-free or sufficiently higher SNR environments. On the other hand, the autocorrelation function method is well known as a robust pitch estimation method. Since the autocorrelation function tends to average the additive noise components, the method is capable of extracting the relatively accurate pitch from the noisy speech signals. In the very lower SNR environments, however, the autocorrelation function method also estimates inaccurate fundamental frequencies [2,3].

In speech enhancement, a large number of methods based on subspace principles have been developed [5-7]. In the subspace methods, the observed noisy speech signal is decomposed into the signal dominant subspace (signal subspace) and the noise counterpart (noise subspace) by applying the eigenvalue decomposition (ED) or singular value decomposition (SVD). Then, the enhanced speech signal is obtained by using these subspaces, e.g., orthonormally projecting the observed noisy signal onto the signal subspace. In recent literature, it has been shown that the subspace methods are effective even in the pitch extraction of the speech signals corrupted by the noise [3,4]. In practice, however, the subspace methods are computationally heavy since ED and SVD involve complex algebraic calculations, especially when the long analysis frame is used.
In this paper, we propose a novel method for extracting the fundamental frequency of speech signals buried in the noise. In the proposed method, we adopt the multiple signal classification (MUSIC) algorithm \[8,9\] to reconstruct the spectrum buried in additive noise. The MUSIC algorithm is one of the well known eigen-based subspace methods. This paper also presents a method for reducing the computational complexity within the MUSIC algorithm.

2 APPROACH

2.1 The Model

Consider an \( N \)-sample observed noisy signal vector \( y = [y(0), y(1), \ldots, y(N - 1)]^T \) (\( T: \) a vector transpose) composed of \( P \) sinusoids modeled as

\[
y = x + n = Sa + n,
\]

where the signal vector \( x = \{s(\phi_0), s(\phi_1), \ldots, s(\phi_{P-1})\} \), \( s(\phi_k) = [1, e^{j2\pi\phi_k}, \ldots, e^{j2\pi\phi_k(N-1)}]^T \), and \( n \) is the additive noise signal vector. In this paper, \( n \) is assumed to be Gaussian with zero-mean and variance \( \sigma_n^2 \), and uncorrelated with \( x \).

2.2 Eigenvalue Decomposition of Autocorrelation Matrix

Consider the autocorrelation matrix of the signal. Since \( x \) and \( n \) are uncorrelated with each other, the autocorrelation matrix \( R_{yy} \) of \( y \) is given by

\[
R_{yy} = E[yy^H] = R_{xx} + R_{nn},
\]

where \( E[\cdot] \) denotes the expectation operator, \( H \) denotes the Hermitian transpose, and \( R_{xx} \) and \( R_{nn} \) are respectively the autocorrelation matrices of \( x \) and \( n \) expressed as

\[
R_{xx} = E[xx^H] = SAS^H,
\]

\[
R_{nn} = E[nn^H] = \sigma_n^2 I,
\]

where the diagonal matrix \( A \) is given by

\[
A = E[aa^H] = \frac{1}{N^2} \text{diag}([X(\phi_0)]^2, [X(\phi_1)]^2, \ldots, [X(\phi_{P-1})]^2). \tag{8}
\]

In this paper, we consider a special case in which \( x \) is composed of \( P = N \) sinusoids and those frequencies \( \phi_k \) \((k = 0, 1, \ldots, N - 1)\) are given by

\[
\phi_k = \frac{k}{N}, \quad (k = 0, 1, \ldots, N - 1)
\]

as the discrete Fourier transform (DFT). Above assumption indicates that the row elements of \( S \) in (2) are equally spaced in angle around the unit circle in the complex plane. If \( \phi_k \) are assumed as in (9), the following properties of \( S \) are obtained:

\[
|S| \neq 0,
\]

\[
SS^H = NI.
\]

Therefore, the Hermitian transpose of \( S \) is written as

\[
S^H = NS^{-1},
\]

and we obtain

\[
R_{xx} = SA(NS^{-1}) = S(NA)S^{-1}. \tag{13}
\]

Since \( A \) is diagonal matrix, (13) indicates ED of \( R_{xx} \), namely, the eigenvalues \( \lambda_k \) \((k = 0, 1, \ldots, N - 1)\) and corresponding eigenvectors \( v_k \) \((k = 0, 1, \ldots, N - 1)\) of \( R_{xx} \) are respectively obtained as

\[
\lambda_k = \frac{|X(\phi_k)|^2}{N}, \tag{14}
\]

\[
v_k = s(\phi_k), \quad (k = 0, 1, \ldots, N - 1). \tag{15}
\]

Above equations imply that the autocorrelation matrix has a one-to-one correspondence between the eigenvalues and the power spectrum of the signal and that the eigenvectors of the autocorrelation matrix are equivalent to those of any signal. Therefore, in ED of the autocorrelation matrix, the eigenvalues are obtained by applying DFT to the signal and that the calculation of the eigenvectors is not required. Both the eigenvalues and eigenvectors of the autocorrelation matrix are obtained, the MUSIC algorithm is available for reconstructing the spectrum of signal buried in noise.

2.3 MUSIC Algorithm \[8,9\]

Let \( x \) be the signal composed of only \( P \) \((< N)\) sinusoidal components \( s(f_l) \), where the frequencies \( f_l \) \((l =
0, 1, ··· , P − 1) are unknown. Then, since the rank of $R_{xx}$ is P and the autocorrelation matrix is non-negative definite and Hermitian, the eigenvalues $\mu_k = \lambda_k + \sigma_n^2 \ (k = 0, 1, ··· , N − 1)$ of $R_{yy}$ are obtained as

$$\begin{cases} 
\mu_0 \geq \mu_1 \geq \cdots \geq \mu_{P−1} > \sigma_n^2 \\
\mu_P = \mu_{P+1} = \cdots = \mu_{N−1} = \sigma_n^2.
\end{cases} \quad (16)$$

This indicates that the eigenvectors of $R_{yy}$ can be partitioned into two disjoint subsets, namely, the first set associated with the P largest eigenvalues spans the signal subspace, whereas the second associated with the $N − P$ smallest eigenvalues spans the noise subspace. Since the signal and noise subspace are mutually orthogonal, the MUSIC spectrum of $y$ is given by

$$P_{\text{MUSIC}}(f) = \frac{1}{\sum_{k=P}^{N−1} |s^H(f)v_k|^2}, \quad (17)$$

where $\{v_p, v_{p+1}, \cdots, v_{N−1}\}$ are the eigenvectors of $R_{yy}$ corresponding to the noise subspace. Since $P_{\text{MUSIC}}(f)$ is sharply peaked at $f_l$, the estimated frequencies $f_l \ (l = 0, 1, \cdots, P − 1)$ corresponding to $x$ can be obtained by simply taking the P peaks of the MUSIC spectrum.

On the other hand, from (14), the eigenvalues are associated with the power spectrum of the signal. Therefore, the $N − P$ smallest eigenvalues, i.e., those corresponding to the noise subspace, are directly associated with the $N − P$ smallest frequency components of the signal. Furthermore, from (15), the eigenvectors spanning the noise subspace are obtained as the sinusoidal signal vectors whose frequencies are equivalent to those of the $N − P$ smallest frequency components of the signal. If the frequencies are given by (9), the elements of the sinusoidal signal vectors are equally spaced in angle around the unit circle in the complex plane. Therefore, the denominator in (17) is obtained as follows:

- The frequency $f$ is given by (9) and equal to one of those of the $N − P$ smallest frequency components of the signal

$$\sum_{k=P}^{N−1} |s^H(f)v_k|^2 = N^2. \quad (18)$$

- $f \notin \{\phi_0, \phi_1, \cdots, \phi_{N−1}\}$

$$\sum_{k=P}^{N−1} |s^H(f)v_k|^2 > 0. \quad (20)$$

These equations indicate that the MUSIC spectrum given by (17) has the peaks at the frequencies which are equal to those of the P largest frequency components of the signal. Hence, instead of calculating (17), the frequencies $f_l \ (l = 0, 1, \cdots, P − 1)$ are directly obtained from the power spectrum of $y$.

2.4 Restoration of Spectrum And Pitch Detection

The frequencies $f_l$ estimated by the MUSIC algorithm are then used to restore the power spectrum of $x$. It is well known that the orthonormally projecting $y$ onto the signal subspace leads to reduce the noise components and can recover the spectrum of the signal corrupted by noise [8,9]. In practice, however, the orthonormal projection is computationally heavy, especially the analysis frame length is long. Therefore, we instead propose a method for restoration of the spectrum without performing such orthonormal projection operations.

In the proposed method, only P largest frequency components $Y(f_l) \ (l = 0, 1, \cdots, P − 1)$ of $y$ are simply extracted since $x$ is composed of only P sinusoidal components, and then, the noise components in $Y(f_l)$ are eliminated. As stated above, the eigenvalues of the autocorrelation matrix are associated with the power spectrum of the signal and the eigenvalues $\mu_k$ of $R_{yy}$ are given by $\mu_k = \lambda_k + \sigma_n^2$. Therefore, the estimated spectrum of $x$ is obtained as

$$|X_{\text{MUSIC}}(f_l)| = \sqrt{|Y(f_l)|^2 - N\sigma_n^2}. \quad (21)$$

Finally, the estimated spectrum $|X_{\text{MUSIC}}|$ is transformed to a time domain data via the inverse DFT (IDFT) as

$$x_{\text{MUSIC}} = \text{IDFT}[X_{\text{MUSIC}}]. \quad (22)$$

The peak of $x_{\text{MUSIC}}$ is appeared at the pitch location as the autocorrelation function. Therefore, the fundamental frequency is determined by using peak picking [10] from $x_{\text{MUSIC}}$.

3 SIMULATION RESULTS

In the simulation study, the signals used were the speech uttered by three male and two female speakers. Each utterance is “Sakura ga saita” in Japanese,
Table 1: Comparison of the calculation time (normalized to the cepstral method = 1)

<table>
<thead>
<tr>
<th>Method</th>
<th>Calculation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cepstral</td>
<td>1</td>
</tr>
<tr>
<td>Autocorrelation Function</td>
<td>1.01</td>
</tr>
<tr>
<td>Conventional MUSIC</td>
<td>241.74</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>1.45</td>
</tr>
</tbody>
</table>

The observed noisy signal was divided into 256-points frames. The each frame was Hamming windowed, and then the proposed pitch estimation method was applied. The extracted fundamental frequencies were compared with the true frequencies obtained by applying the cepstral method to the noise-clean speech signals. For the evaluation of the performance, in terms of the averaged absolute error rate was considered [4].

The averaged absolute error rate was given by

$$Err = \frac{1}{M_V} \sum_{k=0}^{M_V-1} \left| \frac{f_M(k) - f_T(k)}{f_T(k)} \right| \times 100 \, \%,$$

(23)

where $f_T(k)$ and $f_M(k)$ are respectively the true and estimated fundamental frequencies at $k$th frame, and $M_V$ is the number of frames where speech is present. In the simulation, the determination of speech presence was achieved by manual inspection of the noise-clean speech signals.

Table 1 shows the comparison of the calculation time. As in the table, it was shown that the calculation time is drastically reduced by the proposed method. Fig. 1 shows the comparison of the averaged absolute error rate. As in the figure, it was seen that the proposed method estimates most accurate frequency among four methods at SNRs lower 0dB.

4 CONCLUSION

In this paper, we have proposed a novel method for pitch extraction of the speech signals. In the simulation, it has been shown that the proposed method can estimate more accurate fundamental frequencies than the conventional pitch estimation methods with saving the computational complexity within the MUSIC algorithm.