Assessment of Disordered Voices Using Empirical Mode Decomposition in the Log-Spectral Domain

A. Kacha¹, F. Grenez², J. Schoentgen²³

¹Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Algeria  
²Laboratory LIST, Université Libre de Bruxelles, Brussels, Belgium  
³National Fund for Scientific Research, Belgium

akacha@ulb.ac.be, fgreneze@ulb.ac.be, jschoent@ulb.ac.be

Abstract

Empirical mode decomposition (EMD) algorithm is proposed as an alternative to decompose the log of the magnitude spectrum of the speech signal into its harmonic, envelope and noise components and the harmonic-to-noise ratio is used to summarize the degree of disturbance in the speech signal. The empirical mode decomposition algorithm is a tool for the analysis of multi-component signals. The analysis method does not require a priori fixed basis function like conventional analysis methods (e.g. Fourier transform and wavelet transform). The proposed method is tested on synthetic vowels and natural speech. The corpus of synthetic vowels comprises 48 stimuli of synthetic sounds [a] that combine three values of vocal frequency, four levels of jitter frequency and four levels of additive noise. The corpora of natural speech comprise a concatenation of the vowel [a] with two Dutch sentences produced by 28 normophonic and 223 speakers with different degrees of dysphonia.

Index Terms: Disordered voices, empirical mode decomposition, harmonic-to-noise ratio.

1. Introduction

Objective measures obtained from acoustic analysis of speech are of great importance for clinical evaluation of voice disorders because the analysis is noninvasive and provides a severity index of the disorder which enables clinicians to monitor the progress of patients and document quantitatively the perceived degree of hoarseness. Despite the number of acoustic markers that have been proposed in the literature to characterize the speech of dysphonic speakers, finding global descriptors of voice function and voice quality is still an issue. One of the manifestations of voice disorders is the lack of periodicity in voiced speech produced by dysphonic speakers. Aperiodicities may be caused by additive noise owing to turbulence and modulation noise owing to external perturbations of the glottal excitation signal, as well as aperiodicities due to intrinsically irregular dynamics of the vocal folds. As a consequence, the energy of the harmonic structure of the spectrum is decreased in favour of that of the nonharmonic structure. Several acoustic markers used to assess vocal fold function reflect the deviation of the speech waveform from the perfect periodicity. For instance, jitter and shimmer are frequently used to measure perturbations produced by the variations in the fundamental period and amplitude, respectively.

Most techniques for estimating vocal aperiodicities have been applied to steady fragments extracted from sustained vowels. The widespread use of sustained vowels is due to the technical feasibility of the analysis rather than clinical relevance. Recent approaches for vocal aperiodicities estimation in continuous speech are based on generalized variogram [1] and cepstral analysis [2]. The generalized variogram-based method does not require the signal to be locally periodic or that the average period length has to be known a priori [1] which makes it very effective for tracking cycle-to-cycle aperiodicities in any speech sound produced by any speaker. Segmental signal-to-dysperiodicity ratio (SDRSEG) that summarizes the dysperiodicities has been shown to correlate strongly with the degree of perceived hoarseness. The method proposed in [2] is based on the cepstral analysis. The real cepstrum of a given signal is defined as the inverse Fourier transform of the log of its magnitude spectrum. Cepstral peak prominence (CPP) is an acoustic cue that has been used for connected speech fragments. It summarizes indirectly the degree of disturbances via the size of the first harmonic of the cepstrum of a speech frame. It has been shown to correlate strongly with perceived hoarseness, even though the detection of the first harmonic may be error prone for severely hoarse speakers [3], [4]. In this presentation, we propose an approach based on the empirical mode decomposition (EMD) algorithm [5] to decompose the log of the spectrum magnitude of the speech signal into its harmonic, envelope and noise components. The harmonic-to-noise ratio is used to summarize the degree of disturbance in the speech signal.

The remainder of the paper is organized as follows. Speech data and perceptual ratings are described in Section 2. Empirical mode decomposition algorithm is introduced in Section 3. Results based on both synthetic and real speech signals are presented in Section 4. Finally, to conclude, remarks are given in section 5.

2. Corpora and perceptual ratings

Corpora used in the present study include synthetic sustained vowels as well as natural speech.

2.1. Synthetic vowels

Sustained vowels have been generated by a synthesizer that involves models of glottal area and airflow through the glottis. The ability of the synthesizer to mimic natural speech has been demonstrated in the framework of several experiments [6]. The corpus (corpus 1) comprises 48 stimuli of 1-second synthetic sounds [a] that combine three values of vocal frequency, four levels of jitter frequency and four levels of additive noise. The vocal frequencies are 100 Hz, 120 Hz and 140 Hz. The jitter and additive noise have been fixed based on the independent advice of one phoniatrician and one therapist so that the stimuli are perceived as covering the full ranges of grade (G0-G3), roughness (R0-R3) and breathiness (B0-B3) on the GRB(AS) scales.
Eight speech therapists and one phoniatrician have perceptually evaluated the synthetic vowels according to perceived “grade” (G), “roughness” (R) and “breathiness” (B) with four degrees per scale with the degree 0 corresponding to a normal sound and the degree 3 corresponding to a severely perturbed sound. Each score has been averaged over the nine judges. The grade (G) provides a measure of the global quality of the voice and is used in the present study to evaluate the performance of the proposed method for voice disorder assessment.

2.2. Natural speech

The corpora comprise the vowel [a] (corpus 2), a concatenation of the vowel [a] with two Dutch sentences (“Papa en Marloes staan op het station. Ze wachten op de trein.”) (corpus 3) and a concatenation of the vowel [a] with the voiced segments extracted from the two sentences (corpus 4) [7]. The stimuli produced by 28 normophonic and 223 speakers with different degrees of dysphonia have been sampled at 44100 Hz. Five judges have carried out the perceptual evaluation. Each judge has rated the item “grade”, (G) of the GRABS scale, from 0 (normal) to 3 (severe). The “grade” refers to the overall perceived abnormality of the speech stimuli, which have been the concatenation of the two (complete) sentences with vowel [a] (corpus 3). The five perceptual scores per stimulus have been averaged. The recordings and evaluation have been made at the Sint-Jan General Hospital, Bruges, Belgium.

3. Methods

3.1. Empirical mode decomposition

The empirical mode decomposition (EMD) algorithm is a tool for the analysis of multi-component signals. The analysis method does not require a priori fixed basis function like conventional analysis methods (e.g. Fourier transform and wavelet transform). It has been proposed initially in [5] to analyse data from nonlinear and nonstationary processes like ocean waves and has found applications in many fields such as geophysics, biomedical signal processing and speech processing.

The EMD algorithm decomposes adaptively a given signal \( s(t) \) into oscillation modes namely the intrinsic mode functions (IMFs) extracted from the signal itself. Each IMF component has a zero-mean value and only one extremum between zero-crossings. The IMFs are obtained via the iterative sifting process, which involves the following steps:

1. Initialize the algorithm: \( j=1 \), initial residue \( r_{0,j}(t)=s(t) \) and fix the threshold \( \delta \)
2. Extract local maxima and minima of \( r_{j-1,j}(t) \)
3. Compute the upper envelope \( U_j(t) \) and lower envelope \( L_j(t) \) by cubic spline interpolation of local maxima and minima, respectively
4. Compute the mean envelope
   \[ m_j(t) = \frac{U_j(t) + L_j(t)}{2} \]
5. Compute the \( j \)th component \( h_j(t) = r_{j-1,j}(t) - m_j(t) \)
6. \( h_j(t) \) is processed as \( r_{j,j}(t) \). Let \( h_{j,k}(t) = h_j(t) \) and \( m_{j,k}(t) \), \( k = 0, 1, \ldots \), be the mean envelope of \( h_{j,k}(t) \) (\( k \) denotes the number of sifts), then compute

\[
SD_h = \frac{\sum_{t=0}^{T} \left| h_{j,k+1}(t) - h_{j,k}(t) \right|^2}{\left( h_{j,1}(t) \right)^2} < \delta
\]

7. Compute the \( j \)th IMF as \( IMF_j(t) = h_{j,d}(t) \)
8. Update the residue \( r_j(t) = r_{j-1,j}(t) - IMF_j(t) \)
9. Increase the sifting index \( j \) and repeat steps 2 to 8 until the number of local extrema in \( r_j(t) \) is less than 3.

Each IMF is a narrowband AM-FM component that can be characterized by its instantaneous frequency.

The signal can be reconstructed exactly by summing all the \( J \) IMFs and the residue

\[ x(t) = \sum_{j=1}^{J} IMF_j(t) + r_J(t) \]

For the corpora used in the experiment, the number of IMFs per frame varies between 8 and 12.

3.2. EMD-based speech components separation

A voiced speech frame \( x(t) \) can be modeled as a periodic source component, \( e(t) \) convolved with the impulse response of the vocal tract, \( v(t) \) [8]:

\[ x(t) = e(t) * v(t) \]

where * denotes the convolution.

Windowing the signal frame \( x(t) \) and taking the Fourier transform magnitude gives

\[ |X_v(f)| = |E_v(f) * V(f)| \]

Where \( X_v(f) \), \( E_v(f) \) are short-time magnitude spectra of the windowed speech frame and windowed excitation signal, respectively and \( V(f) \) is the frequency response of the vocal tract.

Taking the logarithm changes the multiplicative components into additive components:

\[ \log|X_v(f)| = \log|E_v(f)| + \log|V(f)| \]

From (6), it is observed that the log magnitude spectrum is the sum of two spectral components: \( |E_v(f)| \), the log magnitude spectrum of the windowed excitation signal and \( |V(f)| \), the spectral envelope due to the filtering characteristic of the vocal tract. Because of the presence of aspiration noise at the glottis, the excitation spectrum itself can be regarded as composed of two parts: the first part is a regularly spaced series of harmonics having a decreasing magnitude with frequency and the second part is an irregularly distributed noise.

The log magnitude spectrum can be considered as composed of a slowly varying (with respect to frequency) contour due to the contribution of the vocal tract, a series of harmonics characterized by a periodic structure and an irregular and rapidly varying part due to noise at the glottis. The EMD algorithm yields a tool that enables to separate the three components of the log magnitude spectrum. Indeed, the EMD algorithm acts as a filterbank [9], so that the decomposition of the log magnitude spectrum via the EMD algorithm results into several oscillating components (IMFs) that can be clustered in three classes by a simple thresholding operation and each class of components is assigned to some part of the log magnitude spectrum. Let \( f_j \) be the mean frequency of the
jth-IMF component of the log magnitude spectrum obtained via the EMD algorithm. The different IMFs are clustered in terms of their mean quefrecies as follows:

Class 1: $f_j < \theta_1$: IMF $j$ belongs to the envelope part
Class 2: $\theta_1 < f_j < \theta_2$: IMF $j$ belongs to the harmonic part
Class 3: $f_j > \theta_2$: IMF $j$ belongs to the noise part

where $\theta_j$, $j=1, 2$ are some thresholds that depend on the average fundamental frequency $f_0$ of the speech signal and they are fixed empirically.

Each part of the log magnitude spectrum is estimated by summing the IMFs belonging to the corresponding class. Experiments have shown that the optimal thresholds are $\theta_1 = 0.3/f_0$ and $\theta_2 = 4/f_0$.

As an illustration of the effectiveness of the decomposition algorithm, Figure 1 shows the different estimated components of the log magnitude spectrum of a frame of 320 ms of length taken from a sustained vowel [a] produced by a normophonic speaker.

### 3.3. Baseline correction

An examination of the estimated harmonic (Figure 1-c) shows that the baseline dips slightly towards negative values at low frequencies. Here, the baseline is the inter-harmonic contour. The residue contour (Figure 1-e) follows the baseline closely at high frequencies and deviates slightly above the baseline at low frequencies. Consequently, in order to straighten out the baseline, a correction is carried out. The baseline correction is summarized as follows:

i) The deviation from the 0-dB axis is estimated as the lower envelope of the harmonic part by locating local minima and interpolating linearly between them.

ii) The estimated lower envelope is low-pass filtered.

iii) The estimated lower envelope is subtracted from the harmonic part and added to the spectral envelope to obtain their respective corrected parts.

The baseline correction procedure has been applied to the above harmonic component. Figure 2 shows the uncorrected harmonic part, the estimated deviation and the corrected harmonic part.

### 3.4. Acoustic marker

The acoustic marker used to summarize the amount of aperiocities within an utterance is the harmonic-to-noise ratio. For a given utterance, the analysis interval is divided into $L$ frames and the HNR is computed as the average of the $HNR_i$ ($i=1, \ldots, L$) of the $L$ frames:

$$HNR = \frac{1}{L} \sum_{i=1}^{L} HNR_i$$

(7-a)

$$HNR_i = 10 \log \left[ \frac{\sum_{k=0}^{M-1} H^2(k)}{\sum_{k=0}^{M-1} N^2(k)} \right], i=1, \ldots, L$$

(7-b)

with $H(k)$ denoting the magnitude spectrum of the harmonic part and $N(k)$ the magnitude spectrum of the noise and $M$ is the number of frequency points.

The frequency band involved in the computation of the HNR has been limited to 4 kHz.

**Figure 1:** Decomposition of the log magnitude spectrum of a 320 ms speech frame of sustained [a] into three components via the EMD algorithm. (a) Log magnitude spectrum. (b) Envelope component. (c) Harmonic component. (d) Noise. (e) Sum of the three components superposed to the estimated envelope.

### 4. Results and discussion

The effect of frame length on the correlation of the HNR values with the degree of perceived grade has been investigated for different frame lengths by carrying out experiments on synthetic [a] as well as on natural speech. It has been found that the strongest correlation is achieved for a frame length of 200 ms. The frame length has been set to this value accordingly.

The empirical mode decomposition-based approach for HNR estimation has been applied to synthetic [a] and natural...
speech. Table 1 gives Pearson product moment correlations of the HNR values with average score of grade for the different corpora. The last two lines of Table 1 give the correlations obtained by using generalized variogram based on the full-band SDRSEG as an acoustic cue and the CPP-based method (correlations corresponding to corpus 1 are not available). It can be seen that a strong correlation is achieved for the synthetic [a] ($\rho=0.85$). For natural speech, the strongest correlation ($\rho=-0.71$) is observed for the concatenation of [a] with the sentences (corpus 3, for which the perceptual ratings had been carried out). These correlations are slightly higher in absolute value than those based on the SDRSEG and CPP.

Table 1. Pearson’s correlation coefficients between average grade scores and acoustic markers HNR, SDRSEG and CPP for the four corpora.

<table>
<thead>
<tr>
<th></th>
<th>Corpus 1</th>
<th>Corpus 2</th>
<th>Corpus 3</th>
<th>Corpus 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>HNR</td>
<td>-0.85</td>
<td>-0.65</td>
<td>-0.71</td>
<td>-0.67</td>
</tr>
<tr>
<td>SDRSEG</td>
<td>—</td>
<td>-0.63</td>
<td>-0.70</td>
<td>-0.65</td>
</tr>
<tr>
<td>CPP</td>
<td>—</td>
<td>-0.56</td>
<td>-0.70</td>
<td>-0.63</td>
</tr>
</tbody>
</table>

Figure 3 displays the estimate HNR versus the average perceived grade scores for synthetic [a] (corpus 1). As shown, the estimate HNR is linearly decreasing as the average perceived grade score increases. The values of the estimated HNR range from 6 dB to 26 dB.

5. Conclusions

In this presentation, the empirical mode decomposition algorithm has been proposed as an alternative to decompose the log magnitude spectrum of speech signal into spectral envelope, harmonic part and noise. The method is compared to the cepstrum-based method. Harmonic-to-noise ratio has been used as an acoustic cue to evaluate the overall quality of the disordered voices produced by dysphonic speakers. Experimental results show that the EMD-based approach results in a high correlation between HNR estimates and average perceived grade scores for synthetic [a] as well as for natural speech including sustained [a] and a concatenation of [a] and two Dutch sentences. It has been found that the EMD-based HNR estimation method is equal to that based on the generalized variogram and cepstral prominence peak methods in terms of correlation of the acoustic cue with the average perceived grade scores. In this study, the thresholds used in the algorithm have been fixed empirically. Improvement of the algorithm by incorporating an appropriate method to estimate automatically the thresholds is considered as future work.

6. References