Empirical Mode Decomposition-Based Spectral Acoustic Cues for Disordered Voices Analysis

Abdellah Kacha\textsuperscript{1}, Francis Grenez\textsuperscript{2}, Jean Schoentgen\textsuperscript{2,3}

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

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

Abstract

Cepstral-based acoustic cues for disordered voices analysis have been investigated in a number of studies. It has been shown that cepstral-based acoustic cues such as the harmonics-to-noise ratio (HNR), the amplitude of the first harmonic (R1A) provide acoustic correlates for hoarse voice quality. The aim of this presentation is to investigate an acoustic analysis of speech by means of spectral acoustic cues obtained via empirical mode decomposition (EMD) of the log of the magnitude spectrum of the speech signal as an alternative to the cepstral-based acoustic cues. The spectral acoustic cues investigated in this article are the harmonics-to-noise ratio (HNR) and the amplitude of the first harmonic (H1A). The EMD-based spectral acoustic cues are evaluated on a corpus of synthetic stimuli generated by a synthesizer of disordered voices as well as on a corpus of natural sustained vowels comprising 251 normophonic and dysphonic speakers. The performances of the EMD-based spectral acoustic cues (HNR\textsubscript{EMD} and H1A) are compared to those of cepstral-based acoustic cues (HNR\textsubscript{cep} and R1A). Experimental results show that the EMD-based spectral acoustic cues outperform the cepstral-based acoustic measures in terms of the correlation with the perceived degree of hoarseness as defined as the global quality of the voice and provided by the grade (G) in the GRBAS scale.

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

1. Introduction

Objective measures 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, i.e., the global quality of the voice and provided by the grade (G) in the GRBAS scale. Despite the number of acoustic markers that have been proposed in the literature to characterize the speech of dysphonic speakers, finding reliable and accurate descriptors of voice function and voice quality is still an issue.

Although there are various medical conditions that can affect the voice, most of the disorders originate from the vocal system due to a malfunction of the vocal cords and frequently result in an increase in the aperiodicity of voiced speech sounds. 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 \cite{1,2}. As a consequence, the energy of the harmonic structure of the spectrum is decreased in favor 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 \cite{3}.

Most approaches for disordered voices analysis are based on the periodicity of the vocal folds vibration. These techniques require a stationary portion of the speech signal either for the mathematical model to be valid or for an accurate measurement of the acoustic parameters of the analyzed speech signal. The difficulty to isolate the individual speech cycles and the individual harmonics in dysphonic speech gives rise to a biased acoustic marker of vocal noise.

In a recent study, it has been shown that cepstral-based acoustic cues correlate strongly with the degree of perceived hoarseness. Harmonics-to-noise ratio (HNR), cepstral peak prominence (CPP) and the amplitude of the first rhammonic have been used as acoustic cues for disordered voices analysis \cite{4,5,6}. The HNR summarizes directly the degree of perturbations of the signal by comparing the harmonic content to the noise \cite{7} while the CPP and the amplitude of the first rhammonic summarize indirectly the degree of perturbations via the size of the first rhammonic of the cepstrum of a speech frame \cite{4}. Strong correlations of the cepstral-analysis-based acoustic cues with perceived hoarseness have been observed, even though the detection of the first rhammonic may be error prone for severely hoarse speakers \cite{3}.

In \cite{8}, the empirical mode decomposition (EMD) algorithm \cite{9} has been used to decompose the log of the spectrum magnitude of the speech signal into its harmonic, envelope and noise components. The harmonic-to-noise ratio has been used to summarize the degree of perturbations in the speech signal. In this presentation, the effectiveness of two EMD-based spectral acoustic cues for assessing disordered voices is investigated and their performances in terms of correlation with the perceived degree of hoarseness are compared to those of their counterpart based on cepstral analysis \cite{4,5,7}.

The remainder of the paper is organized as follows. Speech data and perceptual ratings are described in Section 2. Empirical mode decomposition-based method for magnitude spectrum decomposition and spectral objective measures are presented in Section 3. Results based on both synthetic and real speech signals are presented in section 4. Finally, conclusions are given in section 5.

2. Corpora and Perceptual Ratings

Corpora used in the test comprise synthetic /a/ as well as natural sustained /a/.

\textsuperscript{1} Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Algeria
\textsuperscript{2} Laboratory LIST, Université Libre de Bruxelles, Brussels, Belgium
\textsuperscript{3} National Fund for Scientific Research, Belgium
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 [4]. The corpus 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 GRABS(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 sustained vowels

The corpus comprises the vowel /a/ produced by 28 normophonic and 223 dysphonic Dutch subjects with different degrees of dysphonia. The stimuli have been sampled at 44100 Hz. The recording devices and conditions are detailed in [10]. Five judges have carried out the perceptual evaluation. Each judge has rated the item “grade”, “roughness” (G) of the GRABS scale, from 0 (normal) to 3 (severe). The “grade” refers to the overall perceived abnormality of the speech stimuli. 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 [9] 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 x(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(t)=x(t) and fix the threshold δ
2. Extract local maxima and minima of r_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 jth component \( h_j(t) = r_j(t) - m_j(t) \)
6. \( h_j(t) \) is processed as \( r_j(t) \). Let \( h_{j0}(t) = h_j(t) \) and \( m_{j0}(t) = m_j(t) \), \( k = 0, 1, \ldots \) be the mean envelope of \( h_{j0}(t) \) (k denotes the number of sifts), then compute \( h_{j+k}(t) = h_{j+k}(t) - m_{j+k}(t) \) until
   \[ SD_k = \sum_{i=0}^{T} \left| \frac{h_{j+k-1}(t) - h_{j,k}(t)}{h_{j,k-1}(t)} \right|^2 < \delta \]
7. Compute the jth IMF as \( IMF_j(t) = h_{j0}(t) \)
8. Update the residue \( r_j(t) = r_j(t) + IMF_j(t) \)
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, \( \nu(t) \) [7]:

\[ x(t) = * e(t)^* \nu(t) \]

where * denotes the convolution.

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

\[ X_n(f) = |E_n(f) \times V(f)| \]

Where \( X_n(f) \), \( E_n(f) \) 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:

\[ \log X_n(f) = \log |E_n(f)| + \log |V(f)| \]

From (6), it is observed that the log magnitude spectrum is the sum of two spectral components: \( \log |E_n(f)| \), the log magnitude spectrum of the windowed excitation signal and \( \log |V(f)| \), 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 [11], 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 quefrency of the $j$th-IMF component of the log magnitude spectrum obtained via the EMD algorithm. The different IMFs are clustered in terms of their mean quefrencies as follows:

- Class 1: $f_j < th_1$: IMF belongs to the envelope part
- Class 2: $th_1 < f_j < th_2$: IMF belongs to the harmonic part
- Class 3: $f_j > th_2$: IMF belongs to the noise part

where $th_1 = 0.3/f_0$ and $th_2 = 4/f_0$. 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 $th_1 = 0.3/f_0$ and $th_2 = 4/f_0$.

3.3. Baseline correction

The baseline designates the inter-harmonic contour. Because the IMFs are zero-mean oscillating functions, the estimated harmonic spectrum has also zero mean and the baseline is expected to dip below zero at low frequencies, in the vicinity of large harmonics that are positive. The goal of the baseline correction is to straighten out the baseline by carrying out a correction. The residue contour follows the baseline closely at high frequencies and deviates slightly above the baseline at low frequencies.

The baseline correction follows that used in [12] for spectral tilt correction. The correction is carried out in doubly logarithmic coordinates where the envelope of harmonic component is almost a straight line. Firstly, a straight line is fitted to the smallest 60% values of the log harmonic component and secondly, the fitted line is subtracted from the harmonic component 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 1 shows the uncorrected harmonic part and the corrected harmonic part.

3.4. Spectral acoustic cues

The spectral acoustic cues used to summarize the degree of hoarseness of the voice are computed from the harmonic and noise components of the speech signal estimated via the EMD-based approach in the log spectral domain. Two acoustic cues are used in this study: harmonics-to-noise ratio (HNR) and first harmonic amplitude (H1A).

Harmonics-to-noise ratio (HNR): This acoustic marker summarizes directly the amount of aperiodicities within an utterance. For a given vowel, the analysis interval is divided into $L$ frames and the HNR is computed as the average of the HNR, $(i=1, \ldots, L)$ of the $L$ frames:

$$\text{HNR} = \frac{1}{L} \sum_{i=1}^{L} \text{HNR}_i$$  (7-a)

where

$$\text{HNR}_i = 10 \log \left( \frac{\sum_{k=0}^{M-1} H^2(k)}{\sum_{k=0}^{M-1} N^2(k)} \right), i = 1, \ldots, L$$  (7-b)

with $H(k)$ denoting the magnitude spectrum of the harmonic component and $N(k)$ the magnitude spectrum of the noise component and $M$ is the number of frequency points. The frequency band involved in the computation of the HNR has been limited to 4 kHz.

First harmonic amplitude (H1A): The amplitude of the first harmonic in dB provides an indirect measure of the degree of perturbations in the speech signal. The signal is normalized in energy, divided into $L$ frames and the amplitude of the first harmonic is computed for each frame from the harmonic part obtained by EMD. The average of the amplitude of the first harmonic on all frames is used to summarize the degree of perturbations in the speech signal.

Figure 1: Illustration of the baseline correction. (a) Harmonic component in double logarithmic scale before baseline correction. (b) Harmonic component in double logarithmic scale after baseline correction.

4. Results and Discussion

The performances of the HNR and H1A obtained using the EMD-based approach are investigated and compared to their respective counterpart which are the HNR and R1A estimated via cepstral analysis for synthetic /a/ as well as for natural /a/. The development of the EMD-based approach for disordered voices analysis as well as the statistical analysis have been done in Matlab programming environment. Based on our previous investigations, the frame length has been set to 200 ms. Cepstral analysis-based HNR values have been computed with VoiceSauce software presented in [13].

As an illustration, Fig. 2 displays the harmonic component of a synthetic /a/ estimated via the EMD-based approach for two levels of jitter and two levels of additive noise. As observed, the perturbation by of the stimulus by additive noise or jitter results in a decrease of the amplitude of the harmonics.
Table 1 shows the Pearson product moment correlations of the acoustic markers H1A computed via the EMD-based approach and R1A computed via cepstral analysis with average scores of grade for synthetic /a/ as well as for natural /a/. It can be seen that the H1A results in stronger correlations than those achieved by R1A for both synthetic and natural /a/. Pearson product moment correlations of the HNR computed via the EMD-based approach and cepstral analysis with average scores of grade for synthetic /a/ and natural /a/ are given in Table 2. It is observed that EMD-based HNR gives rise to stronger correlations compared to those achieved by cepstral analysis-based HNR for synthetic /a/ as well as for natural /a/.

![Figure 2](image)

**Figure 2:** Harmonic component estimated via EMD-based approach at different levels of (a) jitter and (b) additive noise for synthetic /a/.

Table 1. Pearson’s correlation coefficients between average grade scores and acoustic markers H1A computed via EMD-based approach and R1A computed via cepstral analysis for synthetic and natural /a/.

<table>
<thead>
<tr>
<th></th>
<th>H1A</th>
<th>R1A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>period-asynch.</td>
<td>period-synch.</td>
</tr>
<tr>
<td>Synt. /a/</td>
<td>-0.89</td>
<td>-0.64</td>
</tr>
<tr>
<td>Nat. /a/</td>
<td>-0.69</td>
<td>-0.63</td>
</tr>
</tbody>
</table>

Table 2. Pearson’s correlation coefficients between average grade scores and HNR computed via EMD-based approach and cepstral analysis for synthetic and natural /a/.

<table>
<thead>
<tr>
<th></th>
<th>HNR_{EMD}</th>
<th>HNR_{cep}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synt. /a/</td>
<td>-0.87</td>
<td>-0.85</td>
</tr>
<tr>
<td>Nat. /a/</td>
<td>-0.65</td>
<td>-0.59</td>
</tr>
</tbody>
</table>

Experiments have been carried out on stimuli of synthetic /a/ perturbed by one kind of noise only, i.e., the spectral effects of one kind of noise are not masked by those of the other kind. Pearson’s correlation coefficients between the average grade scores and the different spectral acoustic markers computed via EMD-based approach and cepstral-analysis are given in Table 3. The strongest correlation (ρ_p=-0.96) is achieved by the cepstral analysis-based HNR for stimuli perturbed by additive noise, however, for jitter-perturbed stimuli, EMD-based spectral acoustic cues outperform the cepstral analysis-based acoustic cues.

Table 3. Pearson’s correlation coefficients between average grade scores and spectral acoustic markers H1A and HNR computed via EMD-based approach and R1A and HNR computed via cepstral analysis for synthetic /a/.

<table>
<thead>
<tr>
<th></th>
<th>H1A</th>
<th>R1A synchron.</th>
<th>HNR_{EMD}</th>
<th>HNR_{cep}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add. noise</td>
<td>-0.93</td>
<td>-0.9</td>
<td>-0.92</td>
<td>-0.96</td>
</tr>
<tr>
<td>jitter</td>
<td>0.86</td>
<td>0.8</td>
<td>0.86</td>
<td>0.83</td>
</tr>
</tbody>
</table>

For each fundamental frequency f0, data corresponding to synthetic stimuli have been pooled to form a single sequence. Figure 3 shows HNR values of synthetic stimuli estimated via EMD-based method and cepstral analysis-based approach for different fundamental frequency values. It is observed that cepstral analysis-based HNR estimates are higher than those estimated via EMD-based method and exhibit more variability with fundamental frequency. Experiments carried out have shown that R1A acoustic cue exhibits more variability with the fundamental frequency compared with H1A.

![Figure 3](image)

**Figure 3:** Effect of the fundamental frequency. (a) EMD-based HNR estimates, (b) Cepstral analysis-based HNR estimates.

5. Conclusions

In this presentation, the performances of spectral acoustic cues obtained from the decomposition of the log magnitude spectrum of speech signal via EMD algorithm have been investigated and compared to those based on the cepstral analysis. Experiments carried out on synthetic /a/ and natural /a/ show that the EMD-based approach results in higher correlations between spectral acoustic cues estimates (H1A and HNR) and average perceived grade scores than those achieved by cepstral-based acoustic markers for synthetic /a/ perturbed by additive noise and jitter as well as for natural sustained /a/. Performance analysis of the EMD-based spectral acoustic cues for connected speech is considered as future work.
6. References


