BIVARIATE ANALYSIS OF DISORDERED CONNECTED SPEECH USING TEMPORAL AND SPECTRAL ACOUSTIC CUES

A. Kacha [1], *F. Grenez* [2], *J. Schoentgen* [2,3]

[1] Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Algeria
[2] Laboratory LIST, Université Libre de Bruxelles, Brussels, Belgium
[3] National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

ABSTRACT

The presentation concerns the assessment of disordered voices produced by dysphonic speakers. The empirical mode decomposition algorithm is used 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 (HNR) is used to summarize the overall quality of the disordered voices. The present study aims at improving a previously proposed algorithm by incorporating an appropriate method that estimates automatically the thresholds required by the algorithm without knowledge of the fundamental frequency and combining the temporal acoustic marker named segmental signal-to-dysperiodicity ratio (SDRSEG) with the harmonic-to-noise ratio in order to predict the degree of perceived hoarseness. The performances of the bivariate analysis-based approach for vocal dysperiodicities assessment in terms of correlation of the predicted perceived grade scores with the original perceived degree of hoarseness are investigated using a large corpus comprising concatenations of two Dutch sentences followed by vowel [a].

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

1. INTRODUCTION

Clinical evaluation of voice disorders is routinely based on listener perception of speech. For example, clinicians rate the degree of perceived overall abnormality, called grade, to monitor the voice of patients. This method of evaluation is subjective, i.e. the outcome is listener-dependent. A major drawback of perceptual ratings is intra and inter-judge variability [1][2]. Experiments have shown that to obtain reproducible evaluations, listeners must have substantial experience in voice timbre rating.

In contrast to subjective measures, objective measures are obtained from acoustic analysis of speech. 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. 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 and frequently result in an increase in the dysperiodicity of voiced speech sounds. Dysperiodicities may be caused by additive noise owing to turbulence and modulation noise owing to external perturbations of the glottal excitation signal, as well as dysperiodicities due to intrinsically irregular dynamics of the vocal folds [3][4]. As a consequence of these dysperiodicities, 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.

Most techniques for estimating vocal dysperiodicities 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 proposed for vocal dysperiodicities estimation in continuous speech are based on generalized variogram [5] and cepstral analysis [6][7].

In [8], we proposed the empirical mode decomposition (EMD) algorithm [9] as an alternative to decompose the log of the spectrum magnitude of the speech signal into its harmonic, envelope and noise components. The acoustic cue named harmonic-to-noise ratio (HNR) is used to summarize the degree of disturbance in the speech signal and consequently to evaluate the overall quality of the disordered voices produced by dysphonic speakers. The effectiveness of EMD-based spectral acoustic cues for assessing disordered voices has been investigated in [10] and their performances in terms of correlation with the perceived degree of hoarseness have been compared to those of their counterpart based on cepstral analysis. Experimental results have shown 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.

In the method proposed in [8], the thresholds involved in the algorithm for IMF clustering have been fixed empirically. These thresholds are f0-dependent, so that, the method requires the estimation of the average fundamental frequency for each stimulus. The objectives of the present study are the following: i) improve the algorithm by incorporating an appropriate method that estimates automatically the thresholds without knowledge of the fundamental frequency. ii) Combine the temporal acoustic marker named segmental signal-to-dysperiodicity ratio (SDRSEG) [5] with the HNR in order to predict the degree of perceived hoarseness.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

The remainder of the paper is organized as follows. The EMD-based method for log-magnitude spectrum decomposition is introduced in Section 2. In Section 3, the bivariate analysis based on the SDR and HNR acoustic cues is presented. Speech data and perceptual ratings are described in Section 4. Results based on real speech signals are presented in Section 5. Finally, conclusions are given in Section 6.

2. METHODS

2.1. 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, $v(t)$ [11]:

$$x(t)=e(t)*v(t) \tag{1}$$

where * denotes the convolution.

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

$$|X_w(f)|=|E_w(f)\times V(f)| \tag{2}$$

where $X_w(f)$, $E_w(f)$ 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:

$$\log|X_w(f)|=\log|E_w(f)|+\log|V(f)| \tag{3}$$

From (3), it is observed that the log magnitude spectrum is the sum of two spectral components: $\log|E_w(f)|$, 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, noted $V_{dB}(f)$, due the contribution of the vocal tract, a series of harmonics characterized by a periodic structure, noted $H_{dB}(f)$, and an irregular and rapidly varying part, noted $N_{dB}(f)$, 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 [12], so that the decomposition of the log magnitude spectrum via the EMD algorithm results into several oscillating components (IMFs, intrinsic mode functions) that can be clustered into three classes and each class of components is assigned to some part of the log magnitude spectrum.

In [8] the clustering of IMFs has been accomplished by a simple thresholding operation. Let f_j be the average quefrency of the jth-IMF component of the log magnitude spectrum obtained via the EMD algorithm. The different IMFs have been clustered by comparing their mean quefrencies to fixed thresholds $th_1=0.3/f0$ and $th_2=4/f0$. A drawback of this clustering procedure is that it requires the estimation of the average fundamental frequency of the speech signal which is not possible for all speakers. In this presentation, we propose a procedure for IMFs clustering that does not require the estimation of the average fundamental frequency.

Let $f0_{min}$ and $f0_{max}$ be the possible minimal and maximal average fundamental frequencies, respectively. The IMFs belonging to the harmonic component are determined according to the following algorithm:

1. Find the sets of IMFs having average quefrencies within the ranges $(0.3/f0_{min}, 4/f0_{min})$ and $(0.3/f0_{max}, 4/f0_{max})$

$$\frac{0.3}{f0_{\min}}<f_j<\frac{4}{f0_{\min}}, \quad j=p_0,p_0+1,\cdots,p_1 \tag{4-a}$$

$$\frac{0.3}{f0_{\max}}<f_j<\frac{4}{f0_{\max}}, \quad j=q_0,q_0+1,\cdots,q_1 \tag{4-b}$$

where p_0 and p_1 denote, respectively, the lowest and highest IMF indices the quefrencies of which are within the range $(0.3/f0_{min}, 4/f0_{min})$ while q_0 and q_1 denote, respectively, the lowest and highest indices of IMFs having quefrencies within the range $(0.3/f0_{max}, 4/f0_{max})$.

2. Form all possible candidates of the harmonic component by varying the lowest index between p_0 and q_0 and the highest index between p_1 and q_1 and then summing the corresponding IMFs for each combination of the indices

$$H_{dB}^{pq}(f)=\sum_{j=p}^{q} IMF_j, \quad p=p_0,p_0+1,\cdots,q_0 \quad q=p_1,p_1+1,\cdots,q_1 \tag{5}$$

where the superscript pq indicates the lowest and highest indices of the IMFs used to form a candidate of the harmonic component.

3. Compute the normalized autocorrelation sequence of each candidate of the harmonic component $H_{dB}(f)$ and perform an exhaustive search to find the normalized autocorrelation sequence with the most prominent peak at a nonzero delay. A large peak of the normalized autocorrelation sequence states for a high regularity of the harmonic component. The estimated harmonic component is given by

$$H_{dB}(f)=\sum_{j=p_h}^{q_h} IMF_j \tag{6}$$

where p_h and q_h denote, respectively, the lowest and highest indices of the IMFs that give rise to a normalized autocorrelation with the highest peak at a nonzero delay.

Once the lowest and highest indices of the IMFs of the harmonic component have been determined, the spectral envelope $V_{dB}(f)$ and noise $N_{dB}(f)$ are estimated as

$$V_{dB}(f)=\sum_{j=q_h+1}^{J} IMF_j+r_J(f) \tag{7}$$

$$N_{dB}(f)=\sum_{j=1}^{p_h-1} IMF_j \tag{8}$$

where r_J is the residue of the decomposition.

As an illustration, Figure 1 shows the estimated components of the log magnitude spectrum of a 200 ms frame taken from a sustained vowel [a] produced by a normophonic speaker. The estimated noise appears to be cyclic because noise energy dominates between harmonics locations and signal energy dominates at harmonic locations [13].

2.2. Baseline correction

The baseline is the inter-harmonic contour. The baseline correction is necessary because the IMFs are zero-mean oscillating functions so that when the harmonics are large, the inter-harmonics should cross zero to take negative values for compensation. Indeed, in the estimated harmonic, the baseline dips slightly towards negative values at low frequencies. The estimated envelope follows the baseline closely at high frequencies and deviates slightly above the baseline at low frequencies. The goal of the baseline correction is to straighten out the baseline.

Each computed candidate of the harmonic component is subject to a baseline correction before computing the normalized autocorrelation sequence. The baseline correction follows what is used in [14] for spectral tilt correction. The correction is carried out in double logarithmic coordinates where the envelope of the 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 has been applied to the components shown in Fig 1.

3. BIVARIATE ANALYSIS

The acoustic markers HNR and segmental SDR (SDRSEG) are used as predictor variables to predict the degree of perceived hoarseness. The HNR acoustic cue summarizes directly the amount of dysperiodicities within an utterance. For a given utterance, the analysis interval is divided into K frames and the HNR is computed as the average of the HNR_i (i=1,..., K) of the K frames:

$$HNR = \frac{1}{K}\sum_{i=1}^{K} HNR_i \qquad (9\text{-a})$$

where

$$HNR_i = 10\log\left[\sum_{k=0}^{M-1} H^2(k) \Big/ \sum_{k=0}^{M-1} N^2(k)\right], \mathrm{i}=1,\cdots,\mathrm{K} \qquad (9\text{-b})$$

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.

The acoustic marker SDRSEG is computed in the temporal domain as presented in [5]. Both the speech signal $x(n)$ and the corresponding dysperiodicity $e(n)$ estimated via the generalized variogram are divided into K frames 5 ms length and the SDRSEG is computed as:

$$SDRSEG = \frac{10}{K}\sum_{k=0}^{K-1} \log\left[\sum_{n=kL}^{kL+L-1} x^2(n) \Big/ \sum_{n=kL}^{kL+L-1} e^2(n)\right] \qquad (10)$$

where L denotes the frame length in number of samples.

The degree of perceived hoarseness is predicted as a weighted sum of the HNR and SDRSEG acoustic cues:

$$Hoarseness = a.HNR + b.SDRSEG + c \qquad (11)$$

The constants a, b and c are computed by carrying out multiple linear regression analysis.

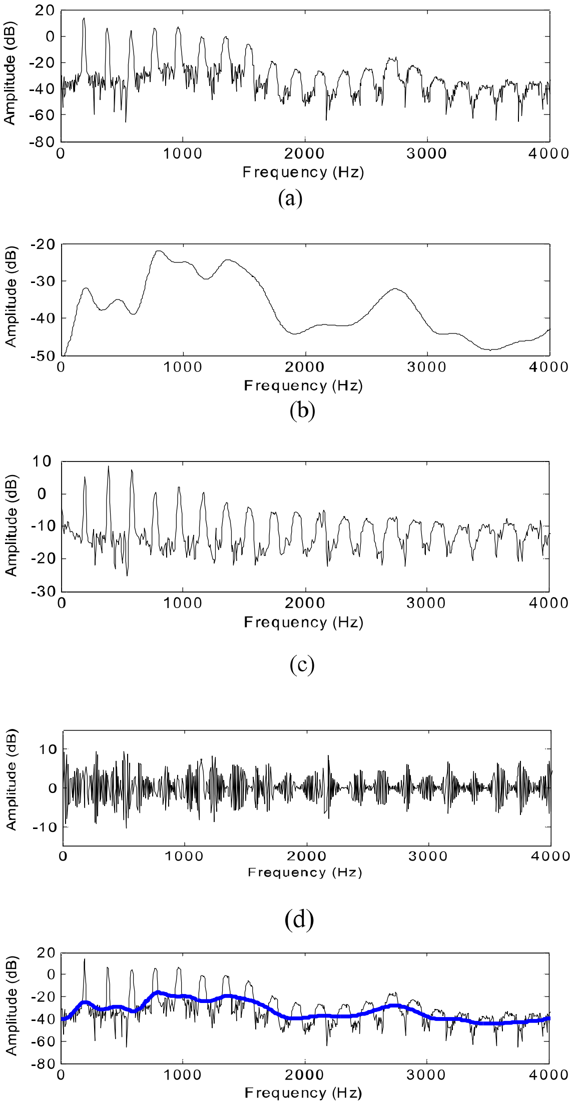

Figure 1: Decomposition of the log magnitude spectrum of a 200 ms speech frame of sustained [a] into three components via the EMD algorithm. (a) Log magnitude spectrum. (b) Envelope component. (c) Harmonic component. (d) Noise. (e) Sum of the three components superposed to the estimated envelope.

4. CORPORA AND PERCEPTUAL RATINGS

The corpus comprises concatenations of two Dutch sentences followed by vowel [a]. Dutch sentences ("Papa en Marloes staan op het station. Ze wachten op de trein.") have been produced by 28 normophonic and 223 dysphonic speakers with different degrees of dysphonia [15]. The motivation behind the structure of the material is to combine both voice contexts (sustained vowels and continuous speech) into one concatenated sample upon which auditory-perceptual ratings and acoustic measures could be completed. The stimuli have been sampled at 44100 Hz. Five judges have evaluated the corpus involving the concatenation of the sentences and vowel [a] perceptually. The five judges are professional voice therapists with at least five years of experience in clinical voice quality ratings. Each judge has rated, from 0 to 3, the item "grade" of the (G)RABS scale. "Grade" represents the degree of hoarseness or voice abnormality. The five perceptual scores per stimulus have been averaged. The recordings and evaluation have been made at the Sint-Jan General Hospital, Bruges, Belgium.

5. RESULTS AND DISCUSSION

The performance of the bivariate analysis based on the SDRSEG and HNR acoustic cues is investigated. Based on our previous investigations, the frame length for HNR estimation has been set to 200 ms.

The empirical mode decomposition-based approach for HNR estimation has been applied to the corpus including concatenations of two Dutch sentences followed by vowel [a]. The possible minimal and maximal average fundamental frequencies $f0_{min}$ and $f0_{max}$ have been fixed to 80 Hz and 250 Hz, respectively.

Table 1 gives Pearson product moment correlations of the HNR as well as SDRSEG values with average scores of grade for the different corpora. The null hypothesis R=0 has been rejected (R_{crit}=0.15, p=0.01). The last column of Table 1 gives the multiple correlation coefficients obtained by carrying out linear regression analysis by combining the HNR and SDRSEG as predictor variables. The null hypothesis R=0 has been rejected (R_{crit}=0.19, p=0.01, F=169.6). It can be seen that the performance of the HNR and SDRSEG alone in terms of correlation with the degree of perceived hoarseness is quite similar. The combination of the HNR and SDRSEG to predict the degree of perceived hoarseness results in a stronger correlation.

Table 2 shows the quartiles of the HNR values computed via the EMD-based approach as well as the quartiles of SDRSEG values estimated via the generalized variogram-based method. It is seen that for high level perturbations (low HNR or SDRSEG values), the EMD-based HNR values tend to be greater than the generalized variogram-based SDRSEG values, however, for small level perturbations, the EMD-based HNR provides lower values than those obtained via the generalized variogram-based SDRSEG as displayed in Fig. 2.

Table 1. Pearson product moment and multiple correlations of the HNR and SDRSEG values with average scores of grade for the corpus comprising concatenations of two Dutch sentences followed by vowel [a].

	HNR	SDRSEG	HNR and SDRSEG
Correlation	-0.70	-0.70	-0.76

Figure 3 displays the average perceived grade scores versus the predicted grade scores for the corpus comprising concatenations of two Dutch sentences followed by vowel [a]. The multiple correlation coefficient between predicted scores and assigned perceived grade scores is R=-0.76 indicating the high predictability of hoarseness scores by means of the HNR and SDRSEG acoustic cues. Bivariate analysis results in an improved performance in terms of correlation of predicted scores with scores of perceived hoarseness over the analysis based on HNR or SDRSEG alone the (individual) correlation of which is R=0.7. The correlation values corresponding to the univariate analysis via the HNR or SDRSEG alone are statistically significantly different from the correlation corresponding to the bivariate analysis (t-test, t=3.63, p=0.01).

Table 2. Quartiles of EMD-based HNR values and generalized variogram-based SDRSEG values, in dB, for the corpus comprising concatenations of two Dutch sentences followed by vowel [a].

	Min	1st quartile	Median	3rd quartile	Max
HNR	5.9	13.2	15.3	17	28.4
SDRSEG	5	15.6	17.5	18.9	24

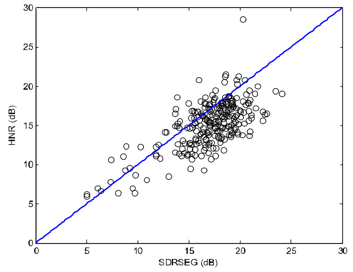

Figure 2: EMD-based HNR versus generalized variogram-based SDRSEG for the 251 samples of the corpus.

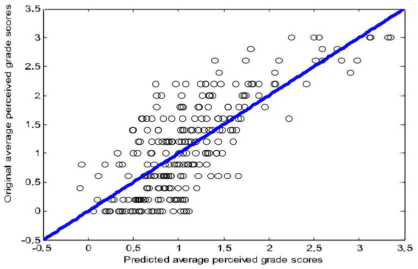

Figure 3: Original average perceived grade scores versus predicted grade scores via the HNR and SDRSEG combination.

6. CONCLUSION

In this presentation, the performance of the bivariate analysis combining the generalized variogram-based SDRSEG and EMD-based HNR acoustic cues has been investigated. The proposed approach has been tested on a corpus comprising 251 normophonic and dysphonic speakers. Experiments have shows that the bivariate analysis results in an improved performance in terms of correlation of predicted grade scores with original perceived grade scores of hoarseness over the univariate analysis based on the HNR or SDRSEG acoustic markers. The proposed method does not require knowledge of the fundamental frequency as an automatic method has been incorporated to estimate the thresholds required by the algorithm.

7. REFERENCES

[1] J. Kreiman and B. R. Gerratt, "Validity of rating scale measures of voice quality, " *J. Acoust. Soc. Am.,* Vol. 104, pp. 1598-1608, 1998.

[2] J. Kreiman, D. Vanlancker-Sidtis, and B. R. Gerrar, "Defining and measuring voice quality," *Workshop on voice quality,* pp.115-120, Geneva, Switzerland, aug. 2003.

[3] J. Schoentgen, "Spectral models of additive and modulation noise in speech and phonatory excitation signals," *J. Acoust. Soc. Am.*, Vol. 113, No 1, pp. 553-562, 2003.

[4] P. Murphy, "Spectral characterization of jitter, shimmer and additive noise in synthetically generated voice signals," *J. Acoust. Soc. Am.*, Vol. 107, No 2, pp. 978-988, 2000.

[5] A. Kacha, F. Grenez, and J. Schoentgen, "Estimation of dysperiodicities in disordered speech," *Speech Comm.*, Vol. 48, No 10, pp. 1365-1378, 2006.

[6] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, and P. Murphy, "Assessment of voice disordered via the first rhamonic," *Speech Comm.*, Vol. 54, pp. 655-663, 2012.

[7] P. Murphy, "On the first rhamonic amplitude in the analysis of synthesized aperiodic voice signals," *J. Acoust. Soc. Am.*, Vol. 120, No 5, pp. 2896-2907, 2006.

[8] A. Kacha, F. Grenez, and J. Schoentgen, "Assessment of disordered voices using empirical mode decomposition," in Proc. *Interspeech 2012*, Portland (USA), 2012.

[9] N.E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. R. Soc. London Ser. A*, Vol. 454, pp. 903-995, 1998.

[10] A. Kacha, F. Grenez, and J. Schoentgen, "Empirical Mode Decomposition-Based Spectral Acoustic Cues for Disordered Voices Analysis," in *Proc. Interspeech 2013*, Lyon (France), pp. 3632-3636, 2013.

[11] G. de Krom, "A Cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech and Hearing Res.*, Vol. 36, pp. 254-266, 1993.

[12] P. Flandrin, G. Rilling, and P. Conçalvès, "Empirical Mode Decomposition as a Filter Bank," *IEEE Signal Proc. Let.*, Vol. 11, pp. 112-114, 2004.

[13] P. –J. Murphy and O. -O. Akande, "Noise estimation in voice signals using short-term cepstral analysis," *J. Acoust. Soc. Am.*, Vol. 121, No 3, pp. 1679-1690, 2007.

[14] J. Schoentgen, M. Bensaid, and F. Bucella, "Multivariate statistical analysis of flat vowel spectra with a view to characterizing dysphonic voices," *J. Speech Lang. Hear. Res.*, Vol. 43, pp. 1493-1508, 2000.

[15] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels," *J. Voice*, Vol. 24, No 5, pp. 540-555, 2010.