Abstract: A short-time spectral analysis of normophonic and dysphonic voices is presented. This analysis has been performed on recordings of both sustained vowels and running speech for comparison purposes. The reported results indicate that pathological voices tend to have a higher concentration of energy in the lowest frequencies (20 to 300 Hz) and less energy in frequencies from 630 to 1,270 Hz. Additionally, pathological voices tend to experience quicker variations in spectral energy from 770 to 1,720 Hz.

Keywords: Speech analysis, spectral analysis, correlation

I. INTRODUCTION

The acoustic analysis of voice for clinical purposes has traditionally been made on sustained vowels [1] and a set of parameters measuring voice instability are of common use in clinical software for voice analysis [2]. However, extrapolating the use of such measures to running speech seldom provides good results [3], since the stationarity assumption only holds for sustained phonations. As for spectral domain, the lack of stationarity can be handled by means of short-time spectral measures. Thus, short-time spectral analysis may be useful for analyzing voice quality in running speech [4]. Within this paper, a short-time spectral analysis of normophonic and dysphonic voices is presented. This analysis has been performed on recordings of both sustained vowels and running speech for comparison purposes.

For the spectral analysis, the speech segments of the processed recordings have been passed through a filter bank so as to split the signal into the 22 first critical bands of the human auditory system, as identified by Zwicker [5]. For each band, the instantaneous energy has been computed and, subsequently, the autocorrelation function of each band energy sequence has been calculated. Both the absolute values of instantaneous energy and the width of the autocorrelation functions have been analysed. The reported results allow identifying relevant differences in spectral energy between normophonic and dysphonic voices. In addition, the width of the autocorrelation function is used as a cue for spectral stability. Results indicate that dysphonic voices tend to be more unstable than normophonic voices but only in bands above the 8th one (over 770 Hz).

II. MATERIALS

Processed voice recordings were taken from the Voice Disorders Database distributed by Kay Elemetrics [6]. Specifically, a subset of 53 normophonic and 173 dysphonic voices was selected [7]. For each voice two recordings were available: one corresponding to a sustained phonation of the vowel /æ/ and another corresponding to running speech, namely a fragment of the rainbow passage: “When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon”. For all recordings, the sampling frequency was 25 kHz and they were normalized in amplitude to have unit mean square value.

III. METHODS

Speech detection was performed on recordings corresponding to running speech. Specifically, a simple detector based on short-time energy and short-time zero-crossing rate was used [8]. Subsequently, both sustained vowel recordings and speech segments in running speech recordings underwent the same process. The first step consisted in the following time-frequency representation.

The short-time spectrogram of a discrete-time signal can be written as:

\[ S_p(k) = \sum_{n=p-L}^{n=p+L} x[n] w[n-pL] e^{-j2\pi(k(n-pL)) / N_{\text{DFT}}} \]  

where \( p \) is the frame number, \( k / N_{\text{DFT}} \) is the normalized frequency, \( L \) is the number of samples between consecutive frames, \( w[n] \) is the framing window which has a length equal to \( 2N+1 \) and \( N_{\text{DFT}} \geq 2N+1 \) is the number of points of the discrete Fourier transform (DFT).

Defining the framing window to be symmetric and if \( w[n] = 0 \; \forall |n| > N \), then (1) can be written as follows:

\[ S_p(k) = \sum_{n=-\infty}^{n=\infty} x[n] w[n-pL] e^{-j2\pi(k(n-pL)) / N_{\text{DFT}}} \]
Since (2) has the form of a convolution, the sequence of values of the spectrogram corresponding to the $k^{th}$ frequency can be written as a convolution followed by decimation by a factor $L$:

$$S_p(k) = S_x[p] = (x[n] * h[k])_{n \text{mod} L} \quad (3)$$

being

$$h[k] = w[k]e^{jk(\Omega \pi / L)} = w[k]e^{j\Omega k} \quad (4)$$

While the usual spectrogram has a common window $w[k]$ for all values of $k$, the formulation in (3) allows defining different windows for different frequency bands, hence $w[k]$ instead of $w[k]$.

For the herein reported work, $w[k]$ were chosen to be hamming windows with odd length. The specific length of each one was selected so that its -3 dB bandwidth matched the width of the $k^{th}$ critical band. The odd lengths allowed integer group delays that permitted subsequent time alignment of all the 22 resulting sequences. Also, each value of $\Omega$ was selected such that $\Omega / f_s / 2\pi$ was equal to the center frequency of the $k^{th}$ band. Fig. 1 provides an overview of the whole scheme.

In a second step, the instantaneous energy, i.e. square modulus, of every filter’s output was calculated:

$$e[k] = |x[n]|^2 \quad (5)$$

Last, the normalized autocorrelation function of each energy signal was computed.

After (5), the band energy decorrelation time was defined as the highest time shift $m f_s$, being $f_s$ the sampling frequency, for which $\rho[m] \geq 0.5$. The concept of decorrelation time is illustrated in Fig. 2.

**IV. RESULTS**

For both sustained vowels and running speech, the instantaneous band energies $e[k]$ were averaged to yield 22 mean band energies per voice record. In the case of sustained vowels, averaging was performed along the full record lengths while for running speech averaging was carried out only along speech segments. Fig. 3 shows the median and the 25th and 75th percentiles of the mean band energies for both sustained vowels and speech segments in running speech. In dysphonic voices, there is a significantly larger portion of energy distributed in critical bands 1 to 4 with respect to the case of normal voices. In the case of sustained vowels, this feature corresponds to a notably lower amount of energy in bands 8 to 10 (770 to 1,270 Hz). In running speech, the difference in energy along the first bands is lower and the spectral range for which dysphonic voices have less energy spans from the 7th to the 14th critical band (630 to 2,320 Hz).
Figure 3. Energy distribution along the 23 first critical bands for sustained vowels (top) and speech segments in running speech (bottom). Continuous lines represent the median value among all recordings, in gray for pathological voices and black for normal voices. Dashed lines indicate the 25th and 75th percentiles.

Within Fig. 4, the median and 25th and 75th percentiles of the band energy decorrelation times are depicted. In running speech, the expectation operator $E[\cdot]$ in (5) has been applied only to speech segments. Shorter decorrelation times indicate faster decays in the autocorrelation function and this, in turn, is a cue of quicker variations in the characteristics of the signal ($e[n]$ in this case). As shown by the graphs, pathological voices tend to have energy variations in high frequency bands significantly faster than normal voices. This difference seems to be relevant in bands above the 15th one, i.e. frequencies above 2,500 Hz, for the case of sustained vowels. However, such trend does not occur equally in the case of running speech. In this case, results indicate a slighter trend of pathological voices to exhibit shorter decorrelation times in bands 8 to 12 (770 to 1,720 Hz) but the distributions of decorrelation times almost completely overlap for bands beyond the 16th (3,150 Hz).

V. DISCUSSION

The reported results indicate that pathological voices tend to have a higher concentration of energy in the lower critical bands (1 to 3). Such fact, especially in what affects the first band (20 to 100 Hz), is related to the lack of periodicity of the voice. In fact, for fundamental frequencies above 100 Hz, the energy corresponding to the first band is only related to inter-period variations. In the case of running speech, articulation is one evident cause for aperiodicity and this is reflected by normal voices having more energy in the lowest frequency bands than in the case of sustained vowels. Typical phoneme durations around 100 ms [9][10] are related to frequency components in the range of hertz or tens of hertz, that is, frequencies corresponding to the lowest bands.

In contrast, pathological voices have similar low-frequency energies in both cases (sustained vowels and running speech); thus, for pathological voices the main cause for lack of stationarity does not seem to be articulation, but pathology itself.

Considering these results, another common feature of both kinds of phonations, which should allow distinguishing between normal and pathological voices, is the ratio of low to high frequency energy; hence a measure of spectral tilt. Spectral tilt has been reported to be a good indicator of breathiness [11]. Herein described results indicate that spectral tilt, measured as the ratio of energy in bands 1 to 3 (20 to 300 Hz) to energy in bands 7 to 10 (630 to 1,270 Hz), should also be a good indicator of dysphonia, both for sustained vowels and for running speech. The use of linear-phase filters in the filterbank of Fig.1 allows such ratio to be computed in short term, at rates up to $f_s$.

As for the band energy decorrelation time, a measure of band energy variability, while for sustained vowels this measure provides a fair distinction between normal and pathological voices for the highest frequency bands, this is not the case for running speech. Only a minor discrimination ability is to be expected for bands 8 to 12 (770 to 1,720 Hz). Yet, from the absolute value of the decorrelation time, a relevant indication can be obtained for the design of short-term spectral processing schemes: frame sequences with a rate below 100 frames/second ($10^2$ seconds/frame) would not be able to adequately capture the energy variability of pathological voices in bands above the 16th (3,150 Hz) for sustained vowels or, alternatively, the 13th (2,000 Hz) for running speech. Furthermore, for higher frequency bands a frame rate of...
at least 1,000 frames per second seems to be required, due to decorrelation times in the order of 1 ms.

ACKNOWLEDGEMENTS

This research work has been financed by the Spanish government through the project grant TEC2009-14123-C04-02. It has also been realized within the framework of COST Action 2103.

REFERENCES


Figure 4. Distribution of decorrelation times of band energy sequences for sustained vowels (top) and running speech (bottom). Figure key is as in Fig. 3.