CEPSTRAL ANALYSIS OF PERCEPTUALLY RATED SYNTHETIC DISORDERED SPEECH STIMULI

A. Alpan¹, F. Grenez¹, J. Schoentgen¹,²

¹ Laboratoires d’Images, Signaux et Dispositifs de Télécommunications,
University Libre de Bruxelles, Brussels, Belgium
² National Fund for Scientific Research, Belgium

Abstract: A number of studies have shown that the amplitude of the first rahmonic peak (R1) in the cepstrum may indicate hoarse voice quality. The cepstrum is obtained by taking the inverse Fourier transform of the log-magnitude spectrum. The goal of the article is to apply cepstral analysis to a perceptually evaluated corpus of synthetic stimuli to learn about the link between the signal properties (fixed by the synthesizer parameters) and the first rahmonic peak. The synthetic stimuli have been generated by a synthesizer of disordered voices that has been shown to generate natural-sounding speech fragments comprising different vocal perturbations. A second objective is to examine the link between first rahmonic peak and perceived breathiness and roughness, link which has not been studied previously. The speech stimuli have been perceptually assessed by nine listeners according to grade, breathiness and roughness. A number of cepstral analysis alternatives have been implemented, including period-synchronous temporal frames and harmonic-synchronous band-limited analyses.

Keywords: cepstral analysis, synthetic disordered speech, first rahmonic amplitude

I. INTRODUCTION

Acoustic analysis has a central place within the context of the assessment of laryngeal function because the speech signal may be recorded non-invasively and it is the basis on which the perceptual assessment of voice is founded. Many voice disorders cause voiced speech to deviate from periodicity. Dysperiodicities may be caused by additive noise owing to turbulent airflow and modulation noise owing to perturbations of the frequency and amplitude of the glottal excitation signal. Dysperiodicities may also be due to intrinsically irregular dynamics of the vocal folds and involuntary transients between dynamic regimes [1].

Several acoustic features that have been used to assess vocal fold function report the deviation of the voiced speech waveform from perfect periodicity. Vocal jitter and shimmer, for instance, are frequently used to summarize perturbations of the voiced speech cycle lengths and amplitudes, respectively. A signal that has shown promise as a global descriptor of voice quality is the cepstrum. Global descriptors designate features that report different voice qualities as patterns rather than focus on narrowly-defined properties of the speech signal.

The cepstrum is defined as the inverse magnitude spectrum of the log-magnitude spectrum [2]. Because the logarithmic power spectrum of voiced speech consists of equally spaced harmonics, a peak occurs in the inverse Fourier transform of this signal (the cepstrum) at a point corresponding to the glottal period [3]. Previous studies have shown that the amplitude of the first rahmonic peak in the cepstrum (R1) is a global descriptor of glottal turbulence noise and modulation noise [4].

Although, it has been frequently observed that increased levels of noise and perturbations in the voice signal decrease R1, a formal description of cepstral peak R1 has been lacking. Murphy has provided a theoretical description of cepstral analysis of voiced speech with aspiration noise, which suggests that R1 is directly proportional to a geometric-mean harmonics-to-noise ratio [5]. He shows that R1 and the geometric-mean harmonics-to-noise ratio (measured spectrally) underestimate the actual geometric-mean harmonics-to-noise ratio when averaged noise levels exceed harmonic levels. Limiting the number of harmonics in the analysis window overcomes this problem and in the case of period-synchronous analysis also alleviates the dependence of R1 on (temporal) window length and F0.

For the present study, a corpus of synthetic sound stimuli has been obtained by means of a synthesizer of disordered voices [6]. It can mimic a wide range of speech source perturbations such as additive noise at the glottis, vocal frequency jitter, vocal shimmer, vocal frequency tremor, amplitude tremor, diplophonia, biphonation and random glottal cycles.

The synthetic stimuli are vowels [a], [i], [u] and transients [ai] and [ia]. Each has been perceptually assessed by nine professional listeners according to grade, roughness and breathiness.

The purpose of the article is to apply cepstral analysis to a perceptually evaluated corpus of synthetic stimuli to learn about the link between the signal properties (fixed by the synthesizer parameters) and the first rahmonic peak. A second objective is to examine the link between first rahmonic peak and perceived breathiness and.
roughness, link which has not been studied previously. A number of spectral analysis alternatives have been implemented, including period-synchronous temporal frames and harmonic-synchronous band-limited analyses.

II. CORPUS

The synthesizer involves models of the glottal area and airflow through the glottis. The time-evolving glottal area is modelled by means of a nonlinear memoryless signal model that transforms a trigonometric driving function into the desired glottal area waveform. One attractive property of the model is that the frequency and harmonic richness of the glottal area are controlled by the instantaneous frequency and amplitude of the harmonic driving function.

The glottal airflow rate is generated by means of an algebraic aerodynamic model involving the glottal area and including interactions between the glottis and the infra- and supra-glottal ducts. The propagation of the acoustic wave through the trachea and vocal tract is simulated by means of concatenated tubes. Wall vibration, viscous and thermal losses as well as acoustic reflection and radiation at the lips and glottis are taken into account. Modulation noise such as jitter or tremor and abnormal voice qualities such as diplophonia, biphonation and irregular vocal cycles are mimicked by means of stochastic or deterministic models of the time-evolving instantaneous frequency or amplitude of the driving harmonics of the glottal area model.

The ability of the synthesizer to mimic natural disordered voices has been demonstrated in the framework of several perceptual experiments [7].

The corpus comprises synthetic sounds [a], [i], [u], [ia] and [ai] which are one second long. The vowel transition has been simulated by evolving linearly the tract area function from one vowel target to the next over an interval of 0.2 s in the middle of the one second interval. Each set is composed of 48 stimuli that combine three values of vocal frequency, four levels of frequency jitter and four levels of additive noise. The vocal frequency values are 100, 120 and 140 Hz. The jitter and additive noise have been fixed based on the independent advice of one phoniatrician and one speech 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. The area function of the vocal tract has been identical for all stimuli of the same vowel category [6].

Eight speech therapists and one phoniatrician have perceptually evaluated each set of synthetic sounds according to perceived “grade” (G), “roughness” (R) and “breathiness” (B) with four degrees per scale: 0 (normal), 1 (feeble), 2 (moderate) and 3 (severe). The scores of G, R and B have been averaged over the nine judges.

III. METHODS

A. First rahmonic amplitude (R1)

Figure 1: Amplitude spectrum and cepstrum of sustained vowel [a] showing the first rahmonic amplitude.

Figure 2: Dashed line: Log-magnitude spectrum. Solid line: Band-limited and offset-removed log-magnitude spectrum for an analysis frame of vowel [a].

• Full-band spectrum

The computation of the amplitude of first rahmonic (R1) involves the following. The amplitude spectra of the hopped Hamming-windowed frames are averaged and the log-magnitude of the average is taken (Fig. 1). The cepstrum is obtained via the inverse Fourier transform of the log-amplitude average spectrum. First rahmonic R1 is located in the vicinity of the quefrency corresponding to the glottal cycle length. The analysis has been period-synchronous, i.e. the lengths of the frame have been multiples of the vocal cycle length.
• Band-limited spectrum

The computation of R1 implicates the same steps as previously. However, prior to the computation of the cepstrum, the log-average spectrum is limited to a fixed number of harmonics and the offset is removed (Fig. 2).

B. Correlation analysis

The correlation coefficients of the amplitude of first harmonic (R1) obtained via the different options (full-band, band-limited) with the average perceptual scores for grade, roughness and breathiness have been computed.

C. Multi-cue regression analysis

The amplitude of first harmonic (R1) has been regressed on the parameters of the synthesizer fixing additive noise, jitter and fundamental frequency. The phonetic category of the stimuli has been taken into account by a dummy variable: 1 for [a], 2 for [ai], 3 for [ia], 4 for [i] and 5 for [u].

IV. RESULTS

A. Correlation analysis

![Figure 3: Correlation of period-synchronous (6 cycles) band-limited R1 with perceptual scores GRB of vowel [a]. The dashed lines correspond to the correlations of full-band R1 with GRB.](image)

Fig. 3 displays the correlations of period-synchronous (6 periods) full-band and band-limited R1 with average perceptual scores for vowel [a]. One observes that R1 is highly correlated with the average scores of roughness and grade and moderately correlated with breathiness. Limiting the spectrum to a feeble numbers of harmonics prior to computing the cepstrum enables improving the correlation with average perceptual scores. In particular, one observes that if the spectrum is limited to only two harmonics a correlation of 0.80 with the average breathiness scores is obtained. This correlation rapidly decreases to 0.55 when increasing the number of harmonics in the spectrum.

The correlations obtained for vowel [i] are slightly smaller but similar to other vowels and vowel-vowel pairs.

B. Multi-cue regression analysis

![Figure 4: Standardized regression coefficients of the additive noise and jitter control parameters predicting R1.](image)

The period-synchronous (6 periods) band-limited R1 has been predicted via the additive noise, jitter and vocal frequency control parameters of the synthesizer and a dummy variable that takes phonetic category into account. Fig. 4 displays the standardized regression coefficients of the additive noise and jitter parameters. Vocal frequency and phonetic category do not contribute to the prediction of period-synchronous R1. Jitter contributes most. However, when the number of harmonics in the band-limited spectrum is smaller than 6, the contribution of pulsatile additive noise exceeds the contribution of jitter. Also, when the number of harmonics in the spectrum increases the correlation of R1 with additive noise decreases (down to 0.35) while the correlation with jitter increases (to 0.85).

V. DISCUSSION AND CONCLUSION

Correlations of full-band R1 with perceived roughness ($\rho \approx 0.85$) and grade ($\rho \approx 0.80$) are good. However, only a moderate correlation is observed with perceived breathiness ($\rho \approx 0.55$).

Limiting the spectrum to a feeble number of harmonics improves the correlation. The largest
improvement is observed for breathiness. Indeed, when the spectrum is limited to 2 harmonics, the correlation increases to 0.80.

One has also observed that jitter contributes most to predicting the period-synchronous band-limited R1 (Fig.4) when the number of harmonics in the spectrum is larger than 6. A possible explanation is that cue R1 mainly reports modulation noise that broadens and decreases harmonic amplitudes and adds spectral sidebands. In the case of homogeneous modulation noise, spectral effects are proportional to the order of the harmonic. When the number of harmonics that is taken into account is feeble at low-frequencies, modulation noise effects are less prominent in R1 and the influence of additive noise, which is harmonic independent, increases. Fig. 5 displays the average spectrum for two vowels [a] with the same additive noise level and low and high vocal jitter levels.

One observes that the correlation of roughness and grade with cue R1 increases (Fig.3) whereas the correlation of R1 with jitter decreases (Fig.4) when the bandwidth of the log-amplitude spectrum decreases. A possible explanation is that experiments reported elsewhere suggest that the perception of roughness is effected both by modulation noise and additive pulsatile noise.

One then observes that period-synchronous (6 periods) harmonic-limited R1 obtained from stimuli perturbed by the additive noise only is very well correlated with the noise level ($\rho \approx 0.90$). This correlation is larger than the correlation obtained for stimuli containing vocal jitter only ($\rho \approx 0.80$).

Also, the synthetic stimuli used here are more natural than in [5]. Indeed, in [5] the purpose of the stimuli was not to mimic natural disordered voice and the noise characteristics also differ between both studies. Aspiration noise in [5] was synthesized by means of a zero-mean white Gaussian noise added to the glottal source, whereas in this study, additive noise is mimicked by means of Brownian noise, the amplitude of which is modulated via an affine function of the glottal airflow rate [7]. Also, in [5] vocal jitter is synthesized through time scaling of glottal waveforms, whereas here it is caused by small random perturbations of the instantaneous frequency of the driving function of the synthesizer.

VI. ACKNOWLEDGEMENTS

This research has been supported by COST ACTION 2103 “Advanced Voice Function Assessment” in the framework of a short-term scientific mission at the University of Limerick, and by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme.

REFERENCES