Analysis of Vocal Dysperiodicities in Running Speech

*F. Bettens(*), F. Grenez(*), J. Schoentgen(**, ***)*

(*) Department Signals and Waves, Faculty of Applied Sciences, CP 160
(**) Laboratory of Experimental Phonetics, CP 110
Université Libre de Bruxelles
50, Av. F.-D. Roosevelt
B-1050 Brussels
Belgium
tel: +32 2 650 3660
fax: +32 2 650 2007
(***) National Fund for Scientific Research, Belgium
jschoent@ulb.ac.be

Abstract

Voice disorders are often characterized by an increase of the dysperiodicities in voiced speech sounds. These dysperiodicities are difficult to track reliably in connected speech uttered by severely hoarse speakers. The object of the presentation therefore is to discuss dysperiodicity analysis methods that are robust with regard to large deviations from strict periodicity and that can be applied to running speech. They do not include a determination of the glottal cycle lengths and can be meaningfully performed on any speech signal whether it is periodic, aperiodic or random. Two experiments are reported that involve sustained vowels as well as connected speech. Results show that perceived degrees of hoarseness of sustained vowels are statistically significantly correlated with computed signal-to-dysperiodicity ratios and that signal-to-dysperiodicity ratios of sustained vowels are statistically significantly correlated with signal-to-dysperiodicity ratios of connected speech.

1. Introduction

Voice disorders are often characterized by an increase of the dysperiodicities in voiced speech sounds. Increased dysperiodicities cause the speaker's voice to be perceived as hoarse. Perceived hoarseness as well as measured dysperiodicities are the effects of multiple causes that include increased vocal jitter and shimmy, vocal amplitude and frequency tremor, additive noise owing to turbulent airflow, random vibrations of the vocal folds, involuntary vibrations of the false vocal folds or aryepiglottal folds and phonation breaks or octave jumps. The great diversity of disorder-related speech phenomena as well as the instationarity of connected-speech signals make that conventional analysis methods may fail on running speech produced by speakers that are moderately or severely hoarse. The reason is that most analysis methods rely on the detection of individual speech cycles or spectral harmonics and are tweaked so as to work well with a limited range of voices. The object of this presentation is to introduce dysperiodicity analysis methods that are robust with regard to arbitrary deviations from strict periodicity and that can be applied to running speech.

2. Problems

The difficulty of analyzing dysperiodicities in (voiced) running speech is that the speech cycle shape and amplitude as well as vocal frequency continually evolve with the identities of the phonetic segments, their onsets and offsets, their intrinsic intensity, as well as accentuation, intonation, declination and vocal loudness. Also, as a rule, the speech signal is sampled before analysis. Sampling and quantization noise is expected to become severe when the glottal cycle length is not exactly equal to an integer number of speech samples. The analysis methods that are presented hereafter explicitly take into account the above phenomena to avoid biasing measured dysperiodicities.

3. Analysis model

A periodic speech signal s(t) is mathematically defined as follows: $s(t) = s(t+T)$, $-\infty < t < +\infty$, with period T a constant. This suggest the following (formal) definition of vocal dysperiodicities: $d(t) = s(t) - s(t+T)$, $-\infty < t < +\infty$. Hereafter, this definition is modified to enable the distinction to be made between changes in running speech that are due to evolving segmental and supra-segmental units and those that are due to voice disorders.

First, the formal analysis interval $-\infty < t < +\infty$ is replaced by an analysis frame of finite length, $0 < t < L$,

to take into account that the speech signal depends on the phonetic identities of the speech segments.

Second, sample s(t) is compared to sample $\alpha s(t+T)$, in place of s(t+T). Coefficient α is a local gain that is the same for all samples within an analysis frame. Its purpose is to compensate for changes in the signal amplitude that are due to intrinsic segmental intensity, onsets and offsets or evolving loudness.

Third, gain α is split among three neighboring samples. That is, sample s(t) is compared to a weighted sum $\alpha_{-1}s(t-1+T) + \alpha_0 s(t+T) + \alpha_1 s(t+1+T)$. The values of the weights are determined once for each analysis frame. The purpose of the running average is to decrease the effects of quantization noise and sampling.

Fourth, within an analysis frame, samples s(t) are compared to weighted samples $\alpha_i s(t+i+T)$ to the right ($T > 0$) and left ($T < 0$), and the smallest differences are kept as markers of the amount of vocal dysperiodicity. The reason is that it is meaningless to compare samples that belong to different phonetic segments, i.e. that are on different sides of a phonetic boundary. Provided that shift T is shorter than a phonetic segment, the combined left-right and right-left comparisons guarantee that a given sample is compared at least once to samples that belong to the same phonetic segment. Taking the minimum difference warrants that the intra-segment difference is retained because inter-segment differences are expected to be larger owing to changes in phonetic identity.

Fifth, length T over which the sample comparison is performed is comprised in an interval (T1 ... T2) that is fixed by the experimenter. Length T is determined anew for each analysis frame. The purpose is to track changes in phonatory frequency that are due to intonation or declination.

4. Methods

The previous analysis scheme comprises three parameters that are fixed by the experimenter. These are the length L of the analysis frame, as well as boundaries T1 and T2. These boundaries have been assigned the values 2.5 and 20 ms respectively [1]. Shift T is therefore comprised in the interval (2.5 - 20 ms), which includes the speaking glottal cycle length of a great majority of subjects irrespective of their gender or age. Length L of the analysis frame has been fixed at 2.5 ms, which is the lower limit T1 of the shift interval. This means that the length of the analysis frame is shorter than or equal to a speaker's glottal cycle length. The analysis is performed by sliding the frame from left to right without overlap.

For each analysis frame position, a pair of shifts T and pairs of gains α_i are computed because each frame is compared to a frame shifted by T to the left and to a frame shifted by T' to the right. To determine the value of the shift, the cross-correlation is computed between the present and lagged analysis frames, with the lag being comprised between boundaries T1 and T2. Shift T is assigned to the lag for which the cross-correlation is maximal.

Coefficients α_i are determined via a long-term predictive analysis. A long-term predictive analysis is similar to a conventional short-term predictive analysis of speech. Short and long-term predictive analyses have different purposes, however. In the case of a long-term analysis, the present sample is predicted by means of a small number of distant samples $\alpha_{-1}s(t-1+T) + \alpha_0 s(t+T) + \alpha_1 s(t+1+T)$. Weights α_i are determined by means of Burg's method. For each analysis frame, the vocal dysperiodicities are assigned to the left-right or right-left long-term prediction error (whichever is smallest).

The reason this analysis scheme is robust is that all operations can be performed meaningfully on any signal, whether it is periodic, aperiodic or random. The analysis does indeed not require estimating the lengths of the glottal cycles. The quantity that is outputted (i.e. the long-term prediction error) has the desirable property of being zero when the signal is periodic and to increase feebly when the signal is disturbed lightly and to increase strongly when the signal is random. The overall degree of dysperiodicity is summarized by means of a signal-to-dysperiodicity (S/D) ratio in dB [1].

Ramachandran et al. and Qi et al. have proposed long-term linear prediction for speech coding and for clinical speech analysis respectively [1,2]. However, the long-term predictive schemes of Qi et al. and Ramachandran et al. are left-right only and therefore not suitable for the direct analysis of speech. The reason is that the speech signal cannot be meaningfully long-term predicted across phonetic boundaries. Qi et al. therefore long-term analyze the residue signal. The residue signal is obtained by processing the speech signal by means of a conventional short-term linear predictive model. Disadvantages are that the residue signal is an artificial signal from which fractions of the vocal disturbances have been removed and the properties of which are not always clinically relevant [3].

5. Experiments

We have carried out two experiments. The first involved 89 normophonic and dysphonic male and female speakers who had sustained vowels [a]. One-second long steady fragments of the vowels were analyzed by the method of Bettens et al. that is explained above as well as by the method of Qi et al [1]. The signal-to-dysperiodicity ratios were then correlated with a perceptual classification according to the degree of hoarseness. The purpose of the experiment was to gauge whether the dysperiodicities that are so isolated increase with perceived hoarseness. The perceptual assessment

of the vowel fragments has been described in [4].

The second experiment involved 22 male and female dysphonic and normophonic speakers who had sustained vowels as well as produced four different sentences that had been matched for number of syllables and grammatical structure. Two sentences were (phonologically) voiced throughout and the other two comprised voiced as well as unvoiced consonants. The utterances were analyzed with the analysis method presented here as well as the one of Qi et al. The objectives of the experiment were indeed to compare both methods (i.e. speech versus short-term residue and bi-directional versus mono-directional analyses) as well as compare the signal-to-dysperiodicity ratios of different speech fragments (i.e. vowels versus connected speech as well as all-voiced versus voiced and unvoiced fragments).

6. Results and discussion

In the first experiment, the rank correlation coefficients of the perceived degrees of hoarseness and computed signal-to-dysperiodicity ratios were equal to -0.80 and -0.71 for the methods of Bettens et al. and Qi et al. respectively. This suggests that these analyses are indeed able to isolate vocal dysperiodicities to which a clinical relevance may be assigned.

The second experiment showed that the values of the signal-to-noise ratio computed by the methods of Bettens et al. and Qi et al. were statistically significantly different. The explanation is that the method of Qi et al. comprises a preliminary analysis stage that discards some of the vocal disturbances. Only those that give rise to signal modeling errors are retained for analysis. Also, the method of Qi et al. may overestimate vocal noise in the vicinity of some phonetic boundaries. The excess noise is due to the mono-directional analysis, which requests the long-term prediction of voiced by means of unvoiced segments and vice-versa.

The results also showed that the signal-to-dysperiodicity ratios computed for sustained vowels (including onsets and offsets) and connected speech were statistically significantly correlated. Similarly, the S/D ratios of sentences that exclusively comprised voiced segments were statistically significantly correlated with the S/D ratios of sentences that comprised voiced as well as unvoiced segments. The explanation is that the signal-to-dysperiodicity ratio is a global measure whose value is mainly determined by vocalic segments. Consonants whether they are voiced or unvoiced, as well as other transients contribute less because they are short. A corollary is that when the aim is to focus on segment onsets or offsets or other transients, which have been conjectured to be increasingly noisy in the case of dysphonic speakers, features must be discovered that are able to describe isolated events in the dysperiodicity trace.

7. References

[1] Qi, Y., Hillman, R.E., and Milstein, C., 1999. "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *J. Acoust. Soc. Am.* 105(4), 2532-2535.

[2] Ramachandran, R., and Kabal, P., 1989. "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Proc.* 37(4), 467-478.

[3] Schoentgen, J., 1982. "Quantitative evaluation of the discrimination performance of acoustic features in detecting laryngeal pathology," *Speech Commun.* 1, 269-282.

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