ESTIMATION OF VOCAL NOISE AND CYCLE DURATION JITTER IN CONNECTED SPEECH

A. Alpan¹, F. Grenez¹, J. Schoentgen¹,²
¹Laboratory of Images, Signals & Telecommunication Devices,
Université Libre de Bruxelles, 50, Avenue F. D. Roosevelt, 1050 Brussels, Belgium
²National Fund for Scientific Research, Belgium

Abstract: The objective is to describe analysis methods that enable tracking vocal dysperiodicities in running speech. Vocal dysperiodicities here refer to deviations from strict periodicity in voiced speech sounds. Two methods are described. They respectively enable the sample-by-sample extraction of vocal noise from the speech signal or the isolation of speech cycles in voiced segments to quantify perturbations of the cycle lengths and amplitudes (i.e. cycle duration jitter and amplitude shimmer). These methods share the property that they are not based on the assumption that the signal is locally periodic and that the average period length is known a priori.

Keywords: Vocal noise, jitter, running speech

I. INTRODUCTION

The objective of the presentation is to describe analysis methods that enable tracking vocal dysperiodicities in running speech. Vocal dysperiodicities here refer to deviations from strict periodicity in voiced speech sounds. The description of vocal dysperiodicities is a common practice in the framework of the clinical assessment of vocal function.

Acoustic descriptors of vocal dysperiodicity are temporal or spectral. Frequently, they are extracted from sustained speech sounds. Privileging steady sounds when analyzing vocal disturbances is a matter of technical feasibility rather than clinical relevance. It is indeed the case that existing clinical voice analysis software is able to deal with sustained sounds only or is known to fail on speech produced by severely hoarse speakers. The reason is that many analysis methods are based on the hypothesis that the analyzed sounds are locally periodic. This is an assumption that is not valid under all circumstances, however [1].

Therefore, we have developed methods that enable estimating vocal dysperiodicities in speech that is not steady and that may be produced by severely hoarse speakers. The methods that are described make possible the sample-by-sample extraction of vocal noise from the speech signal, as well as the isolation of speech cycles in voiced segments to quantify perturbations of the cycle lengths and amplitudes (i.e. cycle duration jitter and amplitude shimmer). Descriptors of vocal jitter and shimmer differ from descriptors of vocal noise in general insofar that they focus on modulation noise exclusively.

Generally speaking, the description of vocal jitter and shimmy is regarded to be meaningful only when the speech segments are pseudo-periodic. At this stage, it is not clear whether these limitations are the consequence of a lack of reliability of existing signal analysis methods or a lack of validity of the extracted vocal cues.

Two methods are described. They share the property that they are not based on the assumption that the signal is locally periodic and that the average period length is known a priori. The first method enables tracking noise (whatever the cause) in any speech sound produced by any speaker.

The second method consists in a multi-resolution analysis of the signal samples in terms of their salience. Sample salience designates the duration over which a signal sample is a maximum. Salience is a relevant signal feature because one observes that signal peaks that are similarly positioned in vocal cycles may have similar saliences even if the peak amplitudes differ widely. This also applies to peaks in cycles the durations of which are perturbed moderately. The salience of signal peaks can therefore be used to detect automatically voiced speech cycles because they display a preeminent peak in the vicinity of glottal closure.

II. METHODS

A. Extraction of vocal dysperiodicities

The method is based on the observation that when in a 2-dimensional graph one reports on the horizontal axis samples of a noise-free periodic signal and on the vertical axis samples that are identically positioned in an adjacent period then all sample pairs \((x,y)\) are located on the bisector of the graph.

In a noisy signal, pairs \((x,y)\) remain in the vicinity of the bisector, as shown in Fig.1. The cumulated distance between pairs and bisector over an analysis frame is a measure of the total signal noise in that frame and the individual distances between each pair and the bisector are sample-by-sample estimates of the noise (whatever its cause).

In practice, a sliding rectangular analysis window of 2.5 ms is used and auxiliary windows are time-shifted to the left and right to minimize the cumulated distance of
all sample pairs to the bisector. The positioning of analysis frames to the left and right of the main analysis window avoids comparing signal fragments that do not belong to the same phonetic segment because the minimum distance is retained as a measure of vocal noise [2, 6].

Before the calculation of the individual and cumulated distances, the within-window signal fragments are energy-normalized and their averages are removed. Energy-normalization enables compensating slow amplitude variations and average-normalization enables removing offsets. Without energy- and average-normalization sample pairs would be aligned on a straight line with a slope different from one and displaced from the origin.

An algebraic formulation of the procedure outline above shows that it is equivalent to the calculation of the variogram of the speech signal involving a current and left- and right-positioned analysis frames. The variogram is minimal for the shift of the auxiliary analysis window that minimizes the cumulated distances to the bisector [5].

To obtain vocal dysperiodicity estimates over a complete signal, the main window is shifted without overlap or gap and the variogram analysis is repeated as often as necessary.

The global ratio involves the log-ratio of the signal and dysperiodicity energies over the whole signal duration. The segmental ratio involves the average of the log-ratio (1) computed for analysis segments of 5 ms. The latter is frequently used to summarize signal degradation owing to lossy coding. The reason is that it is expected to correlate better with perceived loss of signal quality than the global log-ratio [4].

C. Computation of the speech sample salience

The sample salience is defined as the longest interval over which a sample is a maximum. The estimation of the salience consists in considering all possible within-array analysis intervals and noting how often a sample is a maximum within each. Boundary effects are taken into account by rotating N times the samples within an array of length N so that each sample occupies once the left and right boundary positions.

Table 1: Illustration of a multi-resolution salience computation of an array (in bold).

| 1 2 0 4 3 6 1 2 1 1 2 0 4 3 6 1 2 | 1 3 1 5 1 9 1 3 1 |
| 2 1 4 1 9 1 4 2 1 3 |
| 2 1 9 1 4 1 1 4 1 |
| 1 9 1 3 2 1 4 1 4 |
| 9 1 2 1 1 3 1 4 1 |
| 1 6 2 1 3 1 8 1 9 |
| 5 1 1 2 1 7 1 9 1 |
| 2 1 4 1 6 1 9 1 2 |

The calculation of the sample salience involves the following steps. The handling of boundary effects is discussed later.

1. Initialization of all sample saliences to one.
2. Division of the array length N into analysis intervals of length 2. The rightmost interval stops at the rightmost array boundary whatever its length (i.e. 1 or 2).
3. Determination of the maximum within each interval.
4. Assignment of a salience of 2 to the interval maxima.
5. Increase of the interval length by one.
6. Division of the array length N into analysis intervals of length n. The length of the rightmost analysis interval is comprised between 1 and n.
7. Determination of the interval maxima.
8. Assignment of salience n to each interval maximum.
9. Looping back to step 5.
10. Stop when the analysis interval length equals N.

The position of the analysis array within the signal may be arbitrary and the saliences of the samples in the rightmost interval are affected by the anomalous interval lengths. To obtain sample saliences that are less dependent on position, the N samples in the analysis array are rotated N times so that each sample is positioned once at the right and left boundaries, and the sample salience is calculated for each within-array rotation. The final sample salience is the average of the saliences computed for each rotation.

In practice, rotation is carried out by copying the analysis array to the right and shifting the array stepwise from left to right N times. Tab.1 illustrates obtaining the sample salience for an array of length 9. Each line in Tab.1 gives the sample saliences for one array position. The last line gives the final average saliences, which are considered to be independent of the sample positions with regard to the array boundaries.

D. Extraction of the vocal cycle lengths and amplitudes

Preprocessing: The speech signal is low-pass filtered to remove additive noise as well as high-frequency formants. A zero phase filter is used to prevent phase distortion. The cut-off frequency is 900Hz.

Multi-resolution analysis: The cycle positions are determined on the base of the main cycle peaks that occur in the vicinity of glottal closure. These are extracted by computing the salience of each signal sample and discarding those samples that are not peaks.

The main cycle peak sequence is extracted by taking into account the peaks one by one in the order of decreasing salience. For each peak sequence the coefficient of variation of the inter-peak durations is computed. The peak sequence giving rise to a minimal coefficient of variation is retained. The search interval for the minimum is fixed by the frequency band 50Hz to 400Hz in which the average vocal frequency is expected.

Salience analysis is performed twice, once for each polarity of the signal and the polarity giving the smallest coefficient of variation is retained.

E. Corpus

The corpus comprises sustained vowels [a] [i] and [u], as well as four sentences spoken by 22 normophonic and dysphonic speakers. Two of the sentences involve voiced segments exclusively and the other voiced and unvoiced segments. The four sentences are matched grammatically and have the same number of syllables. Seven judges have determined the degree of perceived overall deviation from modal voice (i.e. grade) in the framework of a compared-items paradigm [3].

III. RESULTS AND DISCUSSION

A. Vocal noise

Vocal noise has been extracted by means of the algorithm described in section II.A. Global and segmental signal-to-dysperiodicity ratios have been computed and correlated with perceived degrees of hoarseness (grade). Tab.2 summarizes the Pearson correlation coefficients between perceived degree of hoarseness and global and segmental signal-to-dysperiodicity ratios.

Table 2: Pearson’s correlation coefficients between average hoarseness scores and global and segmental signal-to-dysperiodicity ratios for sustained vowel [a] and sentences S1-S4 obtained via energy-equalized (GV) and energy- and average-equalized variograms (AGV)

<table>
<thead>
<tr>
<th></th>
<th>[a]</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GV</td>
<td>-0.73</td>
<td>-0.71</td>
<td>-0.68</td>
<td>-0.70</td>
<td>-0.69</td>
</tr>
<tr>
<td>AGV</td>
<td>-0.70</td>
<td>-0.72</td>
<td>-0.78</td>
<td>-0.87</td>
<td>-0.79</td>
</tr>
</tbody>
</table>

The results show that, for sustained vowels as well as spoken sentences, the global and segmental signal-to-dysperiodicity ratios correlate with the perceptual ratings. One observes that when energies as well as averages of the signal analysis frames are equalized, the correlation with perceived degree of hoarseness is increased. The increase is more marked for the global signal-to-dysperiodicity ratio. An explanation for this observation is discussed hereafter.

In the speech signal one occasionally observes large-amplitude, low-frequency “pop” noise caused by breath hitting the microphone housing. These parasitic transients are low-frequency and ignored or not perceived by human listeners. The energy of such low-frequency transients may be comparable to the total signal energy, however. The impact of such parasitic low-frequency pops is greater on the global signal-to-dysperiodicity ratio than on the segmental one because the latter dilutes the effects of isolated events by averaging over several segments. A consequence is that segmental signal-to-dysperiodicity ratios correlate better than global ratios with perceived hoarseness.

Average-equalizing the analysis windows removes most of the effects of low-frequency pop noise [2, 6]. A consequence is an increase of the correlation with perceived hoarseness for both segmental and global signal-to-dysperiodicity ratios. The increase is more marked for global than segmental signal-to-dysperiodicity
ratios because the former is more strongly affected by isolated large-amplitude events.

Fig. 2 is a scattergram that shows on the horizontal axis perceptual scores of hoarseness for sentence S3 and on the vertical axis the global signal-to-dysperiodicity ratios, computed by means of the average-equalized and unequalized variograms. Generally speaking, the effect of equalizing frame averages in addition to frame energies is to improve the linearity between perceptual and acoustic cues and to increase the Pearson correlation coefficient, which is a measure of linear correspondence.

One sees that the difference between the two analysis methods increases with the signal-to-dysperiodicity ratio. This is because when frame averages are not equalized the influence of low-frequency pop noise on the global signal-to-dysperiodicity ratio is stronger in clean signals.

![Figure 2: Global signal-to-dysperiodicity ratio (vertical axis) versus perceptual scores (horizontal axis) and linear regression lines for sentence S3. Increasing scores to the right on the horizontal axis correspond to increasing scores of perceived hoarseness, that is, decreasing signal-to-dysperiodicity ratios. The black and white dots correspond to global signal-to-dysperiodicity ratios obtained via average- & energy-equalized and energy-equalized variograms respectively.](image)

B. Cycle duration jitter

Fig.3 illustrates the extraction of cycle lengths via the analysis of peak saliences (sections II.C, II.D). The upper trace is the unfiltered speech signal, i.e. a fragment of vowel [a] sustained by a female hoarse speaker (the degree of hoarseness is 15 on a scale from 1 to 21). The voice is perceived as breathy rather than rough. The second graph shows the peak saliences of the low-pass filtered signal fragment and the bottom graph shows the cycle lengths extracted on the base of the cycle peak saliences and the inter-peak durations.

![Figure 3: Fragment of vowel [a], peak saliences and cycle lengths.](image)

V. ACKNOWLEDGMENT

This research was supported by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme and by COST ACTION 2003 “Advanced voice function assessment”.

REFERENCES