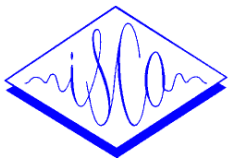

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Determination of Glottal Excitation Cycles for Voice Quality Analysis

Wolfgang J. Hess

Institute of Communications Research and Phonetics, University of Bonn
Poppelsdorfer Allee 47, D−53115 Bonn, Germany
wgh@uni−bonn.de

Abstract. The three-channel pitch determination algorithm (PDA) presented in this paper combines a short-term analysis PDA, which derives fundamental frequency via a periodicity criterion, and two time-domain PDAs that determine the instants of glottal closure according to local signal criteria. The first of these algorithms correlates the speech signal with an estimate of the impulse response of the vocal tract; the second one applies a neural network. The reliability of these time-domain PDAs is increased by constraints on the range of F_0 imposed by the short-term analysis PDA. First results show that the algorithm can be applied for both accurate pitch period determination of running speech and voice quality measurements, particularly the measurement of voice jitter.

1. Introduction

For measurement of jitter and shimmer in voice quality analysis it is necessary to have a pitch determination algorithm (PDA) which is able to track individual laryngeal excitation cycles, i.e., to determine the instant of glottal closure (IGC) or a similarly prominent point within the waveforms of voiced speech.

Usually jitter and shimmer are determined from sustained vowels [a] or [e] recorded in a controlled environment without signal distortions (Titze et al., 1987). Processing such signals is not a severe problem for the PDA as long as the utterance under investigation is fairly regular, i.e., the jitter (and/or shimmer) does not exceed a few percent. However, if the PDA relies upon a periodicity criterion (as most PDAs do at least implicitly), the method breaks down when the voice gets rougher, and measurements on highly irregular voices become unreliable.

In this paper, a three-channel PDA is presented that was designed to make this problem somewhat easier. It applies a combination of a short-term PDA and two time-domain PDAs that detect discontinuities in the signal associated with the point of glottal closure or maximum excitation. Section 2 will describe the individual components of the algorithm; Sect. 3 will discuss interactions between them, and Sect. 4 will present some preliminary results.

Besides voice quality measurement there has been considerable interest in high-precision pitch period determination from other domains of speech processing. For instance, template-based text-to-speech synthesis systems with time-domain processing of the units require the facility of modifying prosodic parameters in natural speech signals; this is usually achieved by one of the well-known PSOLA algorithms. PSOLA needs precisely determined pitch period delimiters that correspond well to the IGC; any error or inaccuracy in these markers will later be audible in the synthetic signal. To support such an application as well, the PDA presented here was not only designed for voice quality measurements, but also to process ordinary good-quality speech signals.

2. The Algorithm

2.1 Rationale for a Multi-Channel Algorithm

Pitch determination algorithms can be crudely categorized into the two categories *short-term analysis* PDAs and *time-domain* PDAs (cf. Hess, 1992). Most time-domain PDAs are able to track the speech signal period by period; not very many time-domain PDAs, however, are able to lock onto the speech signal in such a way that they determine the IGC. Short-term analysis PDAs, on the other hand, perform a transformation on a short frame of speech (typically 25 to 50 ms) which comprises several pitch periods (at least 2) and yields a single, global estimate of fundamental frequency F_0 or fundamental period T_0. Such algorithm is not able to determine individual pitch periods; at the same time, however, short-term analysis PDAs, just because they apply a global criterion for periodicity, are more robust than time-domain PDAs and less sensitive to local distortions which may lead a time-domain PDA into temporary failure. Most PDAs recognize a temporary breakdown; it is their common problem, however, not to be able to decide whether this error is due to the procedure, i.e., a "home-made" failure of the measurement, or due to the signal; i.e. that a signal is being analyzed which is voiced but irregular or aperiodic.

Voice quality measurements require PDAs that precisely determine individual periods and lock onto a certain instant in the waveform, preferably the IGC. Hence a time-domain PDA is mandatory with the additional constraint that it should determine each IGC independently and should depend as little as possible on the periodicity of the glottal pulses. The performance of this PDA, however, can be greatly enhanced when it is supported by a short-term analysis PDA that acts as a supervisor and controls its performance. As a by-product, most short-term PDAs supply a confidence criterion which tells us how "periodic" the signal is. When analyzing ordinary speech this criterion may serve as an indicator for the quality of the measurement or even for a voiced-unvoiced decision. In voice quality analysis, when the utterance is known to consist of sustained vowels, this criterion can be interpreted differently; it gives us − besides the estimate of the fundamental period T_0 − an es-

timate how far the voice under investigation is compatible with the requirement of periodicity. A time-domain PDA alone usually cannot provide such a measure.

Another weak point of most time-domain PDAs is that they get unreliable when they have to cope with a wide F_0 range. In this respect the ones applied here (Sects 2.3,4) are no exceptions. As there are the estimates of the short-term PDA available, however, we can pass these as initialization parameters to the time-domain PDAs in order to restrict their range. In addition the short-term PDA allows us to loosen or tighten the constraints with respect to range and regularity of the estimates of the time-domain PDAs according to the degree of periodicity derived from the short-term analysis quality measure. The two time-domain PDAs for the present implementation have been selected from the literature according to ease of implementation, precision of measurement, and applicability to the analysis of running speech signals.

2.2 The Short-Term Analysis PDA (Indefrey et al., 1985)

This algorithm applies the principle of a double spectral transform with nonlinear frequency-domain processing. This PDA is thus related to the well-known cepstrum and autocorrelation PDAs. A signal frame (typically 30 to 50 ms according to the F_0 range under investigation) is first weighted (e.g., by a Hamming window) and then transformed into the frequency domain by a discrete (fast) Fourier transform (DFT). The spectrum is then distorted by a (memoryless) nonlinear function and transformed by an inverse DFT into a lag-domain representation from which the estimate for the fundamental period $T_0 = 1 / F_0$ is derived.

The frequency-domain nonlinear functions optionally applied in this PDA are 1) squared magnitude (which leads to the autocorrelation function in the lag domain); 2) the logarithm of the power spectrum (which yields the cepstrum in the lag domain); 3) magnitude, i.e., amplitude spectrum; and 4) fourth root of the power spectrum. All these nonlinear functions discard the phase information in the spectrum thus setting the phases of all components to zero. In very simple words, according to the linear source-filter model, speech signals result from the convolution of a pulse train with periodicity T_0 and the (aperiodic) impulse response of the supraglottal system. The pulse train survives the double spectral transform and the frequency-domain nonlinear distortion; with all phases set to zero, however, the first pulse in the lag-domain representation (cepstrum or autocorrelation function or inverse transform of one of the other spectral representations) is fixed to the lag $d=0$, which means that the second impulse has to appear at $d=T_0$. Figure 1 shows an example.

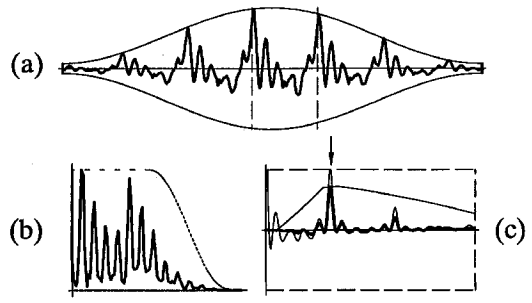

Fig. 1. The short-term analysis PDA – example of performance. (**a**) Signal frame (sustained [a], female speaker; 32 ms) after weighting, window function, one arbitrary period; (**b**) magnitude spectrum (0-2 kHz); (**c**) lag-domain representation (0-12 ms) before and after weighting. Nonlinear function: 4th root of power spectrum

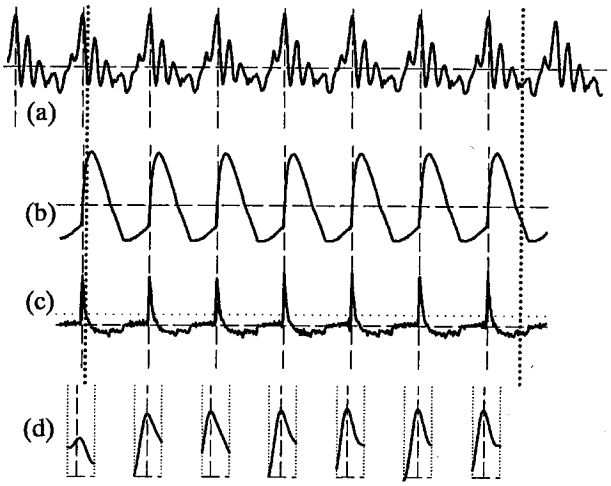

Fig. 2. Reference PDA using the laryngogram (Hess and Indefrey, 1987): example of performance. (**a**) Signal frame (45 ms; same utterance as in Fig. 1); (**b**) laryngogram; (**c**) differenced laryngogram using a first-order differentiator; (**d**) fine positioning of pitch period markers at an increased sampling rate (128 kHz) by local interpolation [time scale increased by a factor of 8]

2.3 Reference Signal – The Laryngogram

The reference contour for the training and the evaluation of the PDA was generated using the output signal of a laryngograph and the algorithm by Hess and Indefrey (1987) to determine the pitch period markers from the laryngogram (Fig. 2). The laryngogram is differenced by a first-order differentiator filter, and the markers are derived from the maxima of this waveform. The marker positions are then refined by interpolation. Within a small interval around each marker the laryngogram is upsampled from 16 to 128 kHz and differentiated. The two steps are simultaneously performed using a combined differentiator-interpolator filter. Since only few samples have to be computed, the computational effort is negligible. The sampling rate increase to 128 kHz yields a measurement accuracy of about 8 µs (or better than 0.5 % at F_0=500 Hz). For jitter measurements at high fundamental frequencies this may not be sufficient; yet it is easy to implement a filter which enables us to realize a higher upsampling factor.

2.4 Time-domain PDA # 1 – Correlation-Based IGC Determination (Cheng and O'Shaughnessy, 1989)

This PDA applies a modified maximum-likelihood principle that gives the maximum likely instant of the beginning of an impulse response of the vocal tract to a glottal pulse. Figure 3 shows an example.

In the source-filter approach the speech signal is the response of the supraglottal system to the pulse train generated by the source, and a pitch period can be regarded as the beginning of the impulse response of the vocal tract. If

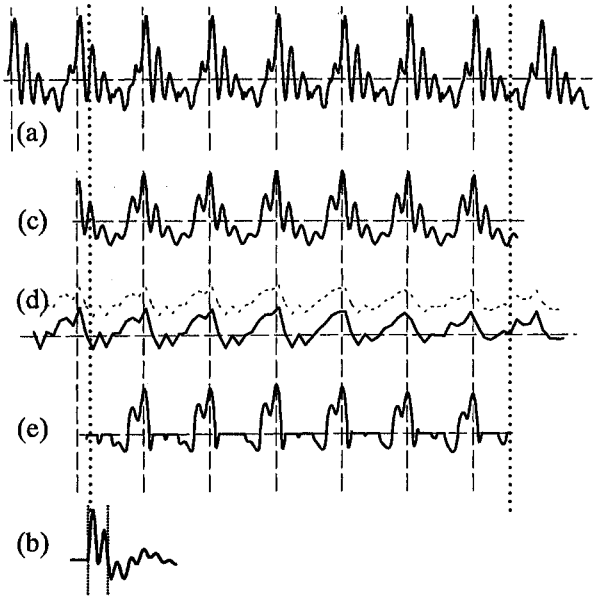

Fig. 3. Determining the IGC via a maximum-likelihood criterion (Cheng and O'Shaughnessy, 1989): example of performance. (**a**) Signal frame (45 ms, as in Figs. 1, 2); (**b**) impulse response $h(n)$ of the vocal tract, estimated by linear prediction; (**c**) correlation function $c(n)$; (**d**) envelope $e(n)$ [dotted line] and high-pass filtered envelope; (**e**) product of $c(n)$ and the filtered envelope

we now compute the correlation function $c(n)$ between the speech signal $s(n)$ and the beginning of the pertinent impulse response $h(n)$ of the vocal tract,

$$c(n) = \sum_q s(n+q)\ h(q)\ ,$$

this function will become a maximum at those instants n which corresponds to an IGC. The impulse response $h(n)$ is estimated via linear prediction.

The principal peaks of the running correlation function $c(n)$ synchronize well with the individual IGCs. Strong formants, however, also show up in $c(n)$. Further processing is thus required to separate the true IGCs from these false alarms. Cheng and O'Shaughnessy suggest to calculate the envelope of $c(n)$, send it through a high-pass filter (with a cutoff frequency around the low end of the F_0 range) and a subsequent half-way rectifier in order to enhance the leading amplitudes at the beginning of each pitch period, and to multiply $c(n)$ by the resulting waveform $d(n)$. The envelope

$$e(n) = \sqrt{[c(n)]^2 + [c_H(n)]^2}\ ,$$

where $c_H(n)$ is the Hilbert transform of $c(n)$, is straightforwardly calculated using a digital Hilbert filter.

This approach has one drawback which does not occur with sustained open vowels but may render this PDA useless when continuous speech is processed. This problem is known from all PDAs that apply some sort of inverse filtering (Hess, 1992). When the formant $F1$ coincides with F_0, the signal will be almost sinusoidal since most of its energy is concentrated in the F_0-$F1$ region. In this case the envelope $e(n)$ becomes a constant which is then suppressed by the high-pass filter, and the waveform $d(n)$ is either zero or fluctuates at random around zero instead of displaying the envelope of a damped formant oscillation, as it is the case when several periods of the $F1$ waveform fit into one pitch period. A way out of this problem is to partly bridge the high-pass filter so that the constant component of the envelope is not completely removed. Of course this weakens the effect of enhancing the leading amplitudes at the beginning of a pitch period. On the other hand, if the short-term analysis enters some reliable estimate of T_0, rather stringent correction routines can remove the unwanted peaks and at the same time preserve correct processing when $F1$ and F_0 take on the same value.

In our implementation the whole calculation is done at a reduced sampling rate of 8 kHz. A 9th-order LP filter supplies the estimate of the impulse response $h(n)$. The calculation of $c(n)$ extends over the first 1.5 to 2 ms of $h(n)$ depending on the measuring range of F_0. The pitch period markers are derived from the product function $c(n)\cdot d(n)$ which is, however, not precise enough to yield the exact marker positions. The correlation function $c(n)$ is thus locally upsampled to 128 kHz, and the final positions of the markers are derived from there.

2.5 Time-domain PDA # 2 – Neural-Network IGC Determination (Howard and Walliker, 1989)

This PDA was orinially designed for a speech-procesing hearing aid. In the original version it applies a four-layer artificial neural network which operates on the rectified output signal of a six-channel filter bank that divides the speech signal into subbands. The network (41 input neurons for each subband, 2 hidden layers with 10 neurons each, one output neuron; layers fully connected) is trained

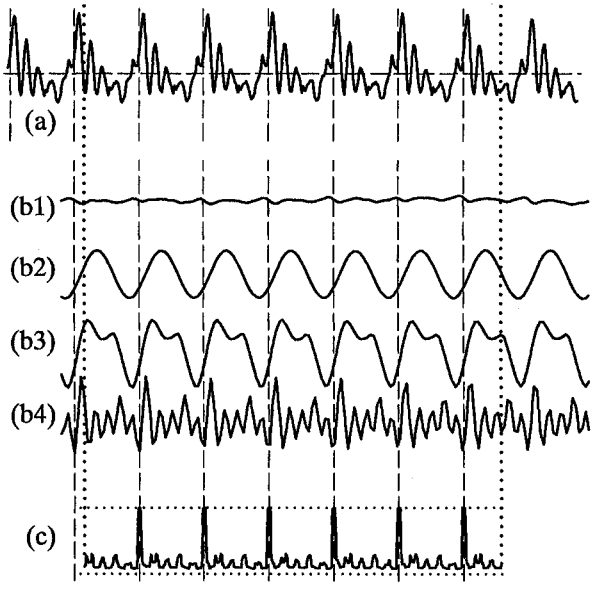

Fig. 4. Determining the IGC via a neural network (Howard and Walliker, 1989): example of performance. (**a**) Signal frame (same as in the other figures); (**b1-4**) subband signals; (**c**) output signal of the neural network. The subband signals were obtained by linear-phase bandpass filters including a Hilbert transform; this makes the design of the filters easier and is advantageous for the network since the subband signals start from a zero crossing at the IGC and are thus less sensitive to amplitude variations

on the differentiated waveform of a laryngograph as reference output (cf. Sect. 2.3).

As Howard and Walliker's approach operates with a sampling frequency of 4 kHz, the 6-times-41-neuron input structure covers a frame of more than 10 ms and relies thus on periodicity, at least implicitly, when T_0 is shorter than 10 ms. Since in voice analysis the PDA must also be suited for processing irregular waveforms, the present approach uses an arrangement of the input neurons that results in a shorter frame. The 85 input neurons are distributed over the unprocessed speech signal (20 neurons, i.e., 1.6 ms) and four subband signals (400-800 Hz with 20 neurons over 1.6 ms; 200–400 Hz with 20 neurons over 3.2 ms; 100-200 Hz, 15 neurons, 4.8 ms; 50-100 Hz, 10 neurons, 6.4 ms). This structure guarantees that every IGC is estimated by itself without reference to some periodicity criterion. Figure 4 shows an example. The network computes the output activation with a sampling rate of 8 kHz. When a peak is detected, the sampling rate is momentarily increased to 128 kHz. Since the neural network is nonlinear, it is not possible to just interpolate the output function; we rather have to propagate interpolated input samples through the neural network.

Like Howard and Walliker's design the PDA is trained on the differentiated laryngogram (after some scaling) of sustained [a]'s and [e]'s as input signals. One problem in this respect is the question of the distance between speaker and microphone during the simultaneous speech and laryngograph recordings. As the laryngograph takes its signal directly from the speaker's larynx this signal is not subject to the delay caused by the acoustic wave propagation from the speaker's mouth to the microphone. At a sampling rate of 16 kHz this makes about one sample per 2 cm. As the response of the neural network is rather sharp (a decrease in the output from 0.9 to 0.5 within one sample is rather frequent), inconsistent training data due to inconsistent delays between speech signal and laryngogram can be detrimental for patterns with high output activation at the IGCs. For low-output patterns which come from anywhere within a pitch period the situation is of course much less critical.

3. Interactions between the PDAs

Both time-domain PDAs in this implementation have been designed to determine the IGC without respect to periodicity. They are thus rather sensitive to strong formants, and a number of correction routines have been implemented that rely on the short-term analysis PDA. First of all, the information that a certain frame shows strong periodicity and that it is part of a longer regular segment, comes from the short-term analysis PDA and its built-in backtracking routine which identifies signal segments with coherent pitch estimates. To avoid bad starting conditions, the short-term analysis PDA moves the entry point for the time-domain PDAs away from the onset of voicing into the middle of a segment which is found periodic. From there the time-domain PDAs proceed with and against time until voicing ceases. Furthermore, if the signal is periodic, the short-term analysis PDA puts constraints on the momentary range within which the individual period lengths are permitted to vary. The more regular the signal is found to be, the more stringent constraints can be imposed. These are realized in form of correction routines that exclude markers with smaller peaks of the underlying decision functions (see Sects. 2.4 and 2.5), but also recheck these functions when no marker has been found in an interval where there should be one.

4. Some Preliminary Results

The algorithm was tested with a set of three sustained vowels [a], [ɛ], [e] from two speakers (one female, one male) where the laryngogram was also available. For both the neural network and the correlation-based PDA about 5 % of the markers were misplaced by more than 5 samples with respect to the laryngogram. For the correlation-based PDA the position of the GCI estimate was slightly dependent (2 to 3 samples) on the vowel. For the female speaker where no high-activation patterns had been included in the training set, the neural network exhibited a tendency toward bifurcation of the principal peak. Further training is necessary, particularly with such input patterns that produce large differences between desired and actual network output. As for jitter determination, the estimates provided by the time-domain PDAs were consistently at least 20 % above those computed from the laryngogram (for instance for [a], male speaker, 0.34 % for laryngogram versus 0.44 % for the PDAs; for [a], female speaker, 0.33 % for laryngogram and 0.59 % for the PDAs). The PDAs thus introduce some noisiness into the measurement which seems almost impossible to be removed. For some utterances of continuous speech which were also tested about 8 % of the markers were off the reference positions by more than 5 samples. The computational effort for the PDAs is about the same for the short-term analysis PDA (two FFTs per frame and lag-domain peak search) and the correlation-based PDA whereas the neural network is slower by a factor of 20. The errors committed by the time-domain PDAs were mostly non-overlapping so that future interactions between the time-domain PDAs will bring further improvement.

More detail work is necessary to optimize this combination of short-term analysis PDA and time-domain PDAs to give good and reliable estimates of the IGCs in running speech. As the first results show, however, this multi-channel PDA will help to improve the measurement of the instants of glottal closure in running speech.

References

Cheng Y.M., O'Shaughnessy D. (1989): "Automatic and reliable estimation of glottal closure instant and period." IEEE Trans. ASSP-37, 1805-1815

Hess, W. J. (1992): "Pitch and voicing determination." In Advances in speech signal processing, chapter 1, ed. by S. Furui and M. M. Sondhi (M. Dekker, New York), 3-48

Hess, W. J., Indefrey H. (1987): "Accurate time-domain pitch determination of speech signals by means of a laryngograph." Speech Commun. 6, 55-68

Howard I.S., Walliker J.R. (1989): "The implementation of a portable real-time multilayer-perceptron speech fundamental period estimator." Proc. EUROSPEECH-89, Paris, 206-209

Indefrey H., Hess, W. J., Seeser G. (1985): "Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain." Proc. IEEE ICASSP-85, vol. 2, paper 11.12

Titze I., Horii Y., Scherer R. (1987): "Some technical considerations in voice perturbation measurements." J. Speech Hear. Res. *30*, 252-260