Effective Implementation of Short-Term Analysis Pitch Determination Algorithms

W.J. Hess
Munich, FRG

1. Introduction

The problem of pitch determination counts among the most delicate problems in speech analysis. A multitude of pitch determination algorithms (PDAs) and devices have been developed; none of them works perfectly (Rabiner et al., 1976). A survey of the state of the art was presented in an earlier paper (Hess, 1982). In this survey the PDAs have been categorized into two gross categories: 1) time-domain PDAs, and 2) short-term analysis PDAs. The time-domain PDAs determine pitch (this term stands for fundamental period, fundamental frequency, or the elapsed time between two consecutive pulses of the voice source) directly from the signal as the elapsed time between consecutive laryngeal pulses. The short-term analysis PDAs, after subdividing the signal into a series of frames, leave the time-domain by a short-term transformation in favor of some spectral domain whose independent variable can be frequency or again time (in the latter case the independent spectral variable is called *lag* in order to avoid confusion).

The short-term analysis PDAs are further categorized according to the short-term transform they apply (Fig. 1). The main possibilities are *correlation, 'anticorrelation'* (i.e., the use of distance functions), *multiple spectral transform* (cepstrum), *harmonic analysis* (frequency-domain PDAs), and *maximum-likelihood* analysis. In the following we will deal only with three examples: 1) autocorrelation (pertaining to the the correlation PDAs), 2) maximum-likelihood, and 3) harmonic analysis.

2. Basic Computational Effort, Spectral Representation and Measurement Accuracy

In general the short-term analysis algorithms perform a short-term transformation of the form

$$X = W x. \tag{1}$$

In this equation, X is the spectral vector, x is the signal vector and W is the transformation matrix which represents the properties of the short-term transformation. For a frame of N samples (the transformation interval) the

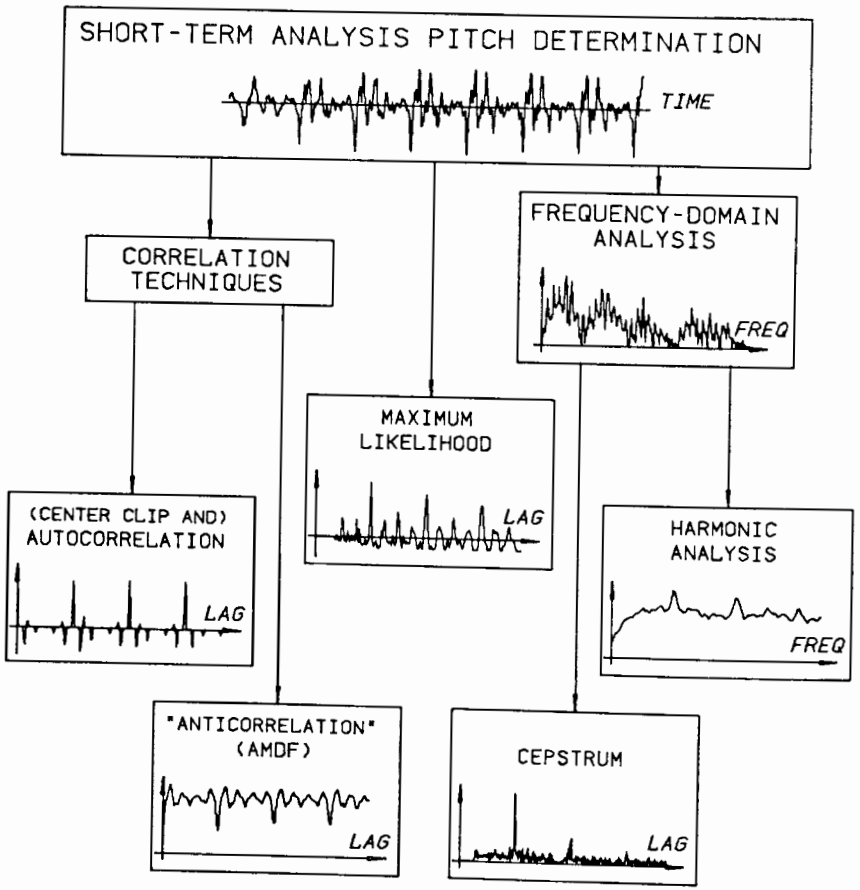

Fig. 1. Grouping of short-term analysis PDAs. Signal: part of the sustained vowel /ɛ/, male speaker, undistorted recording.

basic computational complexity of the short-term transformation is in the order of N^2 if the number of multiplications serves as the basic reference.

In principle the computational complexity given in (1) is valid for all three types of algorithm we are going to deal with. The discrete Fourier transform which is applied in frequency-domain PDAs and (not necessarily but possibly) in autocorrelation PDAs, clearly follows (1). Of course one always tries to use the fast Fourier transform (FFT) whose basic complexity is in the order of N IdN, where Id represents the dual logarithm. If the autocorrelation function is directly evaluated (without using the Fourier transform), it can also be brought into the form (1). The maximum-likelihood PDA involves squaring operations with the same basic complexity.

The question is now how to decrease the computational load when implementing these algorithms. With very complex hardware on-line performance can be obtained even if the algorithms have not been optimized. On the other hand, if the same results can be achieved with reduced effort, it is always worthwhile to think about such implementations. This shall be done in the following, first from a rather global point of view, later in more detail for the three algorithms cited.

The following actions to reduce the computing effort appear possible.

1. Replace multiplications and divisions by additions or table lookups;
2. Replace multiplications by logical operations due to sophisticated preprocessing;
3. Decrease the sampling rate in order to reduce the length N of the transformation interval;
4. Discard redundancies and irrelevance before the spectral transformation, again in order to reduce N;
5. Confine the operating range of the calculation to samples that are actually needed; and
6. Adaptively change the frame length K depending on the current fundamental frequency F_0 and/or the actual value of the trial period p within the short-term transformation.

Actions 5 and 6 are not possible when a FFT is applied; there all spectral samples are computed simultaneously from a constant-length transformation interval.

With respect to the representation of the relevant information on pitch in the spectral domain, the spectrum (in the following the output signal of the short-term transformation will always be labeled 'spectrum' regardless of whether it represents a Fourier spectrum, an autocorrelation function, or a maximum-likelihood estimate) is heavily oversampled. If we limit the measuring range to 50-1000 Hz, then for the PDAs which operate in the lag domain (autocorrelation, maximum-likelihood) a sampling rate of 2 kHz in the spectrum would be sufficient in order to correctly represent the information on pitch, i.e. to satisfy the sampling theorem. For a frequency-domain PDA a spectral resolution of less than 25 Hz would be necessary in order to correctly represent all the harmonics of a signal at the lower end of the measuring range as separate peaks. These crude sampling rates, if applied, would be sufficient to *represent* the relevant information on pitch in the spectral domain, but they are not at all sufficient to *measure* pitch accurately enough. The most critical judge with respect to measurement accuracy is the human ear; data by Flanagan and Saslow (1958) as well as the prevailing theories on pitch perception (e.g., Terhardt, 1979) suggest that errors of less than 0.5% are still perceived. To satisfy this requirement, we would thus need a spectral resolution of less than 1 Hz for a frequency-domain PDA and a sampling rate of more than 50 kHz for a lag-domain PDA. Since all these PDAs involve nonlinear processing (usually squaring and averaging), it is not sure whether interpolation in the spectral domain *after* the nonlinear step will yield correct results. Hence, in order to obtain a reasonable spectral resolution, frequency-domain algorithms usually perform voluminous FFTs

on long transformation intervals (more than 200 ms) which consist of a short frame (30-40 ms) appended with zeros, and a number of autocorrelation PDAs compute the autocorrelation function with the full sampling rate of, say, 10 kHz although they compute it from a speech signal that has been low-pass filtered with a cutoff frequency of 1 kHz. So the basic problem is to find solutions that allow for the spectral resolution necessary for an accurate measurement, but cut down the computational effort as far as possible.

3. The Autocorrelation PDA

The breakthrough in the autocorrelation PDA came when Sondhi (1968) discovered that adaptive center clipping greatly improved the performance of this PDA which hitherto had suffered from a strong sensitivity to dominant formants. Dubnowski et al. (1976) then found that an adaptive three-level quantization did not significantly degrade the performance compared to the signal that was only center clipped. With this three-level quantization, however, it became possible to evaluate the ACF without any multiplications since the input signal of the so quantized signal can only take on values of +1, 0, and -1. It became even possible to replace the adder in the ACF evaluation logic by a simple up-down counter so that for this PDA the problem of computational complexity can be regarded as solved. Actions 4 and 6 from the above-mentioned list, which appear possible and promising, are no longer necessary under this aspect.

4. The Maximum Likelihood PDA

The maximum-likelihood PDA (Noll, 1970; Wise et al., 1976) emerged from the task to optimally separate a periodic component x(n) from Gaussian noise gn(n) in the signal a(n) = x(n) + gn(n) with the finite duration K. The mathematical formulation leads to a comb filter with the trial period p, and the best estimate of pitch is given when p optimally matches the harmonic structure; in this case the energy of the output signal of the comb filter is maximized. Computing the energy of this signal however, involves squaring, and the number of the pertinent multiplications is in the order of P^2_{max}, when P_{max} is the longest period possible within the measuring range. To reduce the computational effort, one can exploit the fact that the only multiplications needed are squaring operations, which can easily be implemented by a table-lookup procedure. Since the maximum-likelihood PDA is rather noise resistant, the input signal can be crudely quantized, and the table can be kept rather small. Another possibility of reducing the computational effort is obtained when one succeeds in replacing the squaring operations by other, less costly arithmic operations, such as the peak-to-peak amplitude of the output signal of the comb filter. This is indeed possible; taking the amplitude instead of the energy hardly affects the performance of the PDA (Ambikairajah et al., 1980).

5. Harmonic Analysis, Frequency-Domain PDAs

All frequency-domain PDAs (e.g. Schroeder, 1968; Martin, 1981) need a Fourier transform to enter the frequency domain. This is preferably done using the FFT although, for special applications, it might be profitable to use the conventional DFT and to compute only a few spectral samples (Duifhuis et al., 1982). In the following, the considerations will be confined to the case where the FFT is used. Let an arbitrary frequency-domain PDA need a spectral resolution of 5Hz. This usually meets the requirements with respect to accuracy since it is mostly possible to obtain the estimate of F_0 from a higher harmonic and thus to reduce the inaccuracy of the measurement due to quantization by the harmonic number of that harmonic. The easiest way to implement such a PDA is to take a segment of the input signal (30-40 ms), apply a suitable time-domain window, extend the segment by zeros, and apply the corresponding FFT to obtain the spectral resolution (a 2048-point FFT would be necessary when the sampling frequency of the input signal is 10 kHz). This procedure however, is in no way optimal with respect to the computational effort. Table I shows the different possibilities of optimization. The basic computational complexity of the FFT is in the order of N Id N (line 'no optimization' in Table I). Using sophisticated programming that avoids multiplication for such samples where the real and imaginary parts of the complex exponential values used in the transform are zero or have a magnitude of 1, the number of multiplications can be reduced by almost 30%. A reduction of 50% is achieved when one takes account of the fact that the input signal is real (and not complex). With programming alone one can thus save 64% of the multiplications.

Further reduction of the computing effort is only possible with additional digital filtering. First, the relevant information on pitch in the speech signal is contained in the frequency components below 2.5 kHz. In contrast to lag-domain PDAs the accuracy of the frequency-domain measurement is not influenced if the sampling frequency of the time-domain signal is reduced by a factor of 2. The computational complexity, however, is reduced by more than 50%.

A last possibility of optimization is given by interpolation in the frequency domain. There are two possibilities of increasing the spectral resolution: 1) appending many zeros to the input signal and applying a long FFT, or 2) appending few zeros to the input signal, applying the shortest FFT possible, and interpolate in the frequency domain using a digital filter with zero phase response (i.e. a linear-phase nonrecursive interpolation filter in a noncausal realization) until the required spectral resolution is achieved. These two possibilities are equivalent as long as the interpolation is performed on the complex spectrum, but they are rather different with respect to their computational complexity. Applying all these optimizations together, in our example, brings down the computational effort by as much as one order in magnitude.

Table I. Comparative evaluation of the effort necessary to compute an FFT spectrum for frequency-domain pitch determination under various aspects of possible algorithmic optimization. Assumed sampling rate: 10 kHz, frame rate: 100 Hz, required spectral resolution: 5 Hz

Optimizing operation	FFT Length	Number of multiplications			Saving (%)
		FFT	Other Operations	Total	
No optimization	2048	45056	–	45056	0
Optimized FFT Programming	2048	32776	–	32776	29
Exploit the fact that the input signal is real: perform spectral rotation, shift imaginary part of the signal against time, and decompose spectrum	1024	20480	2048	22528	50
All programming optimizations	1024	14344	2048	16392	64
Downsampling to 5 kHz	1024	20480	200	20680	56
Downsampling and programming optimizations	512	5942	1224	7166	84
Limit transformation interval to 51.2 ms and upsample spectrum by factor 4 in the frequency domain	512	9016	6144	15160	66
Time-domain downsampling, limitation of transformation interval, and frequency-domain upsampling	256	4096	3272	7368	83
All optimizations applied together	128	1032	3528	4660	89

6. Conclusions

A number of proposals to efficiently implement the short-term transformation in short-term analysis PDAs have been reviewed. The problem of computational effort arises from the fact that, for reasons of measurement accuracy, the spectral function (autocorrelation function, Fourier spectrum etc.) must be heavily oversampled. The proposals range from efficient preprocessing (combined center and peak clipping in an autocorrelation PDA), which avoids multiplications, to the use of signal amplitude instead of energy, and from the use of table-lookup procedures to the optimal combination of the FFT and interpolation by digital filters. If the PDA is carefully implemented, the gain in computing speed can be considerable.

References

Ambikairajah, E., Carey, M.J., Tattersall G. (1980). A method of estimating the pitch period of voiced speech. *Elektron. Lett.* **16**, 464-466.

Dubnowski, J.J., Schafer, R.W., Rabiner, L.R. (976). Real-time digital hardware pitch detector. *IEEE Trans. ASSP*-**24**, 2-8.

Duifhuis, H., Willems, L.F., Sluyter, R.L. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *J. Acoust. Soc. Am.* **71**, 1568-1580.

Flanagan, J.L., Saslow, M.G. (1958). Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am.* **30**, 435-442.

G-AE Subcommittee on Measurement Concepts (1967). What is the fast Fourier transform? *IEEE Trans.* AU-**15**, 45-55.

Hess, W.J. (1983). *Pitch determination of speech signals - algorithms and devices*. Springer, Berlin.

Martin, Ph. (1981). Détection de F_0 par intercorrelation avec une fonction peigne. In: *Actes, 12èmes Journées d'Etude sur la Parole*, Montréal, mai 1981. GALF, F-22301 Lannion.

Noll, A.M. (1970). Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In: *Symposium on Computer Processing in Communication*; ed. by the Microwave Institute; Vol. **19**, 779-797. University of Brooklyn Press, New York.

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, A. (1976). A comparative study of several pitch detection algorithms. *IEEE Trans.* ASSP-**24**, 399-413.

Schroeder, M.R. (1968). Period histogram and product spectrum: new methods for fundamental-frequency measurement. *J. Acoust. Soc. Am.* **43**, 829-834.

Sondhi, M.M. (1968). New methods of pitch extraction. *IEEE Transactions* AU-**16**, 262-266.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research* **1**, 155-182.

Wise, J.D., Caprio, J.R., Parks, T.W. (1976). Maximum likelihood pitch estimation. *IEEE Transactions* ASSP-**24**, 418-423.