PITCH DETERMINATION USING THE CEPSTRUM OF THE ONE-SIDED AUTOCORRELATION SEQUENCE

Climent Nadeu , Jordi Pascual and Javier Hernando

Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya
08034 Barcelona, Catalonia, Spain

ABSTRACT

A new cepstral function, the cepstrum of the one-sided autocorrelation sequence, is presented and applied to pitch determination of speech signals. It shows a performance in terms of pitch period errors significantly better than those of the usual cepstrum and autocorrelation with center clipping algorithms for both clean and noisy speech, due to its remarkable accuracy at non-stationary segments of speech signals and to its noise reduction capability.

1. INTRODUCTION

Pitch is an important parameter of speech signals. Accurate representation of the fundamental frequency and the voiced/unvoiced character of speech is required in many applications: coding, synthesis, speech and speaker recognition, speech training, aids to the handicapped,... However, despite the great amount of techniques developed so far for pitch determination of a speech signal, the problem is "... still far from a general solution" [1,p.521]. Certainly, there exist a number of successful pitch determination algorithms (PDA), as most of those included in the set of the so-called short-term PDAs [1], that obtain reliable and accurate results when the speech signal in the current frame appears as a fairly periodic waveform. However, they often fail for voiced or transitional speech segments showing a rather aperiodic waveform attributable to the type of sound, the type of voice, or a noise contamination.

Particularly, the number of pitch period errors increases significantly at transitions between a voiced sound and another sound that can be voiced, unvoiced or silence, and also in sounds that have a transitional nature. Although these errors can be at least partially corrected by postprocessing, is always preferable to employ an improved basic pitch extractor to reduce the ad-hoc decision making and to avoid the delay that are usually associated to the postprocessor.

One of these algorithms that show good performance for quasi-periodic signals is the cepstrum (CEP) algorithm [2]. However, its ability to separate the source signal (that conveys pitch information) from the vocal tract response fails wherever the speech frame cannot be contemplated as just the result of a linear convolution between both components, as occurs at transitions or non-stationary speech segments, or when the recorded speech signal includes additive noise.

This work was supported by the PRONTIC grant number 105/88

Another successful PDA, the one based on the autocorrelation of the speech signal [3] relies on the fact that the autocorrelation is a measure of similarity between different temporal segments of the signal. So in frames where the speech signal is quasi-periodic, it displays a prominent peak at the pitch period lag. Additionally, the autocorrelation function has a signal-to-(broad band)noise greater than the speech signal, so that the autocorrelation PDA is a fairly robust technique. Nevertheless, the autocorrelation waveform, as the signal waveform, is influenced by both the excitation signal and the vocal tract filter (formants). An useful method of flattening the speech spectrum and, consequently, removing the influence of formants on the autocorrelation consists of using adaptive center clipping [3]. However, center clipping fails to remove the formant structure whenever the signal amplitude largely changes within the frame, as occurs often at transitions. Furthermore, in that case, the similarity between adjacent pitch periods is substantially lost so the height of the autocorrelation peak at the pitch period may be strongly affected.

In this paper, we present a new cepstral function, the cepstrum of the one-sided autocorrelation (COSA), and we show its usefulness for pitch determination. In fact, we propose a cepstral PDA that starts from the autocorrelation sequence in lieu of the speech signal. As will be shown in the following sections, although the COSA pitch determination algorithm does not improve the performance of the ACC (autocorrelation with center clipping) and CEP algorithms in quasi-periodic speech frames, it significantly reduces their pitch period errors at transitional speech segments as well as in speech signals contaminated by noise.

2. THE COSA PITCH DETERMINATION ALGORITHM

2.1 The cepstrum of the one-sided autocorrelation

From the autocorrelation sequence R(n) we way define the one-sided (causal part of the) autocorrelation sequence

$$R^+(n) = \begin{cases} R(n) & n>0 \\ R(0)/2 & n=0 \\ 0 & n<0 \end{cases} \qquad (1)$$

which verifies

$$R^+(n) + R^+(-n) = R(n) , \quad -\infty \le n \le \infty \qquad (2)$$

CH2977-7/91/0000-3677 $1.00 © 1991 IEEE

Its Fourier transform is the complex spectrum

$$S^+(\omega) = \frac{1}{2}\left[S(\omega) + jS_H(\omega)\right] \quad (3)$$

where $S(\omega)$ is the spectrum, i.e. the Fourier transform of $R(n)$, and $S_H(\omega)$ is the Hilbert transform of $S(\omega)$. Due to the analogy between $S^+(\omega)$ in (3) and the analytic signal used in amplitude modulation, a spectral "envelope" $E(\omega)$ [4] can be defined as

$$E(\omega) = |S^+(\omega)| \quad (4)$$

This envelope characteristic, along with the high dynamic range of voiced speech spectra, originate that $E(\omega)$ strongly enhances the highest power frequency bands, i.e. the first formant. Thus, both the higher formants and the noise components lying outside the enhanced frequency band are largely attenuated in $E(\omega)$ with respect to $S(\omega)$; and $E(n)$, the inverse Fourier transform of $E(\omega)$, is less affected by them than $R(n)$.

Additionally, we may define the complex cepstrum of the one-sided autocorrelation (COSA) $C^+(n)$ as the inverse Fourier transform of $\log S^+(\omega)$, i.e. [5]

$$C^+(n) = \mathrm{IFT}\left\{\log S^+(\omega)\right\} \quad (5)$$

Since $R^+(n)$ is always a minimum-phase sequence, its complex cepstrum $C^+(n)$ is also causal and verifies the following relationship [6]

$$\begin{aligned} C^+(n) &= 2C(n), \quad n>0 \\ C^+(0) &= C(0) \end{aligned} \quad (6)$$

where the even sequence $C(n)$ is the (real) cepstrum of $R^+(n)$, i.e.

$$C(n) = \mathrm{IFT}\left\{\log E(\omega)\right\} \quad (7)$$

Due to the causality of both $R^+(n)$ and $C^+(n)$, the following recursive relationship between both sequences can be derived [6]

$$\begin{aligned} C^+(n) &= \frac{1}{R^+(0)}\left[R^+(n) - \sum_{k=1}^{n-1}\frac{k}{n}C^+(k)\,R^+(n-k)\right], \quad n>0 \\ C^+(0) &= \log R^+(0) \end{aligned} \quad (8)$$

which gives a way of obtaining the COSA sequence from the autocorrelation sequence, avoiding the frequency domain and the logarithm.

Let us assume now that the speech signal $x(n)$ whose autocorrelation is $R(n)$ is given by the linear convolution

$$x(n) = h(n) * e(n) \quad (9)$$

where $h(n)$ is the impulse response of a Mth-order all-pole filter driven by $e(n)$ and $e(n)$ is assumed to be a train of impulses for voiced sounds and white noise for unvoiced sounds.

Denoting with $H(z) = 1/A(z)$ the z-transform of $h(n)$ and $S_e(\omega)$ the power spectrum of $e(n)$, it follows that

$$S(\omega) = \frac{S_e(\omega)}{|A(\omega)|^2} \quad (10)$$

It is well known that the causal sequence $R^+(n)$ has the same poles than the signal [7]. Then, denoting with $B(\omega)$ the Fourier transform of the driving function that obtains $R^+(n)$ at the output of the filter $H(z)$, it follows that

$$S(\omega) = S^+(\omega) + (S^+(\omega))^* = \frac{B(\omega)}{A(\omega)} + \frac{B^*(\omega)}{A^*(\omega)} =$$

$$= \frac{B(\omega)\,A^*(\omega) + B^*(\omega)\,A(\omega)}{|A(\omega)|^2} \quad (11)$$

and, from the identification of (10) and (11), it results that

$$S_e(\omega) = B(\omega)\,A^*(\omega) + B^*(\omega)\,A(\omega) \quad (12)$$

so that $B(\omega)$ depends on both $S_e(\omega)$ and $A(\omega)$.

Thus, the cepstrum of $R^+(n)$, does not actually perform a deconvolution between filter and excitation as does the cepstrum of the speech signal. In fact,

$$\log S^+(\omega) = \log B(\omega) - \log A(\omega) \quad (13)$$

and, even though the cepstral component due to $A(\omega)$ is mainly concentrated close to the origin, as happens with the signal cepstrum, the cepstral component due to $B(\omega)$ includes not only the excitation that conveys pitch information but also the speech formants. However, in spite of the cepstrum $C^+(n)$ of the one-sided autocorrelation (which is equivalent to the cepstrum $C(n)$ of $E(n)$ for the pitch determination purpose, due to relation (6)) does not perform a clear deconvolution, for voiced frames it will be able to show a more outstanding peak than the sequence $E(n)$ at the lag corresponding to the pitch period, since the influence of the filter is partially eliminated by the cepstral transformation. From another viewpoint, the logarithm greatly flattens the strong first formant of $E(\omega)$.

2.2 A simple technique to improve the performance of the COSA algorithm

At this point, we should pay attention to a problem associated with the COSA sequence $C^+(n)$. As pointed out by expression (3), the imaginary part of its Fourier transform $S^+(\omega)$ is the Hilbert transform of $S(\omega)$, which shows a deep valley at $\omega=\pi$ due to the intrinsic form of this type of transform. As voiced speech spectra show very low values at high frequencies close to π, the strong decaying behaviour of that imaginary part is usually transferred to the magnitude $E(\omega)$, resulting in a strong peak in the magnitude of $\log E(\omega)$ at $\omega=\pi$. As a consequence, it appears a sample-to-sample oscillation in the COSA sequence which makes difficult the correct determination of the pitch value.

An obvious way of correcting this drawback consists of low-pass liftering the cepstral sequence. However, according to our investigation, liftering is not the best way of removing the disturbing peak at $\omega=\pi$. There exists a computationally inexpensive operation that even improves the results obtained through filtering. It consists of replacing $R(0)$ by $R'(0)$ before the computation of $C^+(n)$, where

$$R'(0) = KR(0) \quad (14)$$

which is equivalent to change the spectrum $S(\omega)$ by $S'(\omega)$, where

$$S'(\omega) = S(\omega) + (K-1)R(0) \qquad (15)$$

Indeed, if K is high enough, the trouble caused by the deep valley of $S_H(\omega)$ at $\omega=\pi$ is removed since the magnitude of $S^+(\omega)$ for frequencies close to π is then almost exclusively influenced by the real part, i.e. the new spectrum $S'(\omega)$.

Moreover, the multiplication of R(0) by K also produces other beneficial effects on the COSA pitch determination algorithm, since it causes a spectral flattening that attenuates formants and it also partially masks noise components at low power frequency bands. Thus, the factor K emphasizes the two above mentioned positive effects of $E(\omega)$.

3. SIMULATION RESULTS

Some simulated experiments were carried out with the aim of evaluating the performance of the COSA pitch determination algorithm and to compare it with the performances of the ACC and the CEP PDAs. For this purpose, we selected two male (FVB,EMG) and two female (AMB,ESP) voices so that the pitch of each one occupies a frequency region that is only partially overlapped with the other three, and in such a way that the global pitch range is approximately 70 - 300 Hz. Every speaker uttered a phonetically balanced Spanish sentence ("El golpe de timón fué sobrecogedor"). The utterances were bandlimited to 3.4 KHz by a 7-pole elliptic filter and sampled at 8 KHz. In order to evaluate the performance of the new PDA noisy conditions, Gaussian white noise was added to each utterance so that the signal-to-noise ratio (SNR) of the resulting signal becomes ∞ (clean), 20, 10 and 0 dB.

For all experiments, the length of each speech frame was 40 ms (320 samples) and the time window was shift 10 ms (80 samples) between successive frames. In our implementation of the ACC algorithm the speech signal is low-pass filtered to 900 Hz with a sharp cutoff linear-phase FIR digital filter and then it is processed according to the compressed center clipping described in equation (5) of [3]. Neither low-pass filtering nor center clipping were employed to implement the CEP and COSA algorithms.

For all the tested PDAs, the pitch value is selected as the lag where the function used to extract the pitch has the absolute maximum within the searching range 2.5 - 20 ms (50 - 400 Hz). The current frame is classified as voiced if that maximum value exceeds a given threshold which is equal to the zero lag value of the function multiplied by a factor r; otherwise, it is classified as unvoiced.

In our study, the accuracy of each PDA was evaluated for each utterance by comparing its sequence of pitch period values and v/uv decisions with a reference pitch contour obtained by means of careful visual inspection. Then, the objective error measures given in [8] were applied to the frame-by-frame differences between the reference pitch and those obtained by the PDAs. A difference of values greater than 1 ms (8 samples) was classified as a gross pitch period error; otherwise, it was classified as a fine pitch period error.

In the following, we will observe the performance of the considered PDAs in term of gross and fine pitch errors; v/uv detection errors will be taken into account subsequently. Tables 1 and 2 show, respectively, the number of gross errors and the standard deviation of fine errors in frames that have been labeled as voiced by the reference pitch contour. Speakers are ordered from low pitch to high pitch. The COSA results are presented for three very different values fo K in order to show the influence of this parameter on the performance of the algorithm.

For clean speech, the performance of the COSA algorithm is clearly better than that of the other two PDAs in terms of both gross and fine pitch period errors. Most gross errors of the ACC and CEP algorithms which were corrected by the

TABLE 1. GROSS PITCH PERIOD ERRORS

Speaker	FVB				EMG				AMB				ESP				GLOBAL			
SNR (dB)	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0
ACC	17	17	17	23	7	6	14	43	0	0	1	38	2	3	9	27	26	26	41	131
CEP	14	19	28	69	8	13	19	45	3	1	6	28	7	7	8	38	32	40	61	180
COSA(k=2)	10	11	17	73	6	8	14	37	1	3	3	11	2	4	4	8	19	26	38	129
COSA(k=8)	7	6	12	47	5	5	5	29	0	1	1	12	2	2	4	7	14	14	22	95
COSA(k=100)	9	10	14	39	4	5	5	26	0	1	1	11	2	2	4	7	15	18	24	83

TABLE 2. STANDARD DEVIATION OF FINE PITCH PERIOD ERRORS

Speaker	FVB				EMG				AMB				ESP				AVERAGE			
SNR (dB)	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0
ACC	1.67	1.68	1.77	1.66	1.30	1.31	1.26	1.48	1.11	1.23	1.12	1.13	1.10	1.10	0.93	1.02	1.29	1.33	1.27	1.32
CEP	1.40	1.45	1.56	1.72	1.55	1.44	1.60	1.72	1.12	1.37	1.34	1.47	0.97	1.15	1.33	1.48	1.26	1.35	1.46	1.60
COSA(k=2)	1.16	1.50	1.35	1.64	1.49	1.68	1.72	1.66	1.07	1.07	1.29	1.48	1.04	1.04	1.21	1.57	1.19	1.32	1.39	1.59
COSA(k=8)	1.44	1.59	1.35	1.40	1.25	1.51	1.51	1.61	0.94	1.04	1.20	1.33	0.99	0.96	1.06	1.30	1.15	1.27	1.28	1.41
COSA(k=100)	1.44	1.51	1.42	1.30	1.12	1.18	1.35	1.64	1.08	0.99	1.23	1.24	1.02	1.04	1.04	1.34	1.16	1.18	1.26	1.38

TABLE 3. VOICED-TO-UNVOICED /UNVOICED-TO VOICED ERRORS

Speaker	FVB				EMG				AMB				ESP				GLOBAL			
SNR (dB)	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0	∞	20	10	0
ACC	15/6	14/1	19/0	55/0	7/7	7/3	20/3	57/0	2/10	1/6	3/3	21/2	0/8	0/8	0/3	19/2	24/31	22/18	42/9	152/4
CEP	31/10	79/1	137/0	173/0	10/8	44/0	93/0	183/0	7/9	21/1	71/0	157/0	6/4	20/1	101/0	146/0	54/31	164/3	402/0	659/0
COSA(k=100)	32/5	34/0	64/0	136/0	13/5	25/0	45/0	132/0	1/11	2/1	3/3	70/0	0/10	1/2	14/0	71/0	46/31	62/3	126/3	409/0

COSA algorithm correspond to transitions. Moreover, the COSA results improve for values of K greater than 2; however, the differences between values so distinct as 8 and 100 are small. The above comparisons are still valid for noisy speech in terms of gross errors, though for very low SNR (0dB) the differences between the ACC and the COSA algorithms becomes smaller. The standard deviation of fine pitch errors of both PDAs is similar for noisy speech.

It is worth noting that the results of the COSA PDA shown in Table 1 could be noticeably improved by weighting the COSA sequence with a slightly increasing function, since many gross pitch errors in male speakers are due to pitch period halvings. For example, using a logarithmic weighting, 6 of the 15 errors for K=100 and clean speech are corrected. Nevertheless, this COSA weighting has not been an object of our investigation since probably its optimization would yield a weighting function excessively fitted to our small data base.

In conclusion, the new PDA is clearly superior to the classical ACC and CEP algorithms in terms of pitch errors, for both clean and noisy speech. Additionally, a large range of K values obtains similar results; however, a large value of K is preferable for very low SNR. Figure 1 depictes the average of percentages of gross pitch errors per speaker as a function of the SNR. The number of voiced frames in the reference pitch contours are, respectively, 174, 198, 158 and 147 (same order than in the table).

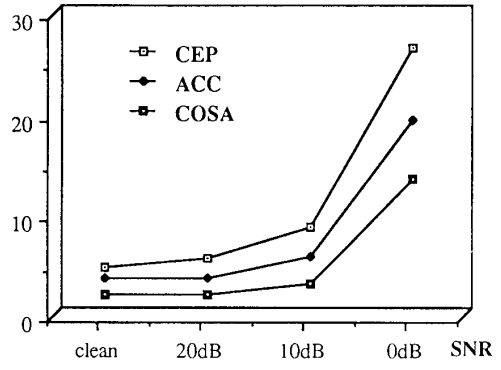

Figure 1. Average of speaker percentage scores of gross pitch period errors for the three considered PDAs.

Table 3 shows the results obtained when the three PDAs are used as voiced/unvoiced detectors just using the relative height of the highest peak. The threshold factor r was set to 0.3 in the ACC algorithm [3] and, for the two cepstral PDAs, values of r yielding the same number of UV-V errors than the ACC algorithm were selected (the number of unvoiced frames of the utterances are, respectively, 28, 24, 29 and 26). Only K = 100 was considered for the COSA detector, since small values of K obtained much worse results.

From the table, we may notice that, even though the COSA detector shows a better performance than the cepstrum PDA, its results are far from the ACC ones for both clean and noisy speech. As for pitch period errors, a weighting function also can obtain better results for our data base. To improve the COSA v/uv detector for noisy speech, a threshold dependent of the noise level should be used. An alternative approach that possibly could also improve its performance for both clean and noisy speech would consist of using more parameters in the decision, e.g. the zero crossing rate of the COSA function.

4. CONCLUSIONS

For both clean speech and speech contaminated with additive white noise, the presented PDA based on the cepstrum of the one-sided autocorrelation has shown a performance in terms of pitch period errors significantly better than those of the cepstrum and the autocorrelation with center clipping algorithms. This better performance is based on its remarkable accuracy at non-stationary segments of speech signals, where it achieves a significant removing of formants, and to its noise reduction capability . However, to be useful as a v/uv detector it would require (as the cepstral detector) a procedure not so simple as the threshold comparison used in the autocorrelation-based detector. Finally, its computational complexity is similar to that of the cepstrum PDA, having as advantage the possibility of obviating the log operation in the frequency domain by doing multiplication-add computations (recursion (8)) in the lag domain.

REFERENCES

[1] W.Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.

[2] A.M.Noll, "Cepstrum pitch determination",*J.Acoust. Soc. Amer.*, vol.41, pp.293-309, Feb. 1967.

[3] L.R.Rabiner, "On the use of autocorrelation analysis for pitch detection", *IEEE Trans. Acoust., Speech and Signal Proces.*, Vol.25, pp.24-33, Feb. 1977.

[4] M.A. Lagunas and M. Amengual, "Non-linear spectral estimation", *ICASSP'87*, Dallas, pp. 2035-8, Apr. 1987.

[5] C. Nadeu, "A simple spectrum estimation technique based on the analytic cepstrum", *V European Signal Proces. Conf.*, Barcelona, pp.465-468, Sept. 1990.

[6] A.V.Oppenheim and R.W.Schafer, *Digital Signal Processing*, Englewood Cliffs, NJ, Prentice-Hall, 1975.

[7] D.P.McGinn and D.H.Johnson, "Reduction of all-pole parameter estimation bias by successive autocorrelation", *ICASSP-83*, Boston, pp.1088-1091, Apr. 1983.

[8] L.R.Rabiner, M.J.Cheng, A.E.Rosenberg and C.A.Mac Gonegal "A comparative performance study of several pitch detection algorithms", *IEEE Trans. Acoust., Speech and Signal Proces.*, Vol.24, pp.399-418, Oct. 1976.