ELSEVIER

Medical Engineering & Physics 20 (1998) 432–442

Medical Engineering & Physics

Acoustic analysis of newborn infant cry signals

Ada Fort [a,b], Claudia Manfredi [a,*]

[a] *Electronic Engineering Department, University of Florence, Via S. Marta 3, 50139 Florence, Italy*
[b] *Information Engineering Department, University of Siena, Via Roma 56, 53100 Siena, Italy*

Received 28 November 1997; accepted 13 May 1998

Abstract

This paper aims at estimating the fundamental frequency (pitch) and the vocal tract resonant frequencies (formants) from newborn infant cry signals. Such parameters are of interest in exploring brain function at early stages of child development, for the timely diagnosis of neonatal disease and malformation. The paper compares a spectral parametric technique and the cepstrum approach, extending previous results. The parametric technique is based on autoregressive models whose order is adaptively estimated on subsequent signal frames by means of a new method. This allows the correct tracking of pitch and formant variations with time. The traditional cepstrum approach is modified in order to follow signal variability. In particular, the cepstrum spectral resolution is improved by applying the chirp Z-transform (CZT) and by adaptively varying the 'lifter' length. The two methods are tested on simulated data, as far as robustness to noise and spectral resolution are concerned, and are then applied to real baby cry data.

Keywords: Autoregressive models; Cepstrum; Medical applications; Parameter estimation; Power spectral density; Speech analysis; Time-frequency representation

1. Introduction

The acoustic analysis of infant cry is used to deduce information on the state of health of new-born babies as well as of children a few weeks old [1,2]. For new-born children, phonation is based on an already developed laryngeal coordination and, therefore, the study of cry signals can lead to the detection of disturbances of brain functions. There are at least two important characteristic parameters of acoustic emission: the period associated with the fundamental frequency f_0, usually named pitch, and the formants. f_0 is related to the quasi-periodic vibrations of the glottis and is given by the average value of the fundamental component in the Fourier spectrum of the utterance, when it shows a quasi-periodic structure (voiced sound). The vocal tract connecting the larynx to the mouth aperture acts as a filter applied to the sounds generated at the glottis. The formants are the resonant frequencies of this filter, which acts on the quasi-periodic source of the acoustic signal from the glottis (for 'voiced' sounds, like vowels), or on the noise source ('unvoiced' sounds), whose origin is turbulence in different parts of the vocal apparatus. The source spectrum is thus a line spectrum (voiced sounds) or a continuous spectrum (unvoiced sounds), to which the transfer function of the vocal tract is applied, along with its resonant frequencies. Infant cry presents both voiced and unvoiced structure. Voiced cry may be classified in the two categories of phonation and hyperphonation, depending on the frequency of excitation, f_0. Such frequency, for infants, may be very high: below 700 Hz for phonation and above 700 Hz for hyperphonation. In fact, the vocal tract of a new-born child is shorter (6–8 cm) and has a different structure with respect to that of an adult. Hence, it is associated with higher resonances and fundamental frequency than those of adults. More details about the speech production apparatus, as well as speech models and properties can be found in Fort et al. [3].

In hyperphonating cries, it is very difficult to separate the excitation spectral contribution from that of the vocal tract, because of the widening of the harmonic structure and possible undersampling of the spectral envelope corresponding to the vocal tract.

Unvoiced cry is called disphonation; in the case of

* Corresponding author. Tel: + 39 55 4796764; Fax: + 39 55 4796767; E-mail: manfredi@die.unifi.it

1350-4533/98/$19.00
PII: S1350-4533(98)00045-9

disphonant cry, the recovery of formant locations and characteristics is quite easy, since the corresponding power spectrum density (PSD) matches the vocal tract frequency response.

Typical values for the first three formants are around 1, 3 and 4.5 kHz. Both pitch variations and formant amplitudes and locations are of interest in new-born cry analysis for pathological detection.

Speech signals are non-stationary and, at best, can be considered quasi-stationary over short time segments (typically 5–20 ms). The statistical and spectral properties of speech are thus defined over short segments. This paper aims at describing and comparing two analysis techniques capable of tracking fast spectral variations, extending the results reported in previous works [3,4]. Specifically, the autoregressive (AR) power spectral density (PSD) technique and the chirp Z-transform (CZT) are considered. The performance of the two methods is compared on simulated data, constructed to have known characteristics. This gives a basis of comparison for deciding the quality of the estimation, since, for real data, the actual f_0 and formant frequencies are not directly known.

The methods are then applied to real baby cry data, collected at the Paediatric Hospital A. Meyer, Florence, Italy. They are relative to premature and low birthweight neonates.

2. Parametric speech analysis

Parametric analysis of acoustic signals is based on pole-zero or all-pole models for the vocal tract response, depending on the signal characteristics. Commonly, for non-nasal signals, an all-pole (auto-regressive, AR) model is used, which assumes that a signal sample is a weighted linear combination of n previous samples described by the equation:

$$y(k) = \sum_{i=1}^{n} a_i y(k-i) + w(k) \tag{1}$$

where $y(k)$ is the signal, $w(k)$ is the error or linear prediction (LP) residual, a_i are the weights applied to the previous signal samples and n is the model order. Passing the signal through the filter $A(z) = 1 - \sum_{i=1}^{n} a_i z^{-i}$ results in the removal of near-sample correlation and produces the LP residual $w(k)$ that contains the pitch information in the speech. On the other hand, $A(z)$ captures the vocal tract information given by the formant amplitudes and locations. This allows separation of the vocal tract information and the pitch information, both of which are contained in the speech signal [5,6].

However, Eq. (1) does not take into account the time delay between input (glottal pulse train) and the output (voice emission). Assuming stationarity, a more realistic model has the following transfer function [5,7]:

$$H(z) = G \frac{z^{-n/2}}{(1 - a_1 z^{-1} - \ldots - a_n z^{-n})} \tag{2}$$

i.e. there are $n/2$ zeros at the origin (the I/O delay). The poles of $H(z)$ (e.g. the zeros of the filter $A(z)$) define the resonant or formant structure of the model. Hence, the model structure that will be used in the present work is the auto-regressive with exogenous variables (ARX) model:

$$A(z)y(k) = z^{-d}B(z)u(k) + w(k) \tag{3}$$

with input $u(k)$ (a delta-pulse train) and output $y(k)$; d is the I/O time delay and $w(k)$ is the system noise component (Gaussian zero-mean white noise). $A(z)$ and $B(z)$ are polynomials in the unit-delay operator z^{-1} of order n_a and n_b, respectively, with A generally monic. In particular, Eq. (2) assumes $B(z) \equiv 1$, $n_a = n$ and $d = n/2$.

The resonances (formants) can be recovered by locating the maxima of the function given by the following equation, relative to the AR part of the model [8]:

$$\mathrm{PSD}(f) = \frac{1}{\left|1 + \sum_{k=1}^{n} a_k e^{-j2\pi fkT}\right|^2} \tag{4}$$

where PSD is the power spectral density and a_k are the $A(z)$ polynomial coefficients.

In this context, the choice of the model order plays a major role: in case of overestimated model order, formant splitting may occur; when the model order is underestimated, phase and temporal relationships between the model and the true system become very unclear. In this case, the model will approximate the magnitude spectrum as well as possible, concentrating on the peaks in the spectrum first. An accepted operational rule for the choice of the model order n is [5]: $n = f_s + (4 \text{ or } 5)$ for voiced speech, $n = f_s$ for unvoiced speech, where f_s is the sampling frequency of the data in kHz.

Several criteria exist for finding the 'best' model order in an objective way. These are usually based on the estimation of a loss function $V_M(\theta_M)$ obtained from a model with p estimated parameters θ_M, where M is the number of data samples upon which the models are fitted, i.e. the data window length. The criterion is that of choosing the model whose order p and parameters minimise, in a least squares (LS) sense, a function of the estimated error variance σ_p^2, which is defined as:

$$\sigma_p^2 = \frac{2V_M(\theta_M)}{M} = \frac{1}{M}\sum_{i-1}^{M} \epsilon_i^2(\theta_M),\ \theta_M \epsilon \Re^p,\ p = 1,\ldots,L \qquad (5)$$

where $\epsilon_i^2(\theta_M)$ represents the squared error between the ith model output and the ith observation, for p varying from 1 to a convenient maximum L. All LS estimation loss functions decrease monotonically with increasing model order. Thus, a penalty term is usually added, which makes the loss function grow as the model order increases. This allows finding the 'best' model order as the one corresponding to the minimum of the criterion. However, none of the classical criteria work well when a short data window is to be selected, as in the present application, i.e. for almost non-stationary speech signals [8]. To overcome this difficulty, in the present work the following automatic model order selection criterion is proposed:

- Let σ_i^2 be the residual variance values, which decrease as p increases: $\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_L^2$, $i = 1,\ldots,L$ for a large enough L;
- select as the 'best' model order p the value of i for which the first $i-1$ σ_i^2 values are 'well separated' from the last $L - i + 1$, according to the following rule, named dynamic mean evaluation (DME) [3]:

$c = 0;$

for $k = 1,2,\ldots,L/2$

$$\text{if } \left| \frac{1}{k}\left(\sum_{i=1}^{k}\sigma_i^2\right) - \sigma_{k+1}^2 \right| > \left| \frac{1}{k}\left(\sum_{i=1}^{k}\sigma_{L-i+1}^2\right) - \sigma_{k+1}^2 \right|$$

then $p = k$ and $c = 1;$

endif;

endfor;

if $c = 0$ (6)

for $k = 1,2,\ldots,L/2$

$$\text{if } \left| \frac{1}{k}\left(\sum_{i=1}^{k}\sigma_i^2\right) - \sigma_{L-k}^2 \right| < \left| \frac{1}{k}\left(\sum_{i=1}^{k}\sigma_{L-i+1}^2\right) - \sigma_{L-k}^2 \right|$$

then $p = L - k;$

endif;

endfor;

endif;

otherwise $p = L.$

In Eq. (6), σ_i^2 is the ith residual variance value, L is the maximum allowed order and c is a Boolean index.

In words, the DME procedure works as follows:

For the residual variance values (rv) in decreasing order: $\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_L^2$, the model order p is selected through an iterative scheme comparing the $(k+1)$th ($k = 1,\ldots,L/2$) rv distance both from the arithmetic mean of the first k rv and from that of the last $L - k - 1$ rv. If the first distance is larger than the second the model order is $p = k$. If no value of k can be found satisfying the above condition, the procedure is repeated starting from the index of the smallest rv, thus giving the model order $p = L - k$. Finally, if none of the above conditions is verified, the model order p is set equal to L, all the rv being of comparable dimension.

The improvement obtained with this criterion over classical methods relies on its deterministic nature, since it is based on a distance measure between the largest variance values and those for which no significant improvement in the model accuracy is gained by increasing the model order. Hence, the criterion also works well for short data frames. This has been successfully experimented in previous applications [3,9].

3. CZT-cepstrum

If a signal $y(k)$ is the convolution of two components $y_1(k)$ and $y_2(k)$, log transforming its z transform $Y(z)$ transforms a convolution of components to a sum. The output of this transform is generally referred to as the complex cepstrum. Linear filtering and inverse transforming the result allows the separation of the two additive components [7].

Since the complex cepstrum of the vocal tract impulse response is concentrated around $n = 0$, the complex cepstra of the vocal tract impulse response and the excitation for voiced speech often tend to occupy somewhat disjoint time intervals. Thus, the cepstral values representing the vocal tract can be extracted from the total cepstrum by means of a linear filter (low-time lifter) that multiplies the low-time values by unity and the remaining values by zero. Similarly, the pitch cepstral value can be recovered by high-time liftering the cepstrum and finding the maxima.

In practice, the speech signal is assumed to be minimum phase, which allows computing the inverse FFT transform of the log magnitude of its transform, i.e. the more widely used real cepstrum.

As already said, the signals under study are characterised by high fundamental frequency (f_0) values, which makes it difficult to separate the vocal tract contribution from the fundamental frequency contribution. Hence, a correct choice of the lifter length is of great importance. In fact, the frequency resolution of the homomorphic analysis depends on the lifter length: a very short lifter smoothes the vocal tract response, while too long a lifter fails to separate the excitation contribution. The lifter length is usually fixed a priori taking into account the local signal characteristics. In the present application, a

low value is required (eight points), because of the high f_0 value of the signals under study.

However, a fixed length may lead to mismatched results. Cepstral analysis can be optimised by choosing the optimal lifter length on each time window in order to achieve the best frequency resolution: this can be obtained by estimating the pitch period T_p by means of the parametric method and accordingly fixing the lifter length ($T_p/2$, in order to exclude undesired 'quefrencies'). In the present work, this approach (named: adaptive cepstrum) was implemented and the results are compared to the traditional approach (fixed cepstrum).

Finally, the time window length is critical, both as far as pitch and formant estimation are concerned. A short time window, required by the almost-stationary signals under study, could prevent recovering the true pitch value. Moreover, it reduces the frequency resolution, giving an almost flat spectral shape, which makes it difficult to determine the presence of peaks.

In order to enhance the cepstral resolution, the spectral analysis algorithm called the chirp Z-transform (CZT) is used, since it allows the computation of samples of the Z-transform at equally spaced intervals along a circular or spiral contour in the Z-plane. In particular, it is possible to compute the Z-transform on a contour passing closer to the pole locations than the unit circle contour, thereby enhancing the peaks in the spectrum and improving the resolution.

In the present work, the choice of the optimal contour is tailored to the varying signal characteristics: it is chosen as the one corresponding to the maximum modulus of the ARX poles estimated on each time window.

4. Experimental results

Because of the non-stationarity of the signals under study, the analysis is performed on consecutive short time windows (about 5–25 ms). The software is implemented with Matlab (rel.5.1) for Windows 95.

The method, applied both to simulated and to real signals, consists of the following steps:

4.1. Parametric analysis

1. Identification of the pitch period T_p (from AR model residuals, Eq. (1)) and generation of the corresponding excitation sequence, i.e. the ARX model input (Eq. (3)). This sequence is a delta-train with period T_p;
2. Identification of the model I/O delay, by minimising the loss function (Eq. (5)) of ARX models (Eq. (3)) with a fixed order ($n = 8$, corresponding to three to four formants) and delay increasing from 0 to the estimated pitch period (procedure DMA, Eq. (6));
3. Automatic selection of the ARX model order by minimising the loss function obtained by varying $p = n_a$ (Eq. (5) and procedure DMA). Steps 2 and 3 thus give the delay and parameters for the 'optimal' ARX model (Eq. (3));
4. Parametric spectral estimation (Eq. (4)) and spectral peak detection by finding PSD maxima locations, i.e. formant positions.

4.2. CZT-cepstrum

5. Cepstrum transform of the signal on each time window;
6. High-time liftering with fixed lifter length (eight points) or variable length ($T_p/2$, with T_p adaptively estimated on each data window in step 1) and pitch estimation;
7. Low-time liftering with fixed lifter length or variable length, as above, to recover the vocal tract response;
8. CZT, evaluated on a circle with fixed radius equal to the largest ARX-pole module (ARX model used in step 3) and formant estimation.

4.3. Simulations

As in Fort et al. [3], two sets of simulated data (voiced cries) were used. The first set was derived by building in the frequency domain an amplitude spectrum given by the product of a slowly varying component (formants) with a fast varying component (pitch excitation). The formant amplitude spectrum was generated as a superimposition of Gaussian curves with mean value F_i selected according to the following expression valid for the lossless tube model:

$$F_i = \frac{c(2i-1)}{4l} \quad i = 1,\ldots,I \tag{7}$$

where l is the vocal tract length (about 6–8 cm for newborn infants), c is the sound velocity and I is the maximum number of formants in $[0,F_s/2]$ (F_s = sampling frequency). The excitation contribution, which corresponds to a quasi-periodic pulse train, is modelled in the frequency domain by a train of Gaussian curves whose width is related to the time duration of the simulated signal and whose periodicity corresponds to the inverse of the time periodicity. The pitch period is fixed equal to 2.9 ms, corresponding to a periodicity in the frequency domain of about 350 Hz. Notice that, for simplicity, quasi-periodicity of human vocalisation is not taken into

account here. The time signal was obtained by applying the inverse discrete Fourier transform.

The second data set was obtained in the time domain by choosing an ARX model (Eq. (3)) with three stable complex conjugate poles. The pole choice was made in order to produce an AR spectrum with peaks located by analogy to real formant positions. To this end, real data underwent a preliminary identification procedure from which poles and parameters of an AR model of fixed order $n = 6$ were obtained. A model of voiced sound is obtained by exciting the resulting model with a periodic impulse train $e(n)$ with varying (decreasing) time period T_p (pitch), in order to simulate the transition from phonated towards hyperphonated cry. The resonances (formants) of the speech signal correspond to the poles of the transfer function, which can be recovered from the position of maxima in the PSD given by Eq. (4).

In both sets of simulations a zero-mean white noise with uniform distribution is added to the signal. The signal-to-noise ratio (SNR) is varied from 4 to 20 dB, in order to test the robustness of the methods against noise. The sampling frequency is set equal to 10 kHz. The time window length varies from 6.4 to 25.6 ms; 256-point FFTs are used for cepstrum computation.

Steps 1–8 were applied to both data sets. Twenty simulations were performed on each window, both for the time domain simulated data and for data generated in the frequency domain, for a statistical evaluation of the performance of the two methods.

The mean estimated pitch value is plotted versus the increasing time window index. The comparison is made among the parametric approach (ARX), the cepstrum with variable lifter length (VCEP) and the cepstrum with a priori fixed lifter length (FCEP).

Figs. 1–4 show the estimated pitch for different window lengths and SNR for signals generated in the time domain. The ARX model has three pairs of complex-conjugate poles near the unit circle, corresponding to resonant frequencies F_1, F_2 and F_3 of 1, 2.5 and 4 kHz, respectively. The input sequence is a delta train, with period decreasing from 2.8 to 1.1 ms (350–900 Hz), corresponding to a transition from phonated to hyperphonated cry. As pointed out in Section 1, a particularly critical situation in the analysis of infant cry occurs when the pitch period is small, and the corresponding excitation spectrum is a slowly varying frequency function.

Figs. 1 and 2 show the performances of the three methods as far as the robustness to noise is concerned. In both cases, a window length of 12.8 ms is used. In Fig. 1 (SNR = 4 dB) the VCEP performs slightly better than the FCEP. Due to the low SNR, both fail to follow the true pitch value below 1.6 ms. When the SNR is increased to 20 dB (Fig. 2), the three methods give comparable results.

The window length is another critical parameter for the cepstral approaches: a reduction to 6.4 ms (Fig. 3) causes both VCEP and FCEP to perform worse than the ARX approach (the VCEP is slightly better): the lowest pitch values are in fact also underestimated for SNR = dB. The results quickly deteriorate as the SNR decreases. Finally, too long a time window is also critical. In this case, several pitch periods are present in the cepstrum, but they are embedded in noise. This often causes a pitch value to be picked which is twice the true value. Fig. 4 is obtained with a window length of 25.6 ms and SNR = 10 dB: in this cae, both VCEP and FCEP heavily overestimate the true pitch value. Results get worse for decreasing SNR.

Thus, it seems that the parametric approach is better able to follow pitch variations for very short signal frames, high pitch values and low SNR values. However, its computational burden is greater than that of the cepstral approaches. Also notice the better behaviour of the VCEP with respect to the FCEP.

For formants, results are reported in a grey-level map ranging from white (no formant found at the specified time-frequency value) to black (all the simulations give the same formant value). Here the ARX approach is compared with the CZT-cepstrum with variable lifter length (CZT-VCEP) and the traditional cepstrum with variable lifter length (VCEP). A comparison between VCEP and FCEP can be found in Fort et al. [3].

In Fig. 5, the simulation results for signals generated in the time domain and for two different SNR values are shown (20 and 4 dB, respectively). When the pitch period is reduced (below 1.8 ms), the VCEP fails to follow the formant positions correctly, while the parametric method tracks them almost correctly. The CZT-VCEP approach has better performance with respect to VCEP at high signal-to-noise ratios, as it succeeds in finding the formant locations even in hyperphonation cases, according to its increased resolution capability. Nevertheless, this method shows a heavy degradation when the SNR decreases. However the results obtained by CZT-VCEP compare favourably with those obtained by Fort et al. [3], where the traditional cepstrum approach was used.

Fig. 6 concerns signals generated in the frequency domain. These results extend those obtained by Fort et al. [3]. Two formants, F_1, F_2, vary their locations in subsequent time windows, in order to evaluate the resolution and tracking capability of the methods. In particular, F_1 varies from 1 to 1.5 kHz, and F_2 from 3.5 to 2.5 kHz. The third formant, F_3, is kept constant and equal to 4.5 kHz. The plots show the estimated frequency values versus the increasing time-window index. The comparison is performed among ARX and cepstral analysis with CZT-VCEP (middle plots) and VCEP (upper plots), for time windows of fixed length (12.8 ms) but increasing noise level (from 20 up to 4 dB).

The figure shows the performance degradation of the methods with respect to SNR reduction. However, the

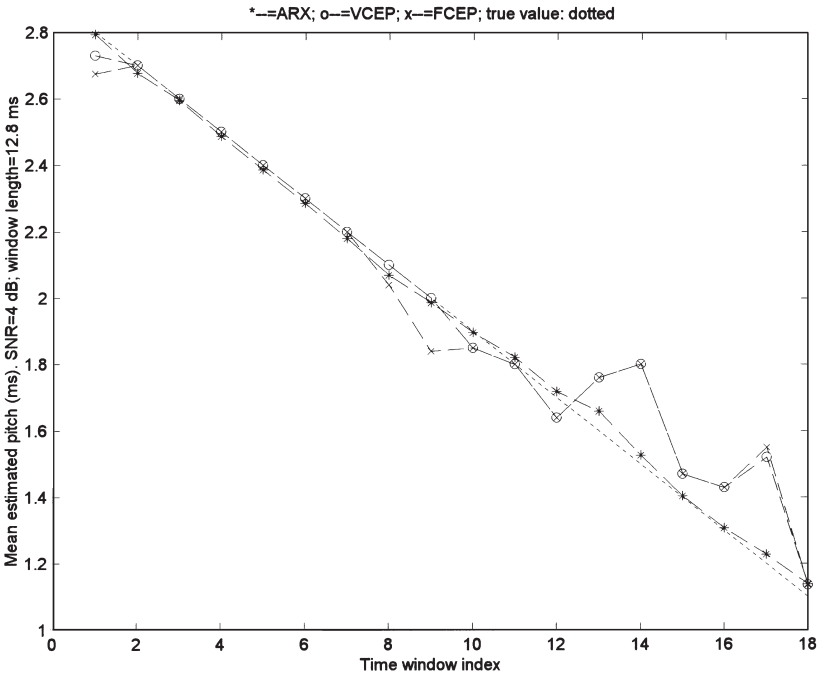


Fig. 1. Time-domain simulated data. Estimated pitch vs. time window index. Window length: 12.8 ms; SNR = 4 dB; *ARX; ○ VCEP; × FCEP; true value: dotted line.

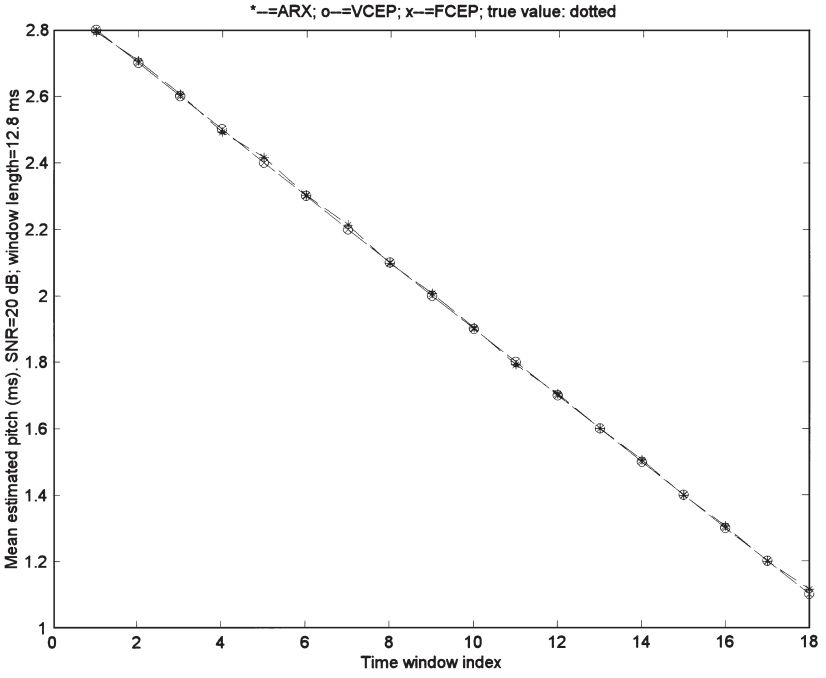


Fig. 2. Time-domain simulated data. Estimated pitch vs. time window index. Window length: 12.8 ms; SNR = 20 dB; *ARX; ○ VCEP; × FCEP; true value: dotted line.

ARX method generally seems more robust against noise. On the other hand, the behaviour of the parametric method is less regular than that of non-parametric methods. A badly selected order gives rise to peak splitting and a consequently false location of formants. Moreover, parametric methods show the tendency to position the formant near the pitch peaks. Finally, when the SNR is very low (Fig. 6, middle-right), the cepstrum also tends to position formants near the pitch peaks.

The bad location of the highest frequency formant F_3, both by the parametric and the non-parametric approaches, may be attributed to its larger bandwidth and smaller amplitude with respect to F_1 and F_2.

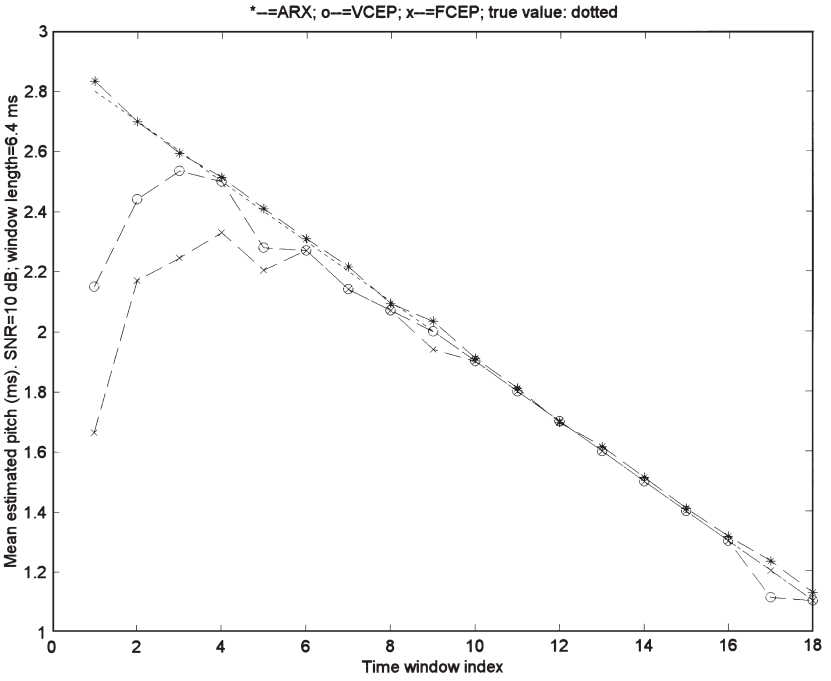


Fig. 3. Time-domain simulated data. Estimated pitch vs. time window index. Window length: 6.4 ms; SNR = 10 dB; *ARX; ○ VCEP; × FCEP; true value: dotted line.

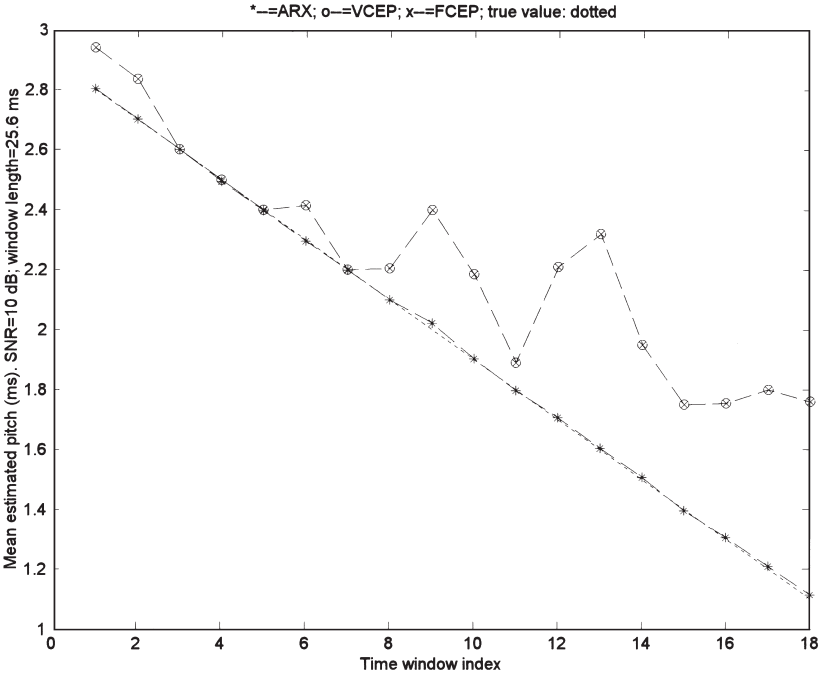


Fig. 4. Time-domain simulated data. Estimated pitch vs. time window index. Window length: 25.6 ms; SNR = 10 dB; *ARX; ○ VCEP; × FCEP; true value: dotted line.

4.4. Real data

Steps 1–8 were also applied to real data obtained from recorded infant cries. Data refer to infants aged from 1 to 2 weeks. They were collected at the Paediatric Hospital A. Meyer, Florence. About 20 registrations were analysed, relative to premature and low birthweight neonates.

The recorded analog signals were A/D converted by using a commercial board, with a sampling frequency of 12 kHz at 16 bits. The analysis was performed on utter-

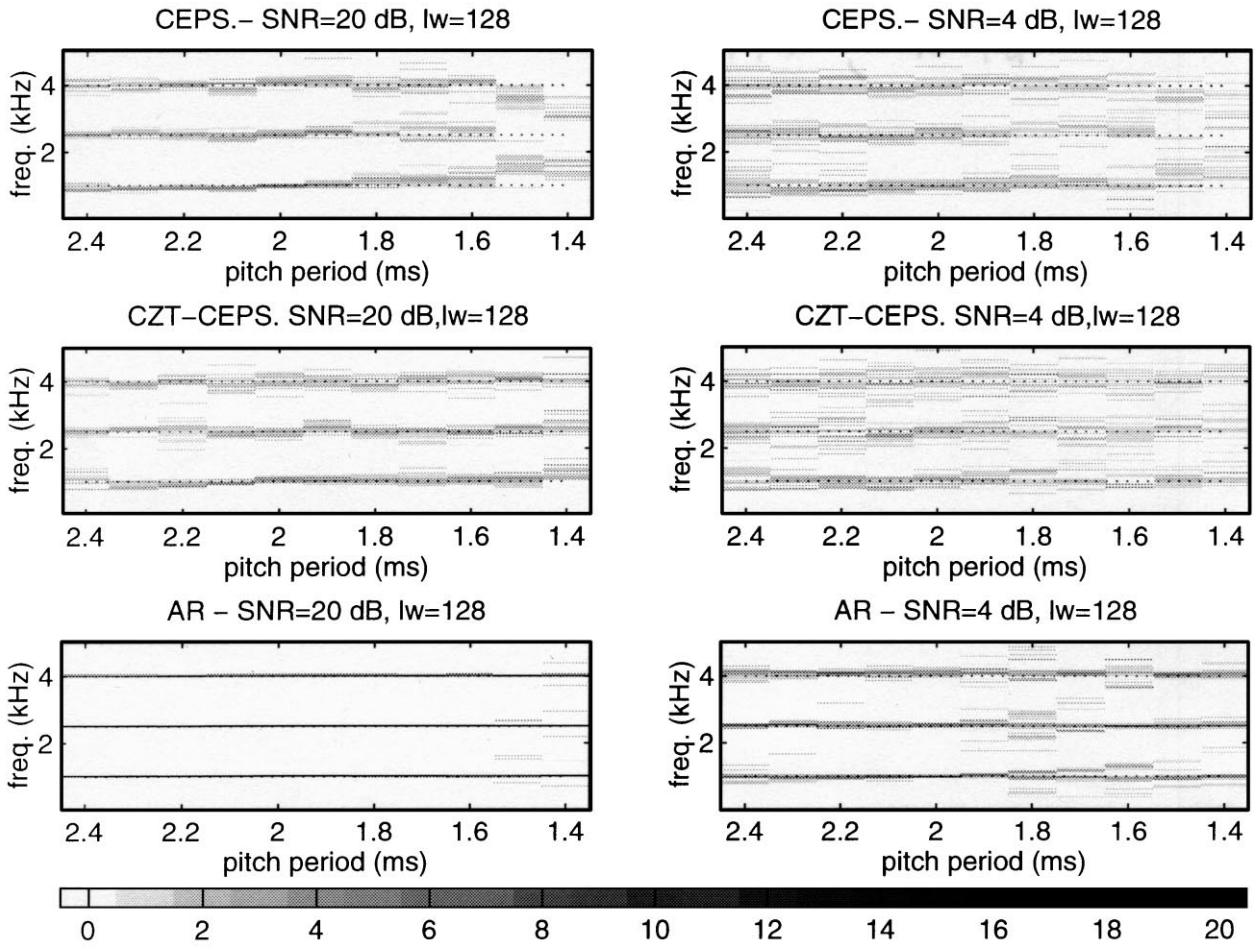


Fig. 5. Time-domain simulated data. Fixed formant frequencies: F_1 = 1 kHz, F_2 = 2.5 kHz, F_3 = 4 kHz. Decreasing pitch period (from 2.4 to 1.4 ms). Window length: 12.8 ms; SNR = 20 and 4 dB. True value: solid line.

ances longer than 0.5 s (cry units). The time window length is 21.3 ms.

Fig. 7(a) shows a typical hyperphonating utterance of a real signal (Fig. 7(b), 5.65–7.33 s). Fig. 7(b) represents the time evolution of the estimated pitch in the transition from a regular utterance towards hyperphonation within the same cry unit. Fig. 7(c) expands the portion of Fig. 7(b) relative to Fig. 7(a). The comparison is made for the parametric method (ARX), the cepstrum with variable lifter length (VCEP) and the cepstrum with fixed lifter length (FCEP). The figure shows the better performance of the ARX approach, which gives more stable pitch values through time. The VCEP sometimes fails to follow the varying pitch value, in particular when abrupt changes occur. The FCEP (with lifter fixed equal to eight points) often incorrectly estimates the pitch, since it is not capable of following its rapid variations.

For the analysed data, the lifter length estimated for the adaptive cepstrum was found to vary between six and 17 points during the utterance. This gives an a posteriori estimate of a 'good' fixed lifter value, corresponding to 12 points. This choice was shown to give fewer variations of the estimated pitch with respect to FCEP and VCEP, but under-estimation of the true pitch value often occurs. Moreover, it is in any case a posteriori deduced from the parametric analysis.

Fig. 8 shows the spectrogram of some utterances of the previous signal. A comparison between the ARX and the CZT-VCEP is reported. The lifter length is estimated according to step 6, i.e. half of the pitch period adaptively estimated on each data window.

The instability of the highest frequency formant location obtained by the traditional cepstral analysis is partially overcome by the enhanced resolution of the CZT algorithm [3].

A CZT-FCEP analysis was also performed. This approach often produces an underestimation of the formant number and a consequent misallocation, giving worse results with respect to the CZT-VCEP.

Also for real data, the formant position is more regular with non-parametric approaches, while parametric methods are sensitive to model order estimation, which causes 'jumps' in the formant location. Moreover, ARX analysis tends to follow the pitch maxima, thus undere-

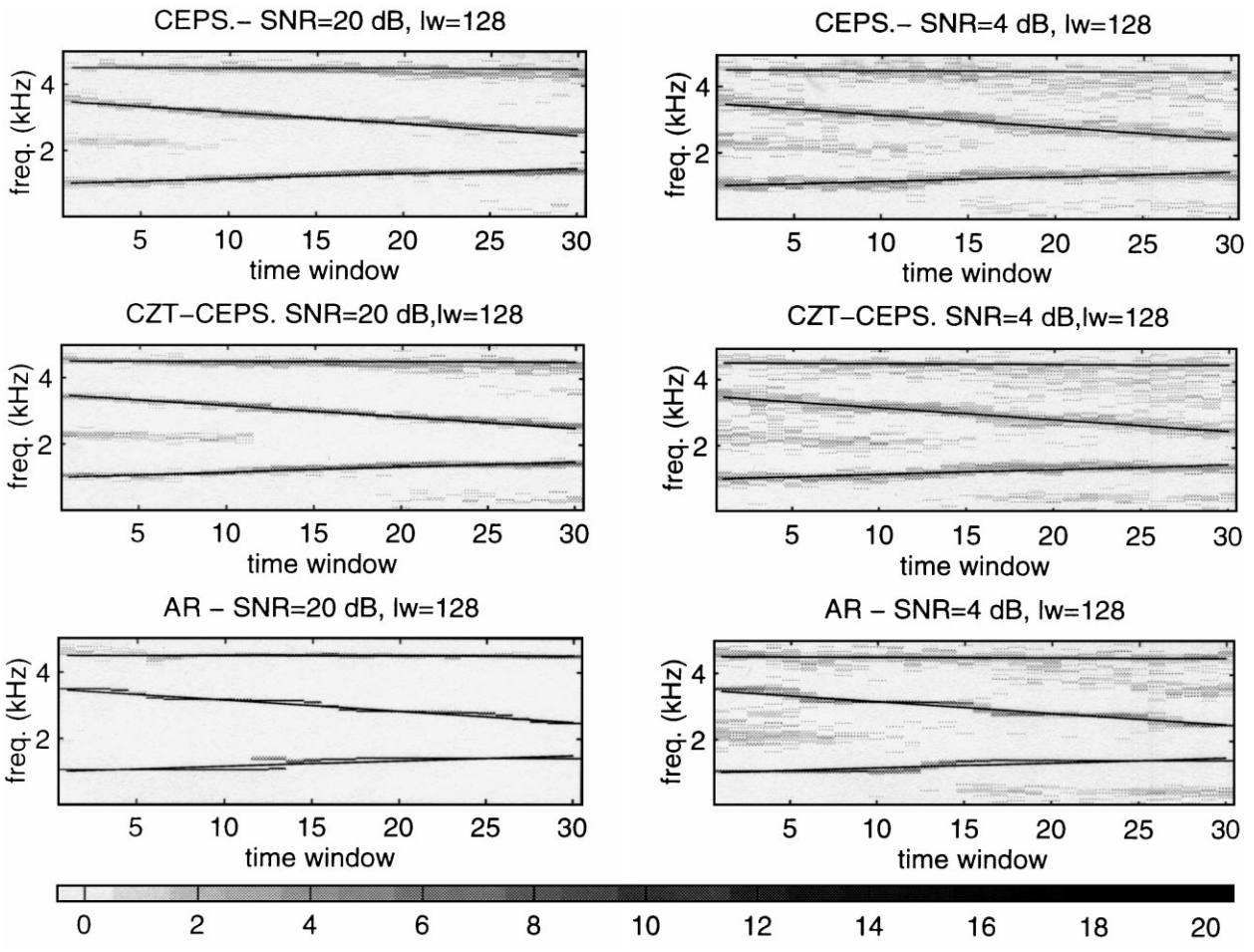
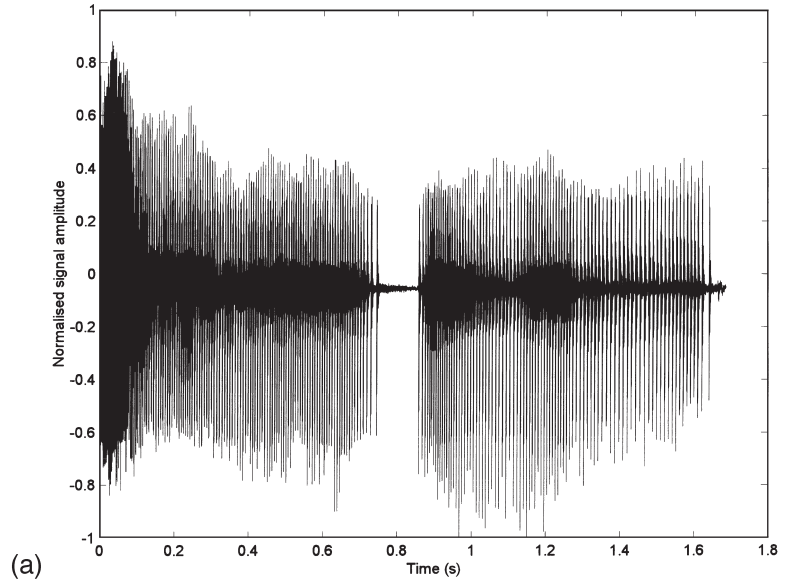


Fig. 6. Frequency-domain simulated data. Varying formant frequencies: F_1 from 1 to 1.5 kHz; F_2 from 3.5 to 2.5 kHz; F_3 = 4.5 kHz. Fixed pitch period (2.9 ms). Window length: 12.8 ms; SNR = 20 and 4 dB. True value: solid line.



(a)

Fig. 7. Real data analysis. (a) Hyperphonating cry utterance (5.65–7.33 s of (b)); (b) Pitch estimation. Transition from a regular utterance towards hyperphonation; (c) Pitch estimation. Enlarged portion of (b) (5.65–7.33 s).

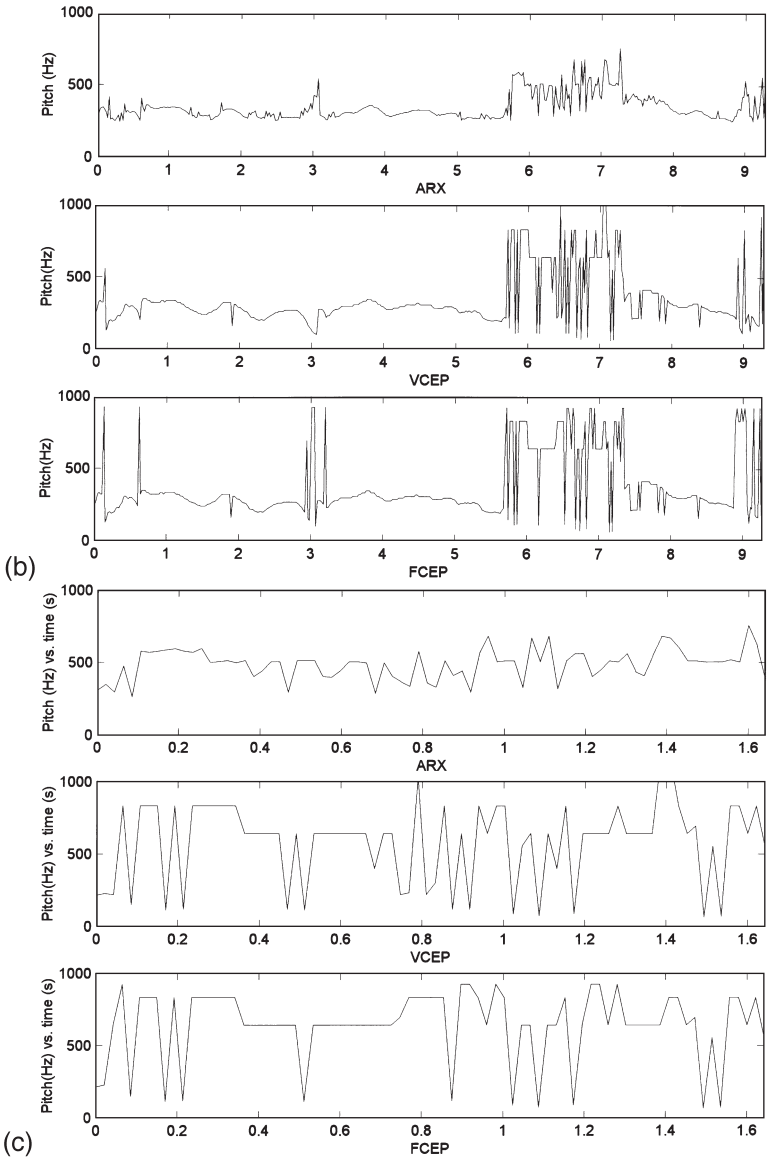


Fig. 7. Continued.

stimating the variation of formant positions. Note that the formants estimated by the CZT-VCEP generally tend to follow those found by the parametric approach, which is thus expected to give more reliable results. As a drawback, in some cases spurious formants are also found by the cepstrum.

Fig. 8(a) shows the spectrogram of a regularly phonated utterance (refer to Fig. 7(b), 0–0.64 s). Fig. 8(b) is relative to the successive utterance of the same cry unit (see Fig. 7(b), 0.64–3.1 s). This signal is characterised by a slight increase in the pitch period (in Hz). The analysis of another utterance of the same cry unit is given in Fig. 8(c). Here, the pitch period is reduced, corresponding to a transition towards hyperphonation (compare to Fig. 7(b), 7.3–9.3 s). Also in this case, particularly critical for the traditional cepstrum, the results obtained by CZT-VCEP are comparable to those obtained by ARX. Hence, the variable lifter length, together with the CZT enhancing algorithm, considerably improves the performance of cepstral analysis in terms of spectral resolution.

5. Final remarks

In the present paper, a comparative analysis of parametric (ARX) and non-parametric (CZT-cepstrum) methods for fundamental frequency and formant estimation is carried out and applied to new-born infant cry analysis, characterised by higher fundamental frequency values and resonances than the vocalisations of an adult. Hence, the signal analysis requires the use of robust and accurate methods capable of following fast signal variations. The choice of the ARX optimal model order, as

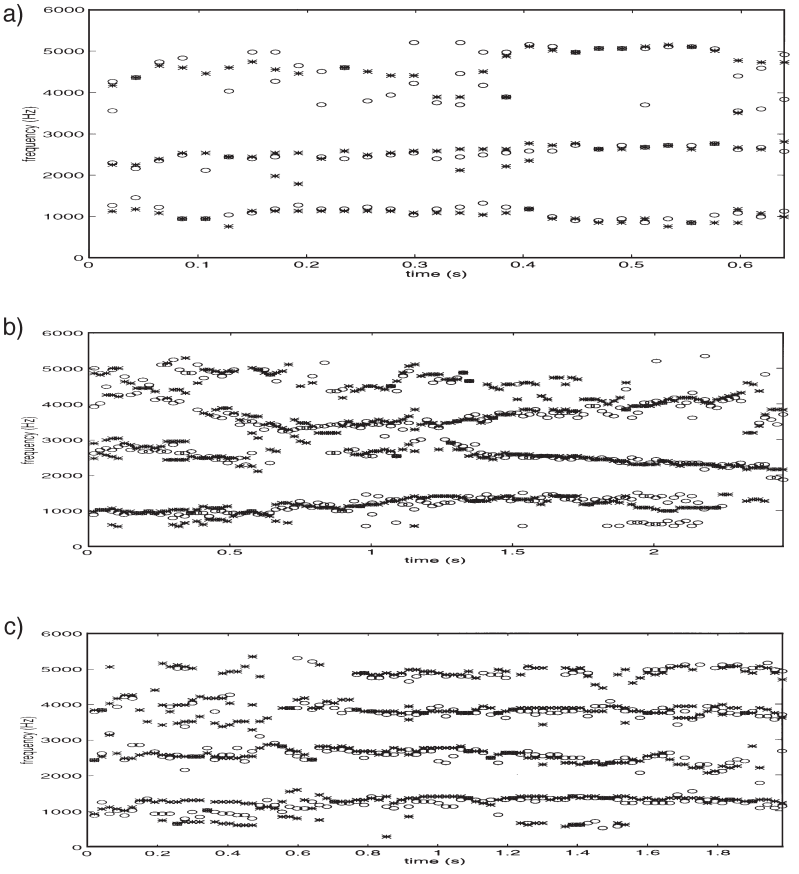


Fig. 8. Real data analysis. Formant estimation. *ARX; ○ CZT-VCEP. (a) Regularly phonated utterance; (b) Reduced pitch; (c) Transition towards hyperphonation.

well as the I/O delay estimation, is performed by means of a new criterion applied to subsequent short time windows, overcoming the difficulties intrinsic to traditional model order selection criteria when applied to short data frames. The cepstrum lifter length is adaptively estimated on each time window, in order to follow the signal variability. The chirp Z-transform enhances the cepstrum spectral resolution, thus providing better formant estimation.

The two approaches are compared both with simulated and real data, in order to highlight the main advantages and drawbacks of each method. It was shown that the parametric approach behaves better than the non-parametric approach, as far as robustness to noise and f_0 tracking capability are concerned, also when short data frames are analysed. On the other hand, the CZT-cepstrum has a smaller computational burden and a more regular behaviour in tracking the evolution of formants.

The results obtained seem promising for biomedical applications as far as early neonatal disease and malformation diagnosis is concerned.

References

[1] Gray L. Signal detection analysis of delays in neonates' vocalisations. J. Acoust. Soc. Am. 1987;82:1608–11.
[2] Kent RD, Murray AD. Acoustic features of infant vocalic utterances at 3, 6, 9 months. J. Acoust. Soc. Am. 1982;72:353–65.
[3] Fort A, Ismaelli A, Manfredi C, Bruscaglioni P. Parametric and non parametric estimation of speech formants, application to infant cry. Med. Eng. Phys. 1996;18(8):677–91.
[4] Donzelli GP, Rapisardi G, Moroni M, Zani S, Tomasini B, Ismaelli A, Bruscaglioni P. Computerised cry analysis in infants affected by severe protein energy malnutrition. Acta Paediatrica 1994;83:204–11.
[5] Markel JD, Gray AH jr. Linear prediction of speech. New York: Springer–Verlag, 1976.
[6] Ramachandran RP, Zilovic MS, Mammone RJ. A comparative study of robust linear predictive analysis methods with application to speaker identification. IEEE Trans. Speech Audio Proc. 1995;3(2):117–25.
[7] Oppenheim AV, editor. Applications of digital signal processing. Englewood Cliffs, NJ: Prentice–Hall, 1978.
[8] Marple SL. Digital spectral analysis with applications. Englewood Cliffs, NJ: Prentice–Hall, 1987.
[9] Fort A, Manfredi C, Rocchi S. Adaptive SVD-based AR model order determination for time-frequency analysis of Doppler ultrasound signals. Ultrasound Med. Biol. 1995;21(6):793–805.