Speech Communication 2 (1983) 37–45
North-Holland Publishing Company

A SYNTHESIS-BASED METHOD FOR PITCH EXTRACTION

K.K. PALIWAL * and P.V.S. RAO

Speech and Digital Systems Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India

Received 9 February 1982
Revised 13 December 1982

Abstract. A synthesis-based method for pitch extraction of the speech signal is proposed. The method synthesizes a number of log power spectra for different values of fundamental frequency and compares them with the log power spectrum of the input speech segment. The average magnitude (AM) difference between the two spectra is used for comparison. The value of fundamental frequency that gives the minimum AM difference between the synthesized spectrum and the input spectrum is chosen as the estimated value of fundamental frequency. The voiced/unvoiced decision is made on the basis of the value of the AM difference at the minimum. For synthesizing the log power spectrum, the speech signal is assumed to be the output of an all-pole filter. The transfer function of the all-pole filter is estimated from the input speech segment by using the autocorrelation method of linear prediction. The synthesis-based method is tried out on real speech data and the results are discussed.

Zusammenfassung. Vorgeschlagen wird ein Verfahren zur Bestimmung der Grundfrequenz von Sprachsignalen nach dem Analyse-durch-Synthese-Prinzip. Das Verfahren erstellt synthetische logarithmierte Leistungsspektren für verschiedene Werte der Grundfrequenz und vergleicht sie mit dem Spektrum des Eingangssignals. Als Vergleichskriterium gilt hierbei die mittlere Betragsdifferenz. Als Schätzwert für die Grundfrequenz des betrachteten Sprachsignalabschnitts wird die Grundfrequenz gewählt, für die die mittlere Betragsdifferenz zwischen dem Spektrum des Eingangssignals und dem synthetischen Spektrum ein Minimum wird. Die Stimmhaft-Stimmlos-Entscheidung erfolgt aufgrund des Wertes der Betragsdifferenzfunktion an der Stelle des Minimums. Zur Synthese der logarithmierten Leistungsspektren wird das Signal als Ausganssignals eines rein rekursiven Digitalfilters angenommen; die Übertragungsfunktion dieses Filters wird von dem betrachteten Abschnitt des Sprachsignals durch Anwendung der Autokorrelationsmethode der linearen Prädiktion gewonnen. Das neue Verfahren wurde anhand natürlicher Sprachsignale erprobt; die Ergebnisse werden im vorliegenden Beitrag diskutiert.

Résumé. Une méthode d'extraction de la fondamentale du signal de parole basée sur la synthèse est proposée. Elle procède par la synthèse d'un certain nombre de spectres logarithmiques de puissance pour différentes valeurs de la fréquence fondamentale, et pour la comparaison de ces spectres avec le spectre logarithmique de puissance du segment de parole analysé. La différence d'amplitude moyenne (AM) entre les deux spectres est utilisée pour la comparaison. La valeur de la fréquence fondamentale qui fournit le minimum de différence AM entre le spectre synthétisé et le spectre du signal d'entrée est choisie comme estimation de la hauteur. La décision voisé/non-voisé est établie sur la valeur de la différence AM au minimum. Pour synthétiser le spectre de puissance, le signal de parole est supposé représenter la sortie d'un filtre tout pôle. La fonction de transfert du filtre tout pôle est estimée à partir du segment de parole par la méthode d'autocorrélation de la prédiction linéaire. Cette méthode basée sur la synthèse est testée sur des segments de parole naturelle et les résultats sont discutés.

Keywords: Speech, pitch extraction, linear prediction.

Introduction

Pitch extraction is an important problem in speech analysis. In pitch extraction of a speech utterance, voiced/unvoiced decisions are made about speech segments and the pitch period (or its inverse, the fundamental frequency) is estimated for the voiced speech segments. The pitch contours of speech utterances are useful in various speech processing applications such as speech analysis-synthesis [1,2], speech understanding [3] and speaker recognition [4,5].

Many methods of pitch extraction are reported

* Present adress: Division of Telecommunications, University of Trondheim, Trondheim-NTH, Norway.

in the literature (see the references given in [6]). But, the problem of pitch extraction is yet to be solved satisfactorily [6,7]. In the present paper, we propose a new method of pitch extraction using a synthesis-based procedure. This method works in the frequency domain. Though a few frequency domain methods of pitch extraction have been developed in the past [8–13], these methods are becoming increasingly popular these days [12–14] due to recent advances made in hardware technology which have made real-time computation of the power spectrum possible.

The synthesis-based method is tried out on real speech data and its performance is compared with that of the cepstrum method of pitch extraction [10]. The cepstrum method is used here for comparison because it can be implemented easily and provides high precision and fidelity when applied to good quality speech. The performance of the synthesis-based method is found to be better than that of the cepstrum method of pitch extraction.

Synthesis-based method of pitch extraction

In the synthesis-based method of pitch extraction, the speech signal is assumed to be the output of an all-pole filter $H(z)$ which is excited either by a periodic impulse train (for voiced speech segments), or by random noise (for unvoiced speech segments). Thus, the power spectrum $S(f)$ of the input speech segment, obtained by taking the discrete Fourier transform (DFT), can be decomposed as

$$S(f) = H(f)E(f), \tag{1}$$

where $E(f)$ is the power spectrum of the excitation signal.

For synthesizing the power spectrum, the input speech segment is assumed to be voiced and its fundamental frequency is hypothesized as being $\tilde{F}_0$. The transfer function $H(z)$ of the all-pole filter is estimated from the input speech segment by using linear prediction analysis [15]. The synthesized power spectrum $\tilde{S}(f, \tilde{F}_0)$ is computed by using the relation

$$\tilde{S}(f, \tilde{F}_0) = H(f)\,\tilde{E}(f, \tilde{F}_0), \tag{2}$$

where

$$\tilde{E}(f, \tilde{F}_0) = \begin{cases} 1, & \text{for } f = \tilde{F}_0, 2\tilde{F}_0, 3\tilde{F}_0, \cdots, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

A number of power spectra are synthesized for different values of $\tilde{F}_0$, the hypothesized fundamental frequency, and compared with the input power spectrum $S(f)$. For comparison, the average magnitude (AM) difference between the input power spectrum and the synthesized power spectrum is used as a measure of dissimilarity between the two spectra. [1] This is defined as follows:

$$D(\tilde{F}_0) = \frac{1}{F_f}\int_0^{F_f} |d(f, \tilde{F}_0)|\,\mathrm{d}f, \tag{4}$$

where F_f is the folding frequency and $d(f, \tilde{F}_0)$ the difference between the two spectra at frequency f, i.e.,

$$d(f, \tilde{F}_0) = \log S(f) - \log \tilde{S}(f, \tilde{F}_0). \tag{5}$$

It may be noted that instead of using the power spectra $S(f)$ and $\tilde{S}(f, \tilde{F}_0)$, the log power spectra are used in eq. (5) to define the difference $d(f, \tilde{F}_0)$. This is done to eliminate the effect of formant structure on pitch estimation.

When the input speech segment is voiced and its fundamental frequency is equal to the hypothesized fundamental frequency $\tilde{F}_0$, the input power spectrum $S(f)$ and the synthesized power spectrum $\tilde{S}(f, \tilde{F}_0)$ have the corresponding harmonic peaks at identical positions. For other values of $\tilde{F}_0$, the harmonic peaks of the two spectra do not coincide. Thus, the AM difference function $D(\tilde{F}_0)$ shows a distinct minimum for voiced speech segments (see Fig. 1). The position of this minimum (denoted by $\tilde{F}_{0\,\min}$) corresponds to the estimated fundamental frequency. For unvoiced speech segments, no such minimum exists. The AM difference function $D(\tilde{F}_0)$ exhibits random fluctuations around the mean value for such segments (see Fig. 2).

For making a voiced/unvoiced decision about a speech segment, the ratio of the mean and minimum values of $D(\tilde{F}_0)$ (which will be referred to as

[1] Instead of using the root mean square difference between the two spectra for comparison, the AM difference measure is used here to save computation time.

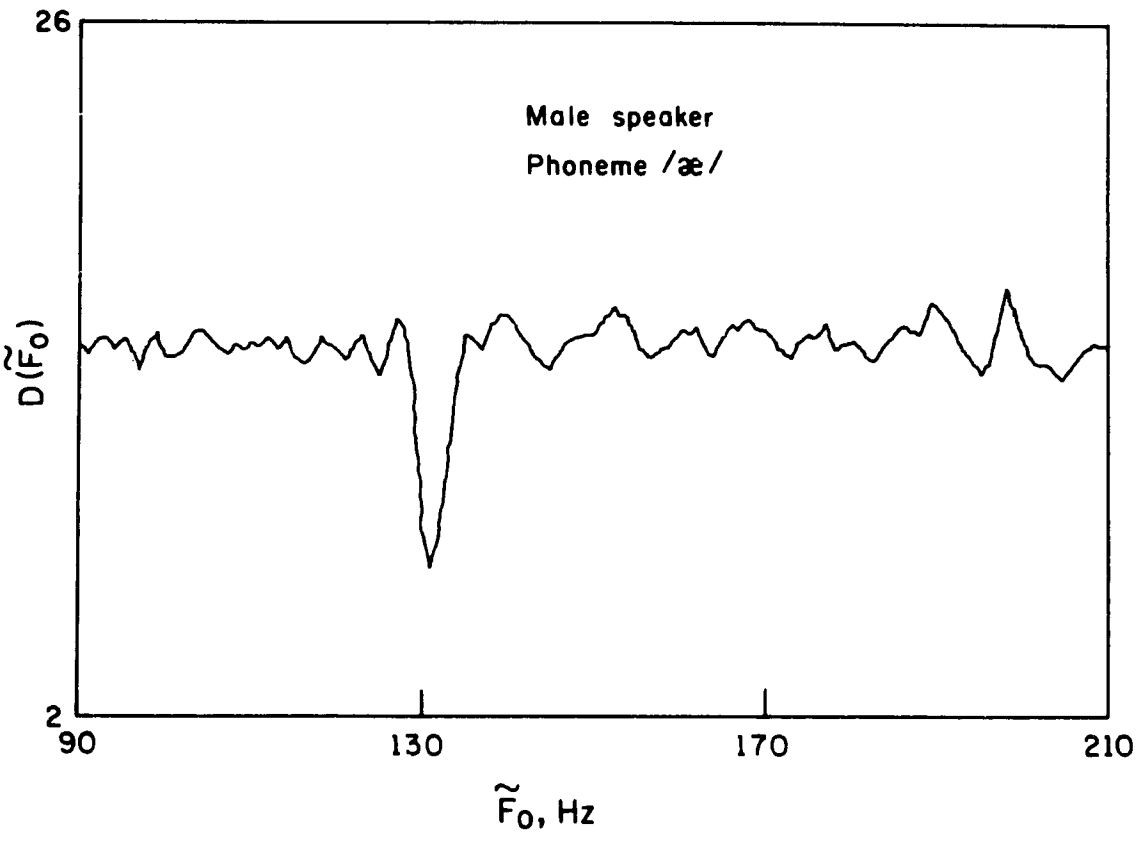

Fig. 1. AM difference measure, $D(\tilde{F}_0)$, as a function of the hypothesized fundamental frequency $\tilde{F}_0$ for a segment of vowel sound /æ/ (male speaker).

the mean-to-minimum ratio hereafter) is computed and compared with a threshold value R_T (which can be fixed empirically from the data). If the mean-to-minimum ratio is greater than the threshold value R_T, the speech segment is considered to be voiced and the value of $\tilde{F}_0$ at the minimum (i.e., $\tilde{F}_{0\,min}$) is taken to be the estimated fundamental frequency. If this difference is less than R_T, the speech segment is considered to be unvoiced.

For computing the AM difference $D(F_0)$, using

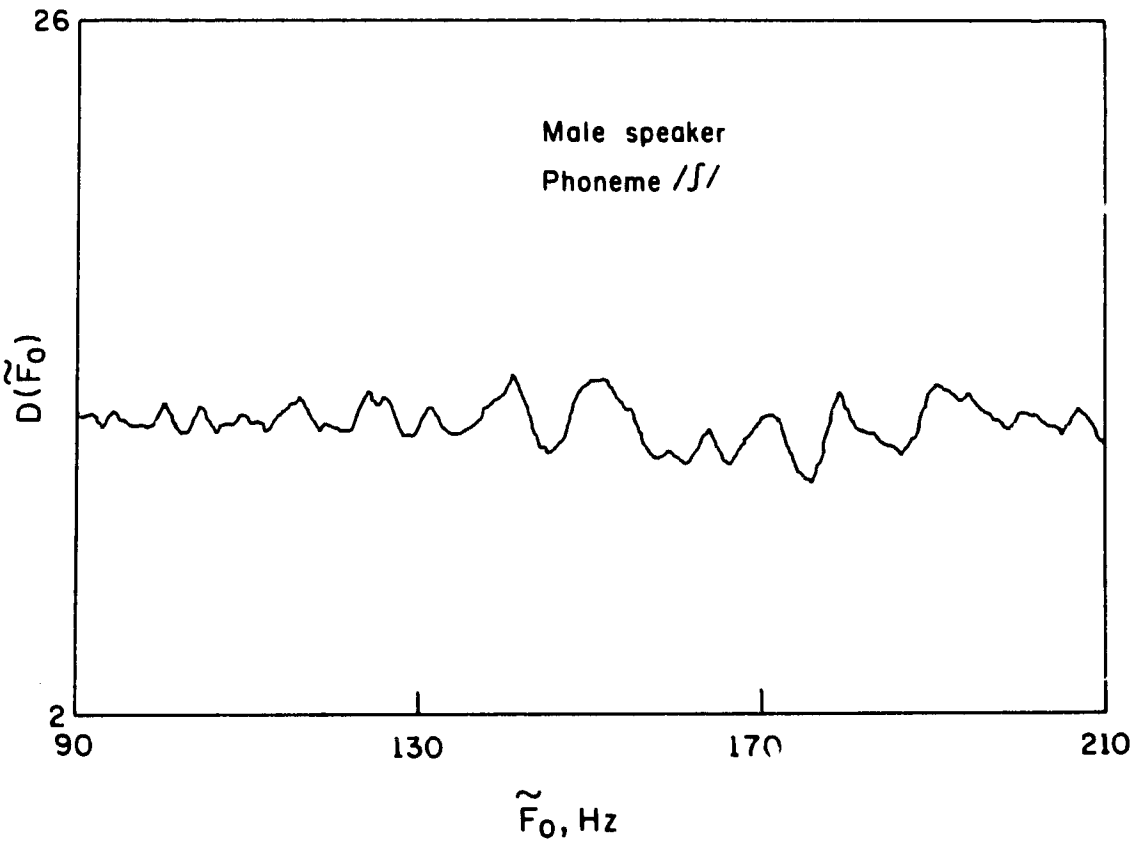

Fig. 2. AM difference measure, $D(\tilde{F}_0)$, as a function of the hypothesized fundamental frequency $\tilde{F}_0$ for a segment of fricative sound /ʃ/ (male speaker).

eqs. (4) and (5), the synthesized power spectrum $\tilde{S}(f, \tilde{F}_0)$ is required for all the values of f in the interval $[0, F_f]$. This involves extensive computation. In order to make the method computationally efficient, we use only the differences at harmonic peaks between $\log S(f)$ and $\log \tilde{S}(f, \tilde{F}_0)$ to compute the AM difference $D(\tilde{F}_0)$, i.e.,

$$D(\tilde{F}_0) = \frac{1}{K} \sum_{k=1}^{K} |\log S(k\tilde{F}_0) - \log \tilde{S}(k\tilde{F}_0, \tilde{F}_0)|, \quad (6)$$

where K $(= F_f/\tilde{F}_0)$ is the number of harmonic peaks in the interval $[0, F_f]$. Since, from eqs. (2) and (3), $\tilde{S}(f, \tilde{F}_0)$ and $H(f)$ are equal at harmonic peaks, eq. (6) can be rewritten as

$$D(\tilde{F}_0) = \frac{1}{K} \sum_{k=1}^{K} |\log S(k\tilde{F}_0) - \log H(k\tilde{F}_0)|. \quad (7)$$

The synthesis-based method of pitch extraction is shown in Fig. 3 in the form of a block diagram. The computational steps of the method are described below.

(1) Take a speech segment $x(n)$, $0 \leqslant n \leqslant N-1$, where N is the number of samples in the speech segment.

(2) Multiply $\{x(n)\}$ by the Hamming window $\{w(n)\}$, i.e.,

$$s(n) = x(n)w(n), \quad 0 \leqslant n \leqslant N-1. \quad (8)$$

(3a) Append the sequence $\{s(n)\}$ with $(512 - N)$ zeroes and compute a 512-point DFT by using a fast Fourier transform (FFT) algorithm. Compute the power spectrum $S(f)$.

(3b) Estimate the parameters $(G, a_1, a_2, \cdots, a_M)$ of the M-pole filter

$$H(z) = G/\left(1 + \sum_{k=1}^{M} a_k z^{-k}\right)$$

from the Hamming-windowed speech sequence $\{s(n)\}$ using the autocorrelation method of linear prediction [15]. The autocorrelation method of linear prediction is used here for analysis because it is computationally more efficient than other methods of linear prediction analysis and always guarantees the stability of the estimated all-pole filter [15]. For computing the power spectrum $H(f)$, the transfer function $H(z)$ is evaluated at 512 equally spaced points on the unit circle by using a FFT algorithm.

(4) Compute the AM difference $D(\tilde{F}_0)$, using eq. (7), for as many values of the hypothesized fundamental frequency $\tilde{F}_0$ as required by the resolution considerations in the interval $[F_a, F_b]$, where F_a and F_b are the minimum and maximum values, respectively, of the expected fundamental frequency for a given speaker.

(5) Find the minimum $(\tilde{F}_{0\,\min})$ in the difference function $D(\tilde{F}_0)$. Compute the mean value $\bar{D}$ of the AM difference function $D(\tilde{F}_0)$ from the relation

$$\bar{D} = \frac{1}{(F_b - F_a)} \int_{F_a}^{F_b} D(\tilde{F}_0)\, d\tilde{F}_0. \quad (12)$$

Compute the mean-to-minimum ratio $(\bar{D}/D(\tilde{F}_{0\,\min}))$ and compare it with the threshold value R_T to make the voiced/unvoiced decision. If the segment is voiced, take the value $\tilde{F}_{0\,\min}$ as the estimated value of fundamental frequency.

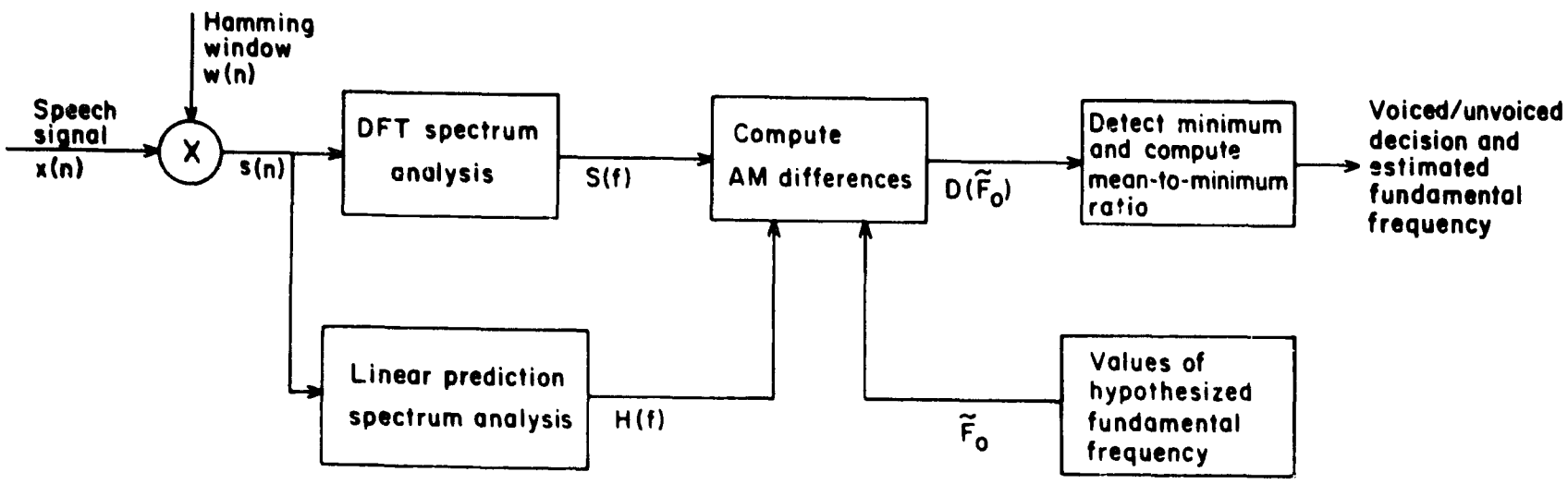

Fig. 3. Block diagram of the synthesis-based method of pitch extraction.

Performance of the method

Speech data used for evaluating the performance of the synthesis-based pitch extraction method consist of five sentences spoken by four speakers (two male and two female), and ten Hindi vowels (/i/, /I/, /e/, /æ/, /ʌ/, /a/, /ɔ/, /o/, /U/ and /u/) spoken 30 times in /b/-vowel-/b/ context by three speakers (two male and one female). The five sentences used here are given below.

(1) The way to save money is not to spend much.

(2) Shut the hatch before the waves push it in.

(3) The odour of spring makes young hearts jump.

(4) Crack the walnut with your sharp side teeth.

(5) He offered proof in the form of a large chart.

Recording of these utterances is done in an ordinary office room. The speech signal is digitized at a sampling rate of 10 kHz by means of a 12-bit analog-to-digital converter. A lowpass filter with a cutoff frequency of 4 kHz is used as a dealiasing filter.

For evaluating the performance of the synthesis-based pitch extraction method, three types of errors are studied: voiced/unvoiced decision errors, gross errors and fine errors. Voiced/unvoiced decision errors include all those cases where voiced speech segments are detected by the pitch extraction method as unvoiced and vice versa. Gross errors are defined as follows. Let $F_{0a}(i)$ is the actual fundamental frequency of the ith segment and $F_{0e}(i)$ the estimated fundamental frequency. If the error $|F_{0a}(i) - F_{0e}(i)|$ is more than 10% of $F_{0a}(i)$, the pitch extraction method is considered to have made a gross error for the ith segment. Otherwise, the error $F_{0a}(i) - F_{0e}(i)$ is considered to be the fine error.

Twenty spoken sentences (5 sentences × 4 speakers) are used for detecting voiced/unvoiced decision errors. The 30 ms steady-state vowel segments excised from the centre of 900 spoken /b/-vowel-/b/ syllables (10 vowels × 30 repetitions × 3 speakers) are used for detecting the gross errors and for computing the root mean square (RMS) value of the fine errors.

The synthesis-based pitch extraction method is used to analyse the speech signal of spoken sentences every 10 ms and each time a 30 ms segment is used for analysis. For synthesizing the log power spectra, a 10-pole filter is used. The values of fundamental frequency are hypothesized at a spacing of 1 Hz within the interval $[F_a, F_b]$ and the AM differences, $D(\tilde{F}_0)$, are computed for these values of $\tilde{F}_0$.[2] Position of the minimum, $\tilde{F}_{0\,\min}$, is detected in the interval $[F_a, F_b]$. For analysing a speech utterance, the interval $[F_a, F_b]$ is initially fixed to range from 80 Hz to 350 Hz to include all the possible values of fundamental frequency of male and female speakers. This interval is used until ten segments are detected as voiced. The average value, $\bar{F}_0$, of fundamental frequency is computed from these ten voiced segments and the interval $[F_a, F_b]$ is then reduced from [80,350] to $[0.6\bar{F}_0, 1.4\bar{F}_0]$ for the analysis of the remaining utterance. This is done to reduce the computation time.

In order to get an idea about the performance of the synthesis-based pitch extraction method, we compare its performance with that of the cepstrum method of pitch extraction. The cepstrum method of pitch extraction is used here for comparison because it is known to provide high precision and fidelity when applied to good quality speech and, hence, has been used in the past as a standard for evaluating the performance of the pitch extraction methods [16–18].

Table 1 gives the number of voiced/unvoiced decision errors made by the synthesis-based method and the cepstrum method for each of the four speakers. Here, the first two speakers are male and the remaining two are female. It can be seen from this table that for each of the four speakers the number of voiced/unvoiced decision errors for the synthesis-based method is consistently less than that for the cepstrum method.

[2] It may be noted that the power spectra $S(f)$ and $H(f)$ computed in the steps (3a) and (3b), respectively, of the previous section have the resolution of 19.53 Hz. But, for computing the AM differences $D(\tilde{F}_0)$, these spectra should have a resolution of 1 Hz which is accomplished here by doing linear interpolation. Linear interpolation can be done either on the complex spectrum or on the log power spectrum. However, we have done it on the log power spectrum to reduce the computation time.

Table 1
Voiced/unvoiced decision errors for the synthesis-based method and the cepstrum method

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Total
No. of segments	2110	1918	2014	1822	7864
No. of voiced/unvoiced errors made by					
(a) Synthesis-based method	80	71	112	74	337
(b) Cepstrum method	86	89	123	102	400

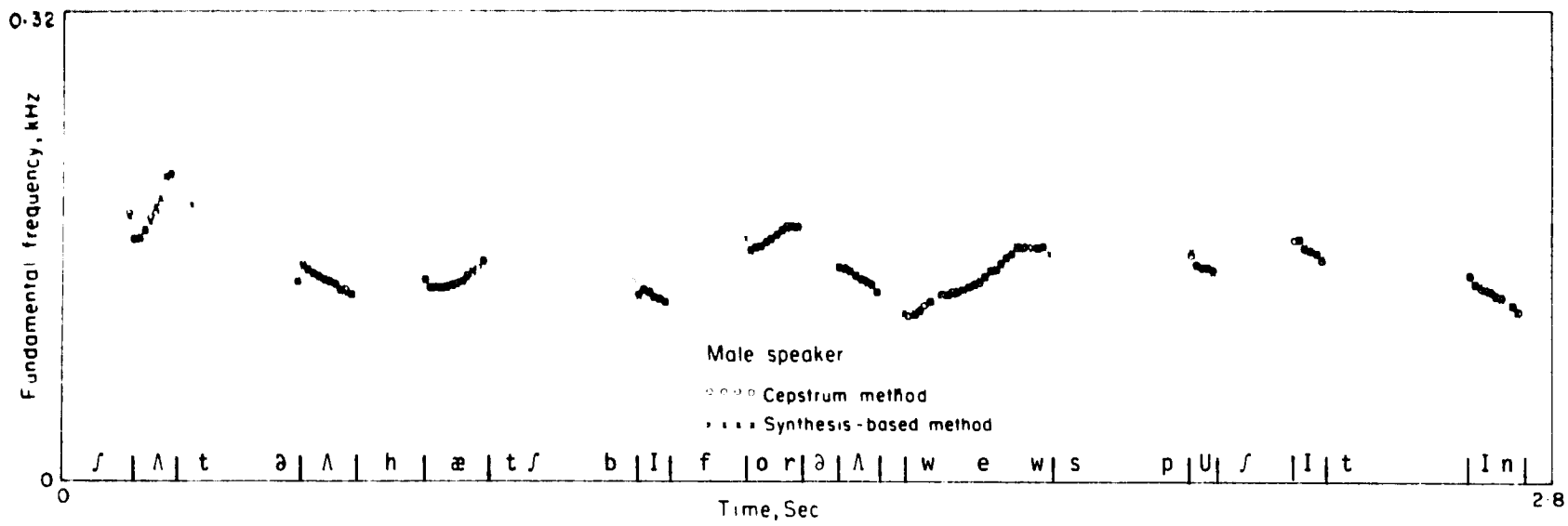

Fig. 4. Pitch contours of the speech utterance "Shut the hatch before the waves push it in" obtained by using the synthesis-based method and the cepstrum method (male speaker).

Also, out of a total of 7864 segments, the synthesis-based method makes wrong voiced/unvoiced decision for 4.3% of the segments, while the cepstrum method makes these wrong decisions for 5.1% of the segments. We show in Fig. 4 the pitch contours of the utterance 'Shut the hatch before the waves push it in' obtained by using the synthesis-based method and the cepstrum method for a male speaker. Similar contours are shown in fig. 5 for a female speaker. It can be seen from these figures that the number of voiced/unvoiced decision errors for the synthesis-based method is less than that for the cepstrum method, specially for the female speaker. Also, the fundamental frequency estimates for voiced speech segments by these two methods are quite close to each other.

Table 2 lists the values of gross error and fine error (RMS) for the synthesis-based method and the cepstrum method. [3] It can be seen from this table that though neither of the two methods give any gross error, the RMS value of fine error for the synthesis-based method is comparatively less than that for the cepstrum method.

So far we have studied the performance of the synthesis-based pitch extraction method for normal speech and found it better than that of the cepstrum method. Now, we shall compare the computation times taken by the two methods. It can be seen from the block diagram of the synthesis-based method, shown in Fig. 3, that the DFT spectrum analysis and linear prediction spectrum analysis of the speech signal can be done in parallel (simultaneously). This was simulated in FORTRAN on a general purpose computer, DEC System 10, and the average computation time taken by the synthesis-based method (971.4 ms) to analyse one speech segment was found to be comparable to that taken by the cepstrum method (978.4 ms).

The synthesis-based pitch extraction method is also tried out on synthetic vowel segments where the fundamental frequency is known a priori. The fundamental frequency estimates of these syn-

[3] The RMS value of fine error (1.1131 Hz) for the synthesis-based method, listed in Table 2, is obtained after doing a 3-point parabolic interpolation around the minimum of $D(\tilde{F}_0)$ detected in the interval $[F_a, F_b]$. When this parabolic interpolation is not done, this value increases to 1.1605 Hz.

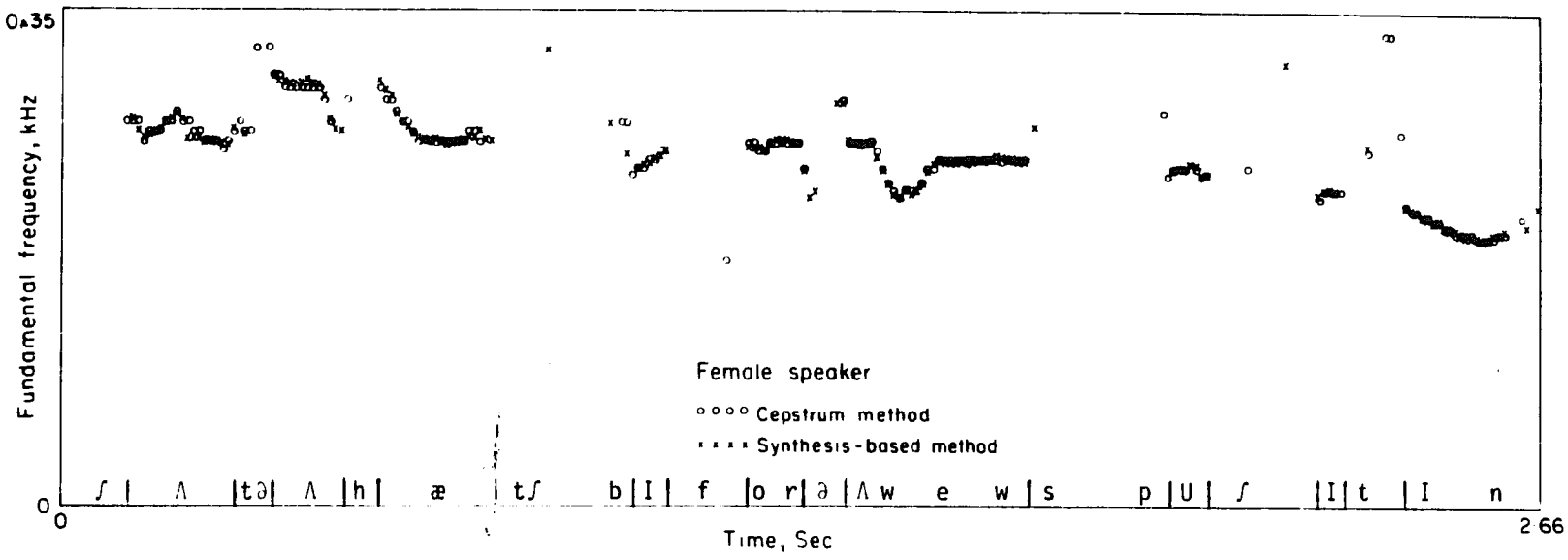

Fig. 5. Pitch contours of the speech utterance "Shut the hatch before the waves push it in" obtained by using the synthesis-based method and the cepstrum method (female speaker).

thetic segments are found to be identical to their true values.

We have also studied the effect of additive white noise, bandpass filtering and reduction in the order of the all-pole filter on the performance of the synthesis-based method of pitch extraction. For this purpose, we use only the first two sentences spoken by two speakers (one male and one female) and all the 900 steady-state vowel segments.

The performance of the synthesis-based pitch extraction method under different noise conditions characterised by signal-to-noise ratio is shown in Table 3. The signal-to-noise ratio in dB is denoted here by SNR and is defined by

$$\mathrm{SNR} = 10 \log \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} w^2(n)}$$

Table 2
Gross error and fine error (RMS) for the synthesis-based method and the cepstrum method

	Gross error	Fine error (RMS)
Synthesis-based method	0%	1.1131 Hz
Cepstrum method	0%	1.2450 Hz

where $s(n)$ is the speech signal, $w(n)$ the white Gaussian noise added to the speech signal and N is the duration of the speech segment. SNR = ∞ means here that no noise is added to the speech signal. It can be seen from the Table 3 that the performance of the synthesis-based method goes down with the addition of noise. However, the method gives less than 10% voiced/unvoiced decision error and no gross error, even at SNR = 15 dB.

The effect of bandpass filtering on the performance of the synthesis-based method is shown in Table 4. A 10th order infinite impulse response digital filter with cutoff frequencies of 300 Hz and 3200 Hz is used here for bandpass filtering. It can

Table 3
Effect of additive white noise on pitch extraction performance of the synthesis-based method

SNR	Voiced/unvoiced decision error	Gross error	Fine error (RMS)
∞	5.1%	0%	1.1131 Hz
30	5.9%	0%	1.1371 Hz
25	6.2%	0%	1.1336 Hz
20	6.9%	0%	1.3214 Hz
15	9.4%	0%	1.5269 Hz
10	[illegible]%	0.33%	1.8991 Hz
5	20.8%	1.67%	2.7458 Hz
0	28.0%	15.00%	3.7359 Hz

Table 4
Effect of bandpass filtering on pitch extraction performance of the synthesis-based method

	Voiced/ unvoiced decision error	Gross error	Fine error (RMS)
Before bandpass filtering	5.1%	0%	1.1131 Hz
After bandpass filtering	10.2%	0%	1.1279 Hz

be seen from Table 4 that the effect of bandpass filtering on the performance of the method is quite severe. The voiced/unvoiced decision errors increase by more than 5% because of bandpass filtering.

The effect of reducing the order of the all-pole filter on the performance of the synthesis-based method is shown in Table 5. When the order of the all-pole filter is reduced from 10 to 2, the voiced/unvoiced decision errors and fine error (RMS) increase, while the gross error remains the same throughout (i.e., 0%).

At the time of reviewing, the referees brought to our notice the spectral comb method of pitch extraction [14]. This method also works in the frequency domain and relies on the correlation between the power spectrum of the speech signal and a spectral comb with teeth of decreasing amplitude and variable teeth interval. The location of maximum of this correlation function is considered in the spectral comb method as the estimated value of fundamental frequency.[4] Though the procedure of voiced/unvoiced detection is not given in the paper on spectral comb method [14], we used the ratio of maximum and mean values of the correlation function to make voiced/unvoiced decision and compared experimentally the performance of this method with that of the synthesis-based method proposed in the present paper. The results of the comparison are shown in Table 6. It can be seen from this table that the synthesis-based method performs comparatively better than the spectral comb method.

Table 5
Effect of reducing the order of all-pole filter on pitch extraction performance of the synthesis-based method

Order of the all-pole filter	Voiced/ unvoiced decision error	Gross error	Fine error (RMS)
10	5.1%	0%	1.1131 Hz
8	5.4%	0%	1.1313 Hz
6	6.0%	0%	1.1403 Hz
4	7.2%	0%	1.1541 Hz
2	7.7%	0%	1.1623 Hz

Table 6
Comparison of synthesis-based method and spectral comb method

	Voiced/ unvoiced decision error	Gross error	Fine error (RMS)
Synthesis-based method	5.1%	0%	1.1131 Hz
Spectral comb method	8.8%	5.67%	1.2898 Hz

Conclusion

A synthesis-based method of pitch extraction is proposed. The method synthesizes a number of log-powder spectra for different values of fundamental frequency and compares them with the log-power spectrum of the input speech segment. The AM difference between the two spectra is used for comparison. The value of fundamental frequency that gives the minimum AM difference between the synthesized spectrum and the input spectrum is chosen as the estimated value of fundamental frequency. The method is tried out on real speech data and its performance is found

[4] It may be noted that the spectral comb method does not eliminate the effect of formant structure on pitch estimation, while the present method eliminates this effect.

to be better than the cepstrum method of pitch extraction. The method is also studied as to its performance under different noise conditions and found to perform reasonably well, even at 15 dB signal-to-noise ratio.

Acknowledgement

The authors are thankful to referees for making constructive comments and to Mr. Dinesh Sharma for translating the paper [14] from French to English.

References

[1] J.L. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, New York, 1972.

[2] J.D. Markel and A.H. Gray, Jr. "A linear prediction vocoder simulation based upon the autocorrelation method", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-22, No. 2, April 1974, pp. 124–134.

[3] W.A. Lea, M.F. Medress and T.E. Skinner, "A prosodically guided speech understanding strategy", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 1, Febr. 1975, pp. 30–38.

[4] B.S. Atal, "Automatic speaker recognition based on pitch contours", *J. Acoust. Soc. Am.*, Vol. 52, No. 6, Dec. 1972, pp. 1687–1697.

[5] A.E. Rosenberg and M.R. Sambur, "New techniques for automatic speaker verification", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 2, April 1975, pp. 169–176.

[6] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg and C.A. McGonegal, "A comparative performance study of several pitch detection algorithms", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, No. 5, Oct. 1976, pp. 399–418.

[7] C.A. McGonegal, L.R. Rabiner and A.E. Rosenberg, "A subjective evaluation of pitch detection methods using LPC synthesized speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 3, June 1977, pp. 221–229.

[8] C.M. Harris and M.R. Weiss, "Pitch extraction by computer processing of high-resolution Fourier analysis data", *J. Acoust. Soc. Am.*, Vol. 35, No. 3, Mar. 1963, pp. 339–343.

[9] M.R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurements", *J. Acoust. Soc. Am.*, Vol. 43, No. 4, April 1968, pp. 829–834.

[10] A.M. Noll, "Cepstrum pitch determination", *J. Acoust. Soc. Am.*, Vol. 41, No. 2, Febr. 1967, pp. 293–309.

[11] E. Terhardt, "Calculating virtual pitch", *Hearing Res.*, Vol. 1, No. 2, 1979, pp. 155–182.

[12] S. Seneff, "Real-time harmonic pitch detector", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-26, No. 4, Aug. 1978, pp. 358–365.

[13] T.V. Sreenivas and P.V.S. Rao, "Pitch extraction from corrupted harmonics of the power spectrum", *J. Acoust. Soc. Am.*, Vol. 65, No. 1, Jan. 1979, pp. 223–228.

[14] P. Martin, "Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne", Actes des 12emes Journées d'Etude sur la Parole, GALF, Montréal, May 1981, pp. 221–232.

[15] J. Makhoul, "Linear prediction: A tutorial review", *Proc. IEEE*, Vol. 63, No. 4, April 1975, pp. 561–580.

[16] T.V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-23, No. 6, Dec. 1975, pp. 562–570.

[17] D.H. Friedman, "Pseudo-maximum-likelihood speech pitch extraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 3, June 1977, pp. 213–221.

[18] D.H. Friedman, "Multidimensional pseudo-maximum-likelihood pitch estimation", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-26, No. 3, June 1978, pp. 185–196.