Speech Communication 3 (1984) 253–259
North-Holland

SHORT COMMUNICATION

A COMPARATIVE PERFORMANCE EVALUATION OF PITCH ESTIMATION METHODS FOR TDHS/SUB-BAND CODING OF SPEECH

K.K. PALIWAL* and A.I. AARSKOG
Electronics Research Laboratory (ELAB), University of Trondheim, Trondheim-NTH, Norway

Received 23 May 1984
Revised 26 September 1984

Abstract. The time-domain harmonic-scaling (TDHS) algorithm provides a computationally efficient method (suitable for real-time implementation) for speech bandwidth compression and expansion. Pitch estimation is an important operation in the TDHS process. In the present paper, we study a TDHS/sub-band coding system for speech operating at 16 kbits/s and investigate the relative effectiveness of five different pitch estimation methods (the autocorrelation method, the cepstrum method, the simplified inverse filtering technique, the average magnitude difference function method and the maximum likelihood method). A formal listening test using 17 human listeners is conducted for their comparative performance evaluation. The average magnitude difference function method was found to be the best pitch estimation method for TDHS/sub-band coding.

Zusammenfassung. Der Algorithmus zur TDHS-Zeitkompression (Time-Domain Harmonic Scaling) stellt ein numerisch leistungsfähiges Verfahren zur Bandbreitenkompression und -dehnung dar (er kann in Echtzeit implementiert werden). Die Bestimmung der Grundfrequenz ist eine wichtige Etappe innerhalb des TDHS-Verfahrens. In diesem Beitrag untersuchen wir ein TDHS-Teilbandkodierungssystem, welches mit 16 kbits/s arbeitet, und vergleichen die Leistungsfähigkeit von fünf Grundfrequenzbestimmungsmethoden (Autokorrelationsmethode, Cepstrummethode, SIFT-Methode, AMDF-Methode, Methode der "Maximum Likelihood"). Ein psychoakustischer Test mit 17 Hörern erlaubt den Vergleich der Leistungsfähigkeit der verschiedenen Methoden. Die AMDF-Methode erwies sich als am besten geeignetes Verfahren im Rahmen eines TDHS-Teilbandkodierungssystems.

Résumé. L'algorithme de transformation d'échelle de fréquence dans le domaine temporel (TDHS) fournit une méthode numériquement efficace (appropriée à une implantation en temps réel) pour la compression et l'expansion de la largeur de la bande fréuentielle de la parole. L'estimation du pitch représente une opération importante de la procédure TDHS. Dans cet article, nous étudions un système de codage TDHS/bandes-partielles de la parole opérant à 16 K bits/s, et nous investigons l'efficacité relative de cinq méthodes différentes d'estimation du pitch (la méthode d'autocorrélation, la méthode du cepstre, la technique du filtrage inverse simplifié, la méthode AMDF et la méthode du maximum de vraisemblance. Un test d'audition formel avec 17 auditeurs a été entrepris pour évaluer les performances comparatives des cinq méthodes considérées. La méthode AMDF se dégage comme étant la meilleure procédure d'estimation du pitch pour le codage par TDHS/bandes-partielles.

Keywords. Speech coding, TDHS algorithm, sub-band coder, pitch estimation.

1. Introduction

The time-domain harmonic-scaling (TDHS) algorithm provides a computationally efficient method (suitable for real-time implementation) for speech bandwidth compression and expansion [1]. The TDHS system (with a 2:1 compression at the transmitting end and a 1:2 expansion at the receiving end) is known to cause only slight degradation in speech quality and, hence, has been used in the past by a number of authors in cascade with other speech coders (such as CVSD coder [2], adaptive residual coder [3], sub-band coder [4–6], adaptive transform coder [4] and ADPCM coder [7]).

All these authors agree that pitch estimation is

* Present address: Computer Systems and Communications Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India.

the most fundamental and crucial operation in the TDHS coding process and the quality of the resulting speech depends very much on pitch estimation accuracy resulting from the pitch estimation method used. But, surprisingly, they have made arbitrary and different choices of the pitch estimation methods in their implementations (for example, the autocorrelation method being used in [1,2,5,6,7], the cepstrum method in [4] and the average magnitude difference function method in [3]). The aim of the present paper is to evaluate in a systematic manner different pitch estimation methods as to their effect on the quality of speech resulting from the TDHS system. Since the TDHS system alone can not bring down the bit-rate sufficiently low for medium band speech coding (9.6 to 16 kbits/s), it is used in cascade with a sub-band coder in the present study.

There are some studies reported in the literature [8,9] where different pitch estimation methods are compared objectively as to their performance in terms of pitch estimation errors (such as gross error, fine error, voiced-to-unvoiced error and unvoiced-to-voiced error). The objective results obtained from these studies are undoubtedly useful for initial selection of a pitch estimation method for the TDHS system. However, the ultimate decision about the pitch estimation method must come from human listeners who only can decide about the quality of speech resulting from the TDHS system. McGonegal et al. [10] have conducted a subjective evaluation of different pitch estimation methods for the LPC vocoder. However, their findings can not be extended to the TDHS system because it differs significantly from the LPC vocoder in terms of two important aspects as far as the pitch estimation is concerned. Firstly, for the TDHS algorithm there is no necessity of making any voiced/unvoiced decisions because any random pitch estimate within a limited range provided by the pitch estimation method for unvoiced segments causes no noticable difference in the quality of the resulting speech [1–6]. Secondly, the effect of pitch-doubling errors causes much less degradation in speech quality for the TDHS system than for the LPC vocoder [1–6,8,10]. Thus, the task of the pitch estimation method for the TDHS system is relatively easier than that for the LPC vocoder. Because of these differences, the results of subjective evaluation of the pitch estimation methods for the TDHS system might be different from the LPC vocoder. This motivated us to undertake this study of comparative performance evaluation of different pitch estimation methods for TDHS/sub-band coding of speech.

The organization of this paper is as follows. We first describe briefly in Section 2 the TDHS/sub-band coder. We then give in Section 3 a brief description of the pitch estimation methods evaluated in the present study. The results of performance evaluation are described in Section 4. Section 5 presents the conclusions of this study.

2. The TDHS/sub-band coder

Since the TDHS system alone can not reduce the bitrate of the speech signal suitable for medium-band coding, it has to be combined with some other speech coder. In the present study, we have used it in cascade with the sub-band coder because of the following two reasons. Firstly, the sub-band coder exploits the redundancy properties of speech different from those used by the TDHS system. (For example it uses the properties of temporal nonstationarity, spectral-formant structure and auditory masking of speech production and perception [11], while the TDHS system exploits the pitch structure of the speech signal through its pitch-adaptive operation.) Secondly, the sub-band coder introduces degradations in the speech signal which are perceptually different from those introduced by the TDHS system. (For example, degradations introduced by the sub-band coder appear in the form of quantizing noise and intermodulation distortion, while those introduced by the TDHS system appear in the form of reverberance [5].) This complementary nature of the two systems allows their cascaded system to have more effective overall compression and perceptually less objectionable degradation in speech quality.

The block diagram of the combined TDHS/sub-band coder is shown in Fig. 1. The coder is used in the present study for transmitting speech at 16 kbits/s. It accepts at its input digitized speech at 128 kbits/s (digitized by 16-bit analog-to-digital

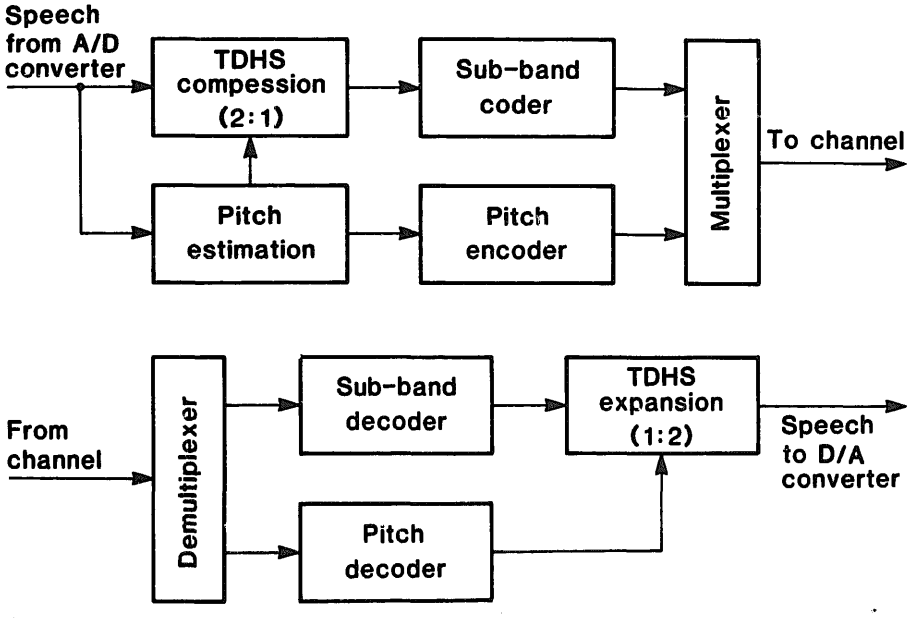

Fig. 1. Block diagram of the TDHS/sub-band speech coder.

converter at a sampling rate of 8 kHz). The digitized speech is analyzed by the pitch estimation method every 16 ms and each time a speech segment of 32 ms is used for pitch estimation. The pitch estimator does not make any voiced/unvoiced decision. For voiced segments, it gives the pitch estimate which is a measure of speech periodicity. For unvoiced segments, the estimated pitch value corresponds to long-term correlation existing in the speech signal, though this correlation might be very low.

The TDHS coder uses these values of pitch estimates to compress speech signal by a factor of 2 in pitch-adaptive manner (i.e., it compresses two pitch periods of speech into one pitch period) [1]. The compressed speech is then encoded by a sub-band coder of the type described in [12]. The sub-band coder consists of a filter-bank of 14 quadrature mirror bandpass filters covering uniformly the frequency range 0-3500 Hz. The sub-band filter outputs are quantisized with an adaptive bit-allocation scheme. For this, these filter outputs are organized in blocks of length 16 ms and a blockwise estimate of power is made for each sub-band. These power estimates are used for normalizing the sub-band outputs and for determining the number of bits to be allocated to each sub-band. The normalized sub-band outputs are quantisized by a Max-quantisizer [13] which assumes these outputs to be Gaussian with zero mean and unit variance.

The TDHS/sub-band coder thus transmits at 16 kbits/s the quantized values of pitch, sub-band powers and sub-band output signals. At the receiving end, a reverse process is used to generate speech signal from the received information. We have also studied another configuration of the TDHS/sub-band coder where the pitch information is not transmitted but extracted at the receiving end from the output of the sub-band decoder. Since this output of the sub-band decoder is a compressed version of the speech signal (which, in addition, has quantization errors and channel distortion), the pitch estimation on this signal is expected to be erroneous. We have observed through informal listening tests that this configuration (with pitch estimation at the receiving end) causes significant degradations in the perceived quality of speech.

3. The pitch estimation methods

There are many pitch estimation methods reported in the literature which are described in detail in a recently published book by Hess [14]. For

the purpose of present evaluation, we have selected the following five relatively well-known pitch estimation methods: 1) autocorrelation method, 2) cepstrum method, 3) simplified inverse filtering technique, 4) average magnitude difference function method and 5) maximum likelihood method. These methods are briefly described below.

1) Autocorrelation method (AUTOC)

Usually a centre-clipping operation is recommended as a preprocessing step in literature for reducing the effect of formant structure on pitch estimation accuracy by the autocorrelation method [15,16]. However, we have shown in our earlier studies [17,18] that centre-clipping deteriorates the pitch estimation performance of the autocorrelation method, specially for noisy speech. So, in the present study we use the autocorrelation method without any centre-clipping. The method uses a 32 ms speech segment with biased short-time autocorrelation estimator with a rectangular window function [19]. The autocorrelation function is computed over a range of 2.5 ms to 15 ms and the position of the maximum in the autocorrelation function is taken as the estimated value of the pitch period.

2) Cepstrum method (CEP) [20]

The speech signal of duration 32 ms is weighted by a Hamming window function and a 256-point cepstrum is computed. The location of the peak in the cepstrum over a range of 2.5 ms to 15 ms is considered as the estimated pitch period.

3) Simplified inverse filtering technique (SIFT) [21]

The 32 ms segment of speech is processed by an 8th order inverse filter and the residual signal obtained as the output of this filter is used to compute the short-time biased autocorrelation function over a range of 2.5 ms to 15 ms. The location of the maximum in this range is taken as the estimated pitch period. For estimating the coefficients of the inverse filter transfer function, an 8th order linear prediction analysis (using the autocorrelation method) is performed on the Hamming-windowed 32 ms segment of the speech signal.

4) Average magnitude difference function method (AMDF) [22]

The average magnitude difference function is computed by cross-differencing the first 17 ms of the speech segment with the whole 32 ms speech segment over a range of 2.5 ms to 15 ms. The pitch period is determined by locating the minimum of this function.

5) Maximum likelihood method (ML) [23]

The likelihood function (given by Eq. (12 b) of Reference [23]) is computed from the 32 ms segment of the speech signal over the range of 2.5 ms to 15 ms. The location of the peak of the likelihood function is considered as the estimated value of the pitch period.

4. Performance evaluation and results

For evaluating the performance of different pitch estimation methods for TDHS/sub-band coding of speech, we conducted a formal listening test. For this, speech data was acquired from two different English sentences, one spoken by a male speaker and the other by a female speaker. The speech signal was digitized at a sampling frequency of 16 kHz by a 16-bit analog-to-digital converter. A lowpass filter with a cutoff frequency of 7.1 kHz was used as an anti-aliasing filter prior to digitization. The signal was then decimated to a sampling frequency of 8 kHz using a Hamming filter with 512 coefficients and 6 dB cut-off frequency of 3.625 kHz. The digitized speech obtained in this manner will be referred to as the "normal" speech hereafter in this paper.

Since we were interested in the TDHS/sub-band coder for mobile telecommunications, we studied here four different types of speech: 1) normal speech, 2) telephone speech obtained by processing normal speech by a linear-phase bandpass filter representing the sending part of the intermediate Reference System [24], 3) tele-

phone speech at signal-to-noise ratio (SNR) = 10 dB obtained by adding a suitable amount of noise recorded from the inside of a moving car to telephone speech, and 4) telephone speech at SNR = 0 dB obtained again by adding the moving car noise to telephone speech.

For subjective evaluation of speech quality resulting from the different pitch estimation methods, we employed here a simple pair preference test where a test pair A-B of the two speech signals was presented to the human listeners. The listeners were asked to respond to either A or B depending on its perceived quality. In a test pair, the two speech signals differed only in terms of pitch estimation methods used in their coding. Listening test was done on 17 subjects in four sessions, each session devoted individually to one of the four types of speech. In each session 40 test pairs (all combinations and orders of two, from the five pitch estimation methods, male and female independently) were presented in random order to each subject through headphones.

The subjective scores of 17 listeners for the five pitch estimation methods are listed separately for each type of speech in Table 1. The maximum score possible for a particular pitch estimation method is 272 and it means that the method is preferred over all the other pitch estimation methods by all the listeners. It should be noted here that the pair preference test used here can only give the relative ranking of the pitch estimation methods, it does not give an absolute measure of speech quality.

The subjective scores (given in Table 1) are used to find the rankings of the different pitch estimation methods. These rankings are shown in Table 2. It can be seen from this table that the AMDF method is the only pitch estimation method which figures in the two top-ranked methods for all the four types of speech. Similarly, only the AMDF and the AUTOC methods figure in the three top-ranked pitch estimation methods for all the four types of speech. From this, we conclude that the AMDF method and the AUTOC method are the first and the second best pitch estimation methods, respectively, for the TDHS/sub-band coding of speech at 16 kbits/s.

A few other interesting observations about the effects of the telephone channel and the additive car noise can be made from Table 1. The telephone channel deteriorates the performance of the ML pitch estimation method most and the CEP pitch estimation method least. On the contrary, the additive car noise deteriorates the performance of the ML method least and the CEP method most.

Table 1
Subjective scores of 17 human listeners resulting from different pitch estimation methods for different types of speech

Pitch estimation method	Subjective score for			
	normal speech	telephone speech	telephone speech at SNR = 10 dB	telephone speech at SNR = 0 dB
AUTOC	133	132	163	176
CEP	132	171	80	15
SIFT	98	121	135	159
AMDF	162	149	161	170
ML	155	107	141	162

Table 2
Subjective rankings of the pitch estimation for the four types of speech

Ranking	For normal speech	For telephone speech	For telephone speech at SNR = 10 dB	For telephone speech at SNR = 0 dB
1	AMDF	CEP	AUTOC	AUTOC
2	ML	AMDF	AMDF	AMDF
3	AUTOC	AUTOC	ML	ML
4	CEP	SIFT	SIFT	SIFT
5	SIFT	ML	CEP	CEP

5. Conclusion

A comparative performance evaluation of the five different pitch estimation methods (AUTOC, CEP, SIFT, AMDF and ML) has been performed for TDHS/sub-band coding of speech at 16 kbits/s. A formal listening test was conducted for subjective evaluation of these methods. The AMDF method was found to be the best pitch estimation method for this purpose.

Table 3
Gross errors and fine errors (in %) made by the five different pitch estimation methods for four different types of speech

Pitch estimation method	Normal speech		Telephone speech		Telephone speech at SNR = 10 dB		Telephone speech at SNR = 0 dB	
	Gross error	Fine error	Gross error	Fine error	Gross error	Fine error	Gross error	Fine error
AUTOC	0.6	1.6	2.5	1.6	7.5	1.8	22.5	2.3
CEP	11.8	1.8	8.1	1.6	30.6	1.8	55.6	2.2
SIFT	11.2	1.7	11.2	1.6	20.0	2.0	41.2	2.1
AMDF	6.2	1.3	7.5	1.4	12.5	1.5	38.7	2.1
ML	5.6	1.7	15.0	1.6	18.1	1.8	33.7	1.9

1. Appendix

Though the aim of the present paper is to study the relative effectiveness of different pitch estimation methods on the subjective quality of speech resulting from the TDHS/sub-band coder, we also present here some objective results about these pitch estimation methods. These objective results are reported here in terms of pitch estimation errors and segmental SNR of the reconstructed speech signal at the receiving end.

As mentioned earlier, the TDHS system does not require voiced/unvoiced decision about the speech segments. It treats all the speech segments (whether voiced or unvoiced) alike and gives the best estimate of the pitch period. Som we have measured here the pitch estimation errors only in terms of gross errors and fine errors for the voiced segments. The gross errors are defined here as those pitch estimation errors where the estimated value of the pitch period differs from the actual value by more than 10% of the actual pitch period [8]. Other errors are treated as fine errors and their absolute normalized values are averaged over all the voiced segments [8]. For evaluation purposes, the voiced segments were detected manually and their actual pitch periods were obtained from the visual display of their waveforms on a Tektronix graphics terminal with the help of a manually controlled cursor. The gross and fine errors for the five different pitch estimation methods are listed separately for four different types of speech in Table 3. It can be seen from this table that the AMDF and AUTOC methods result in the best performance. We also observe that

Table 4
Segmental SNR values (in dB) resulting from the five different pitch estimation methods for four different types of speech

Pitch estimation method	Segmental SNR for			
	Normal speech	Telephone speech	Telephone speech at SNR = 10 dB	Telephone speech at SNR = 0 dB
AUTOC	7.88	7.38	2.98	2.66
CEP	7.82	5.99	1.40	1.93
SIFT	0.02	6.50	−0.80	−1.41
AMDF	9.20	7.04	2.00	−0.49
ML	6.69	4.84	4.26	2.94

performance of the ML method is most severely deteriorated by the telephone channel, while the additive car noise deteriorates the performance of the CEP method most.

In order to measure the segmental SNR, the speech signal was divided into non-overlapping segments each of duration 16 ms. The SNR values were computed individually for each segment and then averaged over the entire speech utterance. The segmental SNR values for the different pitch estimation methods are listed in Table 4 separately for the four types of speech. It can be seen from this table that the AMDF and AUTOC methods result in the best performance for normal and telephone speech. We can also observe that the performance of the ML method is least degraded due to the addition of car noise to the speech signal.

References

[1] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, No. 2. Apr. 1979, pp. 121–133.

[2] D. Malah, "Combined time-domain harmonic compression and CVSD for 7.2 kbit/s transmission of speech signals", *Proc. ICASSP*, 1980, pp. 506–507.

[3] J.L. Melsa and A.K. Pande, "Mediumband encoding using time domain harmonic scaling and adaptive residual coding", *Proc. ICASSP*, 1981, pp. 603–606.

[4] D. Malah, R.E. Crochiere and R.V. Cox, "Performance of transform and subband coding systems combined with harmonic scaling of speech", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-29, No. 2, Apr. 1981, pp. 273–283.

[5] R.E. Crochiere, R.V. Cox, J.D. Johnston and L.A. Seltzer, "A 9.6 kb/s DSP speech coder", *Bell Syst. Tech. Journal*, Vol. 61, No. 9, Nov. 1982, pp. 2263–2288.

[6] R.V. Cox, R.E. Crochiere and J.D. Johnston, "Real-time implementation of time domain harmonic scaling of speech for rate modification and coding", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-31 No. 1, Feb. 1983, pp. 258–272.

[7] M. Copperi, "A robust 4800 bit/s full-band speech coder", *CSELT Rapporto technici*, Vol. 11, No. 2, Apr. 1983, pp. 99–104.

[8] 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.

[9] K.K. Paliwal, "Comparative performance evaluation of different pitch estimation methods for noisy speech", *Acoustics Letters*, Vol. 6, No. 11, May 1983, pp. 164–166.

[10] 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.

[11] R.E. Crochiere, S.A. Weber and J.L. Flanagan, "Digital coding of speech in sub-bands", *Bell. Syst. Tech. Journal*, Vol. 55, No. 8, Oct. 1976, 1069–1085.

[12] T.A. Ramstad, "Sub-band coder with a simple adaptive bit-allocation algorithm: A possible candidate for digital mobile telephony?", *Proc. ICASSP*, 1982, pp. 203–207.

[13] J. Max, "Quantizing for minimum distortion", *IRE Trans. Inf. Theory*, Vol. IT-6, Mar. 1960, pp. 7–12.

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

[15] M.M. Sondhi, "New methods of pitch determination", *IEEE Trans. Audio Electroacoust.*, Vol. AU-16, June 1968, pp. 262–266.

[16] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, No. 1, Feb. 1977, pp. 24–33.

[17] K.K. Paliwal, "Effect of spectral flattening on the pitch estimation performance of the autocorrelation method for noisy speech", *Acoustics Letters*, Vol. 7, No. 5, Nov. 1983, pp. 74–76.

[18] K.K. Paliwal and A.I. Aarskog, "Some considerations about the use of the autocorrelation method of pitch estimation for TDHS/sub-band coding of speech", in preparation.

[19] K.K. Paliwal, "On the use of autocorrelation method of pitch estimation for noisy speech", *Acoustics Letters*, Vol. 7, No. 4, Oct. 1983, pp. 57–61.

[20] R.W. Schafer and L.R. Labiner, "System for automatic formant analysis of voiced speech", *J. Acoust. Soc. Amer.*, Vol. 47, No. 2, Feb. 1970, pp. 634–648.

[21] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, No. 5, Dec. 1972, pp. 367–377.

[22] M.J. Ross, H.L. Schaffer, A. Cohen, R. Freudberg and H.J. Manly, "Average magnitude difference function pitch extractor", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-22, No. 5, Oct. 1974, pp. 353–362.

[23] J.D. Wise, J.R. Caprio and T.W. Parks, "Maximum likelihood pitch estimation", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, No. 5, Oct. 1976, pp. 418–423.

[24] CCITT Recommendation, "Specifications for an Intermediate Reference System", *Yellow Book*, Vol. V, Rec. P. 48, ITV, Geneva, 1980.