Chapter 5

Classifier Based Voicing

5.1 Introduction

The PS SB-LPC was introduced in Chapter 4, this coder can produce high quality synthetic unquantised speech, however occasional artifacts are produced which limit the speech quality. It is believed that these problems are mainly caused by deficiencies in the pitch and voicing analysis algorithms.

The first section of this Chapter describes the actions taken to find and eliminate these artifacts in order to raise the quality of the synthetic speech produced by the PS SB-LPC. The second section and greater part of this chapter is concerned with the design of a voicing classifier intended to improve voicing decisions in the PS SB-LPC. Both of these parts utilise a graphical user interface (GUI) based tool called the Bit Stream Editor (BSE). This tool allows the user to manually set and save to file analysis values at the encoder of the PS SB-LPC and evaluate their effect upon the decoded speech.

5.2 Bit Stream Editor

The Bit Stream Editor has been developed with the Tcl and Tk programming languages and is based on Snack Toolkit [59] and previous work on the SB-LPC [68]. Snack is a Tcl based tool which also allows executables to be called in the C programming language. Snack has functions for basic sound handling such as playback, recording, file and socket I/O. It has callable functions which allows users to open a input speech file, for example in several audio formats, from disk storage and then view and listen to the speech file. Tcl is a scripting language and is used extensively in GUI and embedded applications, Tk is an open source library of basic elements for building a GUI; the combination of Tcl and Tk GUI toolkit is known as Tcl/Tk. A summary of how Snack and Tcl/Tk elements slot into the BSE programming environment is shown in Figure 5.1.

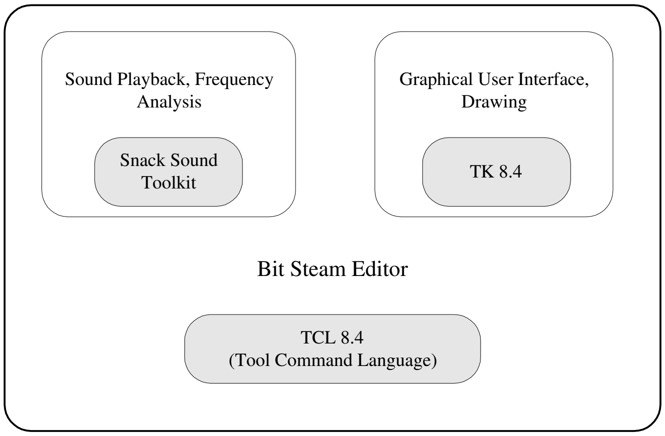

Sound Playback, Frequency Analysis
Snack Sound Toolkit
Graphical User Interface, Drawing
TK 8.4
Bit Steam Editor
TCL 8.4 (Tool Command Language)

Figure 5.1: Bit Stream Editor programming environment

A screen shot of the BSE is shown as Figure 5.2. The user options at the far left are basic functions for dealing with the waveform signal such as open, copy, save and zoom, etc. The call encoder button executes the encoder with initial pitch positions and voicing levels. Functions have been added to Snack so that these values can be loaded onto the screen and edited manually with electronic mouse operation. The PS Encoder mod and PS Decoder mod buttons re-encode and decode with these manually edited pitch and voicing files, producing the output speech shown at the bottom of the plot. A playback toolbar is built into the BSE for the playback of input/output speech. The general structure of the BSE operation is shown as Figure 5.3.

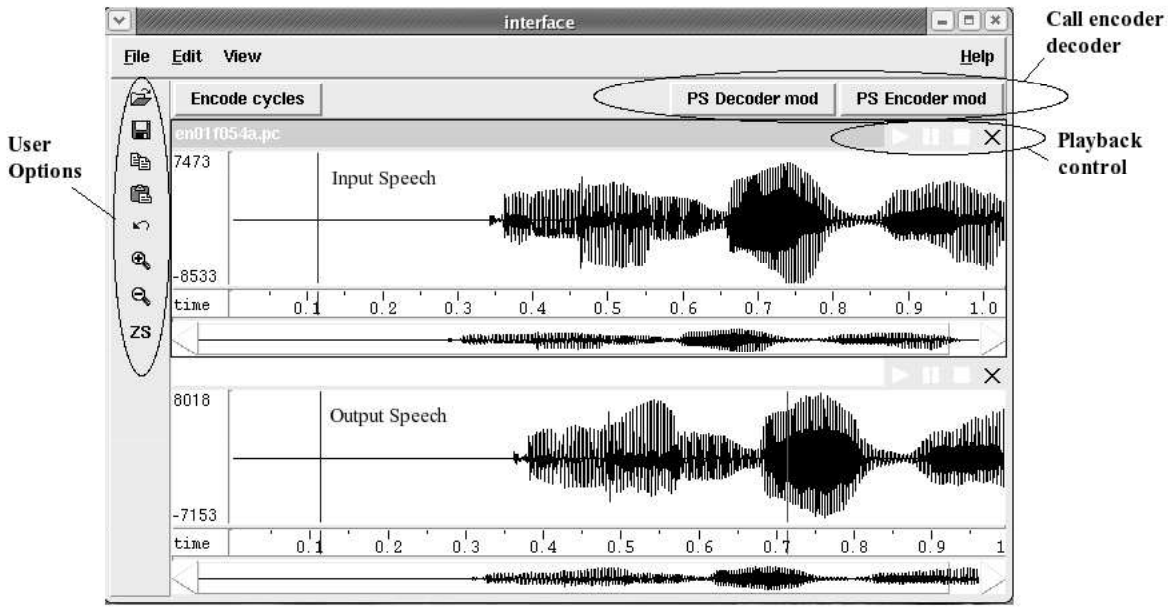


Figure 5.2: Screenshot of BSE

5.3 Utilisation Of Bit Stream Editor

It is believed that the most likely cause of the artifacts that limit the quality of synthetic speech produced by the PS SB-LPC are:

- Pitch cycle detection - occasional errors in the pitch cycle detection process results in incorrect pitch sizes being analysed producing pops and roughness in the voiced synthetic speech. This requires post-processing to be carried out as described in Section 4.5.1.1. However these post-processing routines do not eliminate all pitch cycle size errors.
- Voicing estimator - the voicing estimator is believed to incorrectly select a voicing level which corresponds to a frequency level higher than should be expected for the speech waveform at certain sections. This results in synthetic speech that does not have the same perceptual quality as the original.

The BSE can be used to determine the if pitch and voicing estimation errors are

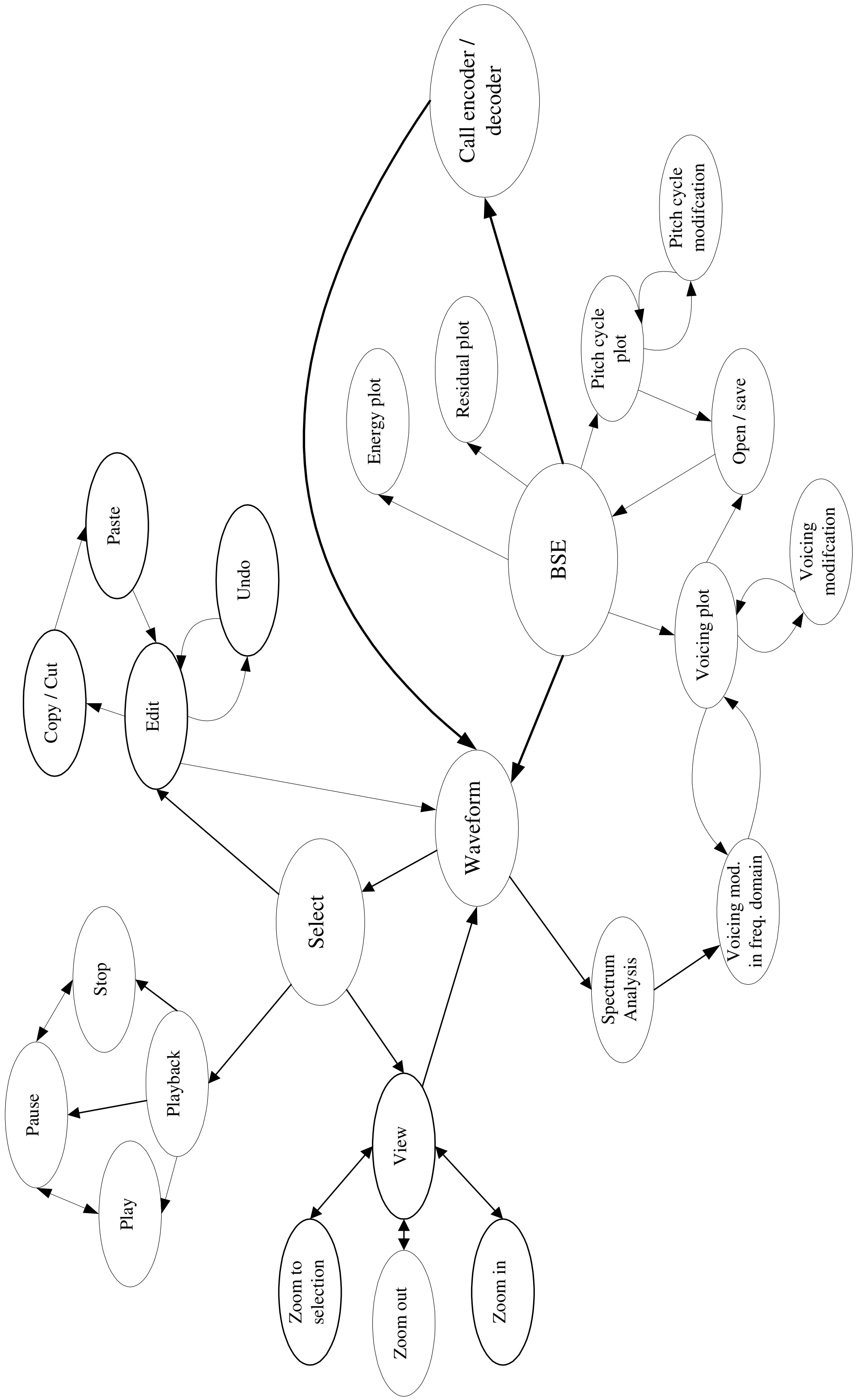


Figure 5.3: General structure of BSE operation

definitely causing the majority of errors heard in the synthetic speech. The aim therefore is to use the BSE to manually set the pitch and voicing values on the screen of the interface through mouse operation. This allow the maximum speech quality of the speech coder to be determined. This operation is summarised in flow chart form as Figure 5.4 and is described in Sections 5.3.1 and 5.3.2.

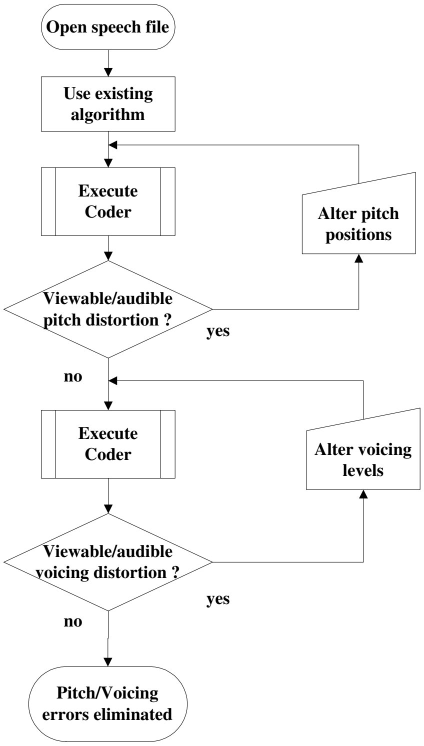


Figure 5.4: Process of removing speech artifacts in BSE

5.3.1 Pitch Cycle Position Editing

All operations take place on 16 kHz PCM files from the NTT database [5] which have been down sampled to 8 kHz to allow for use in the PS SB-LPC. The PS SB-LPC encoder is executed without any smoothing algorithm in place and the pitch cycle sizes written to a text file. These pitch cycle positions can then be plotted onto the BSE window, adjusted and then the encoder/decoder executed with these new positions and the decoded speech evaluated perceptually on the BSE. The Modified

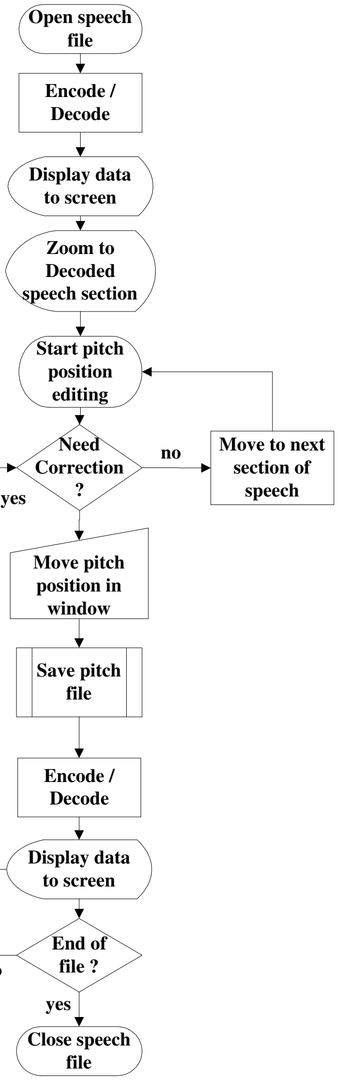


Figure 5.5: Flow chart of pitch cycle position editing

Time Envelope (MTE) signal used by the Trapezoidal Search can be viewed on the Editor as a guide to estimate PCW positions as it is time aligned to the encoded speech. Major modifications to Snack, the BSE and the PS SB-LPC were needed to achieve these goals. This operation is summarised in Figure 5.5. Figure 5.6 shows the occasional pitch cycle detection resulting in harmonic damage to the synthetic speech, this is perceptually displeasing and easily noticed by the listener. It can be seen at (a) and (b) that the pitch sizes variation is not smooth and a sharp change occurs which does not correspond to the input speech which is expected to have a regular pitch

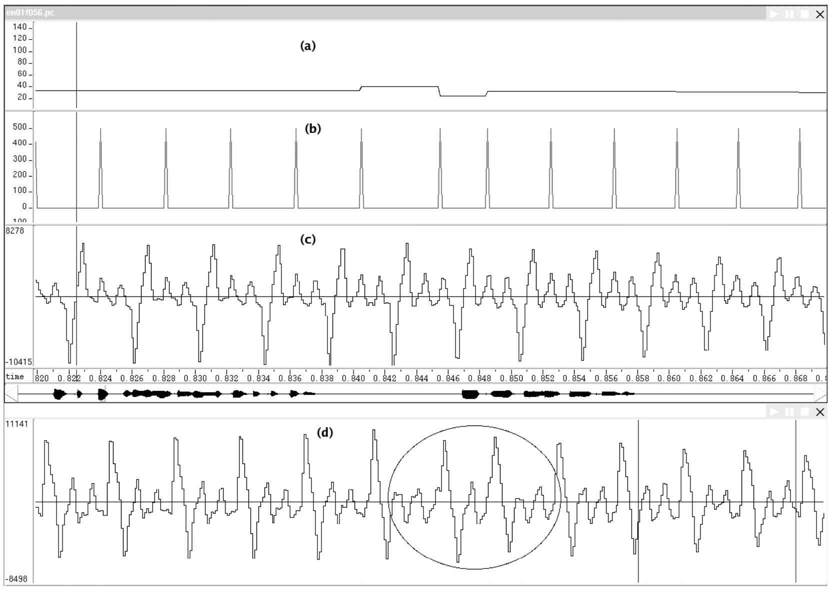


Figure 5.6: BSE with pitch error (a) Pitch sizes, (b) Pitch cycle positions, (c) Input Speech and (d) decoded speech with artifact

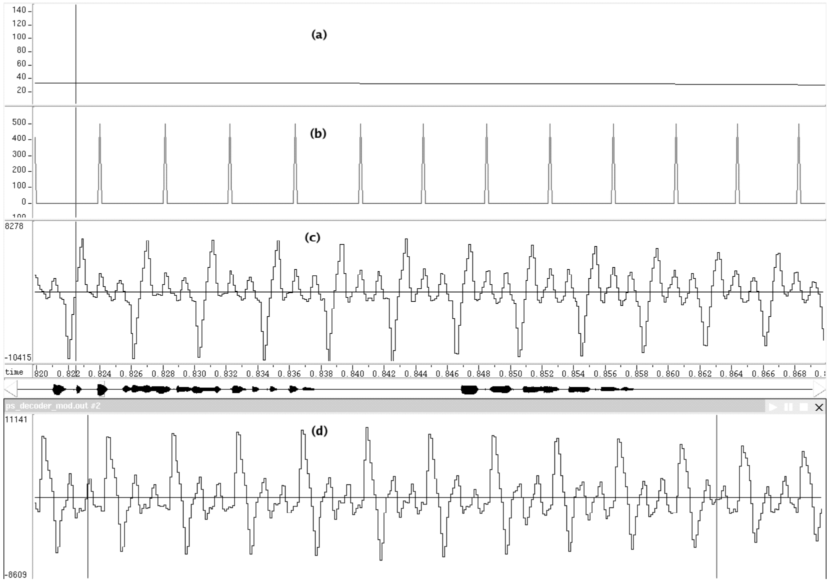


Figure 5.7: BSE with pitch error (a) Pitch sizes, (b) Pitch cycle positions, (c) Input Speech and (d) decoded speech with artifact removed

variation at this point.

Using the BSE the pitch position was manually adjusted to correspond to the neighbouring cycles. This is illustrated in Figure 5.7 where the pitch cycle positions have been manually adjusted and the speech artifact removed. By listening to the output speech on the BSE and manually adjusting all the pitch positions which were causing pitch speech artifacts it was possible to remove almost all errors caused by incorrect pitch cycle positions.

It was found on the BSE that at certain sections where there is little excitation and only resonance from the vocal tract the MTE does not produce a signal with a clear pitch structure. This is illustrated in Figure 5.8 at points (a) and (b), when the MTE is processed by the Trapezoidal Search routine irregular pitch sizes are produced to which post-processing smoothing is applied in the PS SB-LPC.

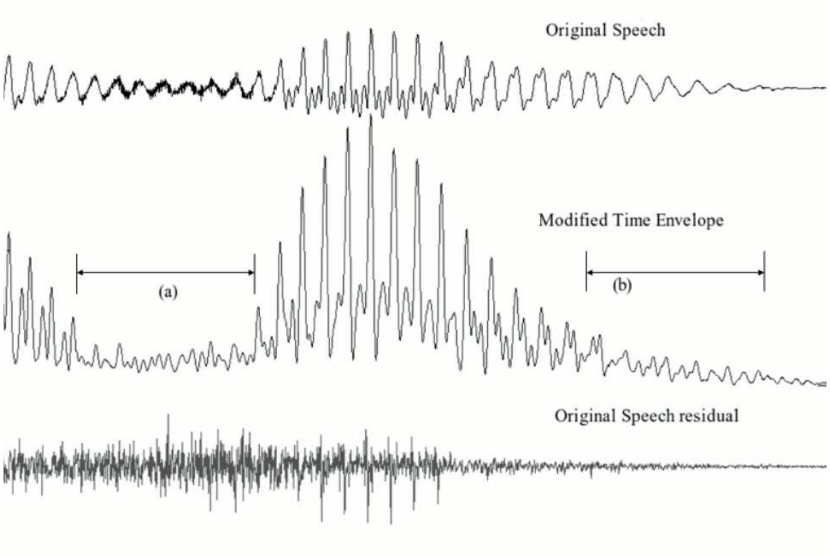


Figure 5.8: Original speech (top), MTE (middle) and original speech residual (bottom)

5.3.2 Voicing Level Editing

The operation of altering the voicing levels in the PS SB-LPC and then viewing/listening to the effect on the decoded speech is very similar to the operation of pitch position editing. This operation is surmised in Figure 5.9. The Snack toolkit upon which the BSE is based provides support for visualisation of speech in the frequency domain.

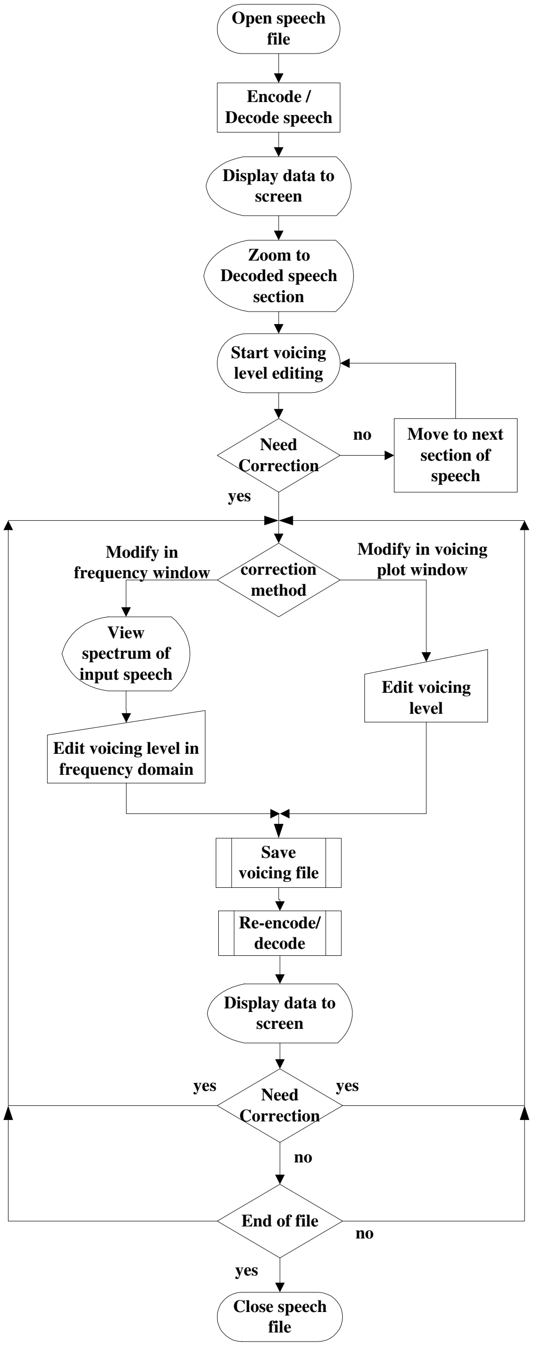
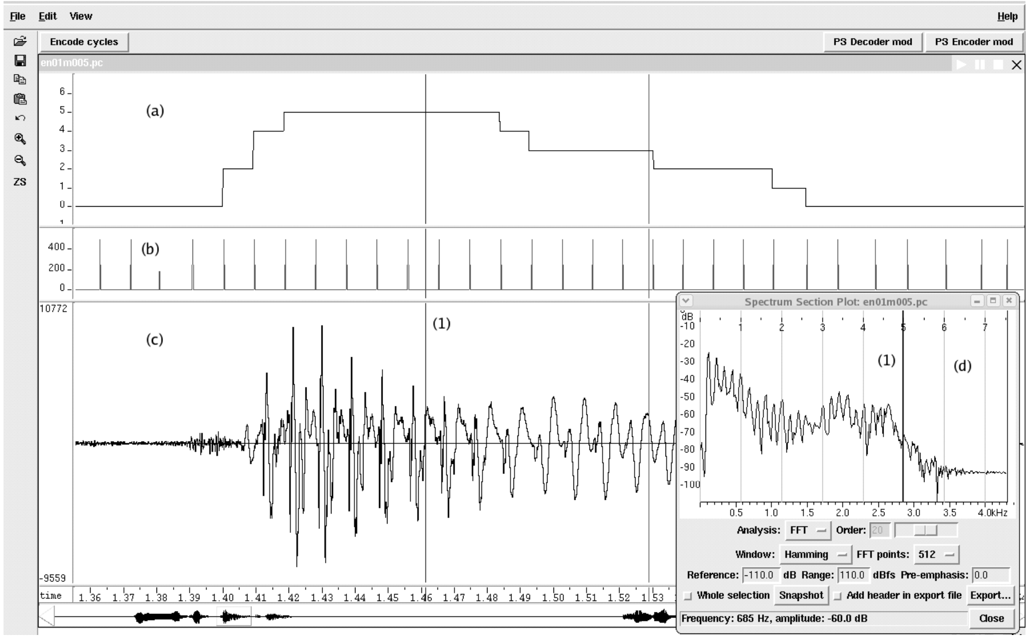


Figure 5.9: Flow chart of voicing level editing

This is illustrated in Figure 5.10, part (a) shows the voicing levels encoded by the PS SB-LPC, these voicing levels have been found from the input speech shown in (c). Part (d) shows the frequency spectrum of the input speech at point 1 in (c). By modifying the BSE the voicing levels in (a) can be manually adjusted through a simple mouse movement or in screen (d) in the frequency domain any changes to the voicing level made in (d) will be updated in (a).

This gives great flexibility as the voicing levels can be set on the BSE both perceptually and in the frequency domain. It was found using the BSE that the current existing voicing algorithm which has been described in Sections 4.5.2.4 and 4.5.2.6 made excellent voiced/unvoiced decisions. However the soft decision voicing levels were generally set too high and at many sections the synthetic speech produced was over voiced and as a result perceptually sounded harsh.



Analysis: FFT Order: 20

Window: Hamming FFT points: 512

Reference: -110.0 dB Range: 110.0 dBfs Pre-emphasis: 0.0

□ Whole selection □ Add header in export file

Frequency: 685 Hz, amplitude: -60.0 dB

Figure 5.10: Modification of voicing levels in spectrum window (a) voicing levels encoded (b) pitch cycle positions (c) input speech and (d) spectrum section of input speech

This over voicing is caused by variations in the band-limited peakiness signal. Initially when each signal is band-limited the highest peakiness value from each of the seven

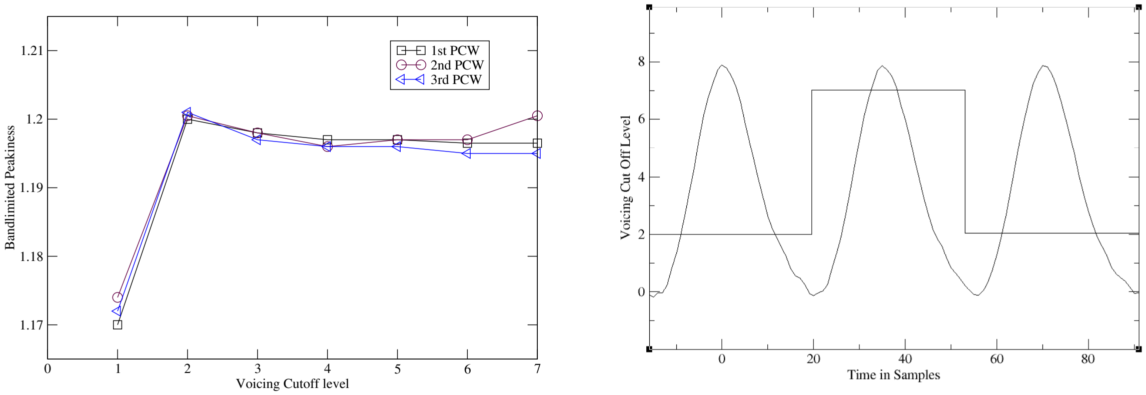


Figure 5.11: Band-limited peakiness values for three PCW (left), shown (right) with original encoded voicing level

cut off frequencies is chosen as the voicing level for that PCW. Figure 5.11 (right) shows three PCW of speech. These PCWs have very similar voicing parameters but the centre PCW is considered to be a voicing level calculation of seven from the voicing determination algorithm currently used by the PS SB-LPC. Figure 5.11 (left) shows the band-limited peakiness values for the three PCWs.

From this figure the first and third PCWs have a maximum peakiness value at a level of two, however although the second PCW varies little over a cut off of two by using the maximum peakiness value the PCW is declared fully voiced. This change in voicing level is not due to the characteristics of the signal but is caused by the voicing calculation itself. Although this variation occurs infrequently to overcome these slight variations that can occur in the peakiness signal a small amount of historical bias was originally introduced.

The peakiness value for the band cutoff frequency selected as the previous PCW voicing level is multiplied by a factor of $1 + \epsilon$. A value of ϵ of 0.07 was found and set experimentally in [63]. If applied to the second PCW in Figure 5.11 (right) it will bias this PCW to have a voicing level of 2. However this solution may remove problems caused by occasional irregular variations in the band limited peakiness but it limits natural variation that can occur in voicing of the speech signal at some sections.

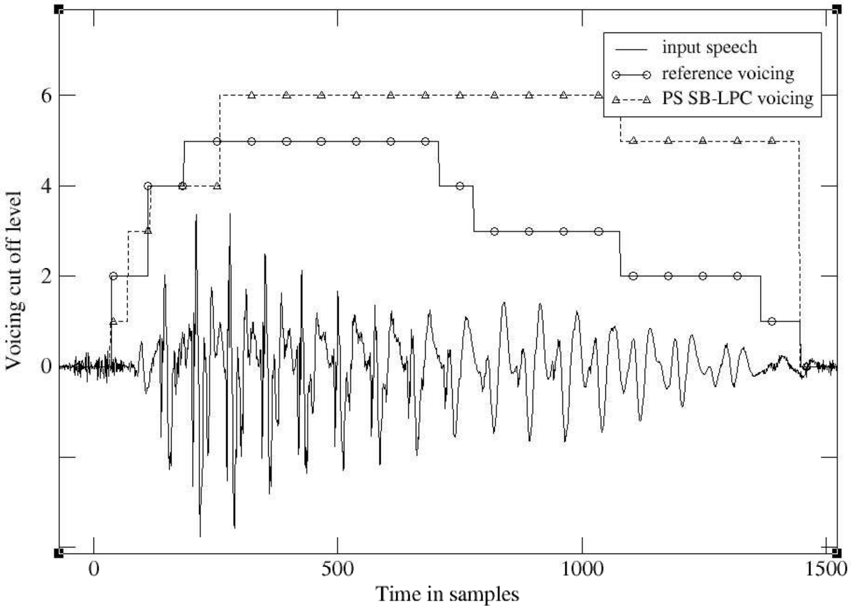


Figure 5.12: Encoded voicing levels in the PS SB-LPC against reference voicing manually set on the BSE

This is illustrated in Figure 5.12 which shows the voicing levels set on the BSE known here as reference voicing and those from the PS SB-LPC through the usual voicing estimation algorithm described in Section 4.5.2.6. For this section of input speech the levels from the PS SB-LPC match closely those of the ideal reference at the start of the speech waveform, however towards the end of the waveform the levels in the PS SB-LPC do not fall to those of the reference, inspection of the band-limited peakiness values determined that this was due to the addition of historical bias, ϵ.

As a result the synthetic speech produced at such sections sounds perceptually over synthetic and does not reflect the natural evolution of the speech waveform. In order to remove the heuristic tuning of the voicing levels which cannot accommodate every input situation we aim to improve upon the voicing level accuracy by using the manually tuned BSE reference voicing levels in a database and utilise vector quantisation techniques to make voicing level decisions. This will be discussed further in the next section.

5.4 Voicing Estimation

5.4.1 Introduction

Standard sinusoidal coders such as MBE and MELP extract parameters at regular intervals; parameter estimation is achieved using the speech waveform falling under an analysis window centered on an analysis point. This procedure assumes the speech to be stationary during the speech segment under analysis. However speech is non-stationary at transitional sections of speech such as onsets, offsets and plosives, which although they only account for a small percentage of speech they are very important perceptually. Therefore it can be considered that a disadvantage of TS coders such as MELP and MBE is that they smooth transitional sections of speech resulting in the loss of fine detail.

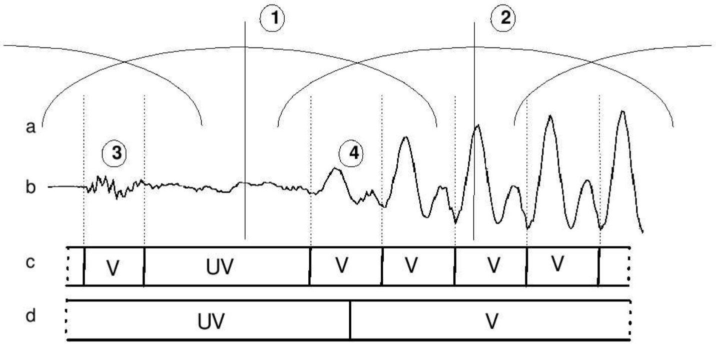


Figure 5.13: Comparison of voicing classification in TS and PS coder a) Windows and analysis points in TS coder, b) Section of speech, c) PS classification and d) TS classification

PS coders such as the PS SB-LPC operate on a per cycle basis and therefore do not smooth transitional sections as they have a shorter analysis window which should result in superior performance over TS coders at such sections. Figure 5.13 demonstrates these differences when applied to the voicing classification of input speech. The TS coder will extract parameters at points 1 and 2, which may be incorrect in its classification (d) of the voicing content of the speech signal at points 3 and 4. At point 3 the segment is too short compared to the length of the window and at point 4 the TS method does not provide the necessary time accuracy at transitions. The PS classification (c) unlike the TS method should correctly classify the first and third speech segments as voiced.

5.4.2 Classic Voicing Estimation Methods

After speech is declared as unvoiced or voiced as described in Section 3.6 using a hard voicing decision, voiced speech is further classified to estimate its actual frequency content. To make this soft decision classic voicing methods such as those in MELP, MBE and SB-LPC typically rely on the comparison of 2 functions:

- A voicing function computed in the speech spectrum with a value per harmonic or frequency band
- A threshold function computed as heuristic function of several speech parameters.

The voicing threshold is necessary as the performance of the voicing function is not sufficient [67]. Several speech parameters are used to give an indication of voicing. The main concern with this method is how to generate a threshold function based on these parameters. The next section describes how this is carried out classical sinusoidal speech coders.

5.4.2.1 MBE Mixed Voicing

In the MBE coder harmonic voicing is estimated by comparing the error of a synthetic voiced spectrum $\hat{S}(m, w_o)$ with respect to the speech spectrum $S(m)$ and comparing it against a threshold function for each harmonic band. Using $\hat{S}(m, w_o)$ a voicing measure is computed for each band on which the voicing decision is made. These bands do not have to be single harmonic bands, they can cover a number of harmonic bands. The MBE splits the spectrum in groups of three harmonics and performs the voicing decisions on these groups.

$$D_k = \frac{\sum\limits_{m=a_k}^{b_k} |S(m) - \hat{S}(m, w_0)|^2}{\sum\limits_{m=a_k}^{b_k} |S(m)|^2} \tag{5.1}$$

where w_0 is the selected fundamental frequency and a_k and b_k are the lower and upper boundaries of the decision bands. Each band is declared voiced if its voicing measure is above the threshold function, unvoiced otherwise. The threshold is defined as $\Delta_k(w_0)$

$$\Delta_k(w_0) = (\alpha + \beta w_0)[1.0 - \epsilon(k-1)w_0]M(E_0, E_{av}, E_{min}, E_{max}) \tag{5.2}$$

where $\alpha = 0.35$, $\beta = 0.557$ and $\epsilon = 0.4775$ are the factors that give good subjective quality and

$$M(E_o, E_{av}, E_{min}, E_{max}) \begin{cases} 0.5; E_{av} < 200 \\ \dfrac{(E_o + E_{min})(2E_o + E_{max})}{(E_o + \mu E_{max})(E_o + E_{max})}; E_{av} \geq 200 \text{ and } E_{min} < \mu E_{max} \\ 1.0; \text{otherwise} \end{cases} \tag{5.3}$$

is the adaption factor that controls the decision threshold for voicing decisions. A favourable value for μ is 0.0075. The parameters E_{av}, E_{max} and E_{min} correspond to the local average energy, the local maximum energy and the local minimum energy respectively. These three speech parameters are updated every frame according to [25].

5.4.2.2 MELP Mixed Voicing

The voicing decision in MELP is performed using time domain techniques on bandpass filtered versions of the original speech. The original speech is separated into 5-sub bands using 6^{th} order Butterworth filter, with pass-bands of 0-500, 500-1000, 1000-2000, 2000-3000, and 3000-4000 Hz [48].

The normalised correlation is then computed at the pitch value P for the first band as well as the range P-5,P+5. The maximum of these correlations is then used as the bandpass voicing strength for the first band and the corresponding lag is saved for use in the computation of the bandpass voicing for the remaining bands. The bandpass voicing strength for the other bands is computed again using the normalised autocorrelation at the lag chosen for the first band on the bandpass filtered signal, and also on the time envelope of that signal. The maximum of these two correlations is then taken as the bandpass voicing strength of the considered band.

These bandpass voicing strengths Vbp_i with i equal to 1,....,5 are then biased using the peakiness of the signal. If the signal is very peaky ($Pk >$ 1.6), Vbp_i for i = 1,2,3 are forced to 1. If it is moderately peaky ($Pk > 1.34$), Vbp_1 is forced to 1.0. Finally the voicing decision for each band is made using Vbp_i:

- If $Vbp_1 \leq 0.6$ all bands are declared unvoiced
- If $Vbp_1 > 0.6$, the first band is declared as voiced and each band i is set to voiced if $Vbp_i > 0.6$, unvoiced otherwise
- The voicing combination 10001 is not allowed and is replaced by 10000

The voicing decision itself makes use of two parameters; the normalised autocorrelation and the peakiness of the signal. This voicing decision is shown as Figure 5.14. Although providing good voicing indication, there are cases when both these parameters fail therefore more parameters are needed for reliable voicing determination.

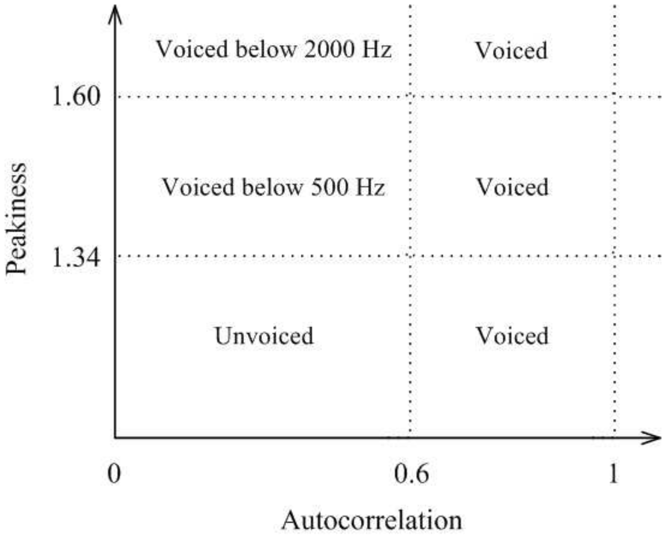


Figure 5.14: Voicing classification technique in MELP coder

5.4.2.3 SB-LPC Mixed Voicing

Coders such as the MBE use a voicing decision for each harmonic or group of harmonics (typically 2 or 3), the SB-LPC coder assumes all bands to be voiced from DC to a certain cutoff frequency and unvoiced above this cutoff frequency. This has the advantage of requiring only a small number of bits to represent the cutoff frequency, 3 bits usually being sufficient. This represents a large saving over the MBE approach which requires up to 12 bits.

The cutoff frequency decision is made by considering the voicing likelihood for each individual harmonic. A voiced band should have a spectral shape similar to the spectral shape of the window used, prior to Fourier Transformation. Unvoiced bands will be random in nature. The voicing likelihood of each band is measured as the normalised correlation between the considered harmonic band and the spectral shape of the window positioned on the harmonic location. The voicing likelihood $V(l)$ for the l^{th} harmonic is given by

$$V(l) = \frac{\left[\sum\limits_{m=a_l}^{b_l} S(m)W(\frac{2\pi}{N}m - lw_0)\right]^2}{\sum\limits_{a_l}^{b_l} W^2(\frac{2\pi}{N}m - lw_0).\sum\limits_{a_l}^{b_l} S^2(m)} \tag{5.4}$$

where S is the Fourier Transform of the speech and W of the analysis window. The value of $V(l)$ is between 0.0 and 1.0 respectively corresponding to fully unvoiced and unvoiced cases. This value is then compared to a threshold function $T(l)$ for each individual harmonic. This threshold calculation is the most important stage during Split Band voicing estimation [67].

The value of $T(l)$ is determined by taking several factors into account. Firstly the lower harmonics are more likely to be voiced so the threshold value is lower for the lower harmonics. Secondly a harmonic is more likely to be unvoiced if unvoiced in the previous frame, so the threshold is raised for harmonics that were previously unvoiced. Thirdly the harmonics are more likely to be voiced if the hard decision voicing metric indicated a voiced signal, therefore the threshold is lowered.

The voicing threshold is biased by using a range of voicing parameters as described in Section 3.6 such as zero crossing and autocorrelation. Thresholds are set for each of these parameters and if triggered the voicing threshold function is biased towards voiced or unvoiced. The pitch value is also used to bias the voicing threshold function. An example of voicing likelihood and threshold function is given in Figure 5.15.

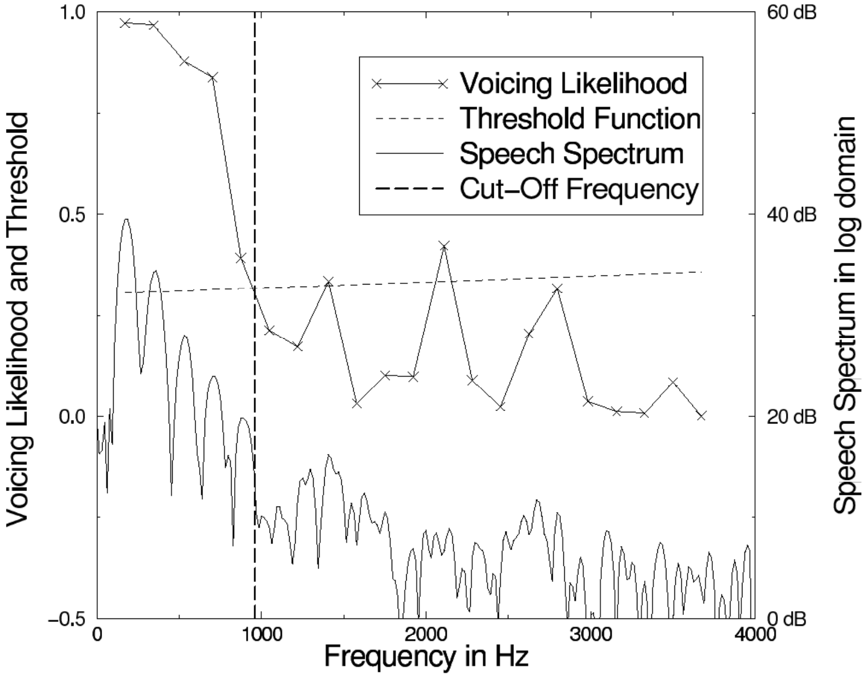


Figure 5.15: Original Speech spectrum with voicing likelihood, threshold function and cutoff frequency [67]

Using a limited number of speech characteristics for the threshold computation does not lead to good voicing determination. In the MBE energy alone is not a reliable enough voicing indication as there can be high energy unvoiced speech sections. In MELP the peakiness factor is not entirely reliable, single peaks can lead to high peakiness, likewise for autocorrelation; in the case of pitch variations, normalised autocorrelation may be quite low when the speech is voiced.

By increasing the number of parameters and other speech characteristics, the SB-LPC improves upon these two coders when it comes to finding a suitable threshold function.

However even in the SB-LPC, this threshold function has to be tuned, through trial and error of several speech parameters which can be difficult and unreliable as every new filtering/noise condition calls for retraining. Given that the voicing has a large impact on speech quality we aim to apply a systematic approach to overcome these disadvantages.

This approach uses the BSE to generate reference hand marked voicing files for a speech database. This database of speech information can be utilised:

- A classifier is trained using these and a number of selected speech parameters
- Vector Quantisation techniques are used to cluster speech parameters and associate each cluster with a threshold function.
- Generate a threshold function computed as giving best classification in the cluster according to the training database.
- The PS SB-LPC encoder stores a list of clusters and a voicing threshold function for each cluster. It compares speech parameters for current cycle to stored clusters and chooses a threshold function associated with best matching cluster.

This technique should be easy to adapt for use in various filtering/noise conditions. As an example if this approach was carried out and implemented in the MELP, the result may be as shown as Figure 5.16, this can be compared to the original MELP approach previously shown in Figure 5.14.

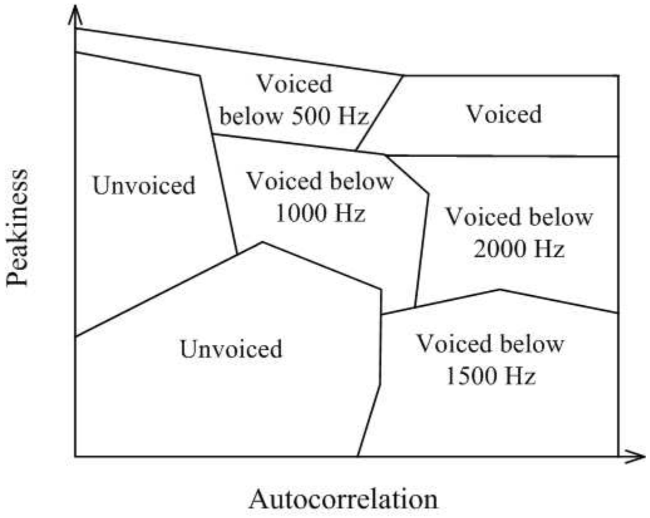


Figure 5.16: Proposed classification technique if applied to MELP voicing

The next section will describe this method in detail where we employ a Codebook classifier technique; during normal PS SB-LPC operation, if the PCW is declared as voiced the peakiness values at each frequency cutoff level are used by the classifier, which has been trained to give the best voicing decision.

5.4.3 PS SB-LPC Voicing Classifier

Using the BSE, for each PCW several speech parameters are stored along with their corresponding reference voicing decision from the manually tuned results and the corresponding peakiness values at each frequency cutoff - this vector is known here as the voicing vector. These voicing vectors therefore contain the voicing information for each PCW, PCWs which have similar speech parameters usually have a similar evolution of peakiness over the seven cutoff frequencies.

Vector quantisation techniques are used to determine which of the N1 cycles over the range of the training database are the closest. Over the length of the training database, 9300 cycles, the N1 closest PCW in terms of the speech parameters are grouped together by finding the minimum distortion between the current test vector and all vectors in the codebook

$$(k_1, k_2, ..., k_{N1}) = argmin \sum_{j=1}^{N1} \left(\sum_{i=0}^{L-1} (X(i) - C_{k,j}(i))^2 . W(i) \right) \tag{5.5}$$

where W is the training weights applied to the speech parameters, the number of weights used is given by L here a value of 9, these are defined in Table 5.1. X is the current test vector and C are the vectors in the database.

Number	Speech Parameters	Weight
1	Energy to peak energy ratio	1.5
2	Peakiness	1.6
3	Correlation	1.4
4	Zero crossing ratio	1.5
5	Low band to full band energy ratio	1.5
6	Pre-emphasis energy ratio	1.5
7	Previous voicing	3
8	P_1	2.1
9	P_{max} - P_{min}	2

Table 5.1: Weights used in training procedure

In Table 5.1 P_1 is the peakiness value at the cutoff level of one and P_{max} and P_{min} are the maximum and minimum peakiness values of the seven cut off levels of each pitch cycle. The remaining weights are those parameters described in Section 3.6. The previous voicing value was normalised by its maximum size of 7 to ensure it had similar values to the other weights. These weighting values were found empirically after extensive trial and error.

Once N1 closest vectors have been found from (5.5) a set of thresholds for the N1 must be found which when input to a algorithm will produce the closest match to the reference voicing. The first step is to determine the optimum value of N1. This illustrated in Figure 5.17 which shows the band-limited peakiness and reference voicing (manual voicing) for a PCW.

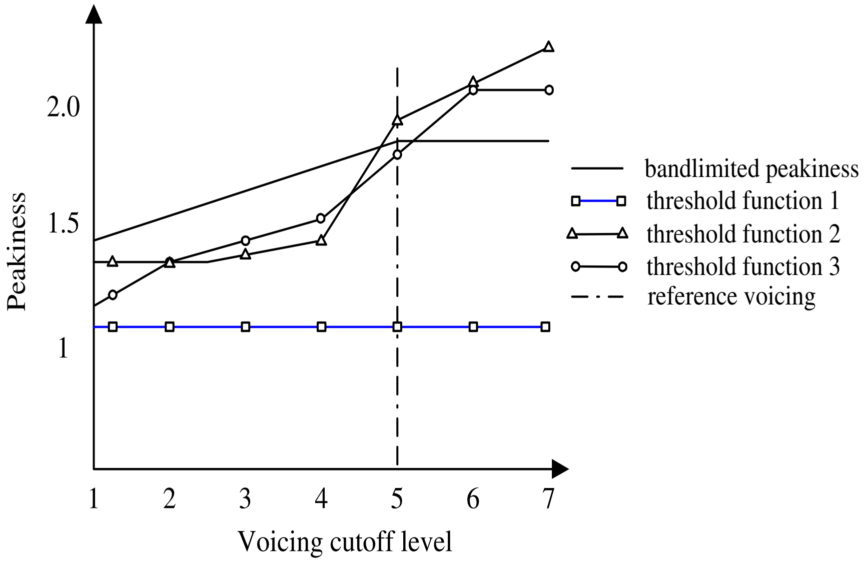


Figure 5.17: Bandlimited peakiness, reference voicing and threshold functions

Also shown are some suggested threshold functions. The aim here is to find a threshold function which will produce a voicing level of 5 - the reference level. In this example threshold function one maybe set too low but threshold functions two and three may both be suitable, many more threshold functions will be suitable also over the number of combinations. As there can therefore, be considerable variation between suitable threshold functions for each of the voicing vectors, N1 voicing vectors were used to generate the one threshold function. It was determined that N1 equal to 25 was optimum.

A step size δ is found from $(P_{max} - P_{min})/7$ of the N1 vectors. The value of the band-limited peakiness at each of the seven cutoff frequencies $P(l)$ is then compared to a threshold function which corresponds to the current threshold at that cutoff frequency. Over every iteration of δ the value of i which produces the greatest value of matching function M is considered to be the calculated voicing. This matching function $M(i)$ is given by:

$$M(i) = \sum_{l=1}^{7}(P(l) - T(l))V_i B(l) \qquad (5.6)$$

For any given voicing level i individual voicing decisions $V_i(l)$ will have +1 (i.e. voiced) up to the cut off $f_c(i)$ and -1 for the higher values (i.e. unvoiced). The matching function is computed at each iteration for every possible voicing step i. For any given i, each voicing level computed correctly i.e. the product $(P(l) - T(l))V_i(l)$ is positive, will contribute to the total sum $M(i)$ proportionally to the difference between the band-limited peakiness value at $P(l)$ and the threshold $T(l)$. Each incorrect voicing level will decrease the total sum. The weighting $B(l)$ is usually set to 1.0 when unvoiced $(T(l) > P(l))$ and higher for voiced, as it is more important perceptually to get the higher voiced levels correctly set.

The matching function M from (5.6) returns one of the voicing levels (1 - 7), this calculated voicing known here as V_{calc} for each threshold iteration is compared to the reference voicing V_{ref} (set in the BSE of Section 5.3.2) for each of the N1 vectors. A score V_{score} is found over all N1 vectors using

$$V_{score} = \sum_{i=0}^{N1-1} (V_{ref} - V_{calc})^2 \qquad (5.7)$$

The lowest value of V_{score} corresponds to the best thresholds for these N1 vectors as it indicates that the calculated voicing is closest to the reference voicing for the N1. The parameters associated with these N1 vectors and their associated thresholds at each of the seven cut off frequencies are then written to a training database. During normal operation of the PS SB-LPC encoder for each PCW the closest N2 vectors in the training database, N2 equal to 25, are found using the same weights as in the training procedure. The associated thresholds values are then averaged at each cutoff frequency and input into (5.6) to determine the voicing level for that PCW.

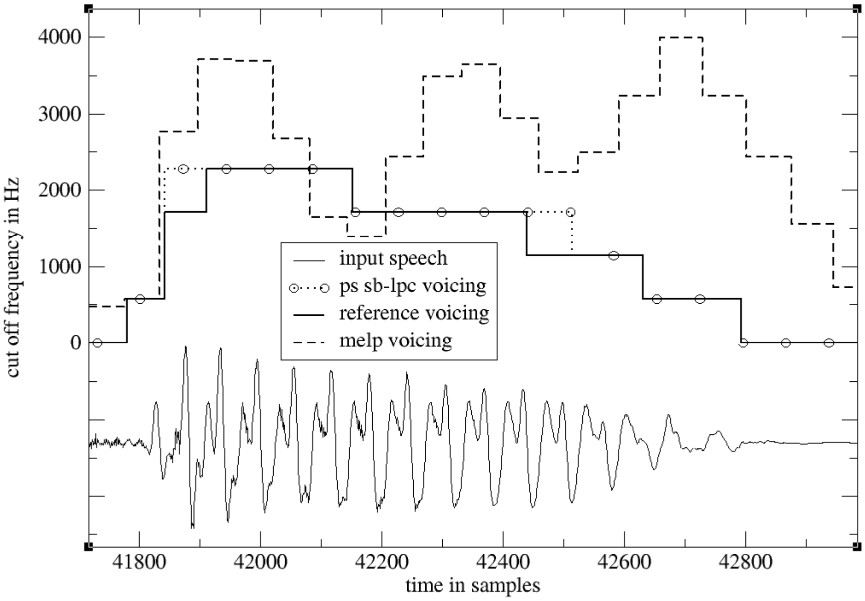


Figure 5.18: Voicing levels in the MELP and PS SB-LPC against reference voicing

We compared the results from the PS SB-LPC classifier method to those employed by the 2.4 kbps MELP method of Section 5.4.2.2 at a 8 kHz sampling rate. The MELP coder makes a voicing decision for each 200 sample frame of speech at the encoder and then interpolates voicing across the decoded frame. The speech files used were not included in the training database. A comparison of the manually set voicing on the BSE (reference) is shown against the PS SB-LPC classifier and MELP voicing in Figure 5.18.

As shown in the figure at the waveform onset and offset where the pitch cycles are changing more rapidly in the time domain, the superior time resolution of the PS SB-LPC more closely follows the reference voicing levels set in the frequency domain. In addition at the waveform centre where there are no PS issues the PS SB-LPC still gives a more accurate result. It was generally found that the classifier method returned a voicing level within one voicing level of the reference voicing.

A comparison was made using cycle sample lengths of the encoded voiced (fully and mixed) and unvoiced decisions of each cycle (identified by Trapezoidal Search) of the two coders. The results are shown as Table 2.2 with silences excluded. The distinction between silence and non silence was based on an energy cutoff level of five. As can be seen the performances of the MELP and PS SB-LPC are comparable when determining voiced but very different for unvoiced speech cycles.

Coder	Speaker	Type	V	UV
MELP	F1	V	97.94	48.96
		UV	2.06	51.04
	M1	V	97.56	32.02
		UV	2.44	67.98
	F2	V	97.15	40.50
		UV	2.85	59.50
	M2	V	96.70	3.80
		UV	2.75	61.57
PS SB-LPC	F1	V	96.70	3.80
		UV	3.30	96.20
	M1	V	96.53	4.13
		UV	3.47	95.87
	F2	V	96.98	3.87
		UV	3.32	96.13
	M2	V	96.80	3.91
		UV	3.20	96.09

Table 5.2: Encoded voicing cycle comparison of MELP and PS SB-LPC against reference voicing. V/UV comparison is made on percentage of cycles according to sample length

This primarily reflected the effect at onsets and offsets where a MELP speech frame contained only one or a few voiced cycles and a large unvoiced component, consequently the MELP decision was to declare such a frame and therefore its constituent cycles as voiced. It is possible that the TS nature of the MELP coder may have forced the designers to bias the voicing decisions towards voiced as it is usually much better perceptually to declare a unvoiced cycle as voiced rather than to declare a voiced cycle as unvoiced.

A good separation of voiced and unvoiced cycles of speech has been achieved. It may be possible to use such an algorithm in a speech activity detection system for end point detection. Endpoint detection is the detection of speech in non-speech events and background noise, it is important in automatic speech recognition (ASR) and speaker verification (SV) environments. Many such systems use pitch detection, zero-crossing rate, autocorrelation and peakiness as part of the detection process. In [43] and [24] robust endpoint detection schemes were introduced for use in noisy environments, both of these methods made decisions on 10ms frames of speech sampled at 8 kHz. Better results may be obtained if such systems were able to make decisions on cycles of speech rather than frames which may contain several cycles.

5.5 Concluding Remarks

In this chapter a GUI tool has been used to determine the cause and effect of distortion in a PS sinusoidal speech coder. A novel technique for accurately measuring the voicing content of a speech signal has also been introduced. This has been achieved by the measurement of phase-spread information contained within individual pitch cycles. The experimental results show that the methods introduced compare favourably with a standard speech coder that operates time synchronously. These methods can be utilised not only by PS but also TS coders to make more accurate decisions on the voicing content of their speech frames.