Pitch Estimation

Sarel van Vuuren

February 4, 1998

Technical Report # CSE-98-005

Anthropic Speech Processing Group
Department of Electrical and Computer Engineering
&
Center for Spoken Language Understanding
Department of Computer Science and Engineering

Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, Oregon 97291-1000

Abstract

This report is about estimation of the pitch of a speaker. It is based on an earlier report by the author [1] and is intended mainly to make the material presented there more accessible. It is intended as a review of a number of pitch estimation techniques with the objective being to highlight issues involved in obtaining a reasonable estimate of the pitch. These issues form the basis for a companion report where a technique using cepstral smoothing will be described.

The first part of the report is a discussion of various pitch estimation algorithms and their relative strengths and weaknesses. The second part elaborates on a well-known technique that is based on the *cepstrum*. The cepstrum itself is typically derived from spectral information obtained using a short-time analysis window. Unfortunately, an apparent weakness of the cepstrum technique is that the length of this analysis window can affect the quality of the pitch estimate deleteriously depending on the pitch of the particular speaker. With the average pitch of a speaker ranging from less than 80 Hz for a male, to more that 300 Hz for a female and as much as 500 Hz for a child, this can be a substantial problem. To alleviate this problem we propose a technique to adapt the length of the analysis window iteratively based on the estimated pitch of the speaker. We provide estimation results for speech from the TIMIT corpus.

Contents

Abstract **i**

1 Introduction **1**
1.1 Model of pitch 1
1.2 Statistics of pitch 2
1.3 Estimation of pitch 2
1.4 Estimation issues 3
1.5 Quality of the estimate 4

2 Background **6**
2.1 Autocorrelation technique 6
2.2 Average magnitude difference function technique (AMDF) 7
2.3 Simplified inverse filtering technique (SIFT) 7
2.4 Cepstrum technique 8
2.5 Parallel processing technique 9
2.6 Pitch estimation with banks of bandpass filter-pairs 11
2.7 Pitch estimation with a neural-net classifier 11
2.8 Post-processing 12
2.9 Estimation error 13

3 Detection of voiced speech **14**

4 Cepstrum technique with variable analysis window length **16**
4.1 Cepstrum 16
4.2 Adapting the analysis window length 16
4.3 Post filtering 18

5 Pitch estimation results **19**
5.1 Smoothed pitch periods of various speakers 22
5.2 Smoothed pitch periods of various speakers with adaptation of the analysis window length 26

6 Discussion **29**

7 Conclusion **31**

A Appendix **32**

Bibliography **35**

List of Figures

1 A simple discrete-time model for the production of a speech signal. The glottal excitation waveform is generated as a first step in generating a voiced sound. 1
2 Time-domain representation of voiced speech. 2
3 Histograms of pitch frequency for male and female speakers. . . . 3
4 Two basic ways in which pitch may be estimated. 4
5 Block diagram of the autocorrelation pitch estimator. 6
6 Block diagram of the average magnitude difference function technique. 7
7 Block diagram of the SIFT pitch estimator. 8
8 Block diagram of the cepstrum pitch estimator. 8
9 Block diagram of the parallel processing technique. 9
10 The peak and valley measurements. 10
11 A bandpass filter-pair. 11
12 Scatter plot of non-speech, voiced and unvoiced speech samples. . 14
13 The pdfs of speech vs. non-speech as a function of (a) magnitude and (b) zero crossings. 15
14 The pdfs of voiced vs. unvoiced speech as a function of (a) magnitude and (b) zero crossings. 15
15 A block diagram for obtaining the cepstrum. 16
16 Cepstrum of a periodic signal. 17
17 Block diagram of the cepstrum pitch estimator. 18
18 Speech signal, zero-crossing rate and average magnitude of U_1. . 20
19 Unsmoothed pitch period and cepstral peak values of utterance U_1. 21
20 Analysis window length not adapted: Pitch period (ms) for U_1. . 23
21 Analysis window length not adapted: Pitch period (ms) for U_2. . 24
22 Analysis window length not adapted: Pitch period (ms) for U_3. . 25
23 Analysis window length adapted: Pitch period (ms) for U_1. . . . 26
24 Analysis window length adapted: Pitch period (ms) for U_2. . . . 27
25 Analysis window length adapted: Pitch period (ms) for U_3. . . . 28
26 Pitch period in the phoneme /ae/ where amplitude modulation is present. 29
27 Pitch period in a nasal /n/. 30

List of Tables

1 TIMIT phonetic labeling. 32

1 Introduction

Robust pitch estimation is important in many areas of speech processing. Pitch estimation is necessary for coding and recognition of speech. For example it is used in modern speech coders [2], technology for the hearing impaired and speaker recognition systems [3]. In this report, depending on the context, *pitch* is used refer to either the *pitch period*, or *pitch frequency*. The pitch frequency is also sometimes referred to as the fundamental frequency.

1.1 Model of pitch

The general idea of pitch estimation or detection is to obtain the period of the glottal excitation waveform. This waveform is the result of the periodic opening and closure of the vocal cords in the glottis while air is forced through from the lungs. This results in a train of alternating high and low-pressure pulses in the vocal tract. The periodic opening and closure happens for voiced sounds only; for unvoiced sounds the air passes through the glottis unrestricted. Fig. 1 depicts a simple discrete-time model for the production of a speech signal. The glottal excitation waveform is generated as a first step in generating a voiced sound. The sequence of high pressure pulses are further manipulated by the

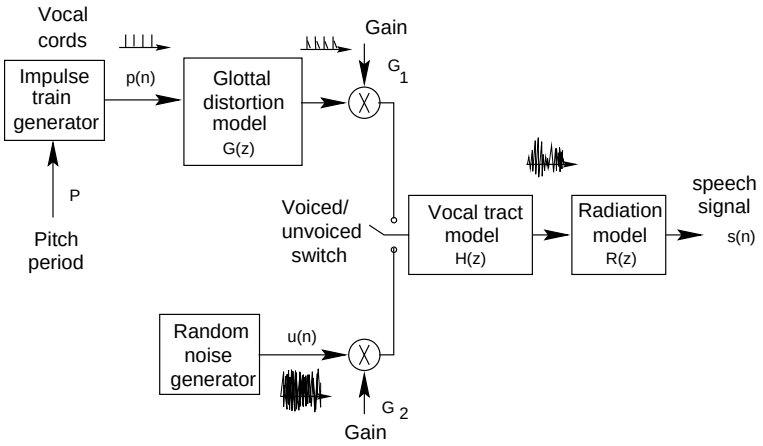


Fig. 1: A simple discrete-time model for the production of a speech signal. The glottal excitation waveform is generated as a first step in generating a voiced sound.

vocal tract and other speech organs. The resultant measured speech signal is modeled as the convolution of the excitation signal with the impulse response of a filter describing the vocal tract and other speech organs.

In the time domain the pitch information in voiced speech is present as quasi-periodic signal excursions, see Fig. 2. The long periods are caused by the excitation (vocal cords) whereas the short periods are caused by the resonant

cavity (vocal tract shape). These periods can generally be labeled by the eye – a method sometimes employed to obtain a reference pitch signal.

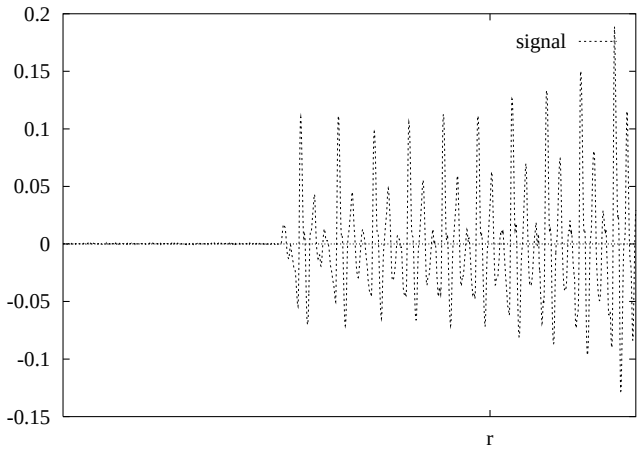


Fig. 2: Time-domain representation of voiced speech.

1.2 Statistics of pitch

Statistical analysis of the pitch frequency on the Switchboard Phase-I corpus [1] indicates a mean of 123 Hz and standard deviation of 16 Hz for adult male voices. The mean for adult female voices of 206 Hz is about twice that of the males, while the standard deviation for females is 23 Hz. Fig. 3 shows normalized histograms of pitch frequency for male and female speakers. What is important to notice from the figure is the wide range of pitch.

1.3 Estimation of pitch

Referring to the model of pitch, a pitch estimator must make a

1. speech or non-speech decision
2. voiced or unvoiced decision (V/U)
3. and estimate of the pitch period in the voiced region.

Several pitch estimation techniques have been proposed to achieve this. Broadly they can be classified into three categories:

1. **Time domain techniques:** Peak and valley measurements, zero-crossings and autocorrelation estimates, often with additional post processing logic. Generally these techniques are noise sensitive.

[1] The Switchboard Phase-I corpus has speech from more than 500 speakers recorded over the telephone.

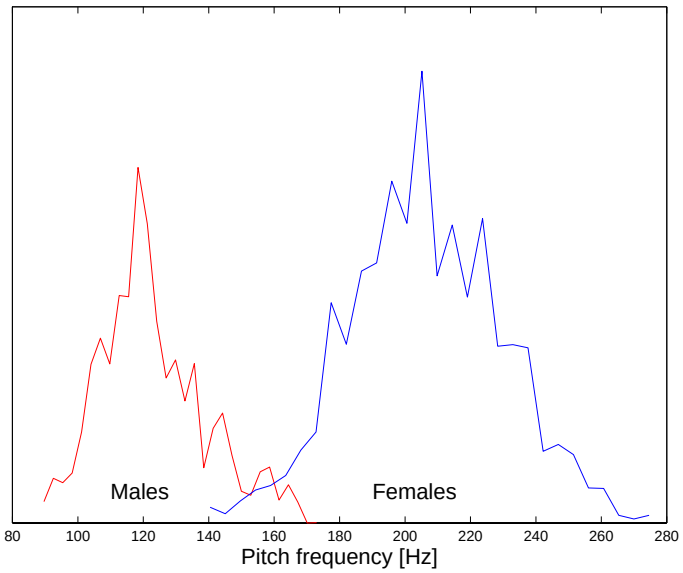


Fig. 3: Histograms of pitch frequency for male and female speakers.

2. **Frequency domain techniques:** Periodicity in the time domain results in useful impulses in the frequency domain at the fundamental and its harmonics. Using short-time analysis to extract the frequency information, these techniques are sensitive to the length of the analysis window so that their useful dynamic pitch range tends to be limited.

3. **Hybrid techniques:** The former two techniques combined, for instance spectral flattening together with autocorrelation. However, this combination does not readily solve the above mentioned weaknesses.

1.4 Estimation issues

As alluded to above, to estimate pitch there are several issues that have to be addressed. Above -mentioned techniques are affected differently by the following issues:

- The glottal excitation waveform is not a perfect periodic train of impulses implying non-stationarity even of the excitation. Pitch vary within a speaker and can sometimes drop or rise significantly. This may happen for example during glottalization or at the end of a phrase.

- The vocal tract transform and glottal excitation are not independent processes, nasals being such a case.

- The vocal tract transform can be assumed stationary only for time periods of less than about 10 ms.

- In the time domain, peak measurements (such as of the instants of excitation) are sensitive to formant structure, while zero-crossing measurements, for instance, are sensitive to noise and signal levels.
- In the frequency domain, telephone systems attenuate the fundamental and its harmonics and introduce both additive and convolutative noise to the speech.
- Low level speech makes discrimination of unvoiced speech more difficult.

1.5 Quality of the estimate

Once the pitch is estimated it is necessary to evaluate the goodness of the estimate. How this is done depends on the application. Fig. 4 depicts two basic ways in which pitch may be estimated.

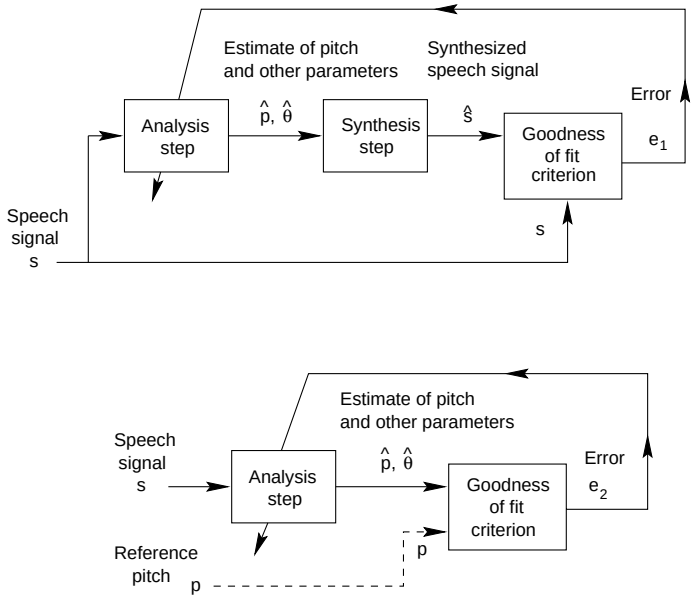


Fig. 4: Two basic ways in which pitch may be estimated.

Depending on the application, pitch estimation may be embedded in one or more steps. In coding it is common to iterate between an analysis step and synthesis step, with some criterion as to the goodness of fit applied after the synthesis step. In this case it is typically to estimate both the position in time and excursion of each glottal pulse. For recognition the synthesis step is usually omitted and only the period or frequency of excitation estimated. Since it is usually easier to obtain a reference of the speech than of the pitch [2] it is generally easier to measure the goodness of fit after the synthesis step.

[2] A laryngraph is sometimes used to estimate the instants of glottal opening and closure.

Accordingly, recognition is often based on rather ad hoc pitch estimates and the goodness of fit evaluated in a rather ad hoc fashion.

While there exist a large number of different pitch estimation techniques, none are clearly superior [4]. Moreover, these techniques have not been thoroughly compared with each other because of their widely differing nature and the difficulty associated with obtaining a reference pitch estimate.

This report draws strongly from an earlier one [1] and is intended mainly to make the material presented there more accessible. While some of the techniques mentioned in this report has since been surpassed by more powerful statistical modeling techniques, they still provide useful information on the issues involved in estimating pitch. As such this report forms the basis for a later one, that details pitch estimation using a cepstral smoothing technique.

Section 2 discusses issues in pitch estimation in more detail and continues to briefly mention and compare a number of algorithms, some classical and others more recent. Among the techniques the trend is towards more statistical approaches utilizing the information contained in large data bases. Notable are classification techniques that try to minimize a global error. A method to measure the quality of the pitch estimates is also introduced in this section. Section 3 discusses an implementation of a statistical approach to detect the segments of voiced speech. A cepstrum-based pitch estimation technique that adapts the length of the analysis window based on the pitch estimate is introduced in section 4. Section 5 presents various results relating to the cepstrum-based pitch estimator. These results are discussed in more detail in Section 6. Section 7 concludes.

2 Background

Rabiner et al. [4] performed in 1976 a comparative study of pitch estimation algorithms for telephone, microphone and wide band recording conditions against a reference pitch signal. The techniques of pitch estimation were regarded as widely representative at the time. These together with more recent techniques are briefly surveyed here. The weaknesses and strengths of individual techniques are highlighted. The survey is intended to give a broad overview of the issues involved in pitch estimation.

2.1 Autocorrelation technique

This is a time domain technique. The speech signal is lowpass filtered to 900 Hz. Fig. 5 shows a block diagram of the technique. Frames of 30 ms length are extracted 100 times per second to account for the 10 ms stationarity of the speech. Each frame is center clipped, according to a heuristic level and for computational efficiency then clipped to values of {-1,0,1}. The autocorrelation function is computed from 2 to 20 ms and then normalized. If the peak in autocorrelation exceeds 0.3 the section is classified as voiced, else as unvoiced. If it is classified as voiced, the position of the peak gives the pitch period. Additional silence or speech classification is done according to the amplitude energy.

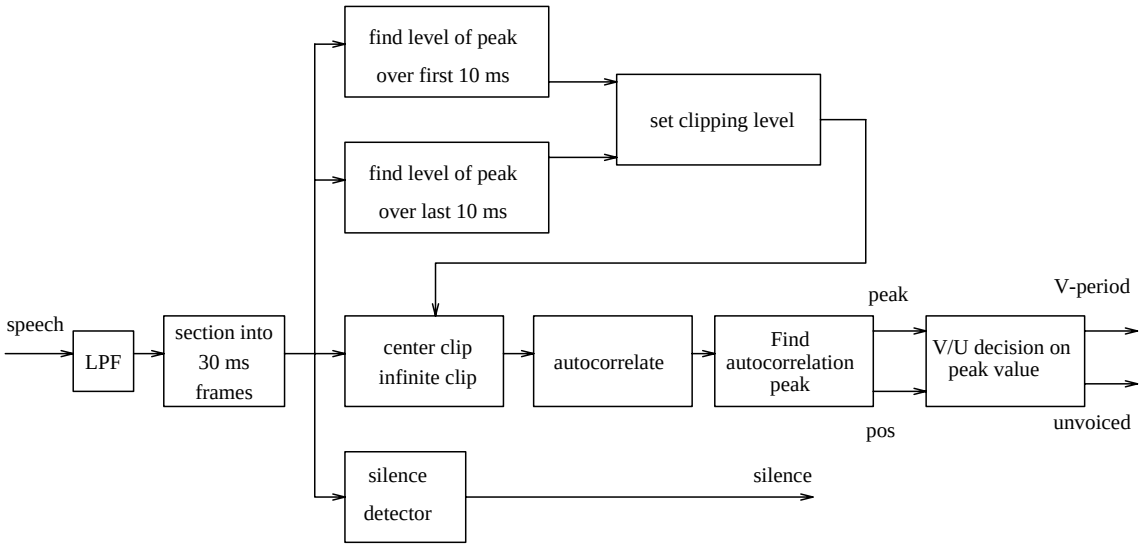


Fig. 5: Block diagram of the autocorrelation pitch estimator.

A confidence estimate in noisy conditions for the voicing estimate is introduced by Krubsack in [5]. Three features are used:

- The RMS energy of a speech segment.
- The normalized value of the maximum correlation over the pitch range.
- The normalized energy of the correlation over the pitch range.

Using the last two features Krubsack shows that voiced and unvoiced speech are partly separated in the plane of these features and that the distance of a speech sample from a decision boundary in this plane can be used as a confidence estimate when noise is present. This statistical approach will later be used in a modified form in the cepstrum implementation described in section 4.

2.2 Average magnitude difference function technique (AMDF)

The number of short-time zero crossings and the short-time energy are used for voiced-unvoiced classification. See Fig. 6. An average magnitude difference

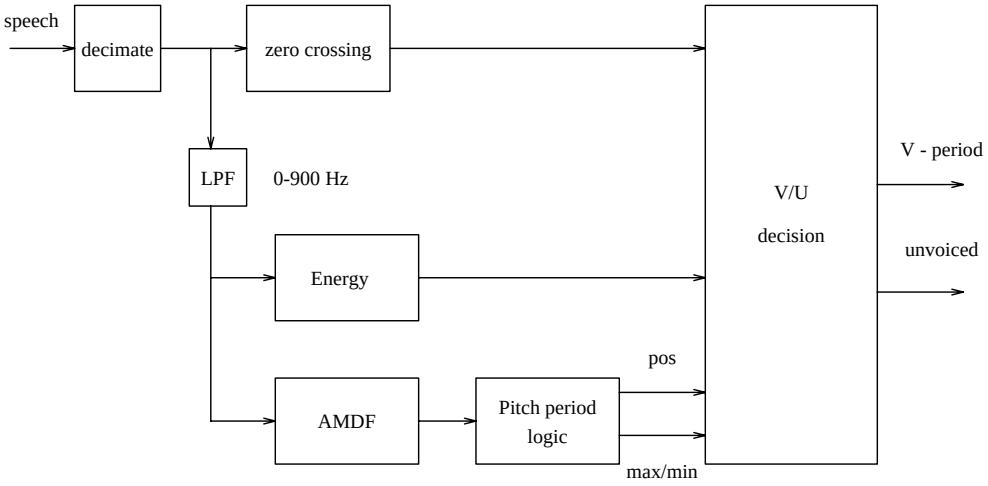


Fig. 6: Block diagram of the average magnitude difference function technique.

function is used in a way similar to the autocorrelation technique. The pitch period is obtained from the location of the minima of this function. This technique is faster than the autocorrelation technique, but 'coarser'.

2.3 Simplified inverse filtering technique (SIFT)

This is also known as the spectral equalization and transform technique. Consecutive 40 ms long speech segments are lowpass filtered to 900 Hz and decimated according to the Nyquist criterion. See Fig. 7. An inverse 4th order LPC filter is used to spectrally flatten the input signal. This partly removes the effect of the vocal tract and leaves the glottal excitation spectrum. This 'error' signal is autocorrelated and interpolated at the peak to determine the pitch. A voiced

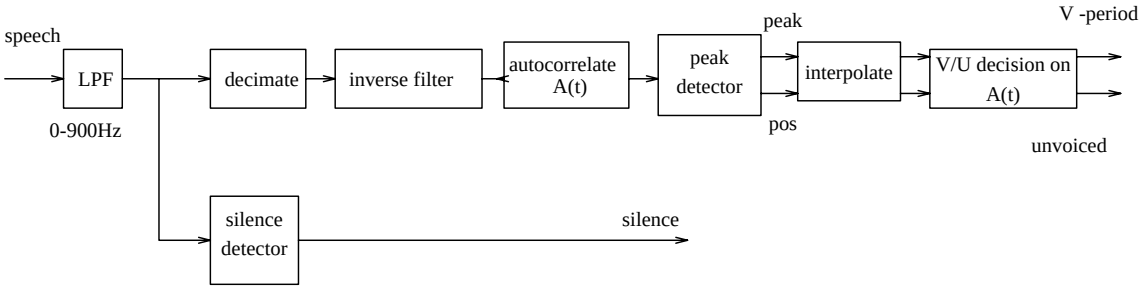


Fig. 7: Block diagram of the SIFT pitch estimator.

or unvoiced decision is based on the amplitude of the peak. A silence detector is also used. Markel and Gray [6] pointed out a weakness of spectral flattening which is that flattening with the spectral estimate for nasals tends to corrupt the pitch information. This is due to the fact that the LPC spectral estimate for nasals is poor because of the presence of zeros [7]. This algorithm requires special bandpass filtering to limit the sensitivity to zeros of the spectral estimate. This can be considered a hybrid technique because of spectral flattening done in the frequency domain and autocorrelation done in the time domain.

2.4 Cepstrum technique

The speech signal is chopped into 32 ms long segments and each segment is weighted with a Hamming window. See Fig. 8. The cepstrum, defined in a

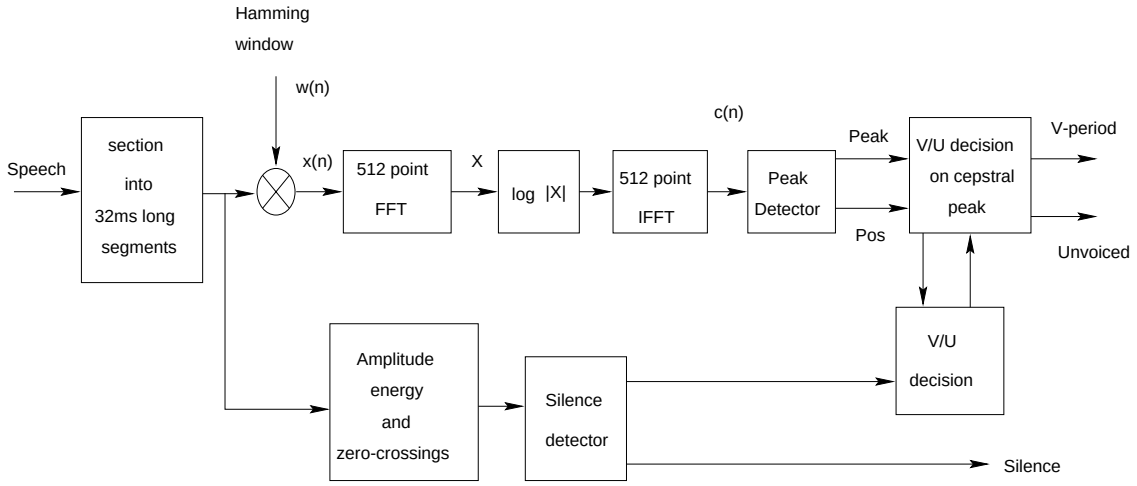


Fig. 8: Block diagram of the cepstrum pitch estimator.

later section, is then computed. As in the autocorrelation technique, if the peak

value in the pitch domain exceeds a certain threshold, the section is labeled voiced, with the peak location determining the pitch. If the peak is below the threshold and the zero-crossing rate is low, the section is classified as unvoiced. Additional silence or speech classification is done according to the amplitude energy. This technique can be regarded as a frequency technique because of the deconvolution that occurs in the frequency domain.

2.5 Parallel processing technique

This time-domain technique [8] has found widespread application in real-time systems. The signal is lowpass filtered to 900 Hz. See Fig. 9. Six impulse function signals denoting peaks and valleys are measured. Pitch periods are obtained from each function. These are then combined to determine the pitch period, while a voiced or unvoiced decision is based on their agreement. Silence detection is also used. The idea is described in more detail below. The outputs

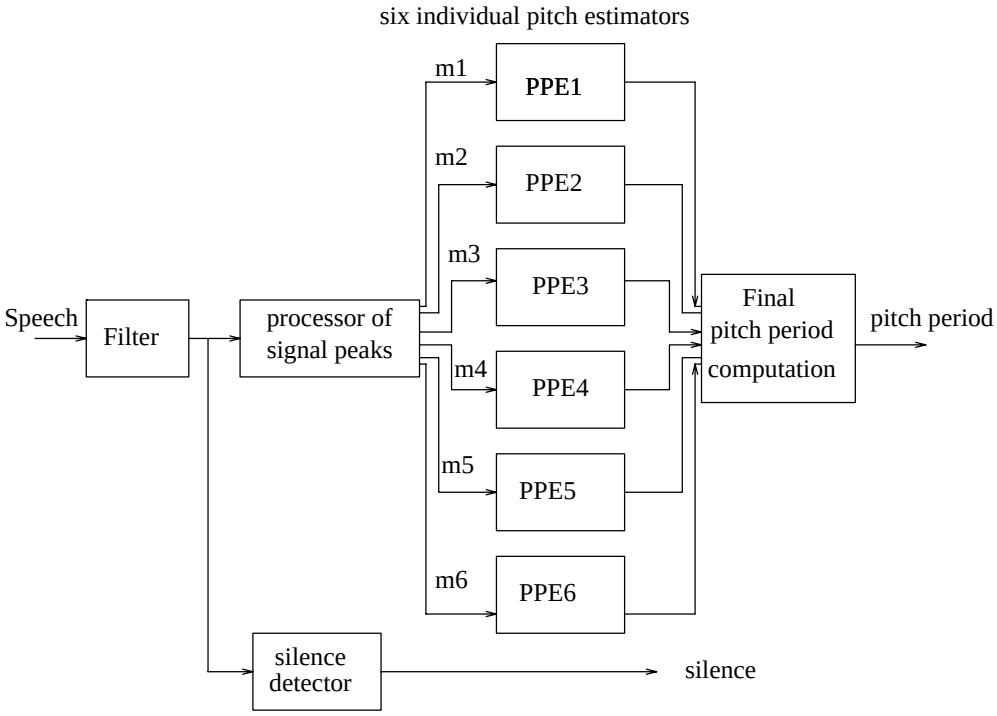


Fig. 9: Block diagram of the parallel processing technique.

of multiple elementary pitch period estimators are combined in parallel. The speech signal is band filtered to between 100 and 900 Hz and peak and valley measurements made. Six different peak-valley functions are generated and for each a preliminary pitch is estimated. The six estimates are combined and a final pitch produced.

The following measurements are made (see Fig. 10): If m_2 or m_6 are negative,

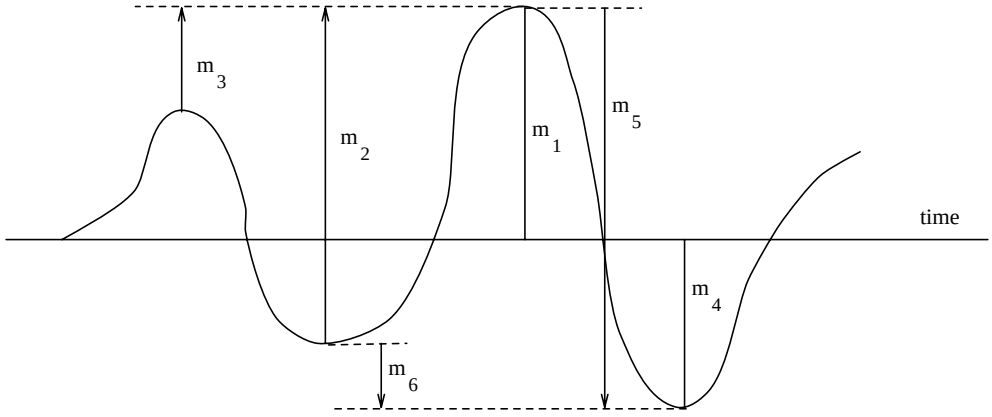


Fig. 10: The peak and valley measurements.

they are set to zero. Each peak is blanked and then decayed.

$$\tau = 0.4P_{av} \qquad\qquad \beta = \frac{P_{av}}{0.695}$$

where β is the blanking interval, τ the decay rate and P_{av} the average pitch period.

A 6x6 matrix with entries from each function is formed:

- The first three rows are the three most recent estimates.
- The fourth row is the sum of the first and second.
- The fifth row is the sum of the second and third.
- The last row is the sum of the first three rows.

The last three rows are in case of 'halving' type errors. Each entry is compared to the 35 others, four times. Each of the four corresponds to a different window length by which the entries can differ while still being taken as similar. A similarity count is made for each entry in the first row and bias is corrected. The preliminary pitch with the highest similarity count is taken as the true pitch. If there is wide disagreement among the preliminary pitch estimates the speech is classified as unvoiced. The algorithm is reported to work well if the pitch frequency is below 220 Hz.

This technique was recently reimplemented by Hassanein [9]. He adapts the window lengths according to the pitch estimate and uses additional zero crossing and energy information to decide whether the speech is voiced or unvoiced. He implemented his parallel processing technique on a TMS320 for a 2400bits/s LPC vocoder.

2.6 Pitch estimation with banks of bandpass filter-pairs

This technique [10] is a frequency domain approach and is based on modern vector quantization techniques. One thousand bandpass filter-pairs are used to extract spectral information relating to the harmonics below a frequency of 1000 Hz. The filter outputs are grouped as a vector and pattern matched to a precalculated code book of reference vectors. Fig. 11 shows one of the filter-pairs. The H_{c+} and H_{c-} are bandpass filters with center frequencies $f_c + \Delta f$

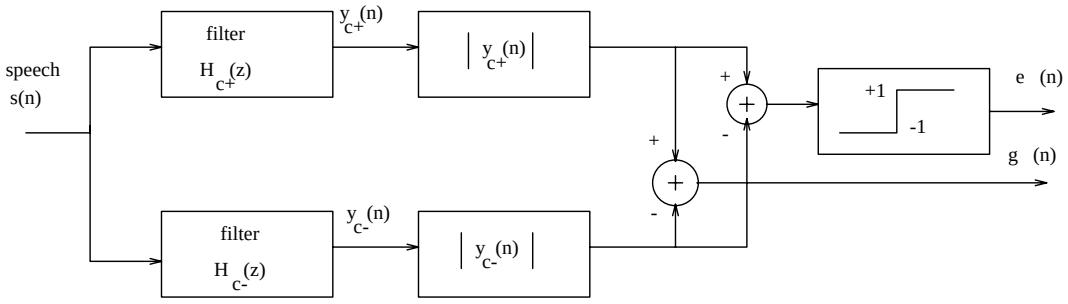


Fig. 11: A bandpass filter-pair.

and $f_c - \Delta f$ respectively. The time average of $e_c(n)$ approaches 1 or -1 according to $f_s > f_c$ or $f_s < f_c$ for input f_s. The time average of $g_c(n)$ becomes low when f_s deviates from f_c. This means that if $f_s > f_c$ the output $e_c(n)$ is positive while $g_c(n)$ is high near each harmonic frequency. The e_i and g_i are smoothed to obtain a slope and power level vector respectively. The pitch are resolved after pattern matching a reference vector (obtained from synthetic data) to the obtained vectors **e** and **g**. Multple candidate frequencies are checked. A poor voiced or unvoiced decision, based on this multiplicity of candidates at higher harmonics, are reported to be the draw back of this technique. Also, pitch resolution is low due to the pattern quantization.

2.7 Pitch estimation with a neural-net classifier

This technique is a time domain technique [11] using a neural network. It is essentially and extension of an earlier 'data reduction' technique described by Rabiner [4]. In the original technique, the speech signal is lowpass filtered and excursion cycles between the zero crossings extracted. Additional energy measurements are made for speech or non-speech and voiced or unvoiced classification. Ad hoc logic is used to place pitch markers on the signal peaks corresponding to pitch. In the neural network system, the 'logic' necessary to place pitch markers is optimized statistically in the form of a neural network using a large speech corpus. Two systems are implemented: one using the waveform samples directly and one using the waveform peaks.

Using the waveform samples
The filtered speech signal is downsampled according to the Nyquist criterion and a number of samples, spaced 0.375 ms apart, are extracted. Invariance is limited when switching from male to female speech. A total of 41 samples are used with the neural network containing 10 hidden units. The lowest attainable error against visually estimated pitch is reported to be 2.5%.

Using the waveform peaks
For neural-net classification on peaks it is critical that invariant features are used. It is known that:

- Pitch peaks are generally larger than neighboring peaks.
- Amplitudes decrease intermediate to pitch peaks.
- Successive peaks are equally spaced.

Information from seven adjacent peaks are used and again a neural-net with 10 hidden units are used. The peak features used are:

- Normalized amplitude.
- Time difference between peaks.
- Correlation of signal between peaks.
- Normalized zero to zero crossing width.
- Negative amplitude between every pair of positive peaks.

The signal is lowpass filtered to 700 Hz. The classification is then a two class problem: label peaks as pitch peaks or not. This technique is therefore similar to the data reduction technique with the basic difference that a neural network is used instead of elaborate decision logic. Non-vocalized parts of the waveform are classified as non-pitch. The peak excursion peaks usually contain most information and the positive peaks are used because they correspond to positive presure at the lips.

Using the waveform peaks, the lowest obtained error rate using peak information against visually estimated pitch is reported to be 2 %. The authors noted that this technique fails on band-limited telephone speech [12]. They report errors due to ambiguous peaks, weak signals and transitions. The neural-net technique shows the trend to intensive use of data, where 20000 training samples were used from 80 different speakers.

2.8 Post-processing

Since estimates of the pitch tend to be noisy, the pitch is smoothed using a median filter. A 5'th order median filter is typically used.

2.9 Estimation error

Rabiner [4] defined four types of errors when comparing a pitch estimate and reference signal of the pitch periods. In the voiced region:

$$e = |\hat{P} - P|$$

where e is the error, $\hat{P}$ the estimated pitch period and P the reference pitch period. The four errors are defined as follows.

1. If the error is more than 1 ms it is classified as a *gross* pitch error. This error is usually due to pitch doubling or formant suppression such as found during nasalization.

2. If the error is less than 1 ms it is classified as a *fine* pitch error. These errors are normally attributed to measurement.

3. Misclassification of the transition from the voiced to unvoiced region is classified as a voiced to unvoiced error (V-U error).

4. Misclassification of the transition from the unvoiced to voiced region is classified as an unvoiced to voiced error (U-V error).

Based on these errors, gross error count, mean of fine pitch error and standard deviation of fine pitch error are useful statistics.

We repeat the conclusions of Rabiner's comparison of the pitch estimation techniques briefly. For more details the reader are referred to [4].

Gross pitch errors The cepstrum technique performed best at low pitch frequencies and overall. The parallel processing technique performed well at higher frequencies.

Fine pitch errors No technique exhibited significant bias so that fine pitch errors were difficult to estimate. The cepstrum technique had the lowest standard deviation at high frequencies, while the parallel processing technique performed less well due to its relatively low time domain resolution.

V-U errors The cepstrum technique had the poorest performance.

U-V errors The cepstrum technique had the best performance.

Generally the cepstrum and parallel processing techniques outperformed the others. No estimator outperformed any other over all types of errors. The time domain techniques did best at high pitch frequencies because more pitch periods were contained in the time window. The frequency domain techniques did best at low pitch frequencies because fewer harmonics of the fundamental were included in the relatively short analysis window.

3 Detection of voiced speech

Before pitch can be estimated it is necessary to decide if the speech signal is speech and then if speech, if it is voiced [13]. This decision is based on a statistical approach. Features on which these decisions can be based include: short-time energy, zero-crossing rate and the residual from eg. LPC-filtered speech. In the following we use short-time energy and zero-crossing rate.

The short time magnitude is computed every 10 ms for a 10 ms frame length. (At a sampling rate of 16000 kHz this means that a frame consists of 160 samples.) The short time average magnitude is defined as

$$M(n) = \sum_{m=n-N+1}^{n} |s(m)|w(n-m)$$

where N is the window length and $w(m)$ a hamming window. Similarly the zero-crossing rate is computed for each frame [14]. The number of zero crossings in a window is defined as

$$Z(n) = \sum_{m=n-N+1}^{n} |(\mathrm{sgn}(s(m)) - \mathrm{sgn}(s(m-1)))|w(n-m).$$

These features are estimated from a large number of speakers in the TIMIT corpus [15] excluding those on which the pitch are to be derived (refer Section 5). Analysis then allows the choice of thresholds and decision boundaries according to probability density. Fig. 12 shows a scatter plot of non-speech, voiced and unvoiced samples as a function of magnitude and zero crossings. It can be seen

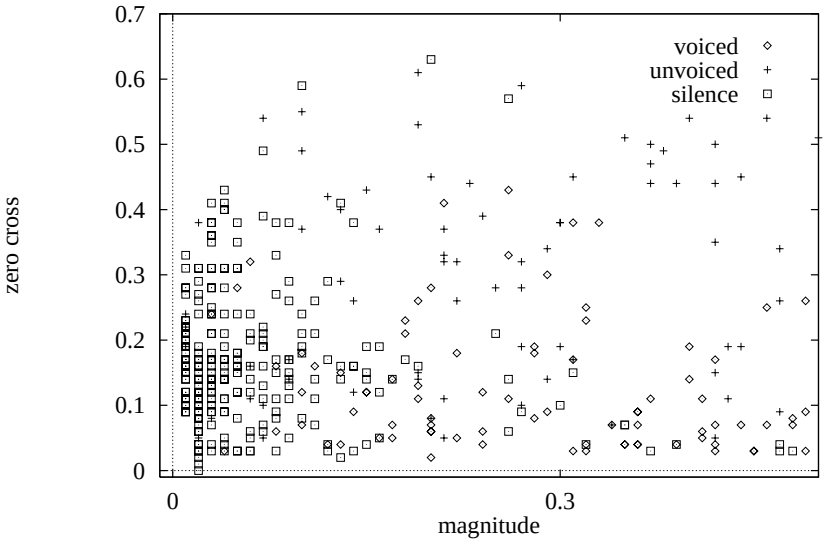
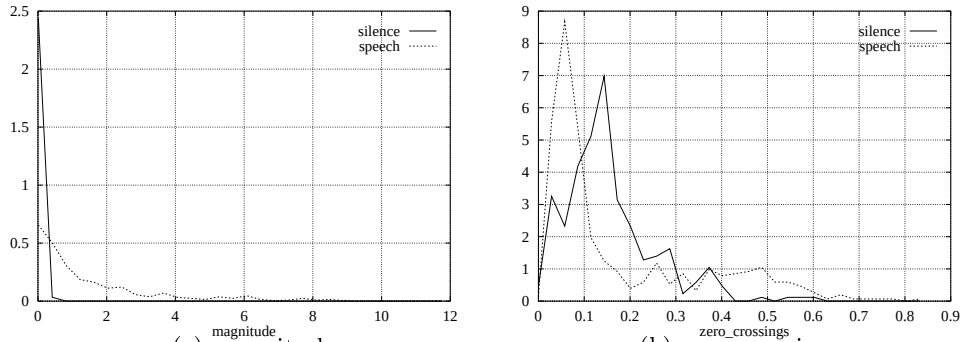


Fig. 12: Scatter plot of non-speech, voiced and unvoiced speech samples.

that a choice for non-speech is to choose non-speech if $Z < -6.25M + 1$ and $Z < 0.45$. The speech or non-speech and voiced or unvoiced decision is made according to the posteriori probability density functions (pdfs) with the choice of classification boundaries as described below.

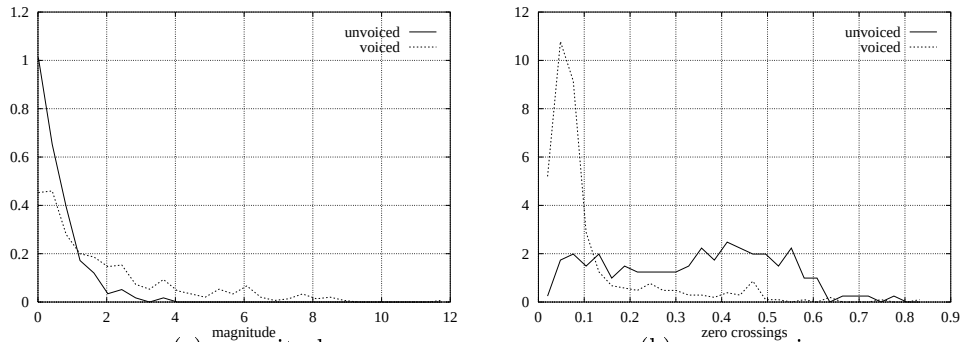
Speech vs. non-speech Fig. 13 (a) shows the pdfs as a function of magnitude. Here the choice for non-speech is $M < 0.14$. Fig. 13 (b) shows the pdfs as a function of zero crossings. Here the choice for non-speech is $0.1 < Z < 0.45$.

Voiced and unvoiced speech Fig. 14 (a) shows the pdfs as a function of magnitude. Here the choice for voiced speech is $1.2 < M$ by ignoring the effect of the zero crossings. Fig. 14 (b) shows the pdfs as a function of zero crossings. Here the choice for voiced speech is $Z < 0.2$.



(a) magnitude (b) zero crossings

Fig. 13: The pdfs of speech vs. non-speech as a function of (a) magnitude and (b) zero crossings.



(a) magnitude (b) zero crossings

Fig. 14: The pdfs of voiced vs. unvoiced speech as a function of (a) magnitude and (b) zero crossings.

4 Cepstrum technique with variable analysis window length

We implement the cepstrum technique of pitch estimation [16, 4, 17, 18]. This technique is generally regarded as being more robust to noise and channel effects as present in telephone speech.

4.1 Cepstrum

In the technique, the cepstrum [16] is used as a means of deconvolving the glottal excitation and vocal tract filter responses according to the model of speech production (refer Fig. 1). Fig. 15 depicts this process.

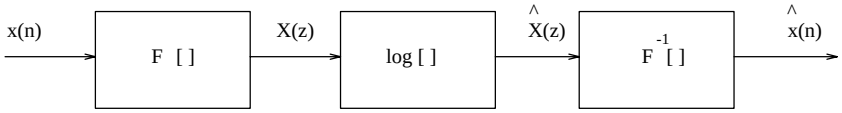


Fig. 15: A block diagram for obtaining the cepstrum.

The speech signal $s(n) = p(n) * h(n)$ is modeled as the convolution of an impulse train $p(n)$ and impulse responses $h(n)$ of the glottal distortion, vocal tract and radiation models. By applying a short-time analysis window $w(n)$ to the speech signal and transforming to the frequency domain

$$X(z) \approx S(z) = P(z)H(z),$$

where $X(z)$ should be interpreted as having been estimated subject to the analysis window. Then in the logarithmic domain

$$\begin{aligned} \hat{X}(z) &\approx \log[X(z)] \\ &\approx \log[P(z)H(z)] \\ &\approx \log[P(z)] + \log[H(z)]. \end{aligned}$$

By transforming back to the time-domain a linear separable representation $c(n) = \tilde{x}(n) \approx \tilde{p}(n) + \tilde{h}(n)$ is obtained. This representation is known as the *cepstrum*. Impulses in the cepstrum correspond to periodicity of the input signal. The location of such an impulse is proportional to the period of the original impulse train and the energy of such an impulse is related to the energy of the original impulse train. The low time information in the cepstrum (periodic components with a high fundamental frequency) corresponds to the vocal tract transfer function while the high time information (periodic components with a low fundamental frequency) corresponds to the pitch. Fig. 16 shows the cepstrum for a speech signal.

4.2 Adapting the analysis window length

It is known that to obtain a reasonable estimate of the pitch period it is necessary for the analysis window to include more than one pitch period [19, 20]. For

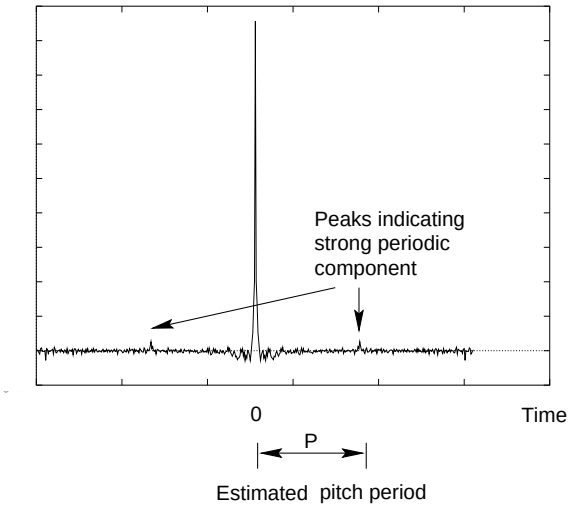


Fig. 16: Cepstrum of a periodic signal.

shorter window lengths, the estimate may be overly sensitive to harmonics of the pitch and pitch halving may occur in the estimate. However, if the length of the analysis window is much longer than two pitch periods then the analysis no longer agrees with the short-time stationarity assumption made of the speech signal and the estimates may become noisy and smeared. In [19, 20] the length of the analysis window was adapted based on previous estimates of the pitch period. Accordingly, here we use a short-time analysis window length of 16 to 64 ms corresponding to a window of 256 to 1024 samples at a 16 kHz sampling rate. This allows the useful dynamic range of the estimator to be increased. The complete pitch estimator is shown in Fig. 17. The pitch frequency is computed between 80 and 500 Hz corresponding to a pitch period of 2 to 12.5 ms or 32 to 200 samples. The cepstrum is computed for every frame and scanned for a peak between 2 and 12.5 ms from the zero time lag point. If the signal is voiced the peak distance is taken as the pitch period. The length of the analysis window is adapted according to the previous 15 estimates of the pitch (15 frames or 150 ms of speech) as follows:

- An initial 140 Hz pitch frequency is assumed and the window length chosen as 512 samples.
- If the average pitch estimate falls below 5 ms then the window length is adjusted to 256 samples.
- If the average pitch estimate rises above 8 ms then the window length is adjusted to 1024 samples.

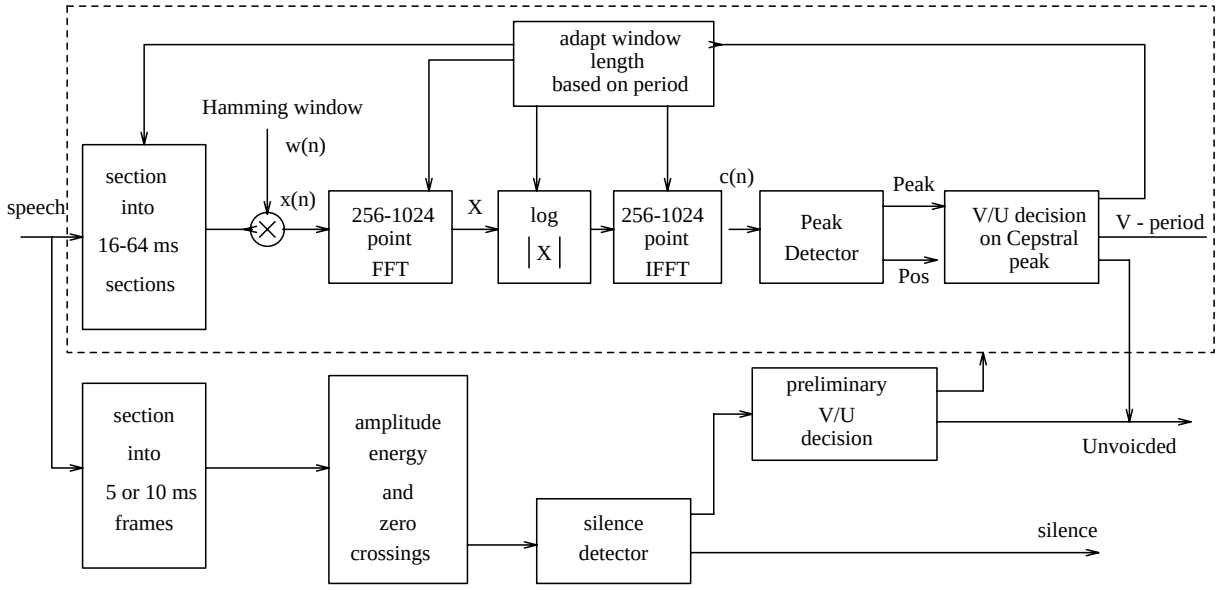


Fig. 17: Block diagram of the cepstrum pitch estimator.

It is assumed that the unadapted pitch estimator will give good enough estimates in the 150 ms of speech so that a threshold shift in the average pitch estimate indicates a valid shift in the pitch. In the few cases that the estimate becomes too noisy the adaptation is restarted. In this case the parameters for the window lengths are initialized to the values necessary for the estimation of a 140 Hz pitch frequency. For computational reasons, we use the Fast Fourier Transform (FFT). This limits the lengths of the analysis window to the three possible values chosen. When the DCT or DFT are used the number of window lengths can in principle be increased.

4.3 Post filtering

The pitch period that is obtained from the cepstrum contains sporadic outliers corresponding to among other, pitch doubling and halving errors. With the variable analysis window length these errors are greatly reduced. To smooth the pitch estimates a median filter of length 5 is used once. This proved to be sufficient, with little smearing. Further smoothing may be necessary if the pitch is to be used in a coder or decoder.

5 Pitch estimation results

Pitch estimation results for processed utterances from three speakers [3] in the TIMIT data base are presented here. Among the speakers the minimum average pitch frequency is 80 Hz and the maximum 300 Hz. Information on each speaker S_i as well as the phonetically rich utterance U_i spoken by that speaker is listed below:

S_1 A female from the Southern part of North America. She is 5'06" high with a low pitched voice of about 140 Hz.
U_1 `'ralph controlled the stopwatch from the bleachers'`

S_2 A male of height 6'00" with a pitch of about 80 Hz.
U_2 `'salesmanship is still necessary but it's a different brand of salesmanship'`

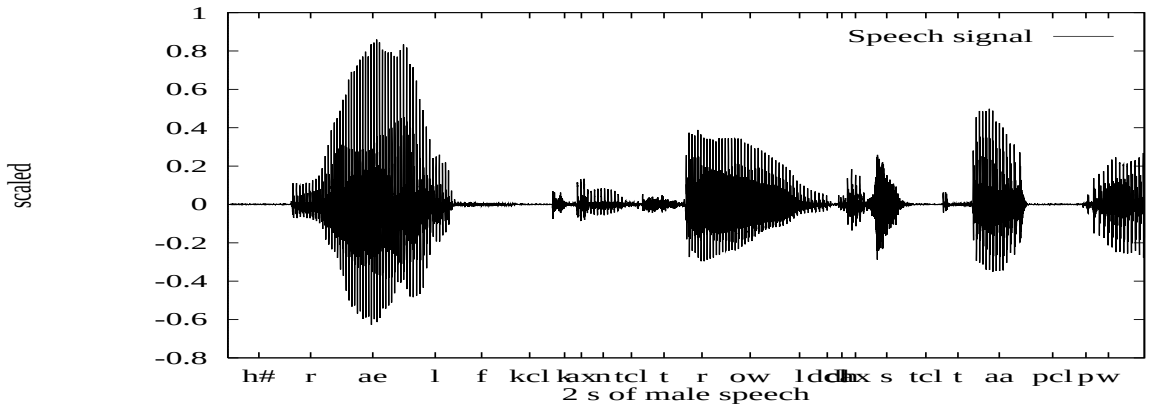
S_3 A female of height 5'4" with a pitch of about 300 Hz.
U_3 `'weatherproof galoshes are very useful in seattle'`

The estimates of the pitch periods of these speakers will be studied in more detail in the following sections.

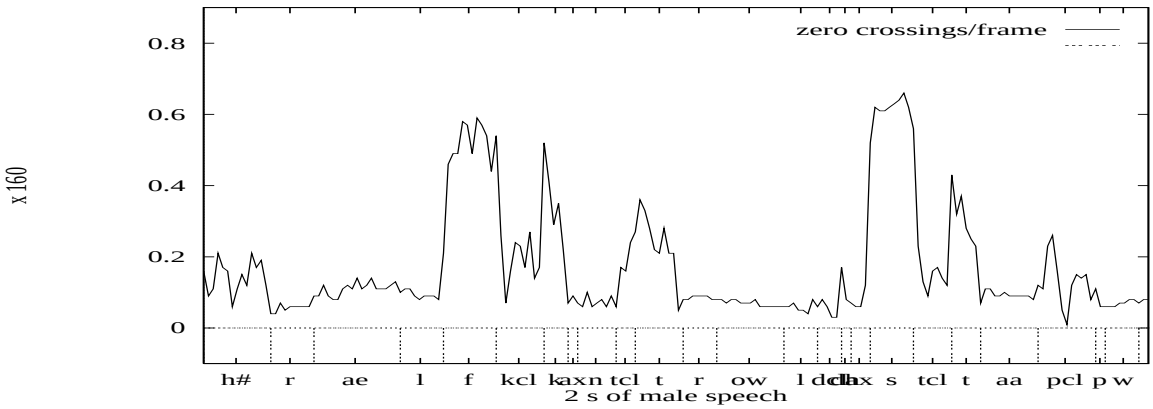
Fig. 18 shows a segment from the labeled utterance U_1 together with the short-time zero-crossing rate and average magnitude. Phoneme labels taken from the TIMIT corpus were used to determine how well the pitch estimation algorithm detected speech or non-speech and voiced or unvoiced segments. The estimated but unsmoothed pitch period for utterance U_1 is shown in Fig. 19. The length of the analysis window was not adapted. The pitch period is plotted in milliseconds. A scaled value of the detected peak in the cepstrum is also plotted. The pitch period is plotted as zero where the speech signal was classified as unvoiced and -1 where the speech signal was classified as non-speech. The plot of the cepstrum values shows that the proposed technique is fairly accurate at discriminating between the voiced and unvoiced sounds. It is seen that although the pitch period is a fairly smooth signal there are some points that lie at half or double the speaker pitch period. This may be due to a too strong harmonic or a weak signal. A too short or too long window length increases these errors.

To reduce some of these errors we smooth the pitch estimates using a length 5 median filter. We also investigate the effectiveness of adapting the length of the analysis window. For comparison purposes, results in the next section pertain to where the length of the analysis window *is not* adapted. In the section thereafter, the length of the analysis window *is* adapted.

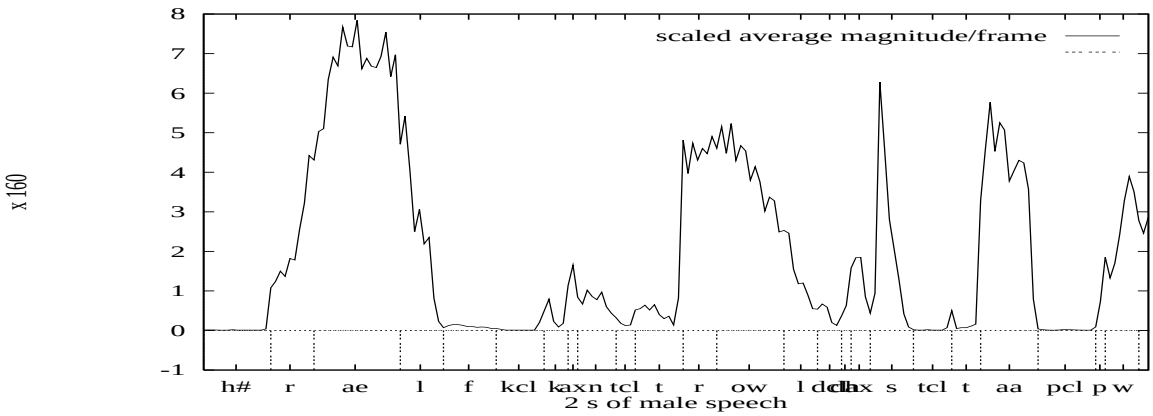
[3] The three speakers were chosen on the basis of the range of pitch frequencies spanned by them. Without loss of generality, the results for only these three speakers are presented here. The estimated pitch of other speakers conformed to the same observations and conclusions reached in this report.



(a) Speech signal of utterance U_1.



(b) Short-time zero-crossing rate of U_1.



(c) Short-time average magnitude of U_1.

Fig. 18: Speech signal, zero-crossing rate and average magnitude of U_1.

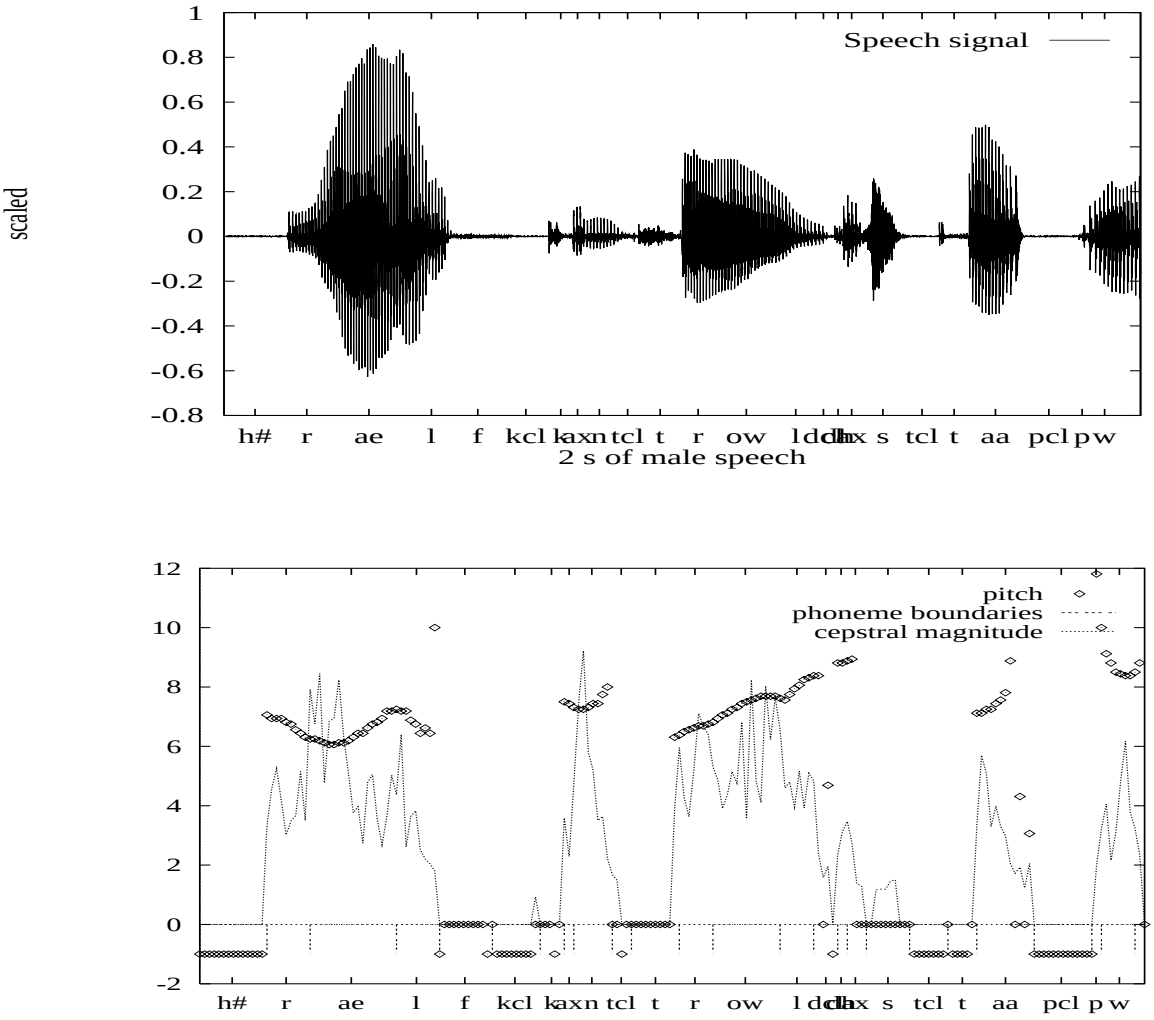


Fig. 19: Unsmoothed pitch period and cepstral peak values of utterance U_1.

5.1 Smoothed pitch periods of various speakers

The results are shown in Figures 20 to 22. Pitch period halving, where the estimated pitch period is 3.5 ms instead of the average 7 ms, is observed in Figures 20 and 21. As will be shown in the next section, this is an affect of the analysis window being too short. In these regions the estimated pitch was actually a harmonic of the pitch.

In U_2, where the first word is repeated again at the end of the utterance, it is informative to observe that the pitch of the same speaker can differ substantially within a single utterance – even for the same word. Also of interest is the misclassified /en/ phoneme at the end of the utterance. Here the zeros in the spectrum suppressed the spectrum and reduced the zero-crossing rate and spectral energy.

Pitch doubling, as visible in U_3, can be attributed to a too long analysis window length. The too long window allows the inclusion of amplitude modulation due to lower formants. The pitch estimate is then wrongly based on the period of this amplitude modulated signal.

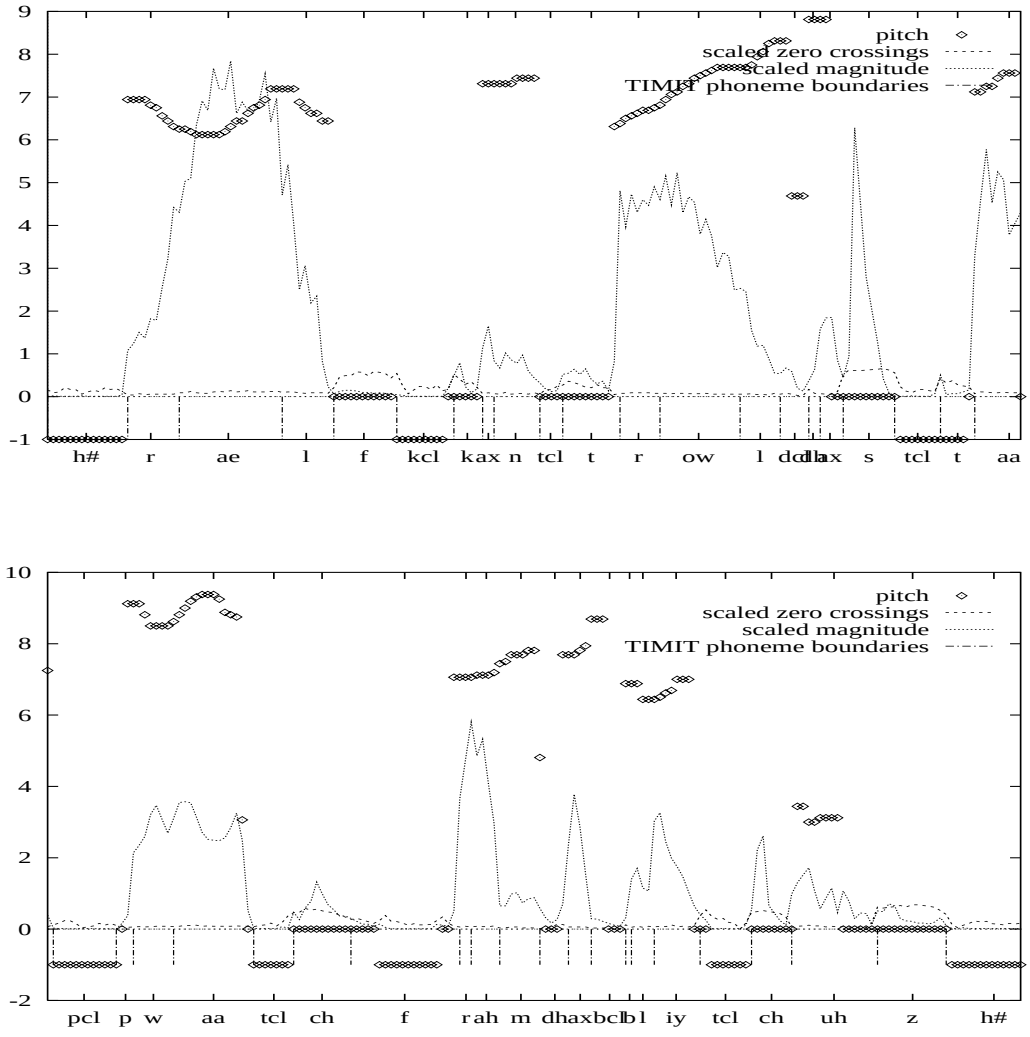


Fig. 20: Analysis window length not adapted: Pitch period (ms) for U_1.

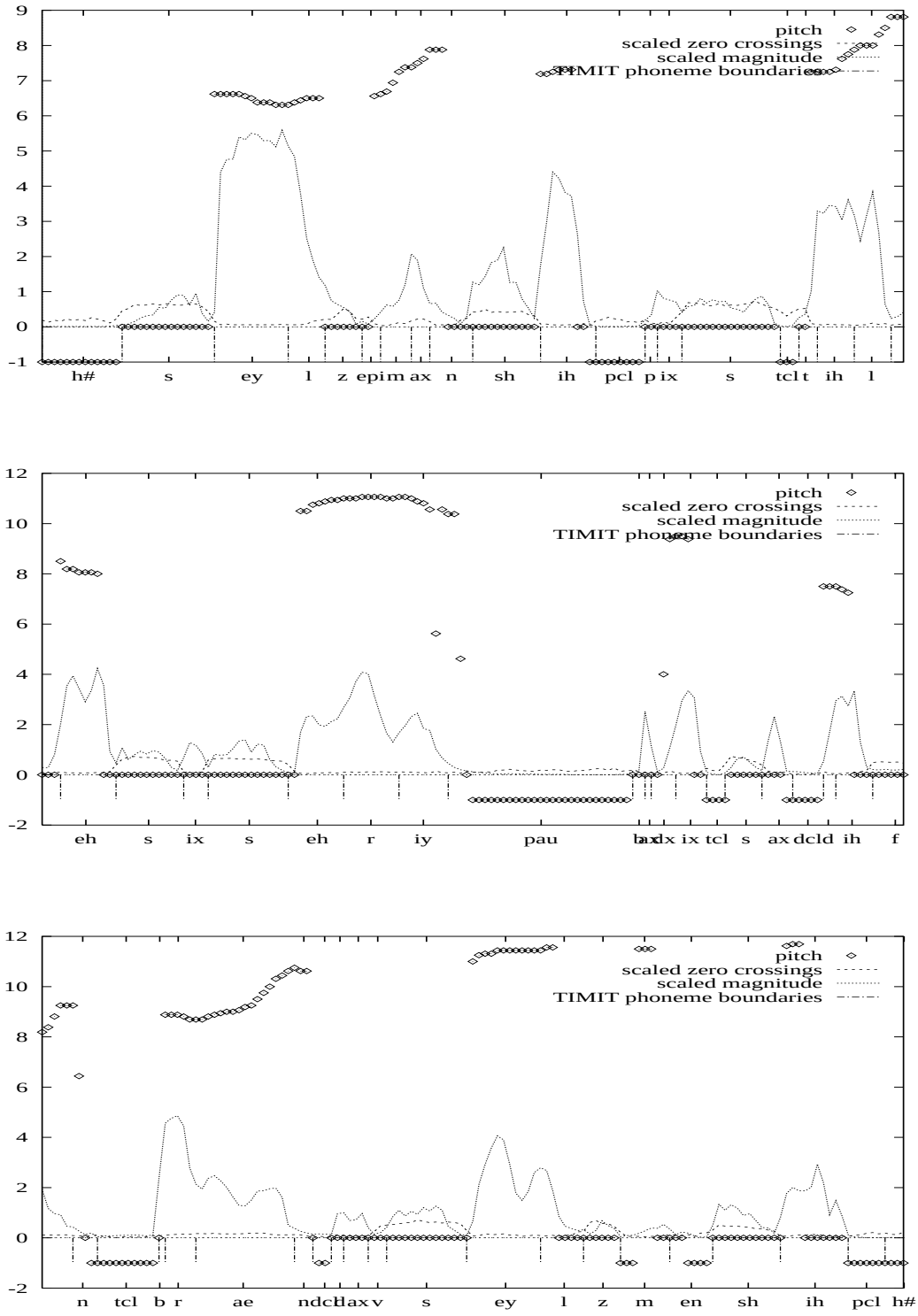


Fig. 21: Analysis window length not adapted: Pitch period (ms) for U_2.

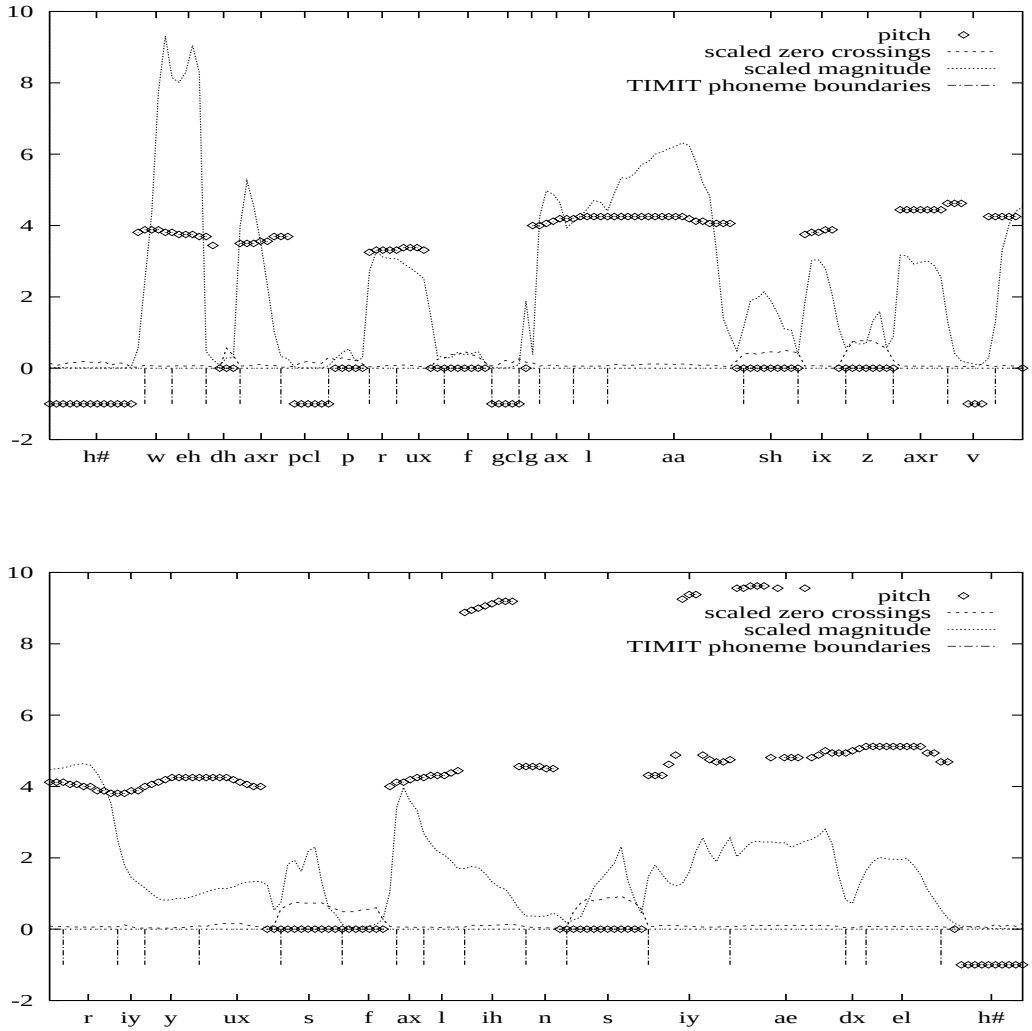


Fig. 22: Analysis window length not adapted: Pitch period (ms) for U_3.

5.2 Smoothed pitch periods of various speakers with adaptation of the analysis window length

Utterances U_1 to U_3 are again used, but with adaptation of the analysis window length. Pitch estimation errors are seen to decrease, with the previously observed errors absent as shown in Figures 23 to 25. The slight increase in

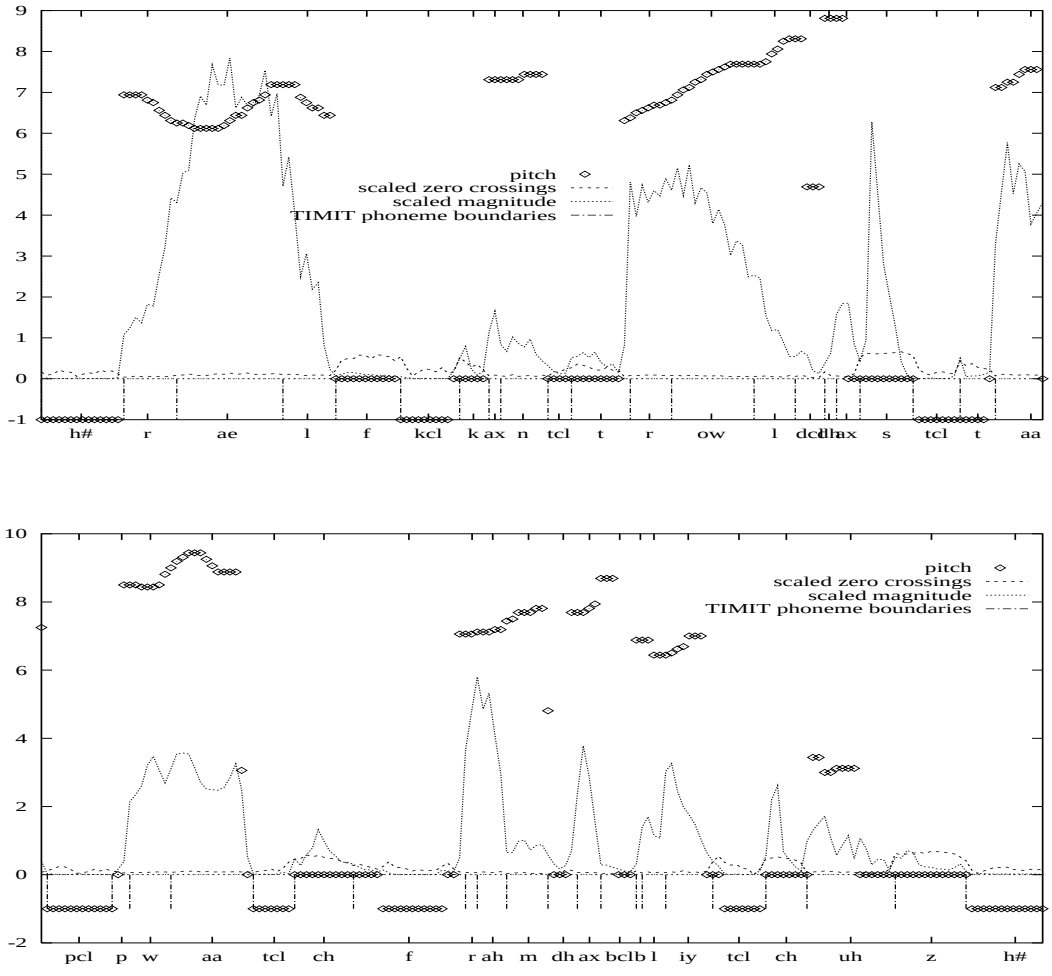


Fig. 23: Analysis window length adapted: Pitch period (ms) for U_1.

misclassification is ascribed to the fact that the cepstrum peak value used in the voiced-unvoiced classification was not adapted in step with adaptation of the analysis window length.

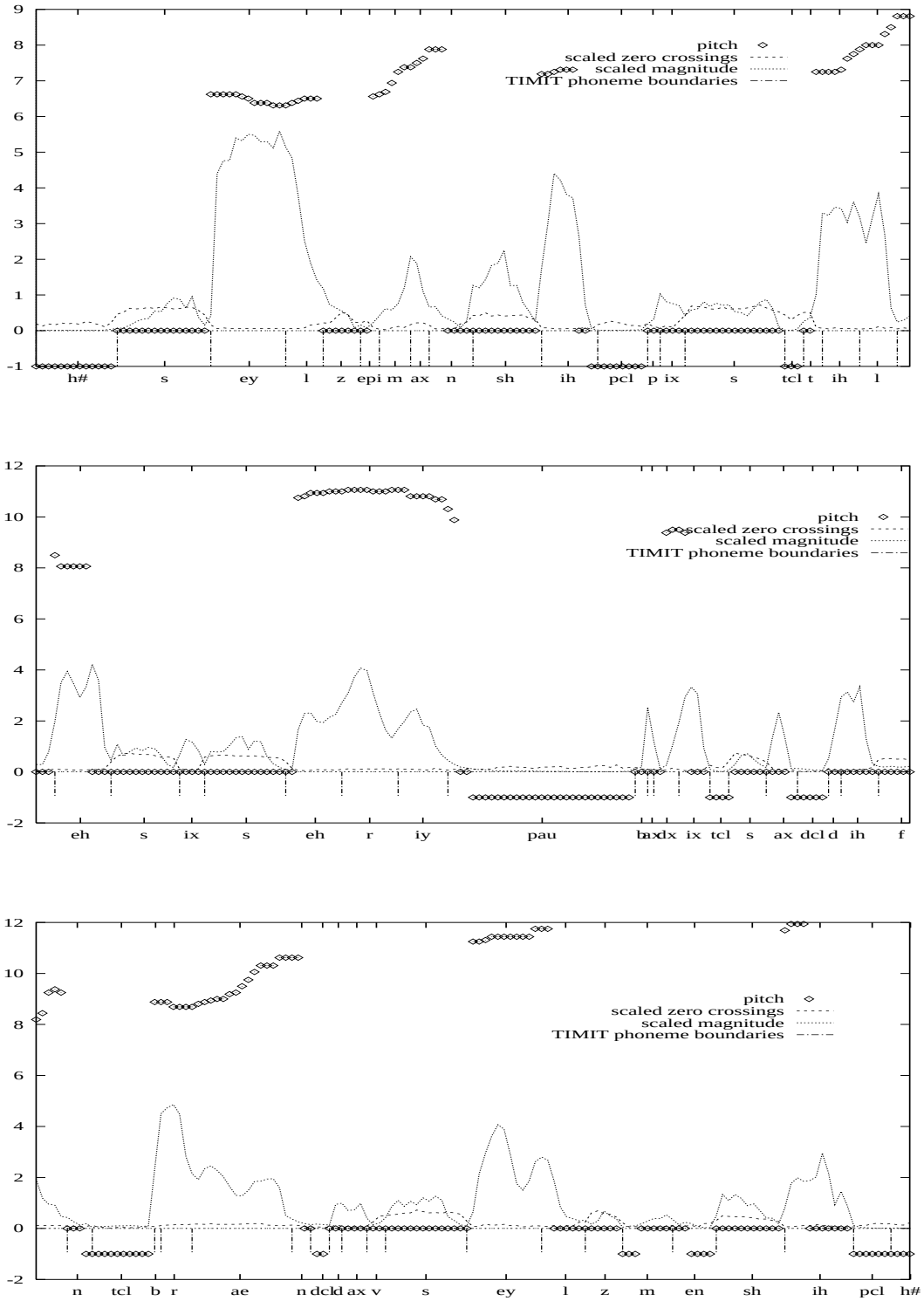


Fig. 24: Analysis window length adapted: Pitch period (ms) for U_2.

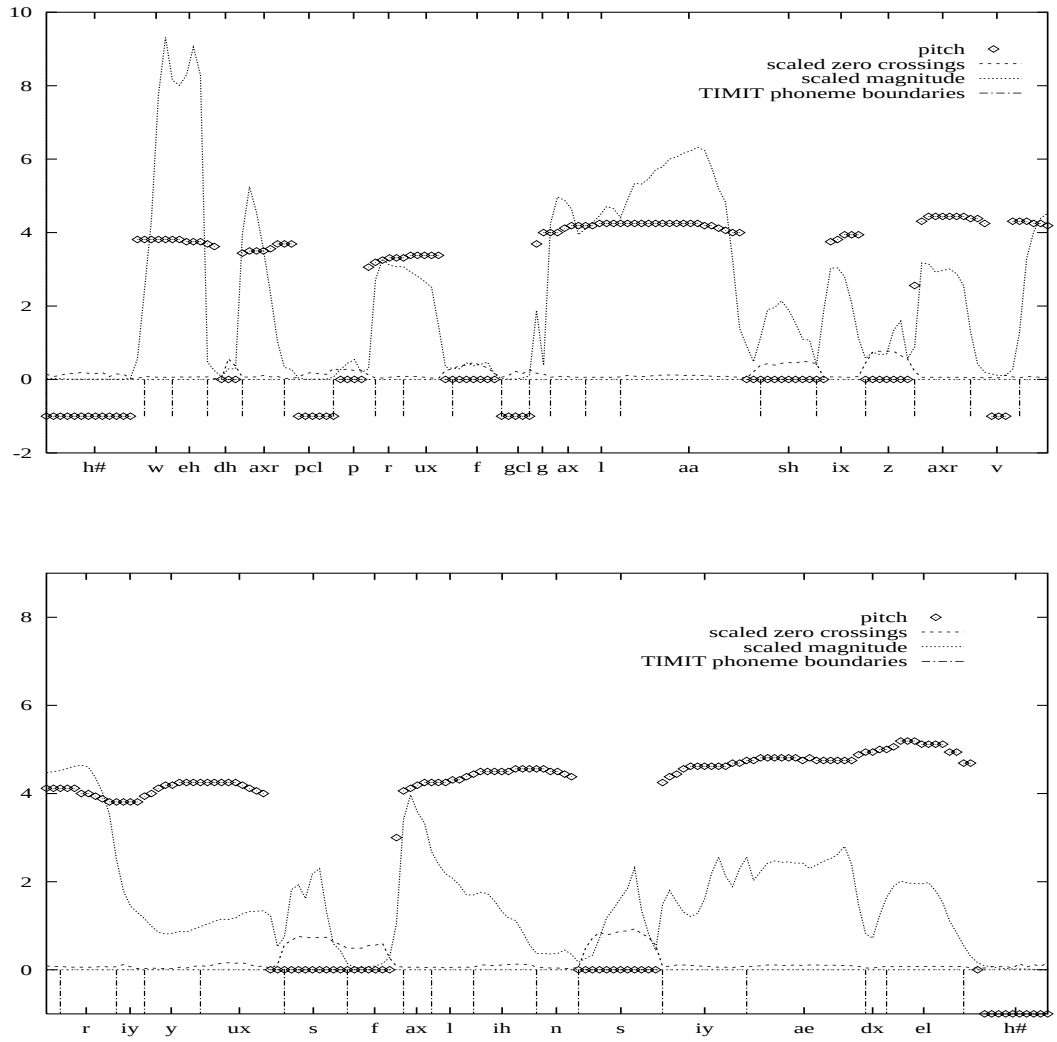


Fig. 25: Analysis window length adapted: Pitch period (ms) for U_3.

6 Discussion

The pitch period doubling observed in U_3 as shown in Fig. 22 with no adaptation of the analysis window length can be attributed to amplitude modulation of the speech signal. See Fig. 26 where the halving of the period can be seen in the

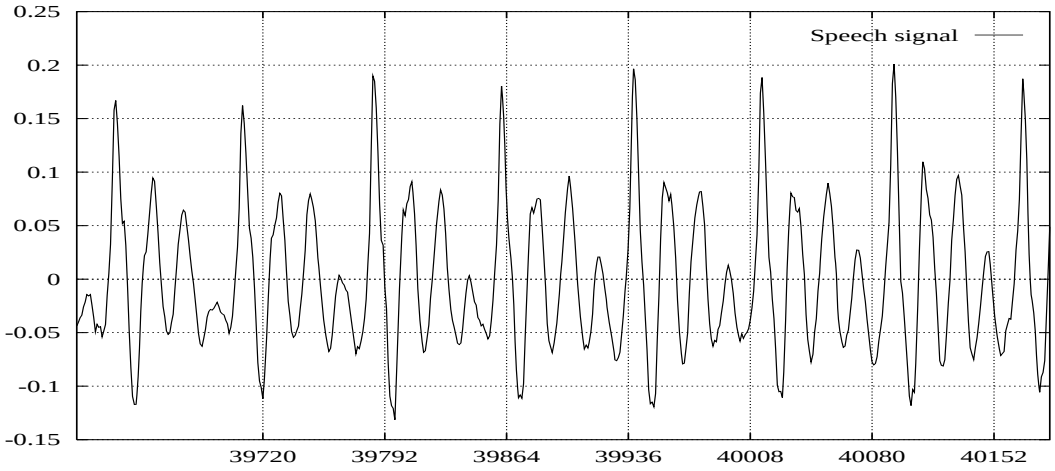


Fig. 26: Pitch period in the phoneme /ae/ where amplitude modulation is present.

middle part of the signal for the /ae/ phoneme. In the figure, it appears that alternating pitch frames are better correlated. The too long analysis window includes the longer period and causes the error. This phenomenon is a problem in most of the frequency domain and autocorrelation type techniques. A peak picking scheme or the proposed analysis window length adaptation helps to alleviate such errors.

Marks [21] pointed out that pitch is difficult to estimate for nasals. We remarked previously that time-domain techniques are in particularly affected by the presence of zeros in the signal due to the nasals. A nasal phoneme with cepstral-estimated pitch is shown in Fig. 27. It can be seen that the peaks in the time signal are indeed very noisy. Doing the estimation in the frequency domain instead, leads to a pitch period of 4.5 ms, which is in agreement with the instants of maximum excitation in the time signal.

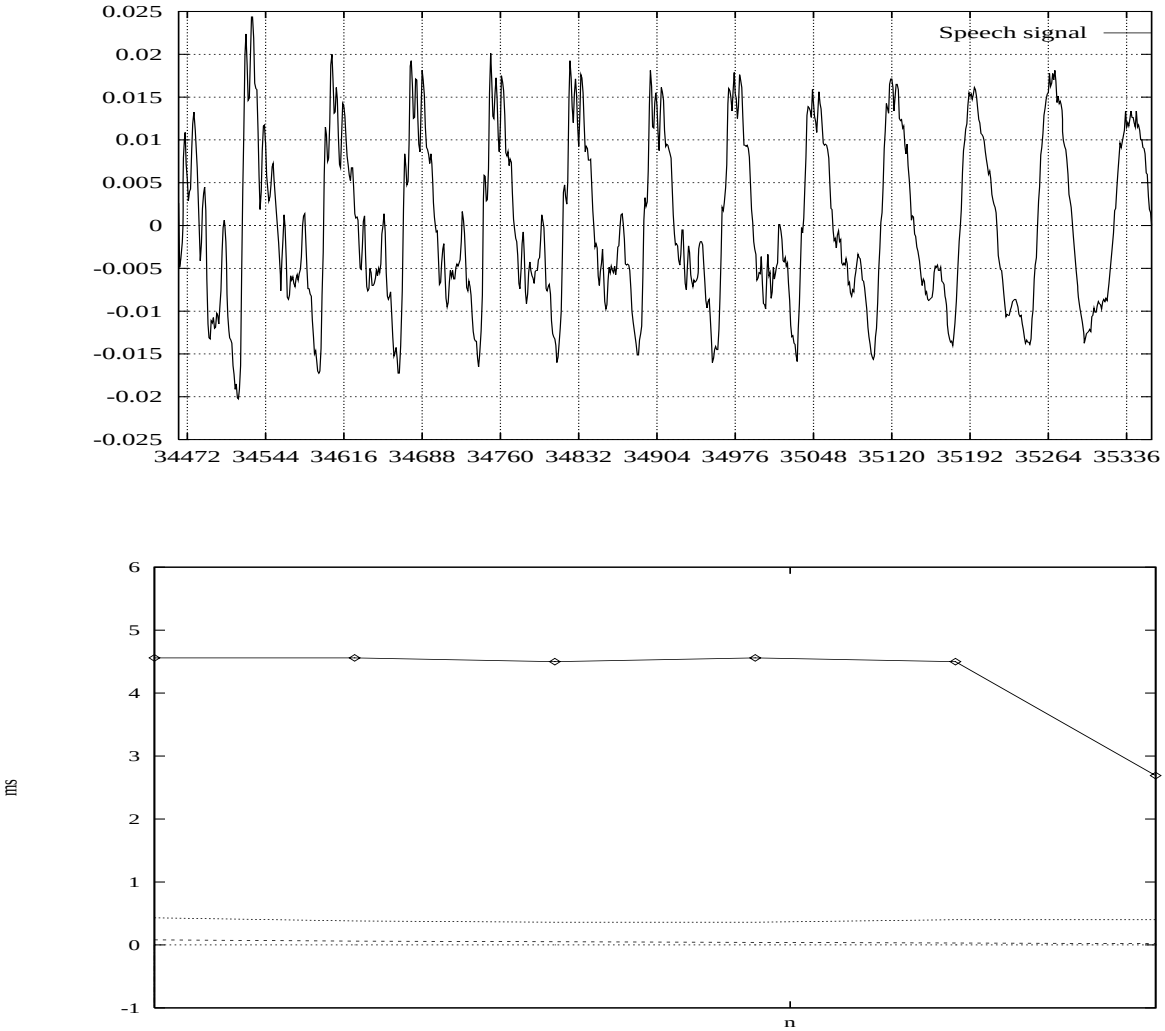


Fig. 27: Pitch period in a nasal /n/.

7 Conclusion

Various pitch estimation algorithms and their relative strengths and weaknesses were discussed. Among these the cepstrum technique of pitch estimation was selected. Provided that the length of the analysis window was adapted this technique performed well on speakers with pitch ranging from 80 to 300 Hz. A number of issues involved in obtaining a reasonable estimate of pitch were discussed. It was shown how pitch estimation in the time domain and frequency domain are affected differently depending on the length of the analysis window. It was shown that adapting the length of the analysis window can improve the estimates of the pitch made by cepstrum-based technique.

A Appendix

type	VOICED	LABEL	EXAMPLE WORD	TRANSCRIPTION
Stops	yes	b	bee	BCL B iy
	yes	d	day	DCL D ey
	yes	g	gay	GCL G ey
	yes	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
	no	p	pea	PCL P iy
	no	t	tea	TCL T iy
	no	k	key	KCL K iy
	no	q	bat	bcl b ae Q
Africates	yes	jh	joke	DCL JH ow kcl k
	yes	ch	choke	TCL CH ow kcl k
Fricatives	yes	z	zone	Z ow n
	yes	zh	azure	ae ZH er
	yes	th	thin	TH ih n
	yes	v	van	V ae n
	no	sh	she	SH iy
	no	s	sea	S iy
	no	f	fin	F ih n
	no	dh	then	DH e n
Nasals	yes	m	mom	M aa M
	yes	n	noon	N uw N
	yes	ng	sing	s ih NG
	yes	em	bottom	b aa tcl t EM
	yes	en	button	b ah q EN
	yes	eng	washington	w aa sh ENG tcl t ax n
	yes	nx	winner	w ih NX axr
Silence	NA	pau	pause	
	NA	epi	epenthetic silence	
	NA	h#	begin/end marker	

Table 1: TIMIT phonetic labeling.

type	VOICED	LABEL	EXAMPLE WORD	TRANSCRIPTION
Semivowels	yes	l	lay	L ey
glides	yes	r	ray	R ey
	yes	w	way	W ey
	yes	y	yacht	Y aa tcl t
	yes	hh	hay	HH ey
	yes	hv	ahead	ax HV eh dcl d
	yes	el	bottle	bcl b aa tcl t EL
Vowels	yes	iy	beet	bcl b IY tcl t
	yes	ih	bit	bcl b IH tcl t
	yes	eh	bet	bcl b EH tcl t
	yes	ey	bait	bcl b EY tcl t
	yes	ae	bat	bcl b AE tcl t
	yes	aa	bott	bcl b AA tcl t
	yes	aw	bout	bcl b AW tcl t
	yes	ay	bite	bcl b AY tcl t
	yes	ah	but	bcl b AH tcl t
	yes	ao	bought	bcl b AO tcl t
	yes	oy	boy	bcl b OY
	yes	ow	boat	bcl b OW tcl t
	yes	uh	book	bcl b UH kcl k
	yes	uw	boot	bcl b UW tcl t
	yes	ux	toot	tcl t UX tcl t
	yes	er	bird	bcl b ER dcl d
	yes	ax	about	AX bcl b aw tcl t
	yes	ix	debit	dcl d eh bcl b IX tcl t
	yes	axr	butter	bcl b ah dx AXR
	yes	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

Table 1: TIMIT phonetic labeling continued.

References

[1] S. H. J. van Vuuren, "Pitch detection." Course project report, University of Pretoria, Pretoria, Oct. 1992.

[2] S. H. J. van Vuuren, "Speech coding in the 4-10kb/s domain." Course project report, University of Pretoria, Pretoria, Aug. 1992.

[3] *Speaker Recognition Workshop Notebook*, NIST, 1997. NIST speaker recognition evaluation on the Switchboard Corpus.

[4] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. Mcgonega, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 399–418, Oct. 1976.

[5] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Transactions on Signal Processing*, vol. 39, pp. 319–329, Feb. 1991.

[6] J. D. Markel and A. H. Gray, *Linear prediction of speech*. New York: Springer, 1976.

[7] S. H. J. van Vuuren, "Detection of nasals in speech." Course project report, University of Pretoria, Pretoria, Aug. 1992.

[8] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Am.*, vol. 46, pp. 442–448, Aug. 1969.

[9] H. Hassanein and B. Bryden, "Implementation of the gold-rabiner pitch detector in real time environment using an improved voicing detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 319–320, Feb. 1985.

[10] T. Funada, T. Suzuki, and L. Yu, "A pitch extraction method using a bank of bandpass filter-pairs," *Speech Communication*, vol. 9, pp. 203–216, 1990.

[11] E. Barnard, R. A. Cole, M. P. Vea, and F. Alleva, "Ieee transactions on signal processing," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 298–307, 1991.

[12] E. Barnard. Personal communication.

[13] D. G. Childers, M. Hahn, and J. N. Larar, "Silent and voiced/unvoiced/mixed excitation (fourway) classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1771–1773, 1989.

[14] Y. Lau and C. Chan, "Speech recognition based on zero crossing rate and energy," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 320–323, Feb. 1985.

[15] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, pp. 351–356, 1990.

[16] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice-Hall, 1978.

[17] F. Wang and P. Yip, "Cepstrum analysis using discrete trigonometric transforms," *IEEE Transactions on Signal Processing*, vol. 39, pp. 538–541, Feb. 1991.

[18] H. Indefrey, W. Hess, and G. Seeser, "Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain - preliminary results," in *ICASSP*, (Tampa, FL), pp. 415–418, 1985.

[19] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, pp. 634–648, Feb. 1970.

[20] L. R. Rabiner, "On the use of autocorrelation for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 24–33, Feb. 1977.

[21] J. A. Marks, "Real time classification and pitch detection," in *COMSIG*, pp. 1–6, 1988.