Internal Use Only

TR-I-0138

Experiments in pitch extraction

Alain de Cheveigne

(ATR自動翻訳電話研究所)
ATR Interpreting Telephony Labs.

1990.2

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

Experiments in pitch extraction.

Contents:

Experiments in pitch extraction.

Introduction

1)Aims

The aims of this study are threefold:

• Develop a methodology.
Many pitch extraction methods have been proposed. None is error-free, and it is often difficult to analyze the cause of failure. Error rate, as usually used for evaluation, counts the number of times the algorithm crosses the border between success and failure, and is therefore rather crude. It gives no indication of how *close* the algorithm comes to failure or success. The periodicity measure proposed here is more sensitive and offers better insights.

• Develop a parallel with auditory perception models.
The aim is to apply knowledge from time-domain models of auditory processing to the speech pitch extraction task. This is also an indirect test of the effectiveness of processing of such models.

• Develop a reliable speech pitch extraction method.
This is not the primary aim. Much research effort has already been invested in this problem (Hess 1983), and yet it has not been satisfactorily solved (e.g. Vaissière 1989). It would be unrealistic to expect this particular study to succeed where so many others have failed.

2) Shift-and-compare methods

• Pitch defined as periodicity

The pitch of speech can be defined as the frequency of the *periodic* vibration of the vocal cords. More generally, the pitch of a sound is related to the *periodicity* of the sound waveform. It is therefore natural to characterize pitch using the basic definition of periodicity (invariance by translation by the period and multiples).

Extraction methods that function according to this definition can be termed "shift-and-compare" methods. The comparison step need not involve a numerical waveform: it can be made on transduced patterns (such as nerve firing patterns) or abstracted representations (such as event markers or zero-crossings).

Cassic methods of this type are autocorrelation and AMDF (Average Magnitude Difference Function), and their variants (Ross et al. 1974, Un and Yang 1977, Ney 1982, Hess 1983, Bailly 1986), and Seneff's Generalized Synchrony Detector (Seneff 1985).

In contrast to shift-and-compare methods, spectral methods characterize periodicity by weighting the signal with sine functions and integrating (Fourier transform). At a general level, there is an equivalence between both approaches (Ney 1982). In

practice, it may be easier to understand how the aperiodicities of real speech affect the algorithm if it works in the time (or lag) domain.

• What can go wrong?

The following figure shows a typical AMDF function (defined further on).

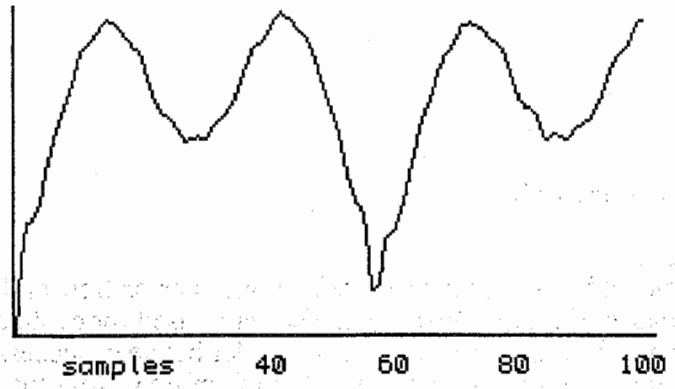


Fig. 1: AMDF function for a vowel.

The sharp dip near 60 samples indicates the period. The AMDF pitch detection method relies on the position of this dip to indicate the period. If the signal were perfectly periodic, the value at the dip would be zero. Real speech is not perfectly periodic, so the value at the dip is often relatively high. Also visible in the figure is a blunt dip at half the period, due to a strong second harmonic component. It can easily happen that such a spurious dip becomes *lower* than the period dip. In this case the algorithm fails.

The improvements discussed hereafter all have the same basic aim: to lower the period dip relative to the spurious dips, to failure of the algorithm.

3) Link with hearing models

Speech pitch extraction methods often rely on speech *production* models. Perception-based methods are less common. Those that have been proposed are mostly inspired by so-called "place" or "pattern matching" theories of pitch perception (Hess 1983).

Pattern matching auditory theories are being questioned recently, because the rate-vs-place representation that they assume fails to show up in recordings of auditory-nerve fibers, and because simpler time-domain processing models are adequate to account for performance (Moore 1982, Møller 1983, Lyon 1984, de Cheveigné 1986). The work reported here is based on this competing notion of time-domain central auditory processing of neural patterns.

Some of the difficulty of pitch extraction is due to the fact that fast transitions of vocal tract shape mask the fundamental periodicity of glottal vibrations. However it often happens also, even in normal speech, that glottal pulses occur at irregular intervals and with irregular amplitude as in the following figure.

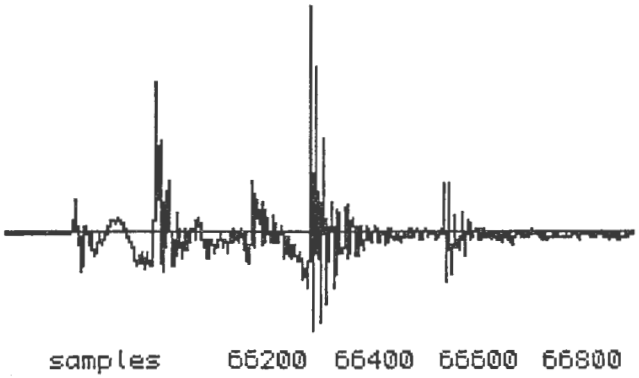


Fig. 2: The "tta" portion of the japanese word "motta" at the end of a sentence (MYI_SD_J02).

The first pulse is the release of the "t", the four following ones are apparently glottal pulses. Even if it were possible to reliably extract glottal pulses from this signal, it is not clear how a pitch could be assigned according to the *production*-based definition of pitch. This does not necessarily imply that there is no pitch to be *perceived*: very short stimuli such as click pairs, and pure tone pulses with as few as 3 periods have been shown to have a pitch that can be discriminated with precision (Moore 1973). A pitch extraction method based on a realistic auditory model might give us an indication of what pitch, if any, is heard in a case such as this.

In applying an auditory perception model to speech processing, it is important to specify precisely what aspects of the model are to be retained.

Here we use three ideas:

• Shift and comparison.

The hypothesis is that pitch perception relies on the comparison of neural patterns elicited by a sound with *delayed* versions of the same patterns, according to a mechanism similar to the cross-coincidence mechanism that has been demonstrated for binaural localization (Yin et al. 1989).

This translates, in terms of extraction, to the use of methods such as AMDF or ACF that perform the same sort of "shift-and-compare" operation on the speech signal.

• Amplitude normalization.

Adaptation mechanisms limit the dynamic range of neural firing patterns (although they also enhance certain transients).

Reduction of the dynamic range translates as amplitude normalization.

• Splitting over a filter bank.

Sound entering the ear is split into different channels by cochlear filtering. According to Møller (1977b), the purpose of this filtering is to prepare the signal for subsequent *time-domain* processing. Several pitch perception models are based on this idea (Licklider 1956, 1959, 1962, van Noorden 1982, Moore 1982, Lyon 1984, de Cheveigné 1986).

There are two possible advantages to be gained. The first is that, if interfering signals (noise or other voices) are present, the signal-to-noise ratio may be higher within certain channels than within others. Restricting attention to such channels might allow easier pitch extraction. The second possible advantage is that filtering may reduce the interaction of different partials. Small phase changes between partials, due for example to the mistuning of a partial, can result in relatively large waveform differences. Such a mistuning is apparently common (McAdams 1989). This idea translates, in terms of speech pitch extraction, to parallel processing of multiple channel outputs of a filter bank.

The parallel drawn here between extraction methods and perception models is quite loose. A detailed similarity between the processing we use and the corresponding auditory processing is not essential, even if we may try to reproduce some details (ex: filter impulse response shapes, see appendix).

Methods

1) Strategy

The basic strategy is to start from a well-known method (AMDF), make modifications, and compare them with the original, and between themselves.

As pointed out above, the AMDF algorithm fails if a "spurious" dip due to harmonics is lower than the "period" dip. Improvements should be judged according to how successfully they *deepen* the period dip relative to the spurious dips. For practical reasons it is easier to use a slightly different criterion that makes use of a "periodicity measure" defined below.

Another possible error is to mistake a dip at a *multiple* of the period for the period dip. Here we make no attempt to avoid such errors because: a) they are basically unavoidable, given the definition of periodicity, and: b) post-processing can take care of them.

AMDF

The AMDF function is defined as:

$$A(n) = \sum_{i \in \text{window}} |s_i - s_{i+n}|$$

where n is the lag in samples (Ross et al. 1974). The AMDF performs a comparison between a fixed reference window and a sliding window, by summing up the sample-to-sample differences. It is also possible to use a sliding reference window so that both windows remain symmetric relative to the analysis point.

2) Mean-normalized AMDF

The AMDF varies with signal amplitude. One way of removing this dependency is to calculate the "Mean-normalized AMDF":

$$B(n) = A(n)\ n \Big/ \sum_{i=1}^{n} A(i) \quad \text{for } n \neq 0$$

$$B(0) = 1$$

The AMDF at a given lag is simply divided by its cumulative mean up to that lag. The following is an example of an AMDF function and the corresponding mean-normalized AMDF:

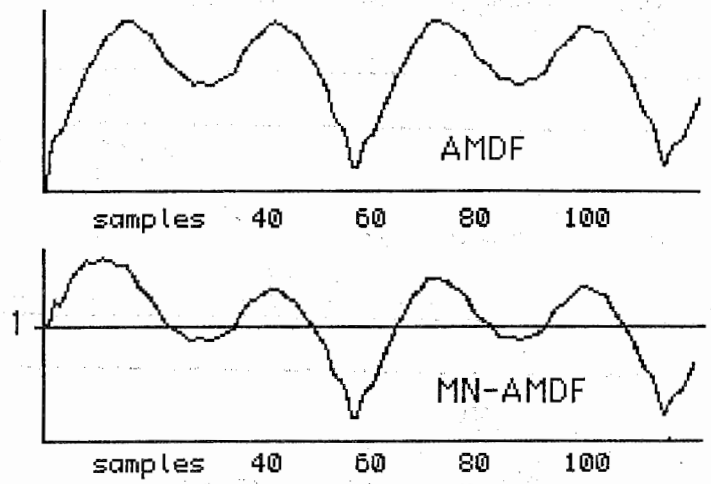


Fig. 3: AMDF (top) and mean-normalized AMDF (bottom).

A dip in the mean-normalized AMDF below 1 indicates that the signal, and its shifted version are *more similar for this shift than for other shifts*, on average.

3) Periodicity measure

The periodicity measure at a dip is defined as:

$$P_n = -\mathrm{Log}_2(B(n))$$

If that particular dip is the period dip, the value of P indicates the *degree of periodicity* of the signal. If there are several competing dips, the periodicity measure for each can serve as a *measure of likelihood* that that dip corresponds to the period. The absolute value of the periodicity measure can serve as a *voiced/unvoiced* criterion. In the following, except when indicated otherwise, the term "periodicity measure" refers to the value at the period. The choice of a base 2 logarithm is arbitrary.

4) Confidence measure

A high value of the periodicity measure does not guarantee that the algorithm won't fail: a "spurious" dip might happen to be even deeper than the "period" dip.

The depth of the period dip can be compared with that of "spurious" dips (defined as dips occurring before 0.8 times the period T) using the "confidence measure" defined as:

$$Q = P(T) - \max(P(n)_{n<0.8T})$$

A negative value indicates that the algorithm would fail at that point.

5) Example

Speech waveform:

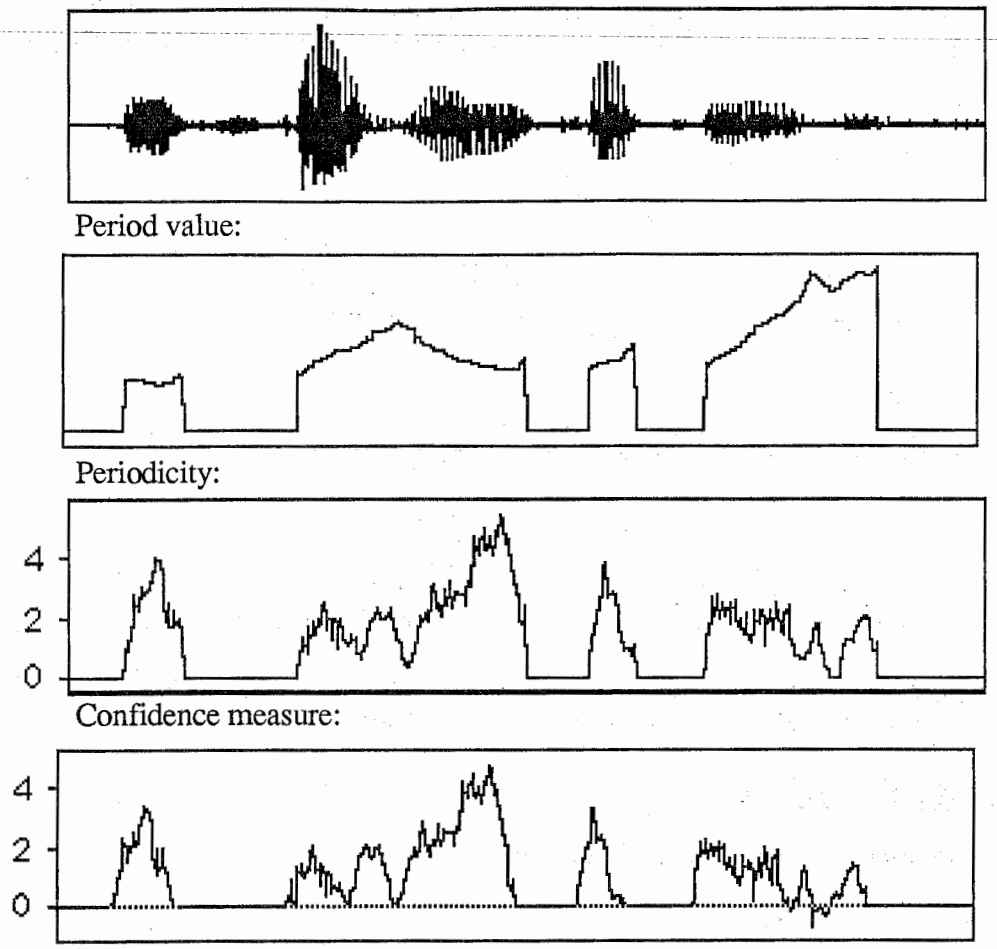


Fig. 4: Waveform, pitch, periodicity and confidence measure.

The confidence measure is very similar in shape to the periodicity measure, apart from a few details (note the negative values). Since the periodicity measure has a "cleaner" definition, and is easier to calculate, we will use it for evaluation purposes.

6) Evaluation methodology

Each "improvement" of AMDF is judged by the effect it has on the periodicity measure. A convenient way to display this is a scatter plot:

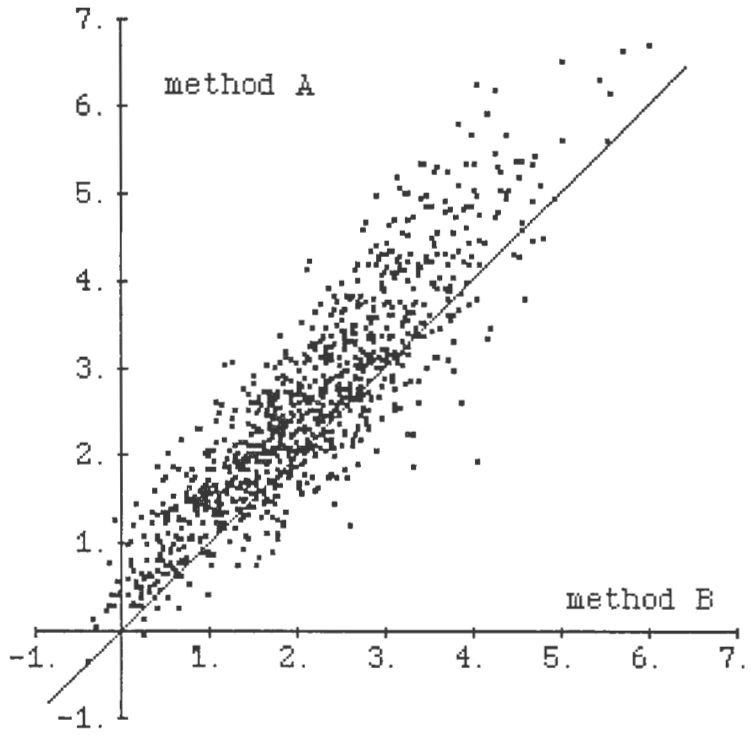


Fig. 5: Scatterplot of periodicity measures between two (hypothetical) AMDF method variants.

Each dot corresponds to a pitch measurement sample. A dot above the diagonal indicates that method A was, for that sample, superior to method B. For legibility, a scatter-plot displays only 1000 samples, chosen randomly.

The improvement can be described more concisely by a "global improvement measure" that calculates the mean improvement, or vertical distance from the diagonal:

$$X^{AB} = \frac{1}{n}\sum_{i=1}^{n}(P_i^A - P_i^B)$$

where n is number of samples and P_i^A represents the periodicity measure at the period for method A.

7) Database

The evaluation method supposes that the "true" value of the pitch period is known.

We chose data from the ATR pitch frequency database (speaker MYI). This labeled database of 503 sentences is described in more detail in Appendix II, and in references listed there. Data consists of the speech data, the manually corrected pitch values, and segmental labels.

For practical purposes, we limited ourselves to a subset consisting of the 20 sentences that caused the highest error rate in the automatic pitch extraction step preceding the manual correction.

The evaluation method requires that the pitch estimate in the database be correct within about 20% of the "true" pitch value. When it became apparent that the methods we

were experimenting were actually *more precise* than the database pitch values, we re-labeled manually the 20 sentences.

This particular database was chosen for convenience, and because it is a standard element of the ATR database. It is certain that a more diversified database would be better for pitch extraction algorithm evaluation.

8) Choice of window size

Integration window shapes are rectangular for simplicity. Window size is chosen according to the following reasoning:

• The aim is to *discriminate* the dip in the AMDF at the period from "spurious" dips at other lags. For this purpose the values of the AMDF must be reliable, in particular, they must not vary erratically with the position of the analysis point.

• The value of the AMDF at a lag τ is actually a sampling of the time function:

$$a_\tau(t) = \int_t^{t+D} |\, s(u) - s(u+\tau) \,|\, du$$

where D is the window size.

• This value can be interpreted as the absolute difference signal

$$d(t) = |\, s(t) - s(t+\tau) \,|$$

filtered by a low-pass filter with a square impulse response of length D. This signal contains a non-null zero-frequency component (the AMDF estimate we seek) on which are superimposed components at the fundamental and harmonics.

• The window length D must chosen large enough to adequately attenuate the fundamental and harmonics in the worst case (the lowest fundamental of the expected range):

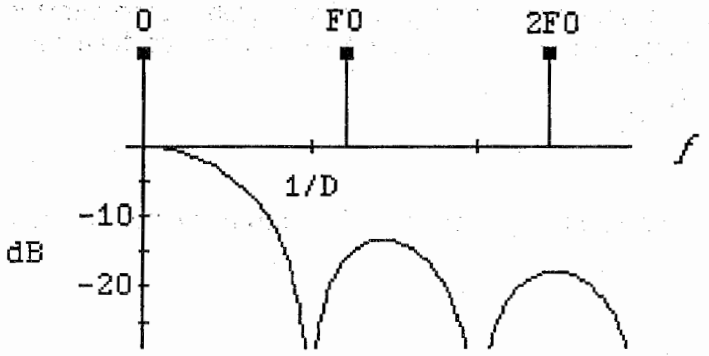


Fig. 6: Line spectrum of absolute difference function d(t) (top), and transfer function of integration window. Non-zero frequency components of d(t) must be attenuated by integration.

It should not be made larger than necessary, to allow tracking of fast transitions.

This reasoning argues for a uniform window size for *all values* of the lag. This runs counter to schemes that adapt the window length according to the lag, on the grounds that shorter lags correspond to higher fundamentals (Fujisaki et al. 1989). The window size can however be reduced *after* the pitch range has been ascertained, in order to improve tracking precision.

Experiments

The following experiments test a number of ideas for "improving" AMDF. The combination of one method with others can either enhance or diminish its effectiveness. Therefore each method is usually tested alone (by comparison to ordinary AMDF), and in combination with another method (by comparison to that other method alone).

1) BP filtering (revcor filter)

• Rationale

Filtering can improve periodicity in two imaginable ways:
- by attenuating spectral regions where periodicity is poor,
- by separating components whose *interaction* degrades periodicity.

A classical technique is to low-pass-filter the signal in order to isolate or enhance the fundamental. There are several difficulties with this approach:
- The cutoff frequency that will accommodate a full range of fundamentals can be difficult to find, even for a given speaker.
- The technique fails if the speech *lacks* a fundamental component (as in telephone speech), In this case it is the interaction of higher partials that creates the periodicity, and one must not search to eliminate them.

An alternative approach is to use multiple channel bandpass filtering. Possible benefits are:
a) Some channels may show enhanced fundamental periodicity.
b) Several channels may show periodicity at *harmonics*, but the total pattern would allow to recognize the fundamental by cross-channel sub-harmonic matching. For example if channel A isolates the second harmonic and channel B the third harmonic, dips in the AMDF of both channels will coincide at the fundamental period:

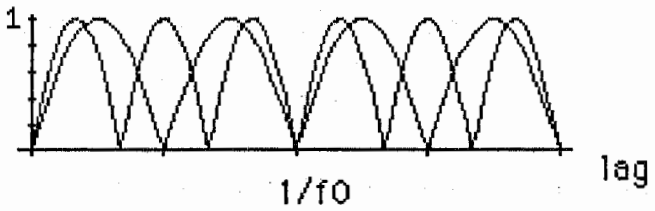


Fig. 7: AMDF for two channels, one isolating the 2nd harmonic, the other the 3rd harmonic of the signal fundamental. Zeros in both channels coincide for a lag equal to the fundamental period.

c) The concept of periodicity restricted to a frequency channel can be of possible use for voice separation, and also for speech synthesis (Fujimura 1968, Rodet et. al. 1988).

The following experiments use 7 outputs of the revcor bandpass filter bank described in Appendix I. The revcor filter provides an approximation of the filtering characteristics of the basilar membrane.

• Channel 1 (111 Hz):

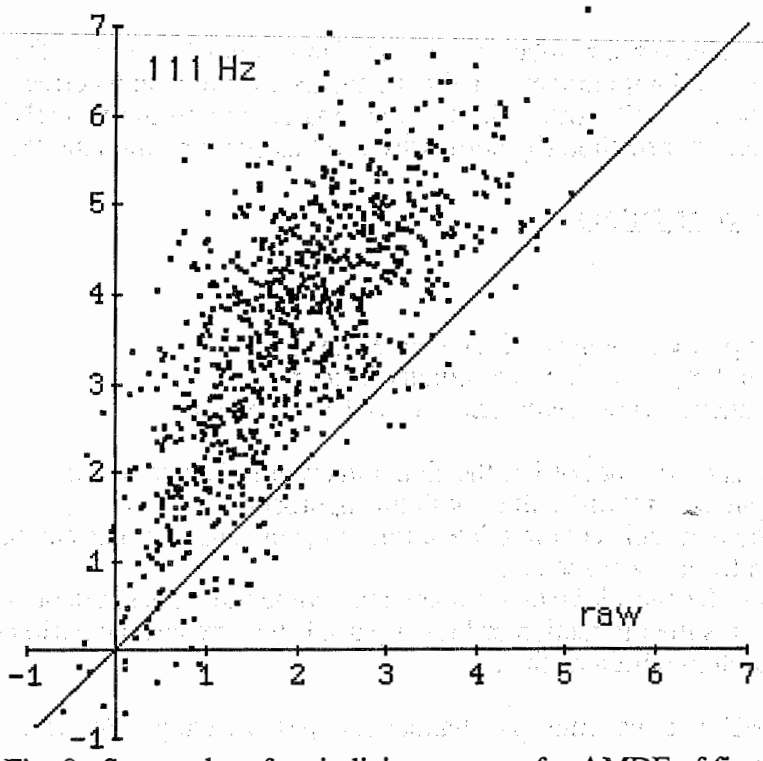


Fig. 8: Scatterplot of periodicity measure for AMDF of first channel output versus AMDF of raw signal.

Most points are well over the diagonal, indicating a clear improvement. Some points fall on the contrary *below*
the diagonal, indicating that this form of filtering can sometimes have negative effects. The global improvement measure (defined above) is 1.6, meaning that AMDF period dips are on the average $2^{1.6} \approx 3$ times deeper for the filtered signal than for the raw signal.

• All channels:

Higher frequency channels show progressively less good performance:

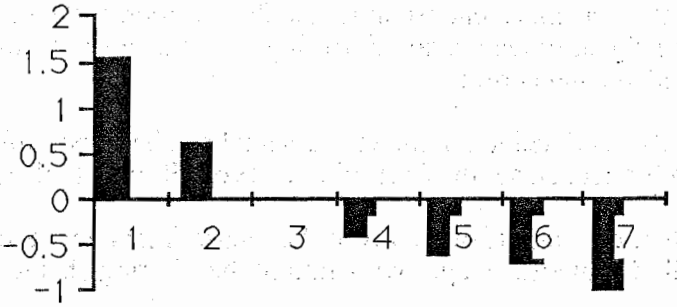


Fig. 9: Global improvement for each revcor channel output.

The filtering of the high channels is detrimental to AMDF, as is visible in this scatter plot for the highest channel:

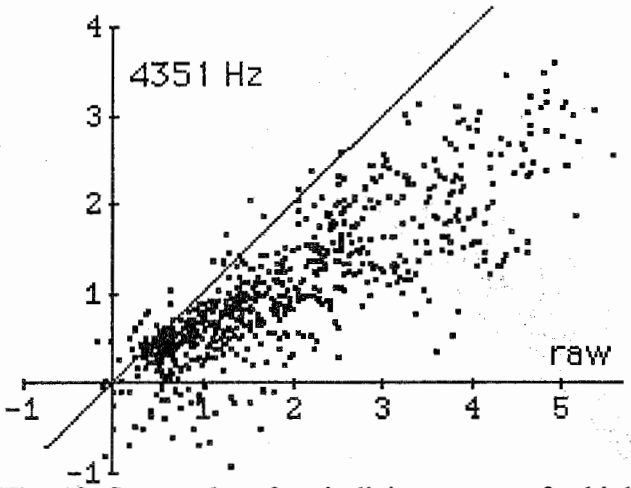


Fig. 10: Scatterplot of periodicity measure for highest revcor channel output versus raw signal.

Revcor channel outputs are used hereafter to test other improvement ideas.

2) Compensation of sampling error

Limited sampling resolution can reduce the depth of AMDF period dips when the fundamental period is not a multiple of the sampling period. To relieve this problem, the AMDF algorithm is modified to calculate:

$$A(n) = \sum_{i \in \text{window}} |d_i(n)|$$

with:

$d_i(n) = s_i - s_{i+n}$ if same sign as $d_i((n-1))$,
$d_i(n) = 0$ if sign is different.

Scatter plot of periodicity values for sampling error compensated AMDF versus ordinary AMDF for the raw signal:

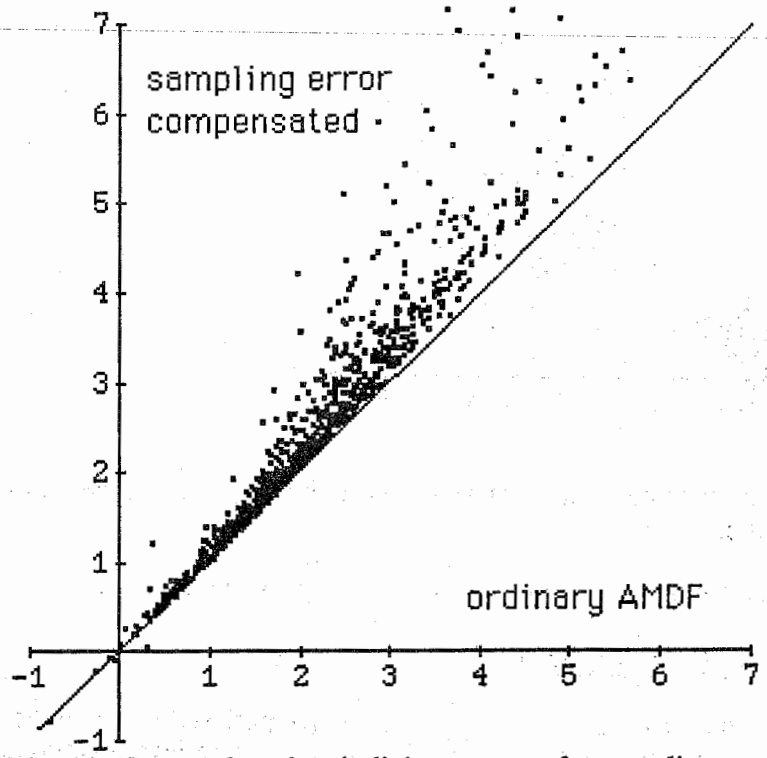


Fig. 11: Scatterplot of periodicity measure for sampling compensated AMDF versus ordinary AMDF, on raw signal.

There is a small but consistent improvement, and in no case a degradation. Unfortunately the improvement is small for small periodicity values, where it would be most needed. This is even more evident in a scatter plot for the first revcor channel:

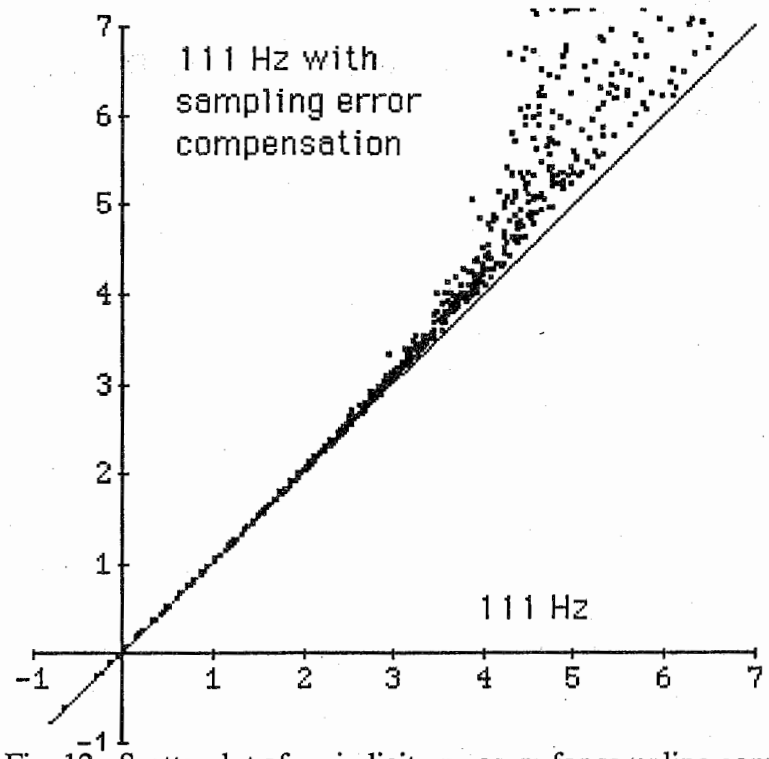


Fig. 12: Scatterplot of periodicity measure for sampling compensated AMDF versus ordinary AMDF, on lowest revcor output channel.

This is easy to understand: AMDF dips for low-frequencies are blunt and therefore relatively insensitive to sampling error.

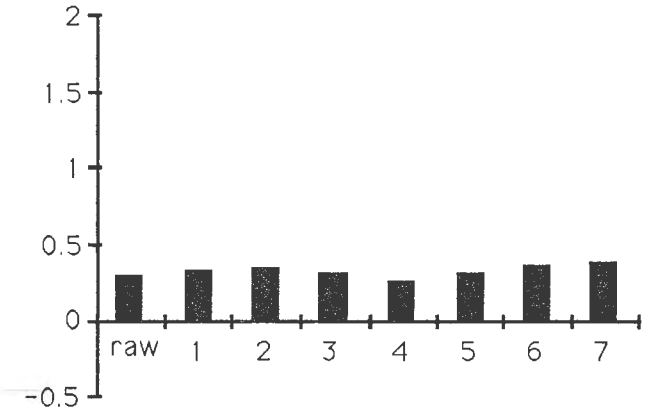


Fig 13: Global improvement measure for raw signal and revcor filtered channels:

3) amplitude compensation

When the speech signal amplitude changes, successive periods will tend to compare badly even if their shapes are similar. The methods in this section aim at attenuating the effects of amplitude change.

a) normalization

The simplest way to compensate amplitude effects is to normalize the signal amplitude. This can be done by dividing each sample of the signal by its amplitude (sum of absolute values) over a window centered on this sample. The window must be chosen long enough to avoid fluctuations of the amplitude estimate following the reasoning outlined previously.

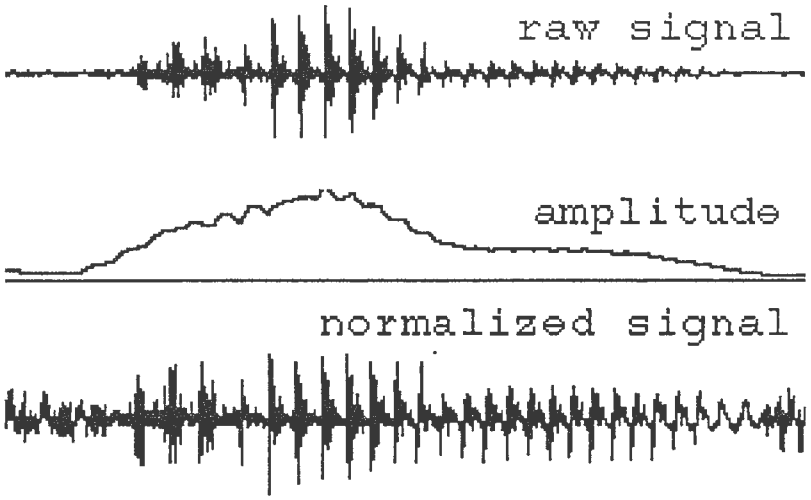


Fig. 14: Raw signal (top), amplitude measure (middle), and amplitude normalized signal (bottom), obtained by dividing (top) by (middle).

Scatterplot of periodicity measure for amplitude normalized and raw signal:

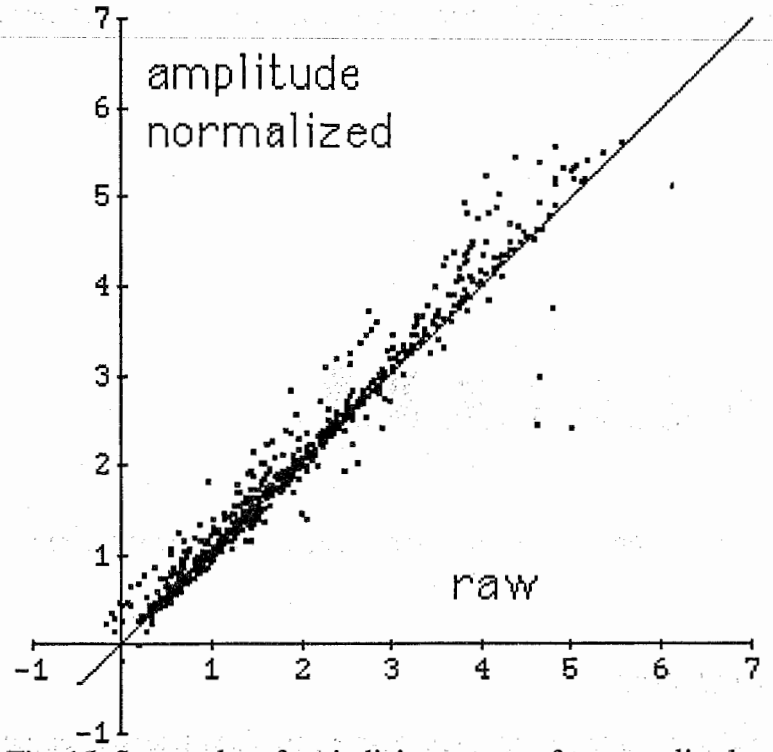


Fig. 15: Scatterplot of periodicity measure for normalized versus raw signal.

The improvement is small, a few points are considerably degraded.
Global improvement measure for raw and revcor filtered signals:

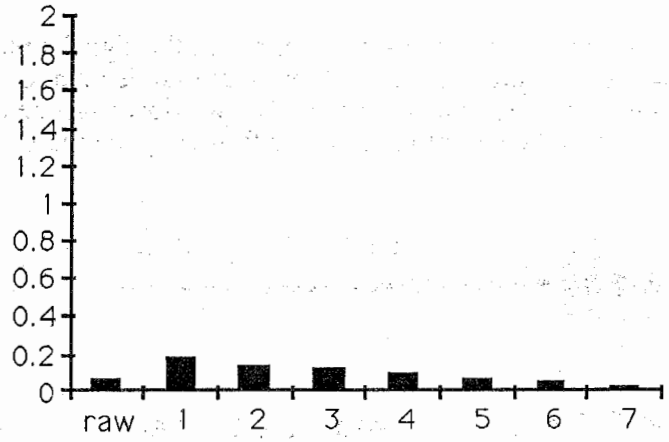


Fig. 16: Improvement measure for raw signal and revcor filtered channels.

b) amplitude compensated AMDF

Normalization distorts the waveform slightly. An alternative is to incorporate the amplitude compensation into the AMDF calculation:

$$A(n) = \sum_{i \in \text{window}} | a_{i+n} s_i - a_i s_{i+n} |$$

where:

$$a_i = \sum_{j=-n/2}^{n/2} | s_i |$$

The results are almost indistinguishable from those obtained with ordinary AMDF on a normalized signal.

Global improvement measure for raw and revcor filtered signals:

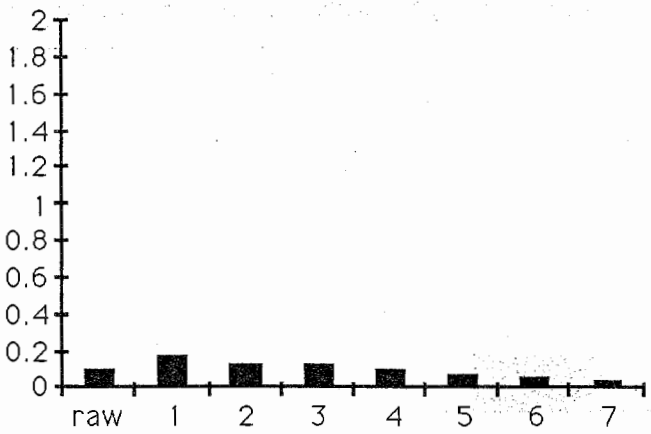


Fig. 17: Global improvement for raw signal and revcor filtered channels.

c) search

The surprising lack of effectiveness of amplitude compensation suggests that somehow the method, or parameters (i.e. window size) might be wrong. To determine the ultimate possible improvement obtainable by amplitude adjustment, we implemented a search algorithm: for each analysis point and lag, the amplitude ratio that give the best correspondence (i.e. minimizes the AMDF at that lag) is searched for using a simple search algorithm. Search is initiated at a ratio equal to the ratio of amplitudes between the two windows.

Scatter plot of periodicity values for search-compensated AMDF and ordinary AMDF, on raw signal:

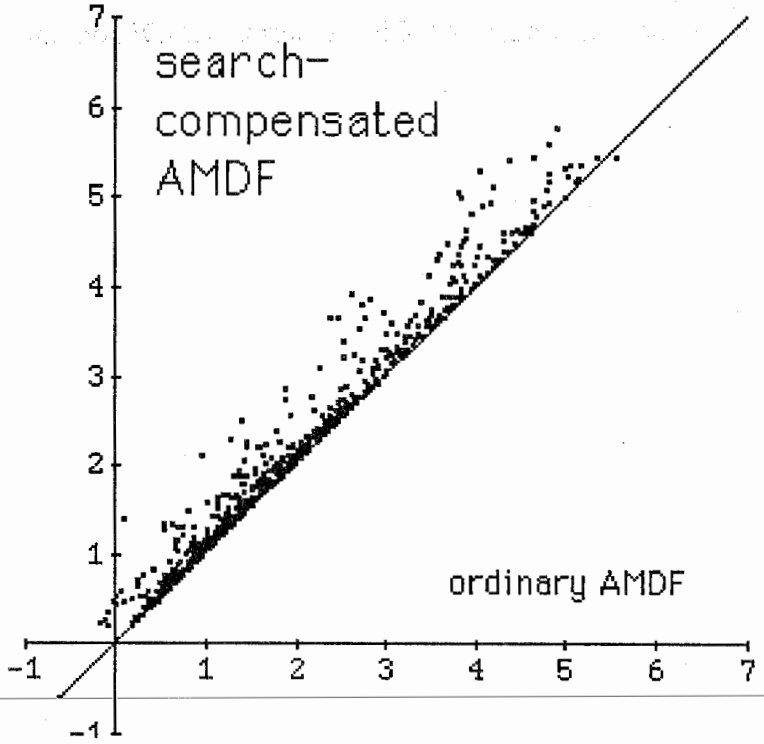


Fig. 18: Scatterplot of periodicity measure for search-compensated AMDF versus ordinary AMDF, on raw signal.

There are practically no values beneath the diagonal, which is normal since the search can only improve the match. The improvement remains small. Revcor-filtered channels show similar results (results for higher order channels are not available):

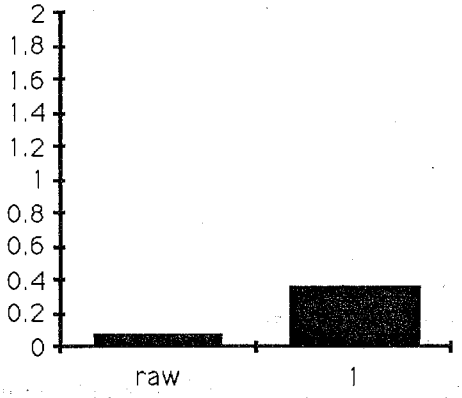


Fig. 19: Global improvement for raw speech and first revcor channel.

d) split-window AMDF:

This idea was suggested by Barry Vercoe of MIT. For each lag, the reference window is compared with the mean of two windows, one advanced and the other delayed with respect to the reference window, by an amount equal to the lag:

$$A(n) = \sum_{i \in \text{window}} | s_i - (s_{i+n} + s_{i-n})/2 |$$

Supposing the amplitude variation is locally linear, the average amplitude of the two sliding windows should equal that of the reference.

Scatter-plot of periodicity for split window AMDF and normal AMDF, on raw signal:

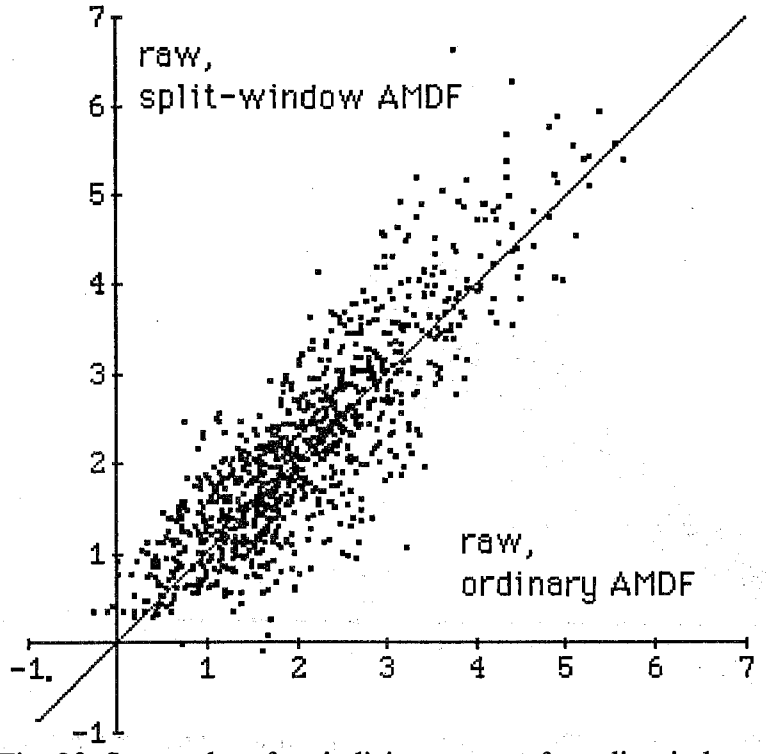


Fig. 20: Scatterplot of periodicity measure for split-window AMDF versus ordinary AMDF, on raw signal.

Performance is overall slightly improved, but many points are degraded, particularly in the critical low periodicity range.

Global improvement measure for raw and revcor filtered signals:

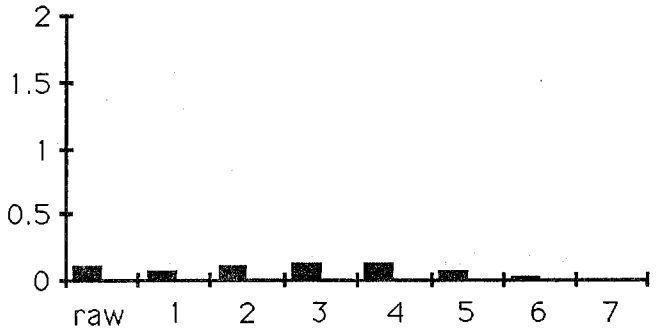


Fig. 21: Global improvemenent for raw signal and revcor filtered channels.

The results of this elegant method are disappointing. This may be due to a degradation of performance at onsets and offsets, where one half of the split window is actually outside the speech data.

e) spectral flattening

This experiment tests an idea similar to that proposed by Stefanie Seneff (1985) to generate a "pitch waveform" for GSD pitch extraction. Seneff's "pitch waveform" is a weighted sum of compressed filter channel outputs that results in a "spectrally flattened" signal.

Here, we add all seven filter channels after amplitude compression, to obtain a crude "spectrally flattened" signal:

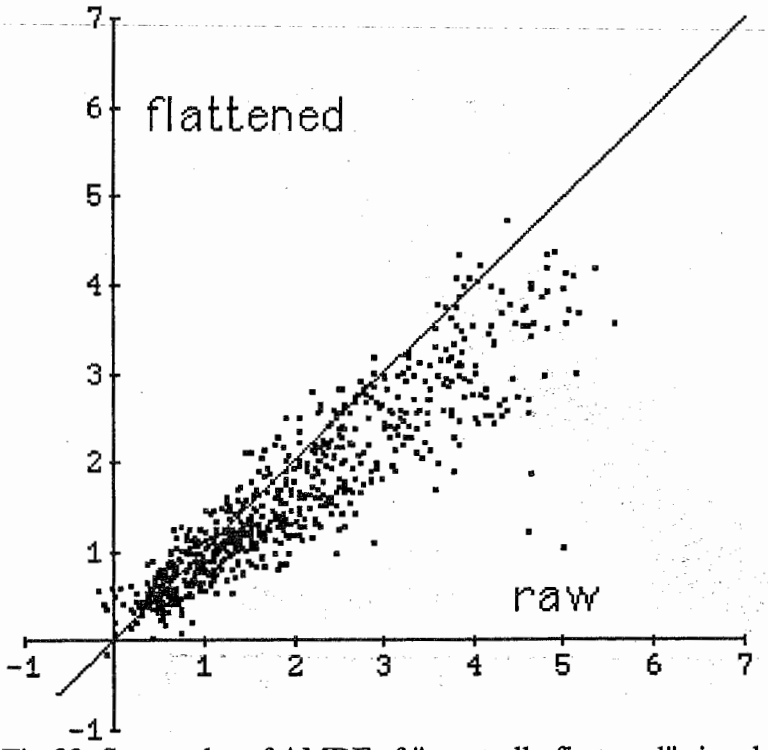


Fig 22: Scatterplot of AMDF of "spectrally flattened" signal versus raw signal.

The result of this experiment is disappointing, but not altogether unexpected: the "flattening" emphasizes the weight of high frequency channels for which periodicity is poor. Global improvement measure:

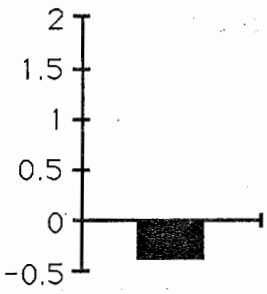


Fig. 23: Global improvement for "spectrally flattened" speech.

4) half wave rectification

Auditory nerve discharge probability functions closely resemble the *half wave rectified* basilar membrane motion at the innervation point.

In addition, nerve fiber discharge synchrony breaks down at high frequencies. The loss of synchrony can be modeled as a gaussian jitter of about 55µs standard deviation, that "blurs" the details of the signal, and acts on PST histogram shapes somewhat like a low-pass filter with a cutoff (-6dB) at about 3 kHz.

This experiment investigates the effect of half-wave rectifying and low-pass filtering the raw signal and revcor filter outputs. Low-pass filtering is done by calculating a moving average over a square window.

• half-wave rectification only:

Global improvement measure for raw signal and revcor channels:

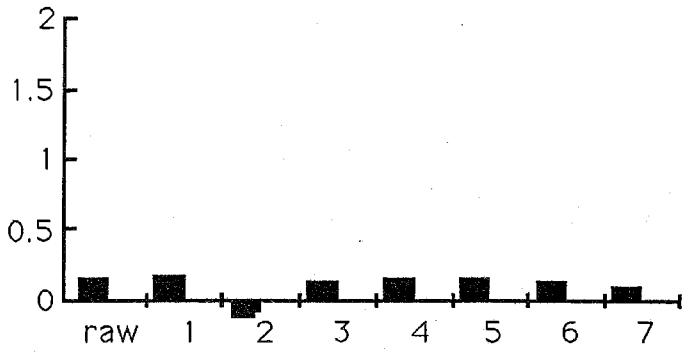


Fig. 24: Global improvement for raw speech and revcor filtered channels.

As might be expected, there is no particular improvement (results for the opposite alternance mirror these).

• 8 point window lpf:

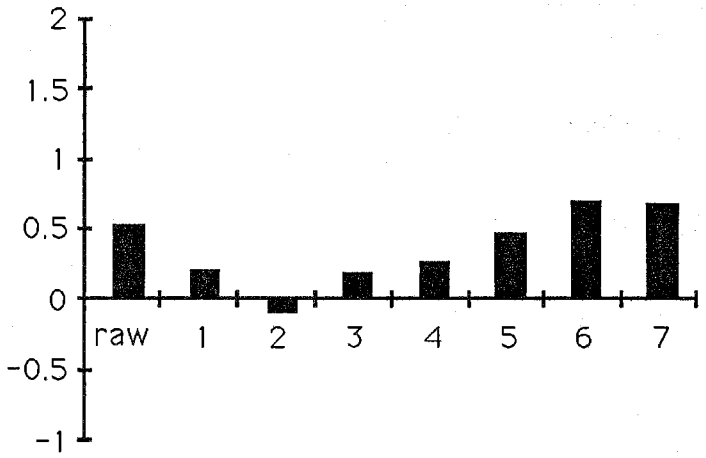


Fig. 25: Global improvement for raw speech and revcor filtered channels.

• 16 point window lpf:

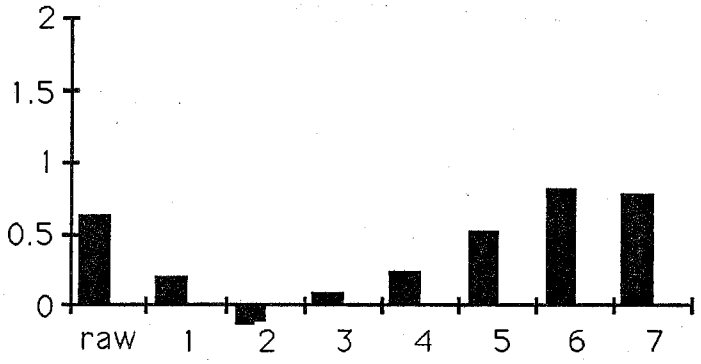


Fig. 26: Global improvement for raw speech and revcor filtered channels.

• 32 point window lpf:

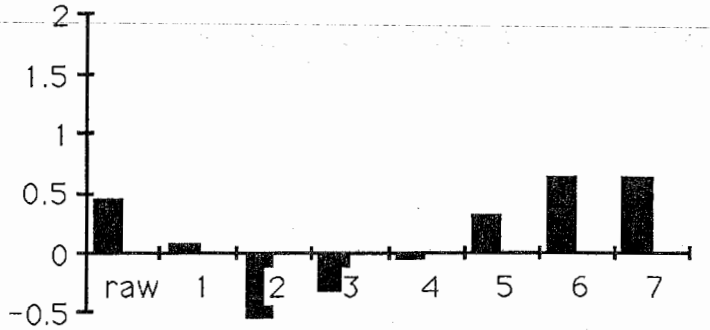


Fig. 27: Global improvement for raw speech and revcor filtered channels.

There is a consistent improvement of the highest channels (which were the most degraded by filtering) as visible in this scatterplot for the highest channel, 16 point lpf:

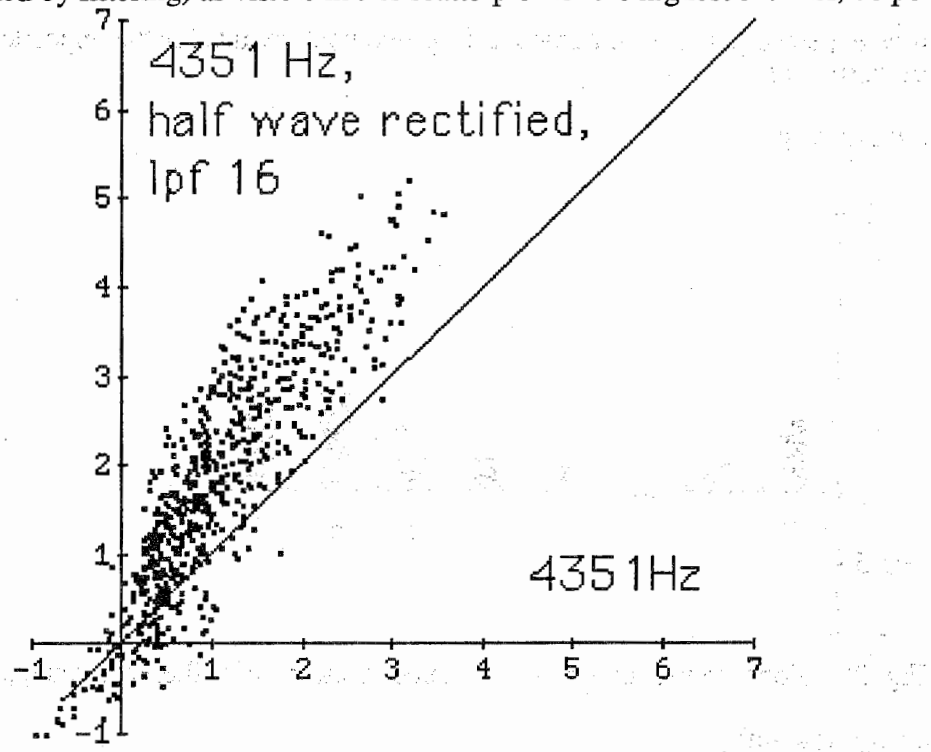


Fig. 28: Scatterplot of periodicity measure for AMDF of low-pass filtered, half-wave rectified 4351 Hz revcor filter channel output, versus AMDF of non-rectified filter output.

Improvement is unfortunately less consistent for the low-periodicity points that have most need for it.

5) Combining information from several channels.

It was suggested in the i ntroduction that periodicity information derived from several filter output channels could be combined to provide a reliable estimate.

There is a wide range of strategies to choose from for combining this information. Here are a few:
- sum the filter outputs after amplitude normalization (as used in Seneff's [1985] pitch extraction method, and described under "spectral flattening" above),
- summing the AMDF patterns over channels ("OR"),

- multiplying the AMDF patterns over channels ("AND"),
- combine channels with weights proportional to their periodicity measures.
- match minima across channels, allowing for a degree of aperiodicity (the ear integrates components that are mistuned by less than 3-8%)

These ideas were not tested, for lack of time. However we did test one basic assumption that they rely on: that different channels do carry different information, or in other words, that all channels do not fail in the same way at the same place. If all channels carry equivalent information, there is no point in combining them.

The following graphs are histograms based on the analysis points where the mean-normalized-AMDF had a global minimum at some lag shorter than the period. A pitch algorithm would fail at such points.

The histograms display the number of such points as a function of the lag at which they occur (expressed as a percentage of the "correct" period). Bin width is 1 %.

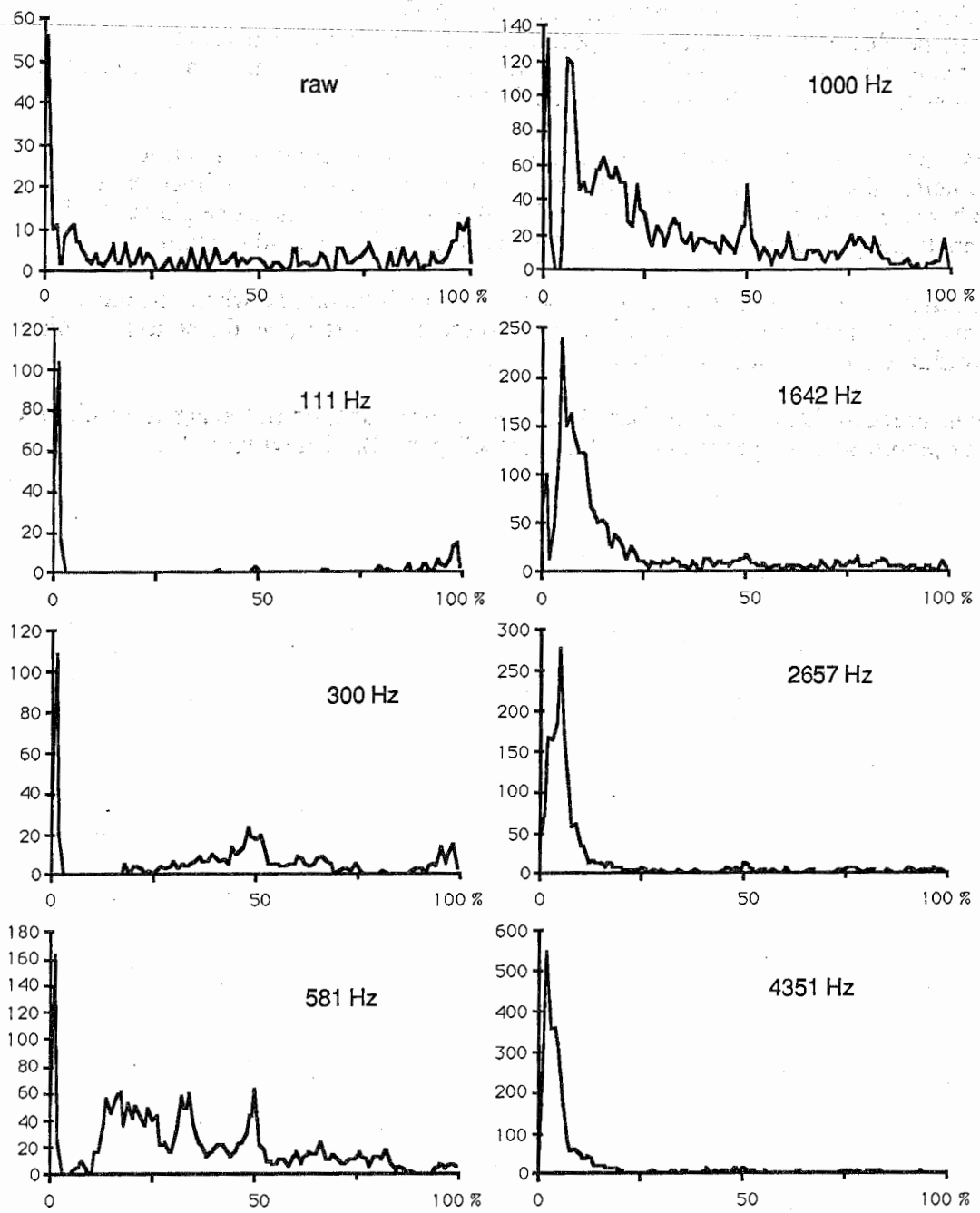


Fig. 29: Histograms of "too-low" pitch errors for the raw signal, and each revcor filter channel. Abscissa is percentage of correct pitch period.

The histogram for the raw signal shows two humps, one near 0%, the other near 100%, as well as values distributed evenly between the two.

The hump near 100% indicates that the mean-normalized-AMDF minimum occurred just below the "correct" period (defined as the lag of the minimum that occurs within 20% of the labeled value in the database). The presence of the hump suggests that the 20% criterion was too severe, or that the database was mislabeled by more than 20%.

The sharp peak near 0% indicates that *no* value smaller than 1 was found. This suggests either that the signal is severely a-periodic at that point, or, more likely, that the minimum occurred *beyond* the allowable range. This also suggests a mislabeling by more than 20%.

The evenly distributed values between the humps correspond probably to a locking to a harmonic

The 111 Hz channel shows the same two humps, but the evenly distributed values are much less common.

The 300 Hz channel shows a broad hump near 50%, indicating a locking to the second (or third) harmonic.

Higher channels show progressively higher counts at shorter lags, indicating a locking to progressively higher harmonics (or to the period of the impulse response of the filter).

It appears that spurious minima occur at different lags in different channels, and therefore that different channels contain complementary information that can be usefully combined. This tentative conclusion needs experimental verification.

Discussion

1) Improvements to AMDF

Filtering through the lowest channel (111 Hz) of the revcor filter bank provided the greatest improvement. Low-pass filtering (not reported here) provided similar improvement. It is likely that such improvement is due to the enhancement of the fundamental component of the speech.

It is tempting to base AMDF pre-processing entirely on such low-pass filtering. Such a move would be unwise for the following reasons:
1) The effectiveness of filtering depends critically on the choice of parameters (cutoff frequency, slope). It is difficult, perhaps impossible, to find a set of parameters that will insure a good performance for the full range of fundamental frequencies.
2) A fundamental frequency component is not always present, nor is it necessary for perception of the fundamental pitch.

In the absence of a fundamental component, periodicity arises from the interaction of higher order harmonics. Useful fundamental periodicity information can in principle be derived from higher channels, provided they are wide enough to allow interaction. Even in the absence of interaction, fundamental periodicity information can also be obtained by pooling AMDF patterns for different channels.

The experiments show that the periodicity of higher frequency channels is severely degraded. This degradation can be partially compensated by half-wave rectification and low-pass filtering, but performance remains low. Combining AMDF patterns across channels (by summation or another scheme) might improve performance, as suggested by the difference in shape of the error histograms of different channels. Unfortunately this has not yet been verified.

Sampling error compensation offers some improvement, as evident in the global improvement measure, but examination of the periodicity measure scatterplots shows that where improvement is most needed (low periodicity), it is small.

Amplitude variation compensation gave poor results. This comes as a surprise: variations in signal amplitude seem an evident cause of aperiodicity. Adjustment of the amplitude ratio between reference window and sliding window by a search technique gave slightly better results, but emphasized the limits of such processing.

The split-window technique was disappointing, particularly as it might have been expected to compensate "linear" timbre changes in addition to just amplitude changes.

Half-wave rectification and low-pass filtering markedly improves the periodicity of higher channels. It would be interesting to compare the effects of positive alternance versus negative alternance rectification. Unfortunately this was not tried.

2) Possible applications of the periodicity measure.

• Pitch estimate "weight".

In speech, the limit between "voiced" and "unvoiced" is not clearly defined. Periodicity of speech that is nominally voiced can be degraded by noise, transitions in vocal-tract shape or irregularity of vocal tract vibration. Most applications require a pitch value to be assigned nevertheless, and will fail if this value behaves erratically.

The usual approach in such a case is to apply post-processing to fill in the gap by continuity. This approach is liable to fail catastrophically if post-processing "locks" on the wrong value.

A possible alternative would be to use the periodicity measure in subsequent processing. For example in error-correction, the periodicity measure allows the algorithm to start continuity tracking from values that are "sure", and to choose against pitch tracks that accumulate a large error. The dynamic programming AMDF method of Bailly (1986) works in a similar fashion.

Another use of the periodicity measure is as a weight, for example in pattern matching of the pitch curve, to de-emphasize portions for which the pitch value is not sure.

Finally, the measure may be of use in itself, as it allows a smooth transition between "voiced" and "unvoiced".

• Periodicity local to a time-frequency zone.

Fujimura (1968) noted that, at a given instant, the periodicity of speech is sometimes restricted to certain frequency zones. Rodet et al. (1988) proposed a similar idea to improve the quality of synthetic speech, by using an excitation waveform that is periodic in some frequency bands and random in others. In both cases, a yes-no periodicity decision is made within each band The periodicity measure applied to the outputs of a filter bank might allow a softer decision.

• Ambiguous pitch.

The vocal cords sometimes vibrate in a truly ambiguous fashion. Similarly, psycho-acoustic experiments show that some sounds have an ambiguous pitch. A possible way to handle such situations is to allow multiple values for the pitch. Adding a periodicity measure to each pitch value allows quantification of its relative salience.

3) Outline of a pitch estimation algorithm

On the basis of these experiments, a pitch extraction method can be outlined as follows:

• Basic algorithm

1) Speech is filtered by a multi-channel bandpass filter bank (revcor or other).

2) Each channel is halfwave rectified for both alternances, low-pass filtered, and amplitude normalized.
3) AMDF is calculated for all channels. The results are combined (in a way yet to be specified: sum, periodicity-weighted sum, periodicity-weighted vote, or other).
4) The output comprises four values:
- first candidate period,
- first candidate periodicity,
- second candidate period (defined as best candidate shorter than first),
- second candidate periodicity.

5) The four values are handed over to an error-correction algorithm.

• Error-correction

The error-correction algorithm uses the second candidate estimates to eliminate "too-long" errors (subharmonics). The first candidate estimates (eventually corrected) serve for pitch tracking. The final output is a pitch/periodicity pair.

• Computational cost

The AMDF calculation is more costly than preprocessing (filtering, etc.), which is relatively inexpensive. The cost is multiplied by the number of channels on which AMDF is performed. However, computation time can be cut if initial estimates are obtained for low channels (on the down-sampled filtered signal), and then calculated only where necessary in higher channels.

• A quick-and-dirty algorithm

A quick-and-dirty algorithm for database marking would be:
1) Filter with a gentle low-pass filter, or a low-frequency wide-band band-pass filter,
2) Amplitude-normalize by dividing by mean of signal over a square window,
3) Calculate AMDF over allowable pitch range,
4) Multiply by an light emphasis function to favor short periods over long ones, to eliminate sub-octave jumps, and find minimum.
5) Correct manually.

Acknowledgements

Roy Patterson kindly provided the GammaTone filter bank software developed by John Holdsworth at the MRC-APU in Cambridge. Part of this work was carried out while the author was supported by an STP fellowship awarded by the Commission of European Communities. The author wishes to thank ATR for its kind hospitality, and the Centre National de la Recherche Scientifique (CNRS, France) and Professor Culioli, Linguistics Department, Université Paris 7, for leave of absence.

References

Bailly, G. (1986) "Detection du fondamental par pretraitement AMDF et programmation dynamique", Proceedings of the JEP (GALF), 285-288. (in French).

Carney, L.H., and Yin, T.C.T. (1988). "Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model", J. Neurophysiol. 60, 1653-1677.

de Cheveigné, A. (1986) "A pitch perception model", Proc. IEEE ICASSP-86, 897-900.

Evans, E.F. (1977), "Frequency selectivity at high signal levels of single units in cochlear nerve and cochlear nucleus", in "Psychophysics and physiology of hearing", edited by E.F. Evans and J.P. Wilson (Academic Press, London), 185-196.

Fujisaki, H., Hirose, K., Seto, S. (1989) "A method for pitch extraction of speech with reduced errors due to analysis frame positioning", Trans. Comm. Speech Res., Acoust. Soc. Japan., SP 89-69, 1-8 (in Japanese).
Fujimura, O. (1968) "An approximation to voice aperiodicity", IEEE Trans. AU-16, 68-72.
Hess, W. (1983) "Pitch determination of speech signals", (Springer Verlag, Berlin), pp698.
Holdsworth, J., Nimmo-Smith, I., Patterson, R., and Rice, P. (1988) "Implementing a GammaTone filter bank", Annex C of the SVOS final report, part A, 1-5.
Langner, G. (1981). "Neuronal mechanisms for pitch analysis in the time domain", Exp. Brain Res. 44, 450-454.
Langner, G., and Schreiner, C.E. (1988). "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms", J. Neurophysiol. 60, 1799-1822.
Licklider, J.C.R. (1956). "Auditory frequency analysis" in *Information theory*, edited by C. Cherry (Butterworth, London), 253-268.
Licklider, J.C.R. (1959). "Three auditory theories" in *Psychology, a study of a science*, edited by S. Koch (McGraw-Hill), vol. I, 41-144.
Licklider, J.C.R. (1962). "Periodicity pitch and related auditory process models", International Audiology 1, 11-36.
Liberman, M.C. (1982). "The cochlear frequency map for the cat: labelling auditory-nerve fibers of known characteristic frequency", JASA 72, 1441-1449.
Loeb, G.E., White, M.W., and Merzenich, M.M. (1983). "Spatial cross-correlation—A proposed mechanism for acoustic pitch perception", Biological cybernetics, 149-163.
Lyon, R.F. (1983). "A computational model of binaural localization and separation", Proc. IEEE ICASSP-83, 1148-1151, reprinted in "*Natural computation*", edited by W. Richards (MIT Press, Cambridge Massachusettts), 319-327.
Lyon, R.F. (1984). "Computational models of neural auditory processing", IEEE ICASSP, 36.1.(1-4).
Møller, A.R. (1977a). "Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli", J. Acoust. Soc. Am. 62, 135-142.
Møller, A.R. (1977b). "Frequency selectivity of the basilar membrane revealed from discharges in auditory nerve fibers", in "Psychophysics and physiology of hearing", edited by E.F. Evans and J.P. Wilson (Academic Press, London), 197-207.
Møller, A.R. (1983). "Auditory physiology", Academic Press, New York, 305p.
Moore, B.C.J. (1973). "Frequency difference limens for short-duration tones", J. Acoust. Soc. Am. 54, 610-619.
Moore, B.C.J. (1982). *An introduction to the psychology of hearing* (Academic Press, London).
Moore, B.C.J., and Glasberg, B.R. (1983). "Suggested formulae for calculating auditory filter bandwidths and excitation patterns", JASA 74, 750-753.
Ney, H. (1982) "A time warping approach to fundamental period estimation", IEEE Trans. SMC 12, 383-388.
van Noorden, L. (1982). "Two channel pitch perception", in *Music, mind, and brain*, edited by M. Clynes (Plenum Press, New York), 251-269.
Patterson, R.D., and Nimmo-Smith, I. (1986). "Thinning periodicity detectors for modulated pulse streams", in *Auditory frequency selectivity*, edited by B.C.J. Moore, R.D. Patterson (Plenum Press, New York), 299-307.
Rodet, X., Depalle, P., Poirot, G. (1988) "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions", ICMC-88 (preliminary draft).
Ross, M.J., Schaffer, H.L., Cohen, A., Freudberg, R., Manley, H.J. (1974) "Average Magnitude Difference Function pitch extractor", IEEE Trans. ASSP-22, 353-362.
Schouten, J.F. (1970). "The residue revisited" in *Frequency analysis and periodicity detection in hearing*, edited by R. Plomp, G.F. Smoorenburg (Sijthoff, Leiden), 41-58.
Seneff, S. (1985). *Pitch and spectral analysis of speech based on an auditory synchrony model*, Thesis, MIT tech. rep. 504.
Un, C.K., and Yang, S.C. (1977) "A pitch extraction algorithm based on LPC inverse filtering and AMDF", IEEE Trans. ASSP 25, 565-572.
Vaissière, J. (1989) "On automatic extraction of prosodic information for automatic speech recognition system", Proceedings Eurospeech 89, vol1, 202-205.
Wever, E.G. (1949). *The theory of hearing*, Dover.
Yin, T.C.T., Chan, J.C.K., and Carney, L.H. (1987) "Effects of interaural time delays of noise stimuli on low-frequency cells in the cat's inferior colliculus. III. Evidence for cross-correlation." J. Neurophysiol. 58, 562-583.

Appendix I.
Choice of filters.

1) Criteria

In the absence of better knowledge, the filters were chosen on the basis of their impulse response shape, according to two criteria:

• short impulse response

This follows the belief that a long impulse response (relatively to the period) might "smear" transitions and mask the signal's periodicity by that of the impulse response itself.

• similarity with physiological data.

Detailed similarity is not essential, since the method is only loosely inspired from an auditory processing model. A relatively detailed discussion is given anyway. Quite precise data is available from reverse correlation measurements (Carney and Yin 1988, Møller 1977a) and modeling (de Boer 1975).

2) Filter shape

The shape of the impulse response of the basilar membrane can be modeled to a high degree of precision as a "revcor" function (Carney and Yin 1988):

$$h(t) = A(t - T_l)^{\upsilon} \exp(-(t - T_l) / T_f) \sin(2\pi F(t - T_l))$$

F is characteristic frequency, T_l is a latency, T_f is the time constant of decay, and υ is a factor that governs the "symmetry" of the response.
A typical response for a fiber at1kHz is:

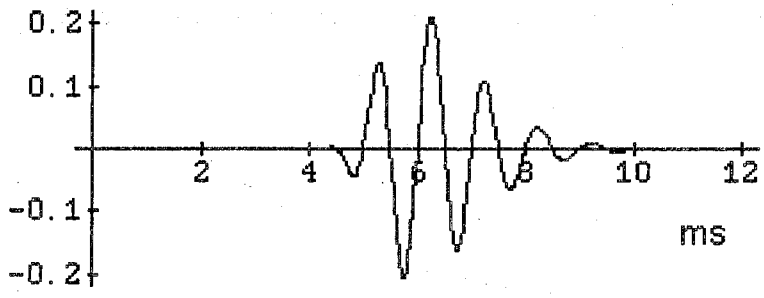


Fig. 1: typical impulse response of a revcor filter similar to the basilar membrane (1000 Hz CF).

Carney and Yin matched the responses of a population of fibers and derived empirical expressions for the parameters as a function of position in the cochlea (distance from apex in mm).

$\upsilon = 5$ (for all fibers)
$T_l = 8.13 \exp(-x/6.49)$
$T_f = 1.3 \exp(-x/2.0) + 0.4 \exp(-x/15.0)$

To relate these parameters to frequency, it is necessary to use the cochlear map proposed by Libermann(1982) that gives frequency in kHz as a function of percent distance from apex:

$$f = 0.456\ 10^{0.021\ d} - 0.80$$

Combining these two sources of information, we can plot the parameters as a function of frequency:

latency:

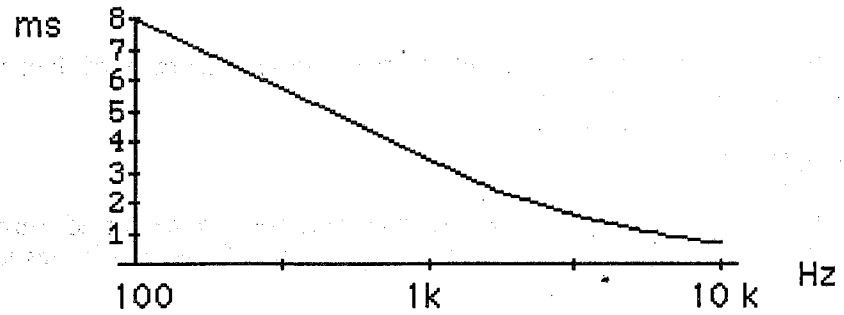


Fig. 2: Latency versus frequency over a population of fibers (Carney and Yin 1988).

time constant:

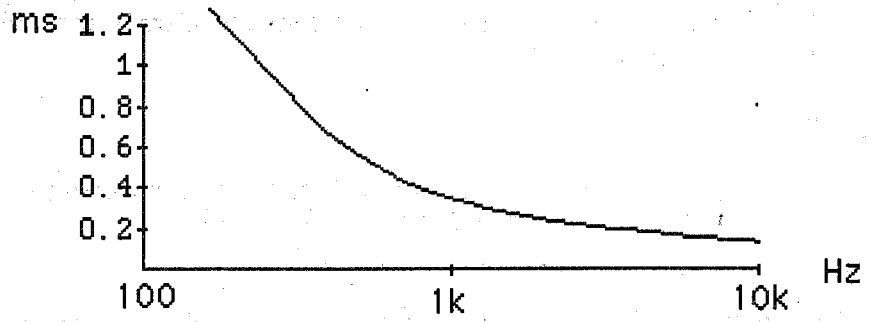


Fig. 3: Time constant versus frequency over a population of fibers (from Carney and Yin 1988).

It is interesting to plot the time constant in terms of *cycles* of the characteristic frequency:

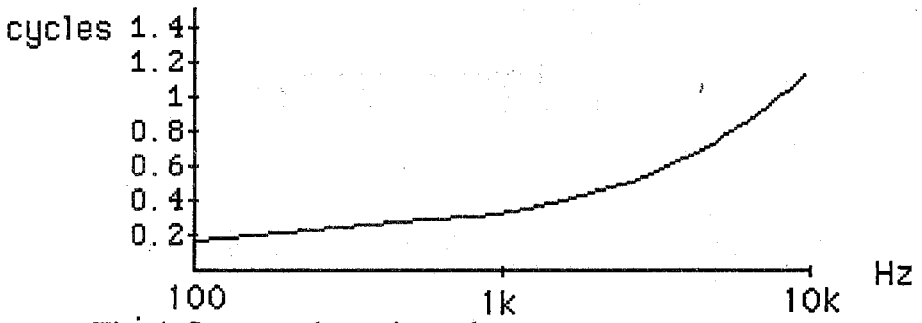


Fig. 4: Same as above, in cycles.

This time constant is remarkably short, meaning that the decay of the impulse response is fast.

The actual nerve fiber values show a scatter between about half and twice these nominal values.

John Holdsworth (1988) has implemented a filter bank based on revcor (or "GammaTone") functions. The parameters of this implementation are based on psychophysical masking patterns (Moore and Glasberg 1983) and differ somewhat from those presented here.:

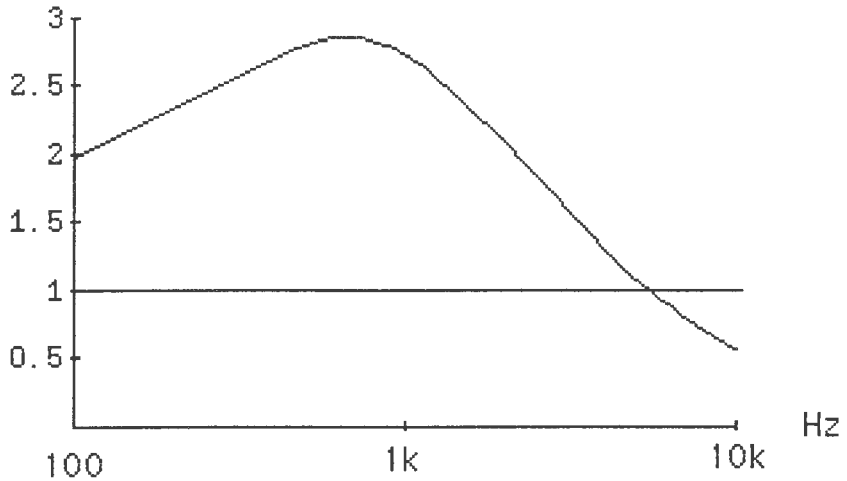
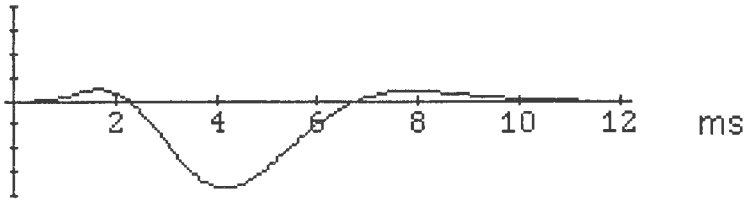


Fig. 5: ratio between the time constant of Holdworth's implementation (for υ = 5) and the value measured by Carney and Yin (1988).

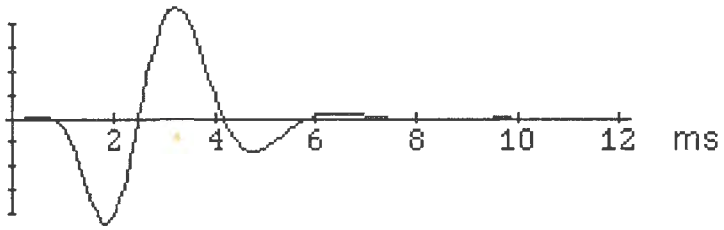
The following experiments use Holdworth's software modified so that:
- υ = 5 (instead of 4),
- the bandwidths are multiplied by 4.

The channel frequency spacing was chosen to correspond to one (widened) bandwidth. Seven channels were used. The impulse responses are plotted hereafter:

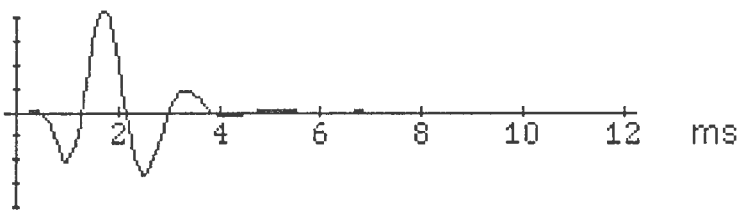
111 Hz:



300 Hz:



581 Hz:



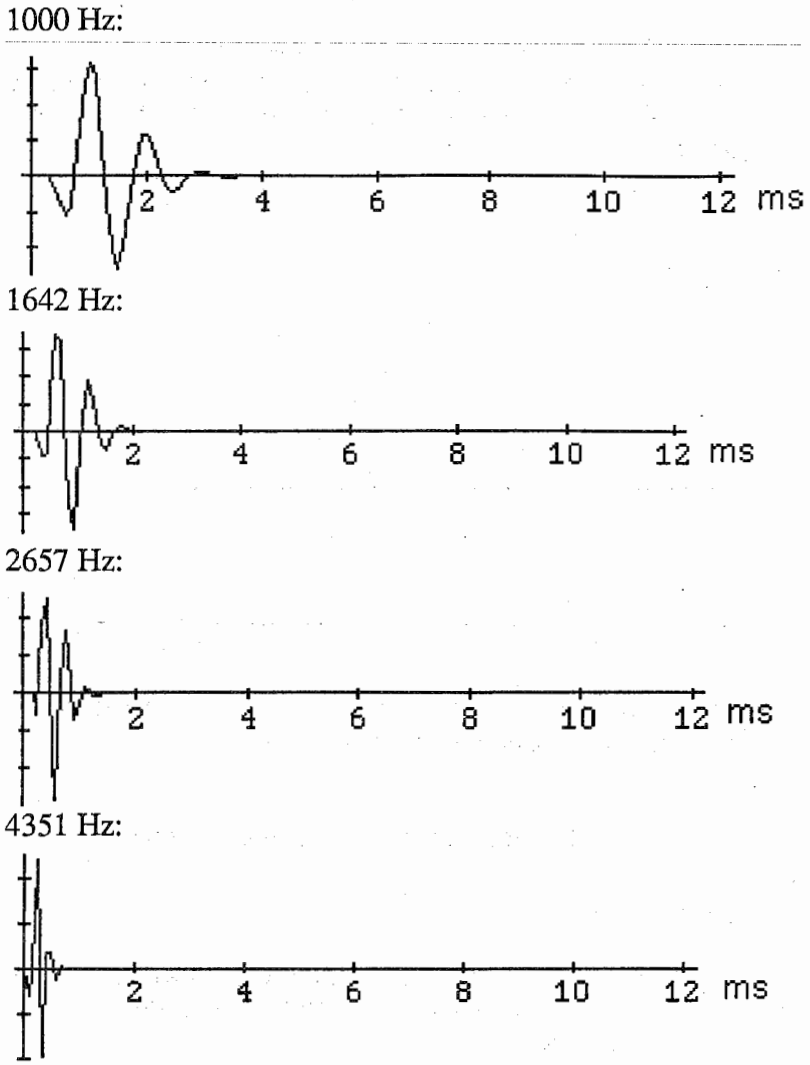


Fig. 6: Impulse response of each channel of the revcor filter bank.

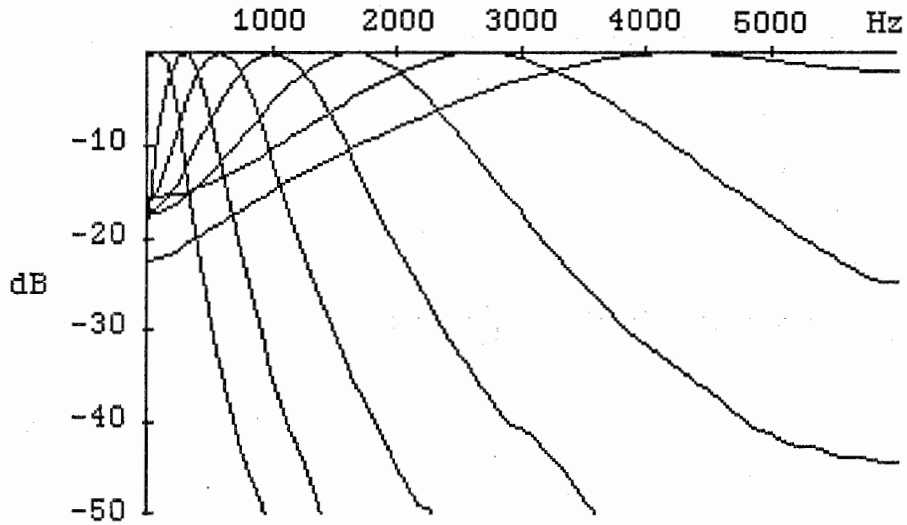


Fig. 7: Frequency responses.

There is no claim that this choice of filter type, bandwidths, range of frequencies and number of channels is optimal for the speech pitch extraction task.

Appendix II.

Statistics of the ATR pitch database (speaker MYI)

This section presents simple statistics of the ATR pitch database (speaker MYI). These provide a description of the data and offer insights as to the factors that cause pitch extraction errors.

I. The ATR pitch database

1) Aim

ATR is presently collecting an extensive speech database for speech recognition and synthesis purposes (Kuwabara et al. 1989).

The (pitch database) described here is more specifically synthesis-oriented. It was designed as a source of speech data labeled with pitch, segmental, and syntactic structure information, for speech synthesis purposes. It was not intended for the evaluation of pitch extraction methods, so the use to which it is put in this study is somewhat an abuse.

2) Speaker and material

The pitch database will eventually comprise data from several speakers, but for the moment only one speaker's data ("MYI", a male professional announcer) has been labeled. The speaker read a set of 503 sentences taken from novels (Abe and Kuwabara 1989, Abe et al. 1989).

3) Extraction method

Speech data was sampled at 12 kHz, 16 bit resolution. Pitch was extracted in two steps:
1) A simple cepstrum method (Abe and Kuwabara 1989) without pre- or post-processing provided an automatic first estimate that was:
2) displayed together with the signal on a specialized pitch editor and manually corrected.

The initial voiced-unvoiced decision was based on an energy threshold, and corrected during manual edition. This threshold was set very low, and all subsequent corrections were made in the *voiced-to-unvoiced* direction.

The automatic extraction method provided a pitch estimate aligned with the *center* of its analysis window, whereas the manual pitch estimator aligns the estimate with the *left* of a measured interval. This discrepancy is of little practical consequence on the estimate, because change is small on the scale of a half period, but it affects the position of voiced-unvoiced boundaries.

4) data format

Data, labels, and raw and corrected period estimates are located in separate directories (resp. DAT, LBL, PIT and MOD_PIT). Period estimates are provided at the rate of one for every 30 data samples (400 Hz sampling rate).

A *zero* estimate marks a non-voiced segment, a *negative* estimate marks a value that has been manually corrected. This allows us to count the occurences of various forms of error.

II. Statistics

The population statistics, pitch value histograms and pitch change histograms are based on the entire database (503 files). All other histograms are based on a limited set of 20 files chosen for their high error rate.

1) Populations

The data format conventions allow us to distinguish four populations of interest:
- voiced: the samples finally judged voiced, after correction,
- substitutions: those of the previous that were initially incorrect,
- insertions: the *unvoiced* samples that were initially judged voiced,
- omissions: the *voiced* samples that were initially judged unvoiced.

A subpopulation of the substitution errors consists of those samples that were initially judged too low.

2) Basic statistics

• Table 1. Number of samples:

voiced	substitutions	insertions	omissions
462528	23832	179775	0

The voiced portion represents 1150 seconds of speech. It is interesting to note the absence of "omission" errors: no portion originally judged unvoiced was subsequently labeled voiced. This is consistent with the very low threshold used in the initial voiced/unvoiced criterion.

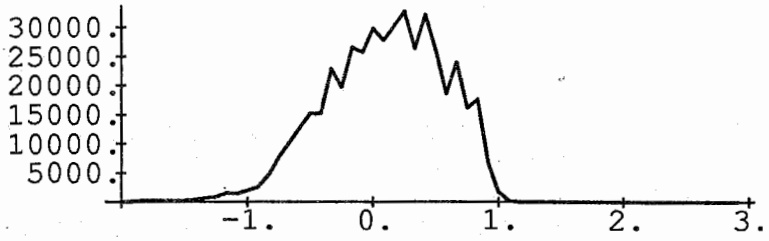
• Table 2. percentages:

substitutions/voiced	5.15 %
insertions/raw data	28.0 %
too low/substitutions	25.4 %

The insertion/raw data rate indicates that 28% of initial pitch estimates were discarded. *Insertion* and *omission* errors both reflect the voiced-unvoiced decision based on an amplitude threshold criterion, and therefore are of limited interest. In the following they are ignored and only *substitution* errors will be discussed.

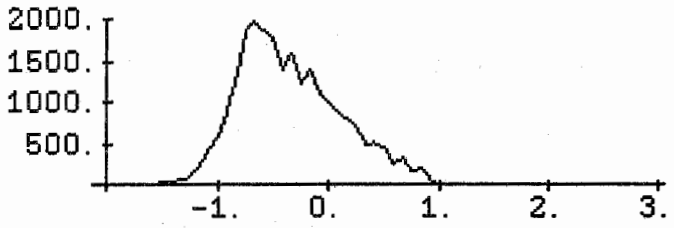
3) Histograms

• Histogram of pitch values for the voiced sample population (abscissa octaves re: 125 Hz, bin size 1/12th octave):

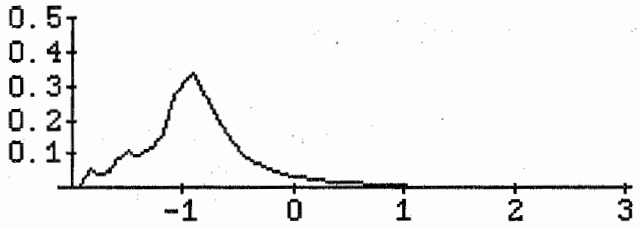


The roughness of the histogram is due to the interaction of the bin sampling with the limited resolution of period values. Apart from this roughness, the shape of the histogram is classic (Howard 1989).

• Histogram of (corrected) pitch values for which substitution errors occured (octaves re: 125 Hz):

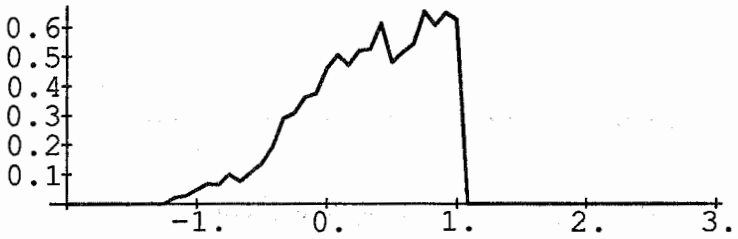


Perhaps more interesting is a probability plot obtained by dividing the second histogram by the first:



The probability of error is much higher for low pitch values. This shows that the automatic extraction method tends to mis-estimate low pitch values. Naturally, such errors will tend to be of the "too high" kind, consistent with the fact that "too high" errors are 3 times more common than "too low" errors (see table above).

• Probability that a substitution error is a "too low" error, as a function of pitch (octaves re 125 Hz):



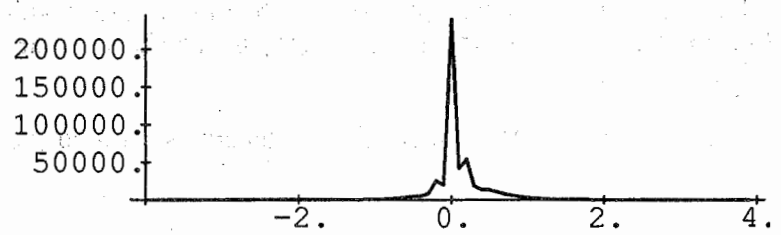
"Too low" errors are naturally more common for high pitch values.

These results together suggest that the automatic pitch extraction method was biased towards high values. Compensation of this bias in the pitch decision criterion might have lowered the error rate.

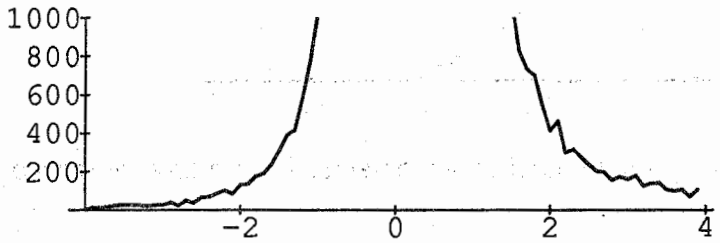
4) Pitch change

Pitch change, at a given pitch sample, is defined as the base 2 logarithm of the ratio of the preceding and following pitch sample values.

• Histogram of pitch change (abscissa 1/12th of octave, bin size 1/120th of octave):

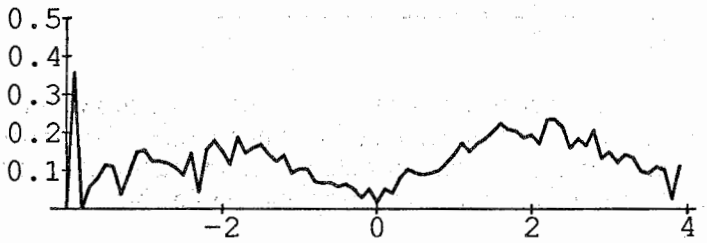


• Same plot with an expanded vertical scale:



Large pitch changes from sample to sample are rare.

• Probability of a substitution error as a function of pitch change (12th of octave):



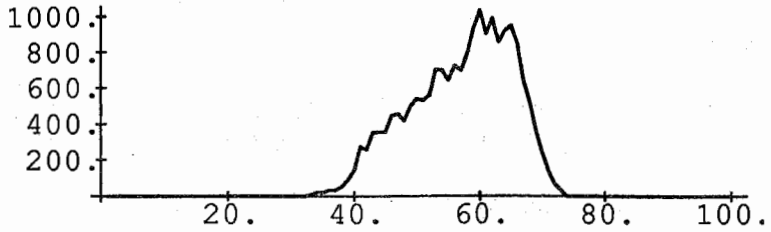
This plot shows that errors are somewhat less likely to occur in regions where the pitch is stable.

5) Amplitude

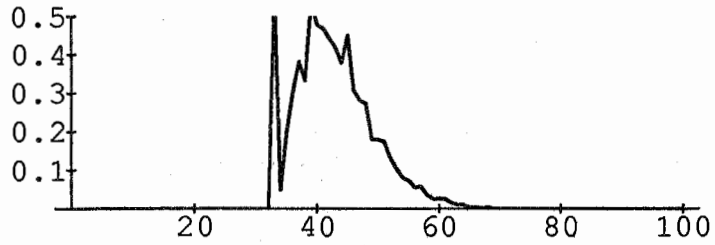
These histograms and all the following are based on a subset of 20 sentences chosen for their high substitution error rate.

Amplitude is mean absolute value over a 384 point (32ms) window.

• Histogram of amplitude values (in dB relative to quantization step):



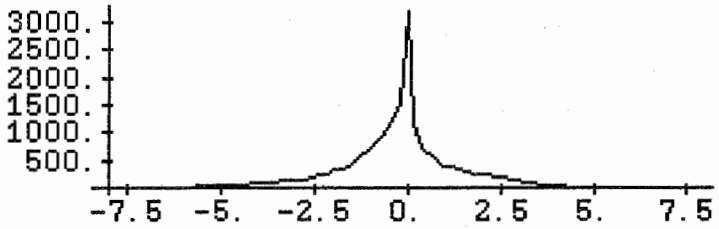
• Probality of a substitution error as a function of amplitude (dB re quantization step):



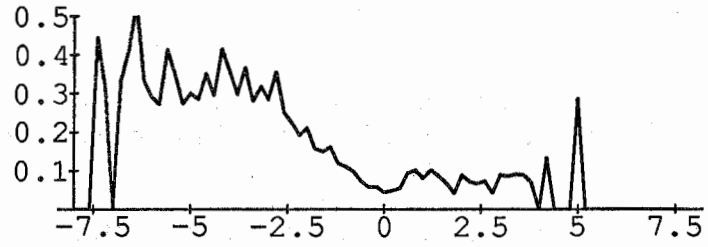
Errors are uncommon when the amplitude is high.

6) Amplitude change

• Histogram of amplitude change between two windows separated by a period (in dB):



• Probability of a substitution error as a function of amplitude change between 2 windows separated by a period (in dB):



Two interesting things to note:
- Errors are much more likely for *decreasing* amplitude than for increasing amplitude.
- Many errors occur even for zero amplitude change.

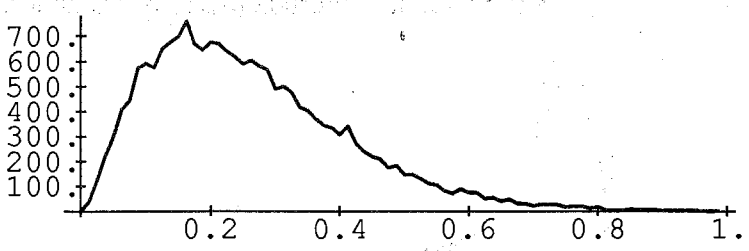
7) Spectral change

Spectral change is calculated in the following way:
- The signal is down-sampled in a 1:4 ratio, and Fourier-transformed using a 64 point hamming window to produce an amplitude spectrum a_i
- The total amplitude A for the window is calculated by summing the spectrum.
- Spectral change is calculated between two such windows separated by a period as:

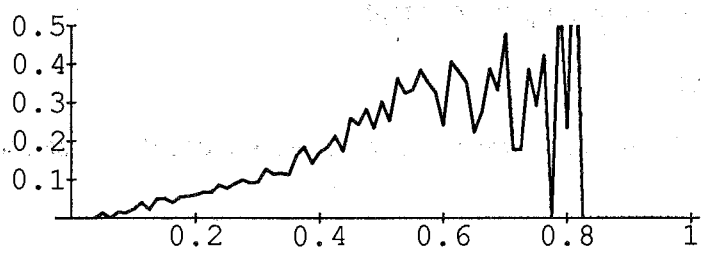
$$d_a = \frac{1}{A_1 A_2} \sum_i |A_1 a_{i2} - A_2 a_{i1}|$$

Since the spectra are normalized for amplitude variations, this measure reflects the change in amplitude spectrum *shape*. Phase differences are eliminated.

• Histogram of spectral change between two windows separated by a period (bin size .01):



• Probability of a substitution error as a function of spectral change:



As might be expected, when the period-to-period spectral change is small the error rate is also small.

8) Signal shape change

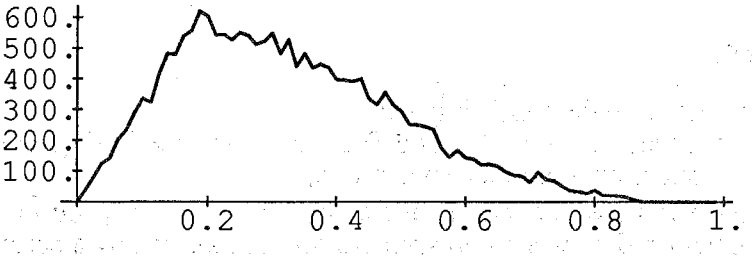
The signal shape change between two windows separated by a period is calculated as:

$$d_s = \frac{1}{S_1 S_2} \sum_i |S_1 s_{i2} - S_2 s_{i1}|$$

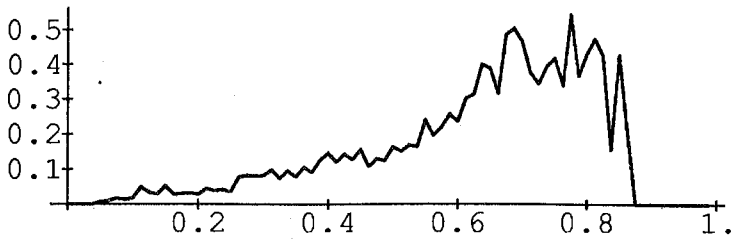
Where S_j is the total amplitude (sum of absolute values) of the signal over window j. Window size is 384 samples. Since the signal is normalized for amplitude changes the measure reflects only waveform shape changes.

Signal shape change differs from spectral change in that it is can be affected by phase changes.

• Histogram of signal shape change between two windows separated by a period:



• Probability of a substitution error as a function of signal shape change:



As expected, the error probability is low when the signal shape change between two windows separated by a period is small.

References

Abe, M., and Kuwabara, H. (1989). "Pitch frequency database on continuous speech", ATR technical report TR-I-0078.

Abe, M., Sagisaka, Y., and Kuwabara, H. (1989). "Integrating linguistic and prosodic information in a continuous speech database", ATR technical report TR-I-0079.

Howard, D. M. (1989) "Peak-picking fundamental period estimation for hearing protheses", JASA 86, 902-910.

Kuwabara, H., Sagisaka, Y., Takeda, K., and Abe, M. (1989). "Construction of ATR Japanese speech database as a research tool", ATR technical report, TR-I-0086.