Speech Enhancement in the Presence of Interfering Music and Noise

Louis Barbeau David Bernardi
Chung Cheung Chu Peter Kabal
Jean-Luc Moncet Douglas O'Shaughnessy

INRS-Télécommunications
3 Place du Commerce
Ile des Soeurs, Que.
CANADA H3E 1H6

January 1987

Rapport technique de l'INRS-Télécommunications no. 87-09

Speech Enhancement in the Presence of Interfering Music and Noise

Abstract

This report summarizes the results of speech enhancement experiments for a signal consisting of speech in the presence of interfering music and noise. Filtering was applied to remove hum and high frequency components. The composite signal was then frequency equalized to flatten the noise spectrum. A reference recording of the same passage of music which interferes with the original recording was obtained and time aligned with the composite recording.

The time-aligned reference music was processed through an adaptive filter and then subtracted from the composite recording. This results in a noticeable reduction and muffling of the music level. While before music cancellation, the music tended to dominate the composite signal, after cancellation the speech has a generally higher level than the music.

A number of other techniques were also investigated. The most successful of these is spectral subtraction. This involves suppressing those frequency components present in the music from the composite signal. This has the effect of suppressing the music, but since the desired speech component also contains the same frequency components, the speech quality is also affected.

The adaptive filtering approach has the least subjective effect on the speech components but does not completely suppress the music. The speech components are considerably more intelligible after music cancellation has been carried out. Spectral subtraction lends a somewhat unnatural quality to the resultant signal, but does render more complete suppression of the music. The speech is slightly muffled. The intelligibility of the speech can be judged to be about the same or better than for the adaptive filtering approach.

Contents

Figures

Tables

Speech Enhancement in the Presence of Interfering Music and Noise

1. Introduction

This report summarizes the results of a speech enhancement project. The signal processing group at INRS-Telecommunications was presented with an audio tape containing speech with interfering music and noise. Using the signal processing facilities at INRS-Telecommunications, we were able to reduce the effects of the interfering signals and to enhance the intelligibility of the speech.

The original tape resulted from electronic surveillance of a room in which two adult males were talking, while a phonograph was playing music. The recording microphone sent the signal over a telephone line to a remote tape recorder. Sources of degradations in the recorded signal are: (1) background room noise, (2) room reverberation, (3) interference (in the form of hum) picked up in the recording process, (4) non-linear distortion, and (5) restricted frequency response (about 300–2500 Hz) due to limitations of the recording path (including the telephone line). However, the principle factor contributing to the lack of comprehension of the speech is the presence of the interfering music in the room. Due to the placement of the microphone, the level of music is in fact higher than that of the speech.

The purpose of this report is to document the techniques used to process the intercept signal. The various techniques used will be discussed and suggestions for future work to develop new algorithms that have potential for better speech enhancement will be given.

1.1 Processing Techniques

The intercept recording was processed using some general equalization and filtering techniques. By examining a section of the tape containing only background noise (no speech or music), a spectral model of the various distortions to which the tape had been subjected was obtained. The intercept signal was then frequency equalized to reduce the variation of the response at different frequencies. This has the effect of reducing some of the resonant qualities of the original recording. In addition, filters to notch out interfering hum components were used.

Our attempts at further processing of the recording centered around subtracting out as much of the music as possible and suppressing the noise. The interfering music is easily identifiable as a popular widely-distributed record. To carry out the enhancement work, digitized versions of both the intercept recording and the reference music were created. The digitization step involved filtering to 5000 Hz (to prevent aliasing) and sampling at 10,000 samples per second with a resolution of 16 bits. This step involved essentially no loss of information of the intercept recording, as could be verified by playback of the digitized signal.

Prior to subtracting the digitized music signal from the intercept recording, the music signal had to be modified to align it properly with the music component of the intercept signal. The music component of the intercept signal had been subjected to room reverberation effects and also to time-scale modifications. These time-scale modifications are due, in fact, to imprecise phonograph speeds and imprecise tape recorder speeds. While these speed differences and variations are small enough that they do not affect the perceived quality, they are large enough to adversely affect the ability to cancel the music components using the reference music signal.

Time alignment of the two signals was accomplished using the waveform display facility in our laboratory. This facility allows for the examination of time waveforms and for the display of frequency components (spectrograms). Initially, the two signals were aligned visually using anchor points which represent obvious points of coincidence. The reference music signal was then "stretched" or "shrunk" using digital interpolation techniques to bring samples into time alignment.

The effects of the room reverberation and the recording path can be largely modelled as linear filters. The aim is to subject the reference music signal to the same filtering actions as the music component of the intercept recording and then to cancel the modified reference music signal from the intercept recording. This is accomplished using adaptive filtering techniques applied to the reference music signal (after time alignment). Since only the reference music signal is filtered, the speech on the tape is not further distorted. Because the ensemble of distortions cannot be completely modelled by the adaptive filter, not all of the music component can be removed by this process. However, there is a noticeable reduction and muffling of the music, which allows the speech components to come to the foreground.

A number of other techniques to enhance the speech comprehension were also investigated. These can be viewed as alternatives to the adaptive filtering technique. Least successful were techniques to track the pitch of the vocal components of the music and to use these to notch out the corresponding components in the intercept recording. Another more successful attempt centered around using spectral subtraction techniques. This involves trying to suppress those frequency components that appear in the reference music, in the intercept recording. This has the effect of suppressing the music components, but since the desired speech sometimes also contains these frequency components, speech quality is also affected.

The adaptive filtering approach has the least subjective effect on the speech components but does not completely suppress the music. The speech components are considerably more intelligible after music cancellation has been carried out. The spectral subtraction technique can lend a somewhat unnatural quality to the resultant signal, but does render more complete suppression of the music. The method allows for a tradeoff between the degree of music suppression and the introduction of tonal noise artifacts. There is a slight muffling of the speech when the music spectrum significantly overlaps the speech spectrum. The intelligibility of the speech can be judged to be about the same or better than for the adaptive filtering approach.

2. Signal Characterization

The intercepted signal has three components: interfering music, a conversation between two people, and background noise. The music component is relatively louder than the speech for most of the recorded segment. The background noise is significant, though the speech is louder than the background noise. The noise magnitude spectrum is shown in the top part of Fig. 5. This was obtained from a portion of the recording with no music or speech present. It has the highest energy in the band between 200 and 1500 Hz.

The music component consists to a large extent of a single singer accompanied by a variety of musical instruments. The top part of Fig. 16 shows the spectrogram of a segment of the intercepted signal with no speech present. The widely spaced series of bands corresponds to the singer's voice, while the denser series belongs to the accompaniment; the overall spectrum, as can be seen, is very rich. The music on this record is for the most part tonal, therefore the spectrum is mainly harmonic (if percussion instruments are excluded).

For the purposes of music cancellation, a recording was also made of the same music that appears as interference on the intercept recording. This reference music signal was recorded from a phonograph record.

Fig. 1 displays a block diagram describing the complete system that models the relationship between the reference music signal and the intercept signal. The upper part of the diagram describes the reference music recording, while the lower part describes the intercept recording. In this figure, each of the boxes tagged with MTR corresponds to a Mechanical Transducing Device, i.e. a phonograph or a tape recorder. A more detailed view of each MTR is given in Fig. 2(a). The three constituents of the MTR's take into account: the frequency response, the non-linear distortion and the time scale or speed errors of these devices. The non-linear distortion in an MTR can be caused by amplifiers or loudspeakers as in the case of the phonograph, or by amplifiers and the non-linearity of the magnetic medium in the case of tape recorders. The time scale errors consist of the wow (0.5–2 Hz), flutter (2–200 Hz), and speed offset error. Not represented in this figure is the electronic and mechanical noise generated by the amplifiers and the transducer.

In the reference music path, MTR_4, MTR_5, and MTR_6 represent respectively a phonograph, the tape recorder on which the music was recorded, and the playback tape recorder used at the input of the digitizing apparatus.

In the intercept signal path, MTR_1, MTR_2, and MTR_3, represent respectively a phonograph, the intercept tape recorder, and the playback tape recorder used at the digitizer input. Note that the reference music signal and the music component of the intercept signal are separated by six MTR devices, each of which can introduce a time scale modification. While motors are the source of most of the speed changes, other sources of time scale errors are those due to tape stretch and the

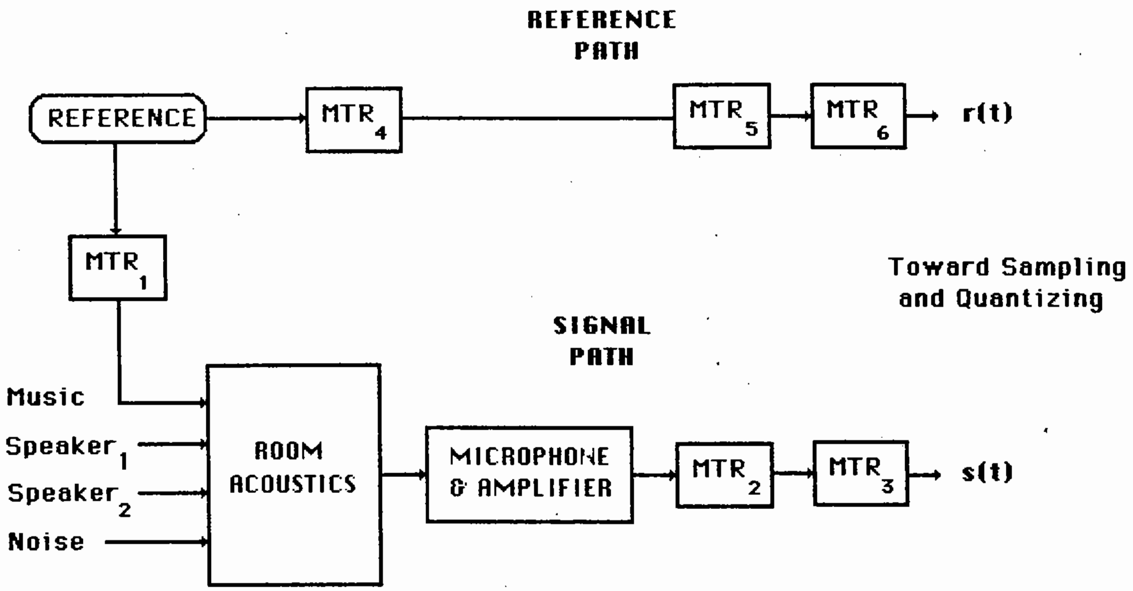
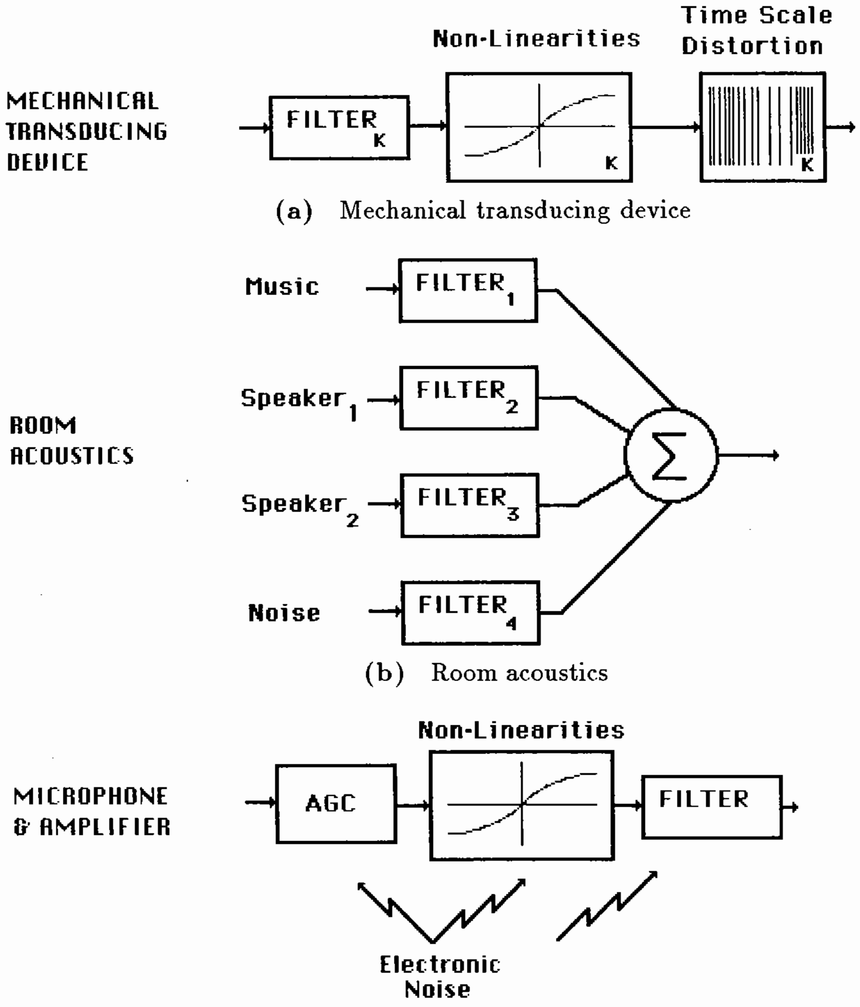


Fig. 1 Block diagram showing the relationship between the reference music signal and the intercept signal

sampling clock at the digitizer. However, it can be anticipated that this last error is negligible with respect to the errors introduced by the motor driven MTR devices. The cumulative time scale error between the digitized reference music signal and the digitized intercept signal is a major problem that must be overcome before synchronous cancellation techniques based on adaptive filtering can be applied.

In the intercept signal path, four different signal sources contribute to the microphone input signal: the interfering music, the two speakers, and the background noise. The model shows filters between the input signals and the microphone input. These represent the reverberant effect of the room acoustics (see Fig. 2(b)). Note that the room acoustics are different for each of the four sources since the geometry of the reflecting paths is different for each of these sources. In addition, the room acoustics change as the speakers move about in the room. This affects primarily the speech, but at a secondary level also affects music.

The microphone and intercept amplifier are modelled as a three-stage sub-system. The filter takes into account the combined frequency response, which is relatively narrow, in the order of 250 to 2500 Hz; see Fig. 2(c). Part of this limited response is due to the telephone line used to transmit the signal from the recording site to the tape recorder. The non-linear transfer function box corresponds to the non-linear distortion, which listening shows to be perceptually important. The Automatic Gain Control, AGC, is a device that tries to optimize the amplifier gain according to the input signal level. It slowly increases the gain when the input level is too low, and rapidly decreases the gain when the level is too high (fast attack, slow decay). The resulting effect of the



(a) Mechanical transducing device

(b) Room acoustics

(c) Microphone and amplifier system

Fig. 2 Details of the signal models

AGC on the program material, when there is a sudden increase in the input signal, here a shout, can be seen on the spectrogram at the top of Fig. 3. The level of the background music decreases dramatically when the shout occurs at around sample 67,000. The behaviour of the AGC, when submitted to a low level signal, is depicted at the bottom of Fig. 3. The light trace is the frequency domain representation of a portion of the signal with no music or speech present, and the dark trace is another portion with background noise only approximately 1.9 seconds later. The rise in gain is in the order of 4 to 5 dB as the AGC tries to compensate for the low input level. On the same figure, we can see that beyond 2500 Hz, there is little variation in level; thus reinforcing our estimate of the upper limit of the frequency response.

Listening to the intercept recording, one can easily make out the music. This music differs significantly in quality from that on the reference music recording in that it has an additional

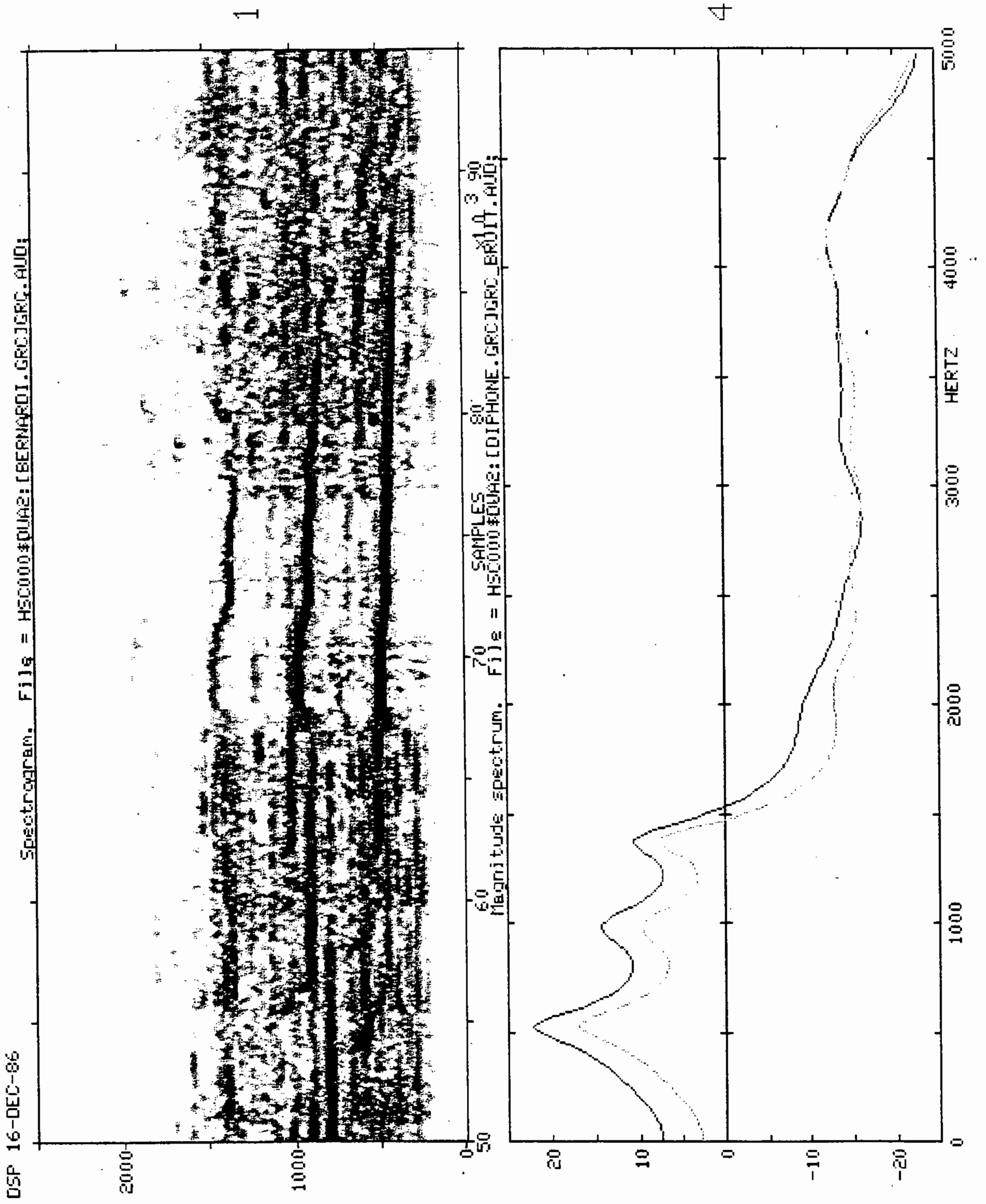


Fig. 3 Automatic Gain Control effects
Top: Spectrogram (narrowband mode) of the intercept signal showing the effect of the AGC
Bottom: Spectrum (20 pole LPC fit) for two frames of background noise alone 1.9 seconds apart

reverberant quality, is severely bandlimited and is subject to distortion (non-linear effects). The speech is also highly reverberant and is typically at a lower level than the music. Even in the gaps between music tracks on the phonograph record, close attention must be paid to be able to comprehend the speech. In the parts in which music dominates, the speech is quite difficult to understand, although with repeated playback of a section it is usually possible to understand a good portion of the speech. In addition, the intercept recording has a resonant quality and the speech components lack presence due to the non-flat frequency response of the overall room environment and the recording device.

3. Preprocessing

For further processing, the intercept recording and reference music recording need to be digitized. In addition, the reference music recording must be time aligned with the intercept recording. These preprocessing steps are described below.

3.1 Sampling

Digitization involves the three steps of analog prefiltering (to prevent aliasing), sampling and quantization. The latter two steps are carried out by an A/D converter. Once digitized and stored as computer files, the signals may be processed. The final product can be played back using a D/A converter.

The digitization equipment is composed of a Digital Sound Corporation DSC-200 audio data conversion system and a DSC-240 input preamplifier and output amplifier. The digitization is carried out to a 16 bit accuracy. The analog prefilters used were Precision Filters model 32-R-LP2. These are 6th order elliptical filters with a 0.1 dB in-band ripple. The 0.1 dB cutoff point was set at 0.415 times the sampling frequency. With this setting, the half sampling frequency is down 18 dB. The recording filter also includes an 80 Hz highpass filter.

The intercept recording was available as a 1/4 inch reel-to-reel tape recorded at 1 7/8 ips. The playback tape recorder was a Revox model PR99. The intercept recording was played back at 3 3/4 ips (the lowest speed available on the Revox machine) and digitized at 20,000 samples/second. This doubling of the sampling rate cancels the doubling of the playback speed. The result is a sampled signal at the equivalent of 10,000 samples/second.

The reference music was available as an audio cassette which had been transcribed from a phonograph record. The audio cassette was played back on a Revox B215 cassette tape deck and was digitized directly at 10,000 samples/second.

Initial experimentation was carried out on a short extract of the intercept signal. This 14.5 second extract contains segments with background noise alone, some speech without music, music without speech, and some music plus speech. This segment of the intercept recording will henceforth be referred to as $s(n)$. The matching segment of the reference music signal will be referred to as $r(n)$.

A third sequence, $b(n)$, was created to satisfy the training needs of some of the noise removal algorithms which will be the subject of Sections 4 and 5. This segment is a 2.5 second, pure noise segment from the intercept recording. This segment is found approximately 3 minutes before the beginning of the $s(n)$ sequence on the intercept recording.

3.2 Digital Filtering

The intercept recording has a limited frequency range. Based on an examination of the response of the intercept recording, filtering to a range between 220 Hz and 3100 Hz is appropriate. Highpass and lowpass filters were designed with these cutoffs. Each filter was as 16th order elliptical design with a passband ripple of 0.1 dB [1].

The first five harmonics of a 60 Hz interference signal also appear in $s(n)$. Those harmonics below 220 Hz are removed by the highpass filter. Notch filters were designed for 240 Hz and 300 Hz. Each notch filter is a 16th order elliptical notch filter with a bandwidth of 8 Hz. The composite response of the notch filters and the highpass and lowpass filters is shown in Fig. 4.

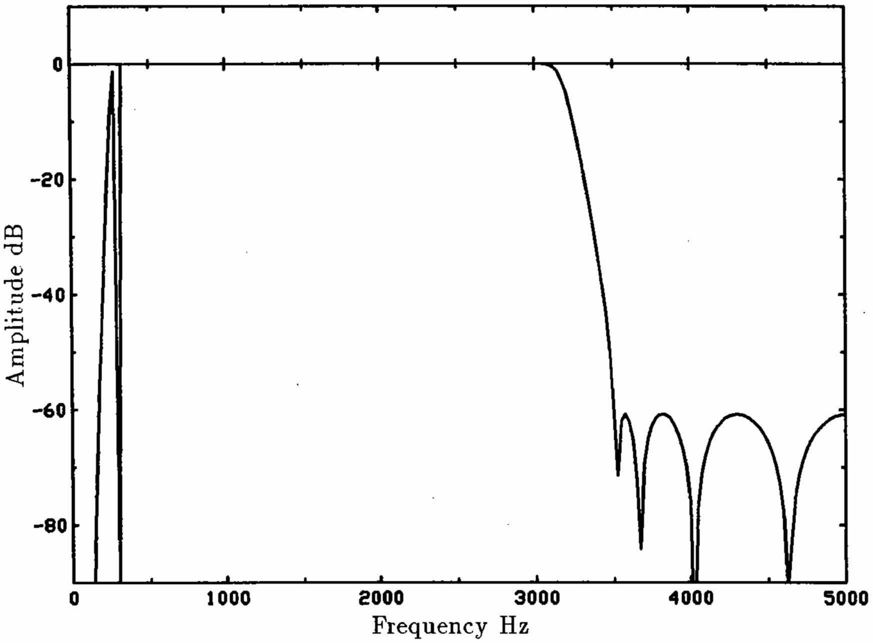


Fig. 4 Frequency response of the composite digital filter

The four filters described above, applied in sequence, make up the digital filtering operation. This filtering is applied to $s(n)$ to form $s_f(n)$ and to $r(n)$ to form $r_f(n)$. The purpose of applying these filtering operations to the reference music signal is to enhance the similarity with the filtered intercept signal and so facilitate the cancellation of the music components of the intercept signal. When required, $b(n)$ is also filtered to form $b_f(n)$. Figure 5 shows the frequency domain representation of $b(n)$ and of $b_f(n)$.

3.3 Temporal Alignment

Adaptive filtering techniques used for cancellation require good temporal alignment to be suc-

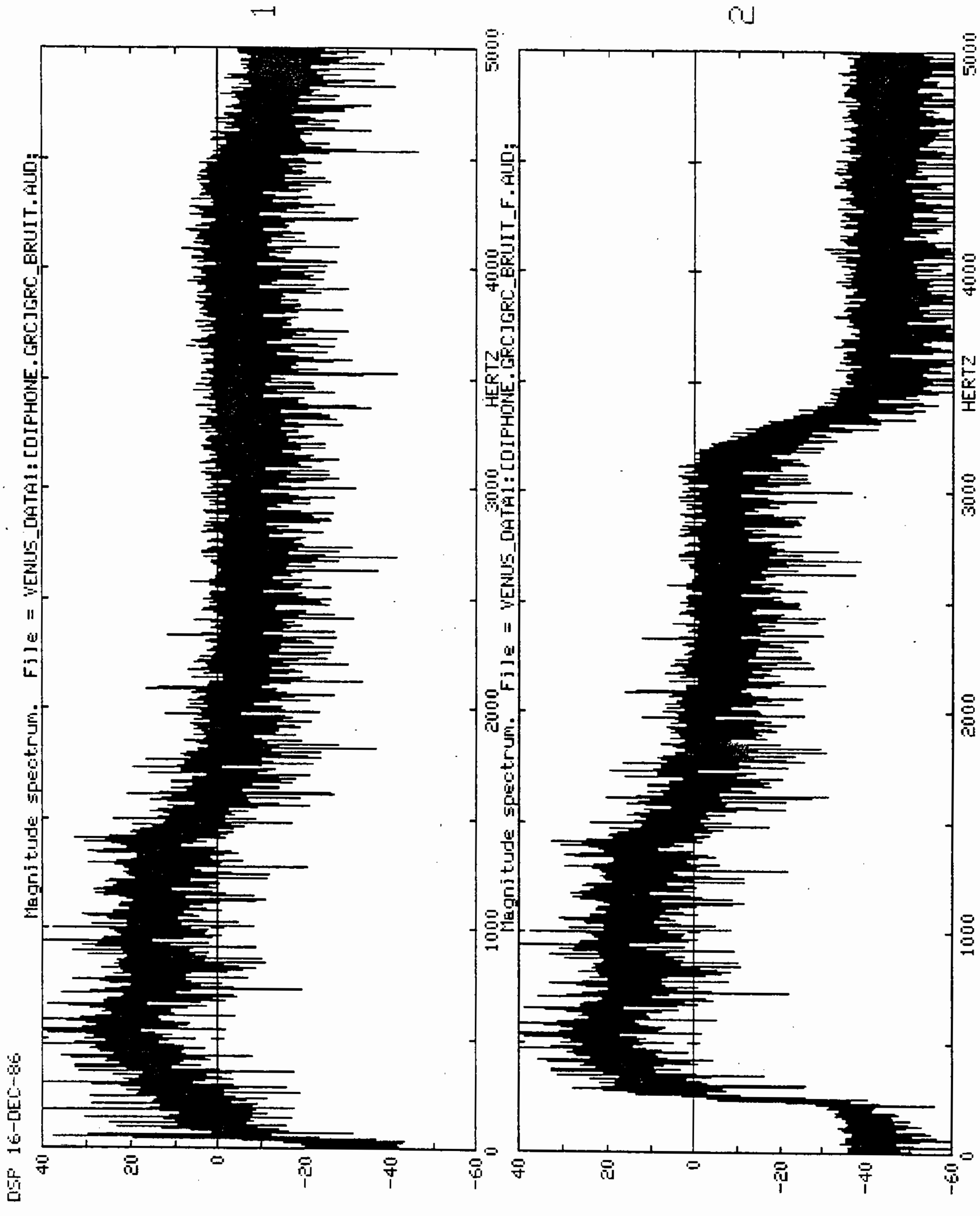


Fig. 5 Background noise spectrum
Top: Before digital filtering, $b(n)$
Bottom: After filtering, $b_f(n)$

cessful.[†] Because of the variability in the speeds of the various recording and playback devices, alignment at only one temporal location is not adequate. The gradual shifting away from synchrony must be compensated for by *stretching* or *warping* one of the signals into time alignment with the other.

3.3.1 Signal Cross-correlation

The signals $r_f(n)$ and $s_f(n)$ can be aligned by first locating corresponding and clearly identifiable points present in both signals and by then shifting $r_f(n)$ so that these points coincide. These events should be selected so that they are very localized in time particularly in $s_f(n)$ where events tend to be temporally smeared due to reverberation effects.

One method of assessing the degree of match is to calculate a cross-correlation between $s_f(n)$ and $r_f(n)$. For this calculation, the cross-correlation is calculated over a block with a block length sufficiently small so that the signals do not drift excessively with respect to one another over the block. The normalized cross-correlation between a signal $x_1(n)$ and a signal $x_2(n)$ is defined as

$$\alpha(k) = \frac{\sum_{i=n_1}^{n_2} x_1(i)x_2(i-k)}{\sqrt{\sum_{i=n_1}^{n_2} x_1^2(i) \sum_{i=n_1}^{n_2} x_2^2(i-k)}} ,$$

where k is the cross-correlation lag. The block length is $N = n_2 - n_1 + 1$. For our purposes, the cross-correlation will be calculated for successive non-overlapping blocks. In this case, $n_2 = n_1 + N - 1$ and n_1 is incremented by N for each block. The top part of Fig. 6 shows a contour plot of the normalized cross-correlation function for $s_f(n)$ and $r_f(n)$ calculated for non-overlapping blocks of 256 samples. The contour delimits all values with normalized absolute cross-correlation values above 0.5. For any given block, the signals are highly correlated for lags contained within the contours. Consider the calculation of the correlation of the signal $s_f(n)$ with itself. One would expect correlation components due to both reverberation and to inherent correlation in the signal itself. Auto-correlation plots are shown for both $s_f(n)$ and $r_f(n)$ in Fig. 7.

Generally, for two adjacent blocks, the correlation functions should be very similar. In the case of the cross-correlation shown, the downward drift of the correlation peaks as time progresses is a measure of the speed differences. From examination of the time waveforms, one can ascertain that the signals are time aligned near block 156 (see also Fig. 6, top). Before this point, $r_f(n)$ lags $s_f(n)$; past this point, $r_f(n)$ leads $s_f(n)$.

[†] Frequency domain techniques also require a temporal alignment, but are generally less sensitive to small misalignments than techniques such as adaptive cancellation which operate in the time domain.

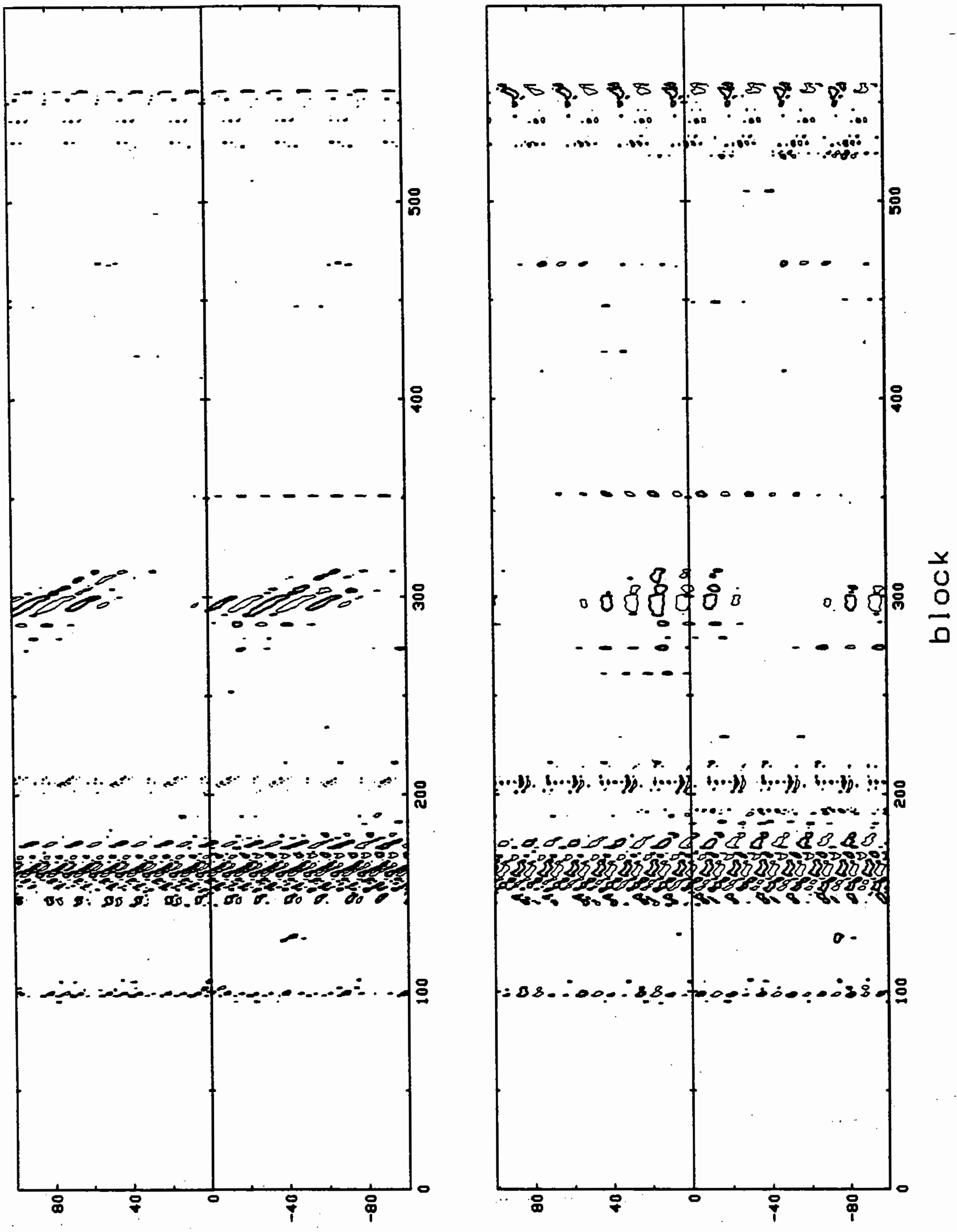


Fig. 6 Contour plots of the normalized cross-correlation. The contours represent the 0.5 level (heavy line) and the -0.5 level (light line). The vertical axis is the lag.
Top: Cross-correlation between $s_f(n)$ and $r_f(n)$ (before alignment)
Bottom: Cross-correlation between $s_f(n)$ and $r_f^a(f)$ (after alignment)

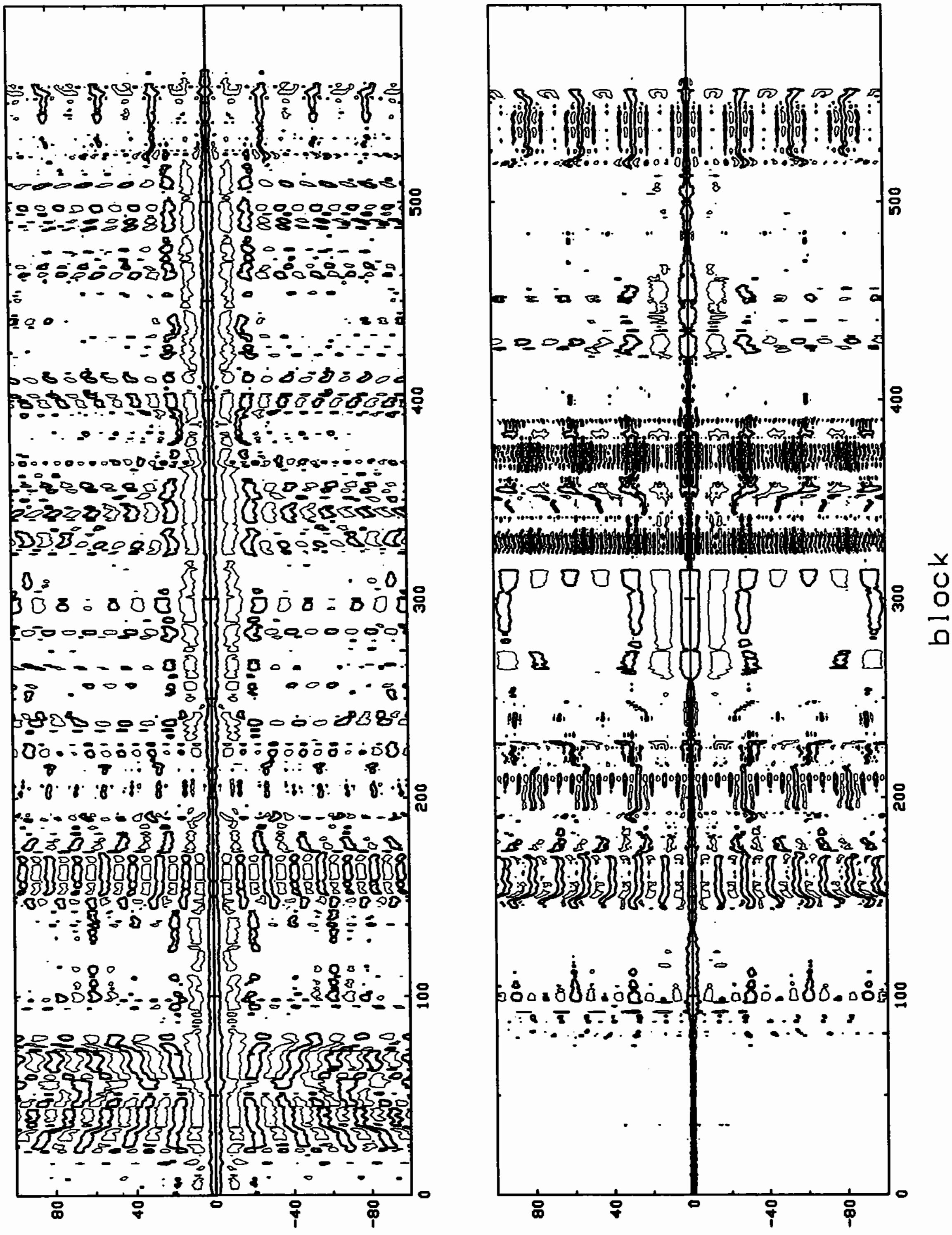


Fig. 7 Contour plots of the normalized auto-correlation. The contours represent the 0.5 level (heavy line) and the -0.5 level (light line). The vertical axis is the lag.
Top: Auto-correlation for $s_f(n)$
Bottom: Auto-correlation for $r_f(f)$

The complexity of the cross-correlation plots show that automatic alignment between the reference and the intercept recordings will be difficult to achieve. In our case, manual alignment was used to find anchor points.

3.3.2 Anchor Points

Given two points of coincidence of the reference and intercept signals, stretching or warping of the refence signal to approximately synchronize the signals over the intervening segment is used. To be effective, the drift must be predominately linear in nature. This is true for the test segment, since (i) the greatest contributing factor to the drift is the differences in the speeds of the various recording and playback devices, and since (ii) over relatively short time periods the ratio of these speeds to one another can be assumed to be constant.

Assuming most of the drift is attributable to a linear component, one can select distant anchor points where local alignment is possible. This is the process that was used for the test segment of the intercept recording. Figure 8 shows a closeup of one of the anchor points used in the alignment and in the estimation of the drift. The temporal alignment seems to be correct to within plus or minus one period or, roughly, to ±30 samples, although the figure illustrates the difficulties in determining correspondences to high accuracy. From these anchor points, the linear component of the drift can be obtained. Over a range of 146,493 samples, the drift amounts to 1089 samples ±60 samples.[†] This translates into a drift of 0.743% ±0.041%. From the cross-correlation plot in Fig. 6, we can check this estimate of the drift by computing, where possible, the slope of a line passing through successive cross-correlation peaks. The slope is about 2.1 samples/block (0.82%) around block 160 and 1.7 samples/block (0.67%) around block 300.

3.3.3 Interpolation/Decimation

The stretching of the reference signal is achieved using interpolation/decimation techniques. Essentially this is a three step process. First the sampling rate of the signal is increased by inserting zero-valued samples between each sample. Filtering is then used to change these zero-valued samples into the appropriate estimates of the interpolated values of a bandlimited process. The last step is to decimate or subsample the increased rate signal. The combination of interpolation/decimation allows for sample rate conversion in which the ratio of the resultant sampling rate to the original sampling rate is a ratio of integer values.

The interpolating filter was designed to minimize the mean-square error in the interpolated signal, given a power spectral model for the input signal. This filter is a linear phase FIR filter. In the actual implementation, interpolation and decimation are combined as a single step to avoid the

[†] The ±60 figure reflects our estimate of the uncertainty in the matching of the anchor points.

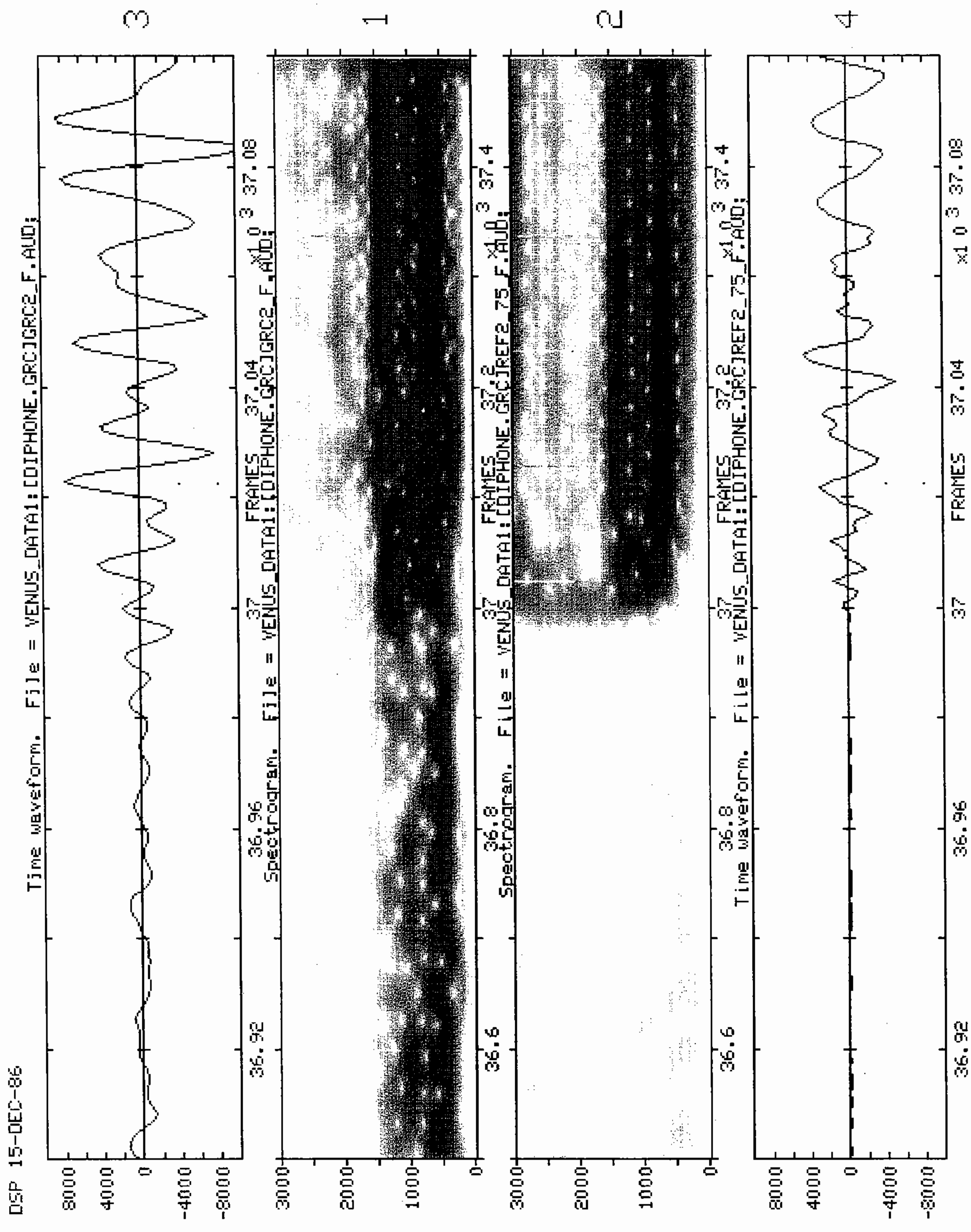


Fig. 8 Waveforms and spectrograms (wideband mode) for one of the alignment points. Abscissa values are in thousands of samples.
Top: Intercept signal
Bottom: Reference music signal

computation of interpolated values which are not needed in the output signal. Our implementation limited the maximum interpolation ratio to be 25. For greater sampling rate resolution two stages of interpolation/decimation were used. For the sample segment the signal was first interpolated by a factor 7, then subsampled by a factor 6 to change the sampling rate by 7/6. Then the resultant signal was interpolated by a factor 22 and subsampled by a factor 19. The overall sampling rate is then changed by $(7 \times 19) \div (6 \times 22)$, corresponding to a speed increase of 0.7576%.

The sampling rate change is performed on $r(n)$. The interpolation and decimation yields the time-aligned signal $r^a(n)$ which is then filtered to produce $r_f^a(n)$. The bottom part of Fig. 6 shows the cross-correlation of $s_f(n)$ with the time-aligned reference signal $r_f^a(n)$. As is evident from this figure, much of the drift has been removed by the interpolation/decimation step.

3.3.4 Time Warping

The simple change of sampling rate as described above is appropriate for the test segment of the intercept recording. However, for the processing of longer segments, anchor points at the beginning and end of the segment are no longer adequate. For the processing of these longer segments a number of anchor points were used. The spacing between anchor points is chosen to be sufficiently small so that the drift is held to less than a fraction of the adaptive filter length (201 samples for most of the processing). The anchor points define points of coincidence. Between anchor points the reference music signal was linearly stretched or shrunk. This requires the availability of values from the reference music signal which occur between samples. These values were determined in two steps. A bandlimited interpolation using an interpolation by a factor 9 was used to find two values which bracket the desired sampling point. Linear interpolation between these bracketing values gives the final output value. The overall process allows piecewise linear changes in the sampling rate.

The interpolation filter was designed to minimize the mean-square error of the interpolated signal for a signal with a model power spectrum which has a cosine rolloff from 4500 Hz to 5000 Hz. The design procedure is a generalization of the procedure described in [2]. The 9 times interpolation rate gives the best performance in the time warp application for filters with the number of coefficients constrained to be less than 400. Fig. 9 shows the frequency response of the interpolating filter.

The accuracy of the interpolation process was determined by changing the sampling rate of the test segment of reference music by a factor corresponding to increasing the length of the signal by 1000 samples over its full length (0.7% change) and then restoring the nominal sampling rate by a second sample rate change process. The signal-to-noise ratio for the restored signal which had been subjected to the interpolation process twice was 45 dB.

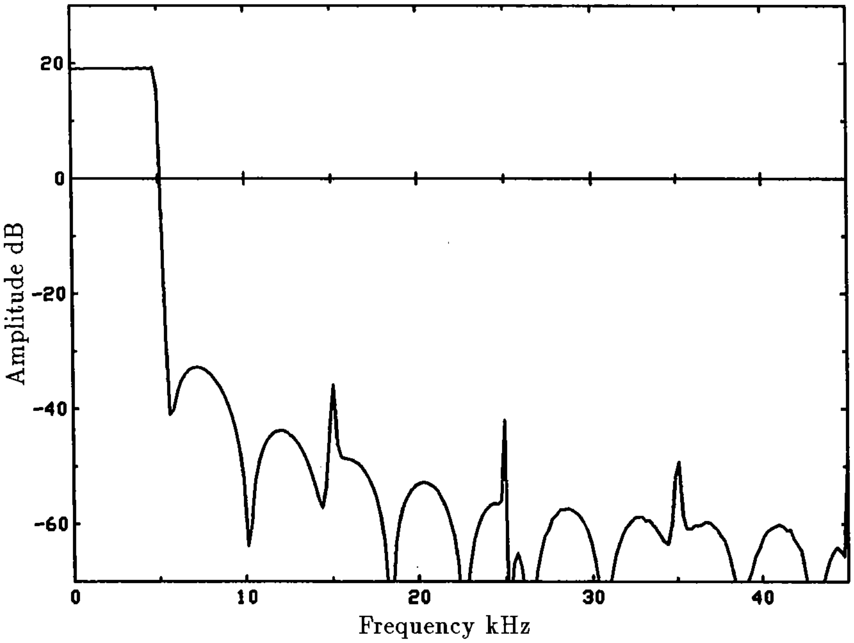


Fig. 9 Interpolating filter

4. Noise Removal

The background noise in intercept recording is present at a level sufficiently high that means to help reduce it were investigated. Two techniques were tried. They are similar in principle and both use a segment of background noise to train the algorithm. The learned noise spectrum is then subtracted from the spectrum of the intercept signal.

4.1 Noise Spectral Subtraction (Method I)

One of the noise removal techniques which was tried is a variation on the spectral subtraction algorithm due to Boll [3]. Boll's technique was designed (and successfully used) for the removal of helicopter noise from speech.

Qualitatively, our version of the noise removal algorithm proceeds as follows:

(1) The average magnitude of the noise spectrum $\mu(k)$ is computed at discrete frequencies. This is done for a segment consisting of background noise only. This is the signal $b(n)$ referred to earlier.

(2) The quantity $\mu(k)$ is removed from the magnitude value $|S_i(k)|$, where $S_i(k)$ is the Discrete Fourier Transform (DFT) of the intercept signal $s(n)$ for a frame. If the result is negative, the result is set to zero. The phase of the original components is retained.

(3) An additional step is applied to reduce the frame-to-frame fluctuations in $|S_i(k)|$.

Let N be the frame length ($N = 256$), $w_H(n)$, a Hanning window of length N and i, the frame number, then

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n + i\frac{N}{2})\, w_H(n)\, e^{-j\frac{2\pi kn}{2N}} \qquad \text{for } 0 \leq k \leq 2N-1 \, .$$

Note that the frames overlap by $N/2$ samples and that each frame is padded with N zeros prior to taking its DFT.

The average magnitude of the noise spectrum, $\mu(k)$ is defined as an M term average of $S_i(k)$.

$$\mu(k) = \frac{1}{M} \sum_{i=I}^{I+M-1} |S_i(k)| \, ,$$

where the segment of $s(n + i\frac{N}{2})$ for $i = I$ to $i = I + M - 1$ is composed of background noise alone.

The spectral function which will remove frequency components is

$$H_i(k) = 1 - \frac{\mu(k)}{|S_i(k)|} \, .$$

This function is multiplied by $|S_i(k)|$. Note that this form can result in negative valued spectral components. These components are set to zero since they represent components heavily corrupted by noise. To this end, define a new rectified filter,

$$H_{Ri}(k) = \frac{H_i(k) + |H_i(k)|}{2} \, ,$$

so that the output spectrum is

$$\hat{S}_i(k) = H_{Ri}(k)\ S_i(k)\ ,$$

where $\hat{S}_i(k)$ is $S_i(k)$ with noise removed.

The final step consists of eliminating frame-to-frame fluctuations attributable to the residual noise. This step takes place whenever

$$|\hat{S}_i(k)| < N(k) - \mu(k)\ ,$$

where $N(k)$ is the maximum value of the k^{th} frequency component over all the noise frames,

$$N(k) = \max_{i \in [I, I+M-1]} |S_i(k)|\ .$$

For those cases satisfying the above constraint, the frame-to-frame fluctuations are removed by setting the "noise-removed" frequency domain representation of the signal $\hat{\hat{S}}_i(k)$ to the "minimum" of $\hat{S}_j(k)$ for $j = i-1, i, i+1$. Specifically,

$$|\tilde{S}_i(k)| = \begin{cases} \min\limits_{j=i-1,i,i+1} |\hat{S}_j(k)| & \text{if } |\hat{S}_i(k)| < N(k) - \mu(k)\ , \\ |\hat{S}_i(k)| & \text{otherwise}\ . \end{cases}$$

The justification for this additional processing is simple. If $|\hat{S}_i(k)|$ is less than $N(k) - \mu(k)$ then that component is likely to be quite noisy so that setting it to the minimum over 3 frames should reduce the noise level. On the other hand, if it is not very noisy then, because of the relative stationarity of $s(n)$ between one frame and the next, the value of the k^{th} component should not have changed significantly between neighbouring frames so that taking the minimum introduces at worst a small error.

Finally, the noise reduced version of $s(n)$ is obtained by taking the inverse DFT of $\tilde{S}_i(k)$ and reconstructing the output signal by summing half-overlapped frames.

When $b(n)$ is used to generate $\mu(k)$ and $N(k)$, the application of spectral subtraction to $b(n)$ or $s(n)$ removes nearly all the noise. Figure 10 shows a spectrogram of the background noise signal $b(n)$ along with its noise reduced version. Similarly, Fig. 11 compares the original and noise reduced versions of the intercept signal $s_f(n)$. In the noise reduced spectrogram for the background noise only, Fig. 10, tone bursts are visible. These tone bursts are perceived as a somewhat annoying tonal noise if present in sufficient numbers. With acclimatization, listeners can to some extent block out this type of degradation to concentrate on the remaining signal. However overall, listeners find this tonal noise more distracting than the original broad spectrum noise.

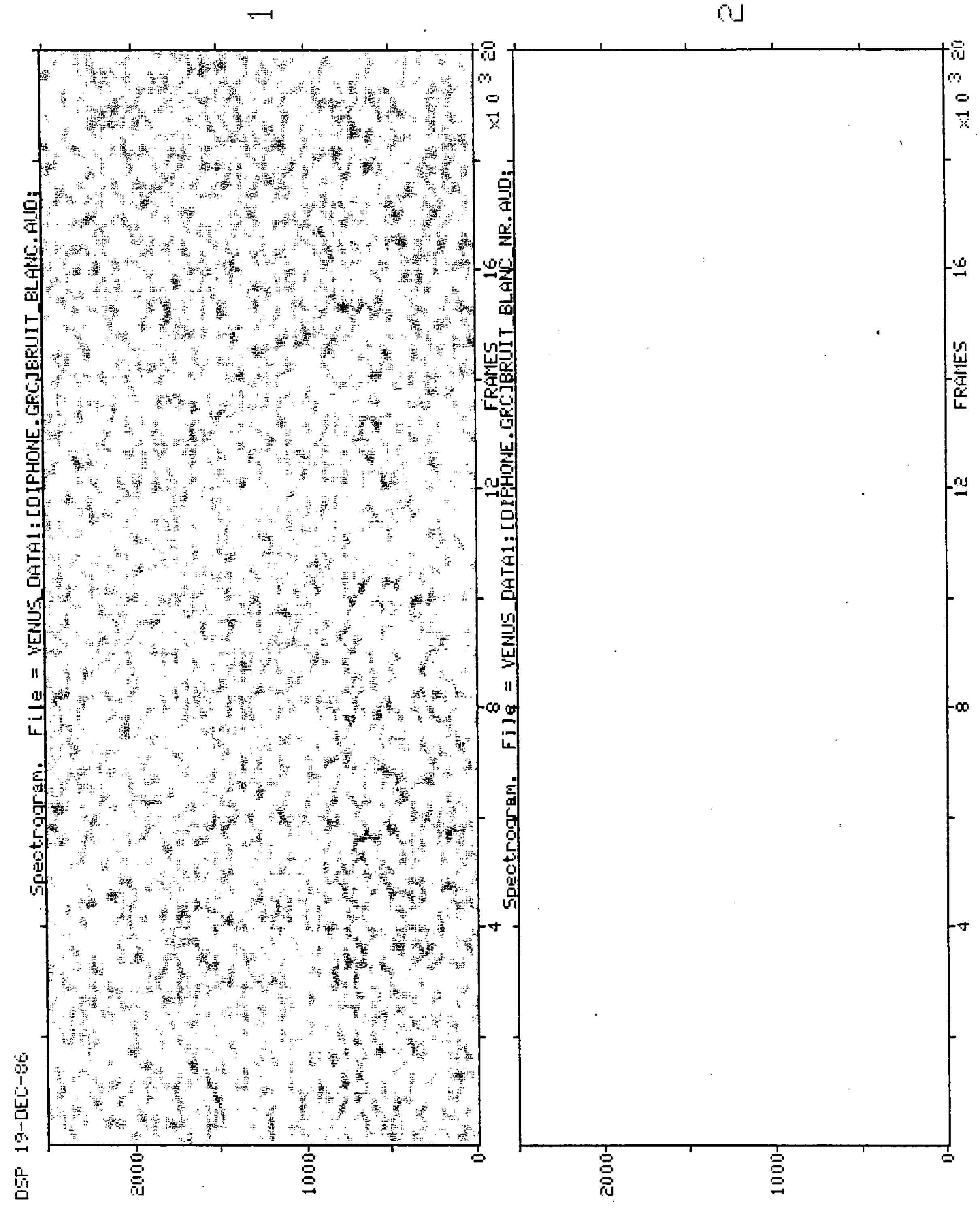


Fig. 10 Spectrograms (narrowband mode) of background noise alone showing the effect of noise removal (Method I).
Top: Original signal $b(n)$
Bottom: Noise reduced version

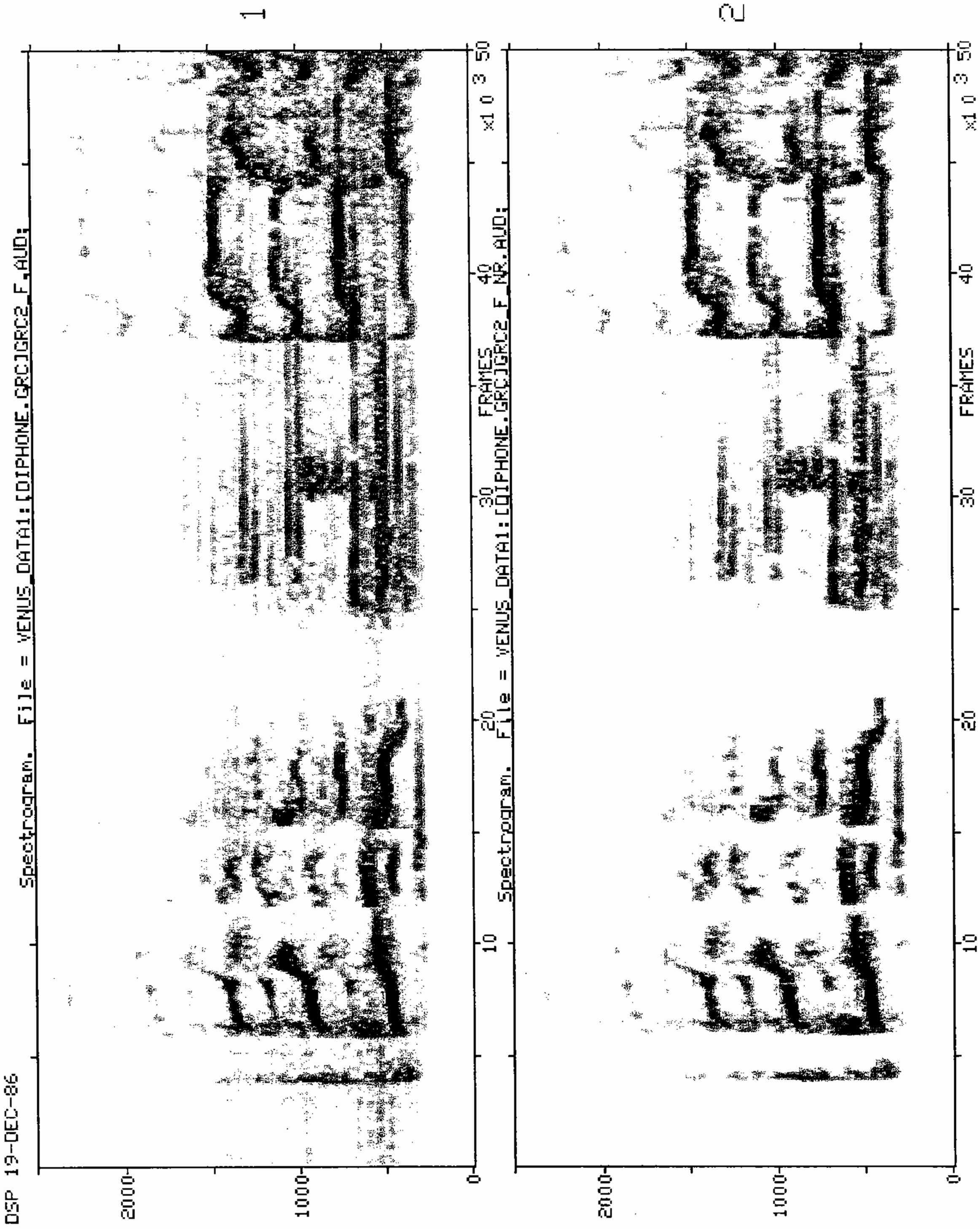


Fig. 11 Spectrograms (narrowband mode) of the intercept signal showing the effect of noise removal (Method I).
Top: Original signal $s_f(n)$
Bottom: Noise reduced version

4.2 Noise Spectral Difference (Method II)

The noise spectral difference process is an alternative means to remove the background noise. The first step of the process is to produce a mean spectrum of the noise contained in this prefiltered signal. This is accomplished by localizing a region in the signal $s_f(n)$ that contains only noise, without speech and music, and where the characteristics of the noise could be considered stationary. This last requirement was not directly satisfied since the bugging technique employed utilized an automatic gain control circuit, AGC, to optimize the recording level. Since we are interested in the portions with speech present, the beginning of the region with background noise alone, where the AGC applies a lower gain, was analyzed to produce a mean noise spectrum.

This mean noise spectrum spectrum was generated by taking the arithmetic mean of 30 amplitude spectra. Phase information is not used in generating the average. Each spectrum is derived from a 256-sample frame, weighted by a 256-point Hanning window, padded with 256 zeros and converted to the frequency domain by a 512-point DFT (FFT algorithm). Those 30 frames represent about 0.75 seconds of background noise. Figure 12 shows this resulting average spectrum. It has the highest energy in the band between 200 and 1500 Hz, with peaks at approximately 550, 900, and 1400 Hz.

The second step in noise reduction is to subtract, frame by frame, the mean noise spectrum from the spectrum of the intercept signal. For this, the intercept signal was segmented into frames of 256 samples, with an overlap 128 samples. The frame of data was then weighted by a 256 point Hanning window, padded with 256 zeros, and transformed to the frequency domain by a 512-point DFT. This produces the spectrum $S(k)$. The noise reduced amplitude spectrum is created as follows:

$$|\hat{S}(k)| = \begin{cases} \beta\mu(k) & \text{if } |S(k)| \leq \alpha\mu(k) \ , \\ |S(k)| - \alpha\mu(k) & \text{otherwise} \ , \end{cases}$$

where $\mu(k)$ is the average noise spectrum. The phase of the resulting spectrum is taken from the phase of the input signal.

The parameters α and β are chosen to reduce the tonal noise, that is produced by the subtraction process [4]. This noise consists of short periodic sounds that continuously change in frequency, thus producing something like a bubbling sound. It is due to the difference between the instantaneous noise spectrum and the mean noise spectrum. The effect of using α is to lower the spectral peaks in the instantaneous noise spectrum, while the effect of β is to rise the level of its valleys. Experimentation led to the following choices, $\alpha = 1.1$ and $\beta = 0.05$.

The resulting spectrum $\hat{S}(k)$ is transformed to the time domain by a 512-point inverse DFT and the final noise reduced signal is formed by overlapping and adding frames.

Figure 13 shows the effect of noise reduction on a 4.5 second segment of the intercept signal. The spectrogram showns that the noise subtraction washes out the very low energy spectral bands but does not seriously deteriorate the components with sufficient energy.

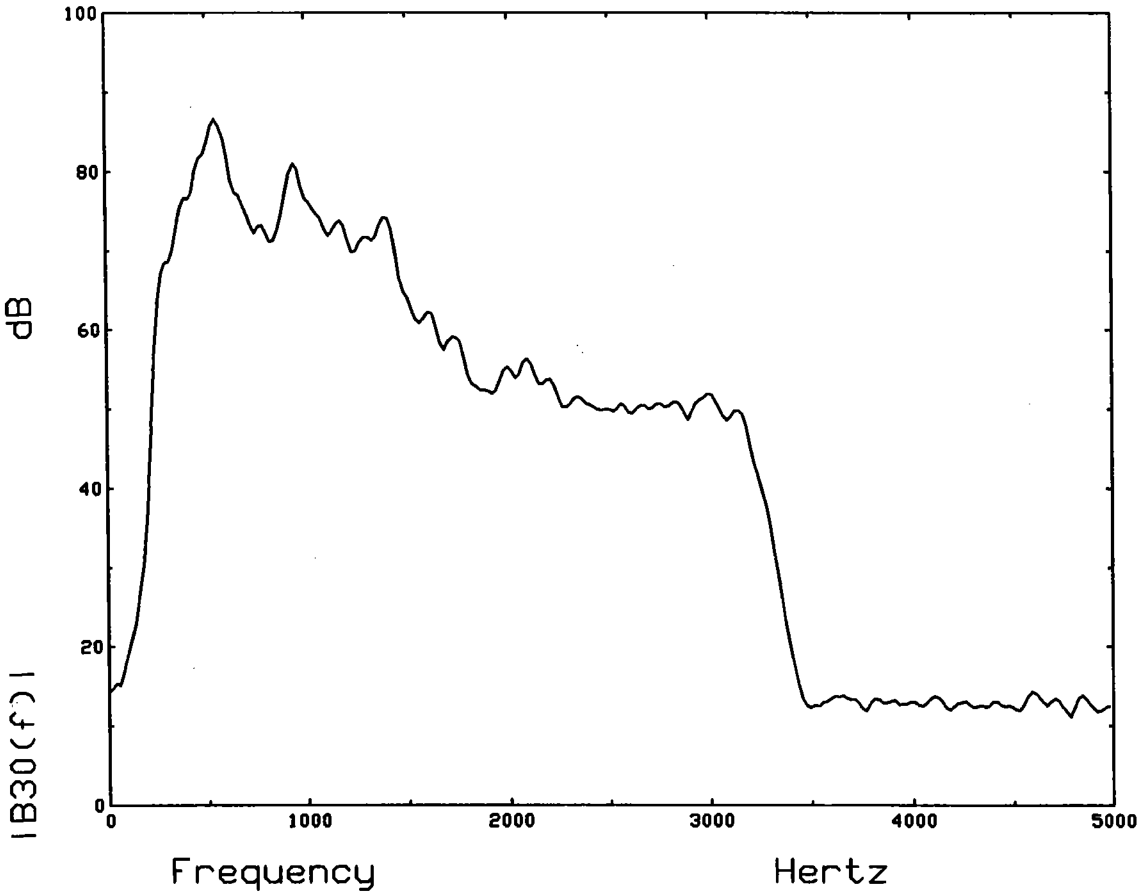


Fig. 12 Mean background noise spectrum

When listening to the resulting signals, one can hear a noise reduction but at the expense of an annoying tonal noise. This last degradation was found too disturbing at this early stage of the process to consider the resulting signal an adequate input signal for the further processing stages.

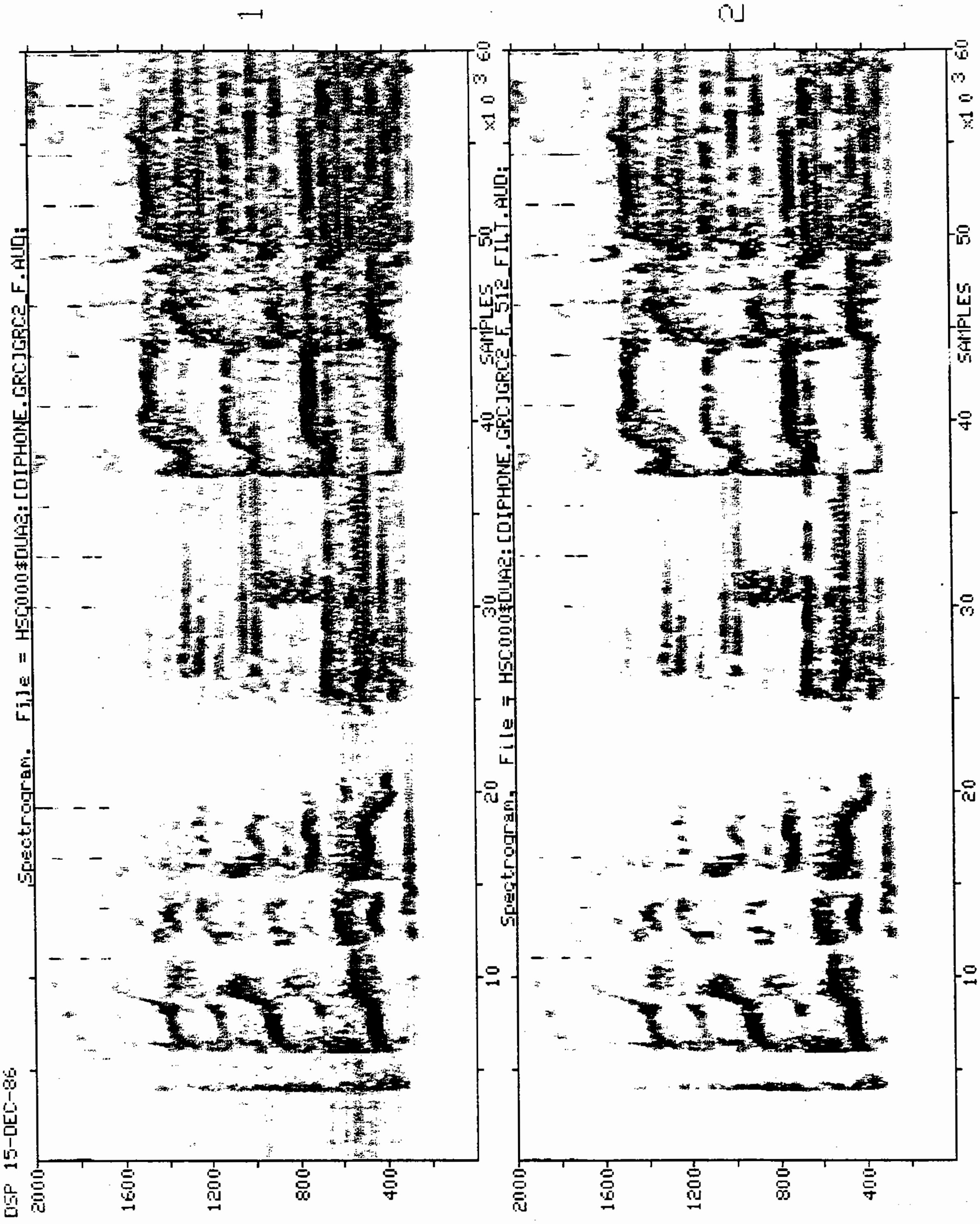


Fig. 13 Spectrograms (narrowband mode) of the intercept signal showing the effect of noise removal (Method II).
Top: Intercept signal before noise reduction
Bottom: Intercept signal after noise reduction

5. Noise Whitening

The process of noise whitening is concerned with the reduction of the perceived effect of the background noise and, as a byproduct, the equalization of the signal. Assuming that the background noise resulted from a process which generates a flat (white) noise spectrum, the spectral colouring that is present in the intercept recording shows the effects of the room acoustics and the recording system. By inverse filtering the signal with the measured spectrum of the noise, the noise spectrum will be whitened. This should render the noise less disturbing, and the overall signal will be equalized.

This first step in the technique is to build an estimate of the mean inverse spectrum of the background noise (see Section 4). To evaluate the mean noise spectrum, 20 frames of the signal $b_f(n)$, each of of 256 samples were inverted and averaged. This produces the average inverse noise spectrum $N_W(k)$, which is shown in Fig. 14.

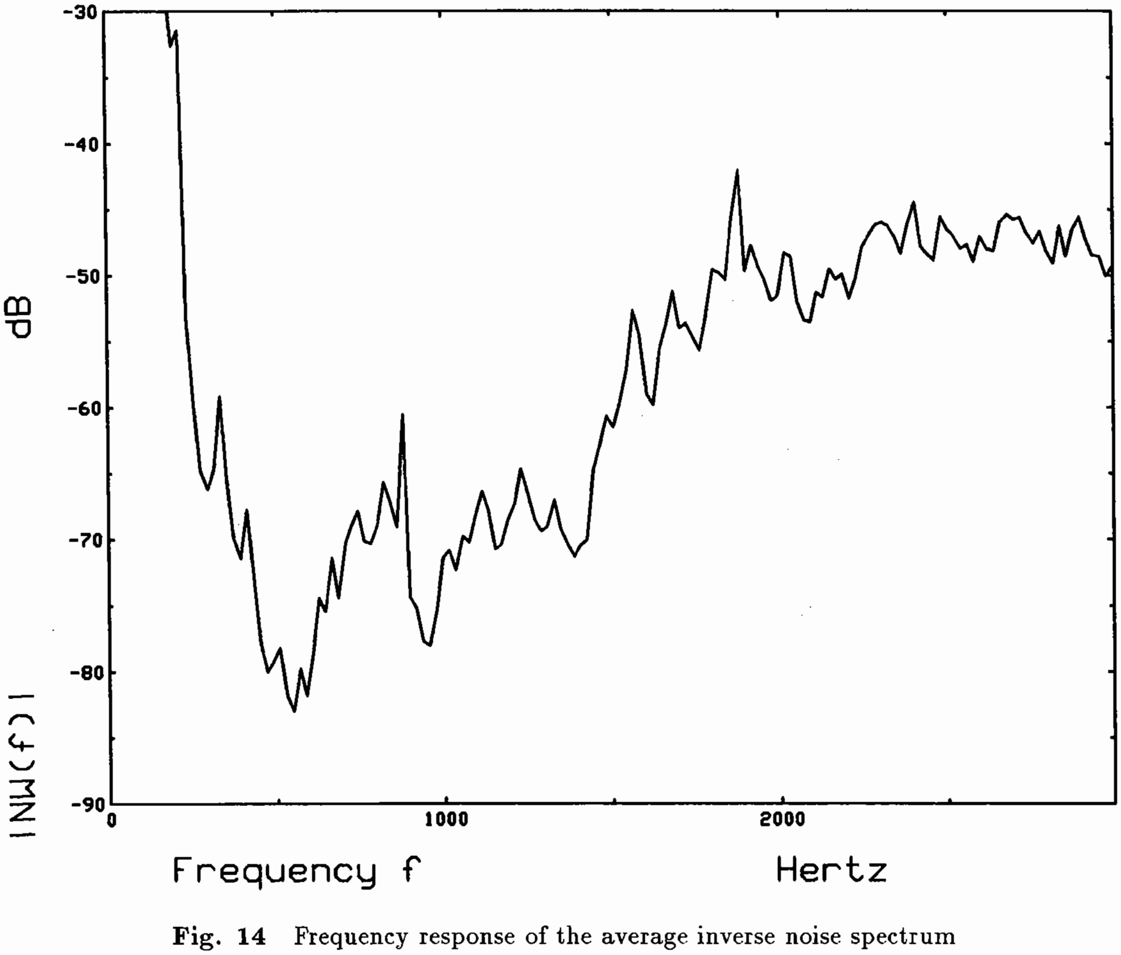


Fig. 14 Frequency response of the average inverse noise spectrum

The second step of the process consists of the modification of the inverse noise spectrum. First, the range of interest is limited to the band from 280 to 2800 Hz. Outside this band, the filter response was set to the minimum in-band value. Second, the in-band response was smoothed over a 5 band window,

$$|N'_W(k)| = \frac{1}{5}\sum_{i=-2}^{2} |N_W(k+i)| \; .$$

As a last step, the passband frequency response was normalized to bring the geometric mean of the in-band amplitudes to unity. This step is used to preserve the in-band gain. The spectrum of this modified filter, $N''_W(k)$, is shown in Fig. 15.

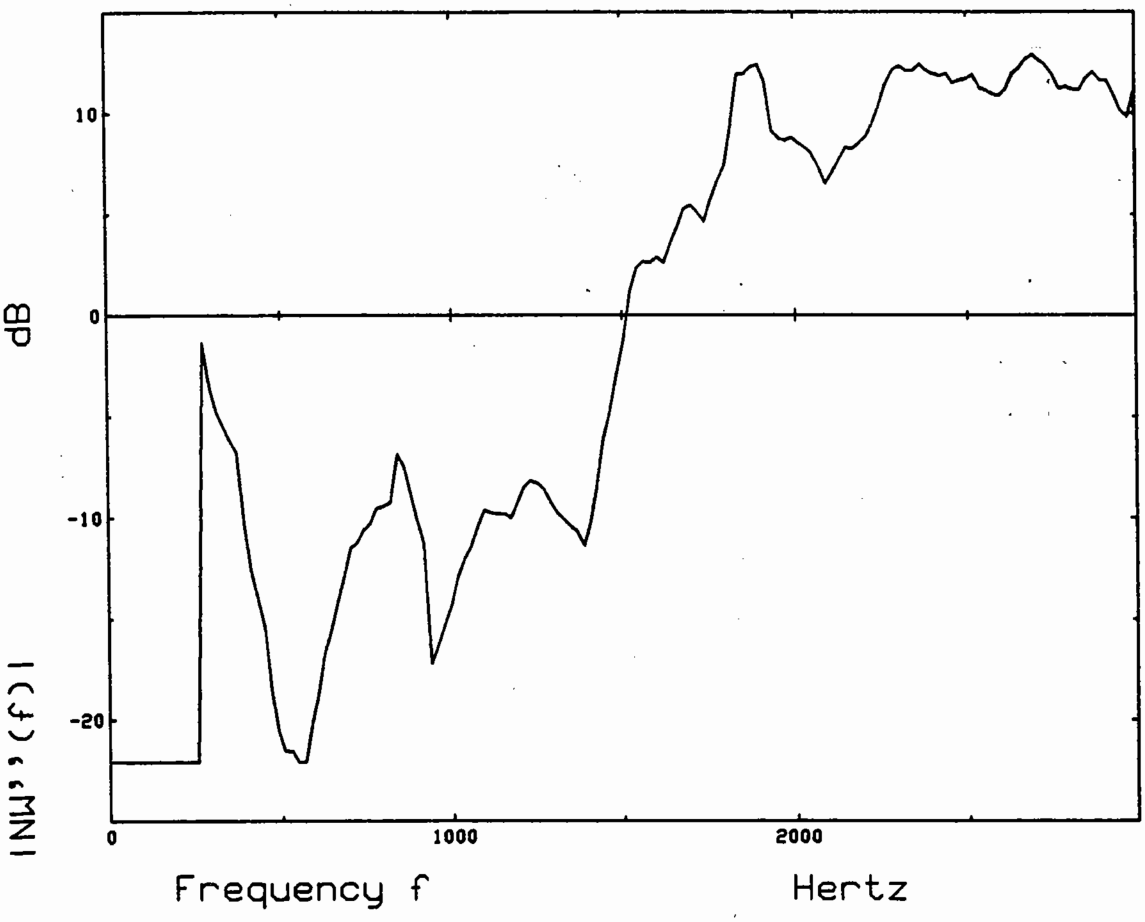


Fig. 15 Frequency response of the modified inverse noise spectrum

The whitening process itself consists of filtering the intercept signal using the inverse noise spectrum. This processing is carried out in the frequency domain by multiplying the spectrum of the intercept signal by the inverse noise spectrum. The spectrum of the intercept signal is formed by segmenting the signal into frames of 256 samples overlapped by 128 samples, multiplying the signal by a Hanning window, padding the result with 256 zeros, and transforming into the frequency domain using a 512-point DFT. The final time domain signal is then formed by an inverse DFT with the appropriate overlapping and adding of frames.

Figure 16 shows a spectrogram of the noise whitened intercept signal. The filtering accentuates the portion above 1500 Hz and deemphasizes the lower spectrum. The overall energy distribution is much more uniform after whitening. When listening to the resulting signal, one can discern the difference in the quality of the signal. Some of the resonant qualities of the original are missing, and moreover the speech component has more “presence”. The boosted high-frequency noise is not annoying; it is barely discernible. The noise whitening step is in essence an equalization operation and as such can be used in conjunction with music cancellation procedures.

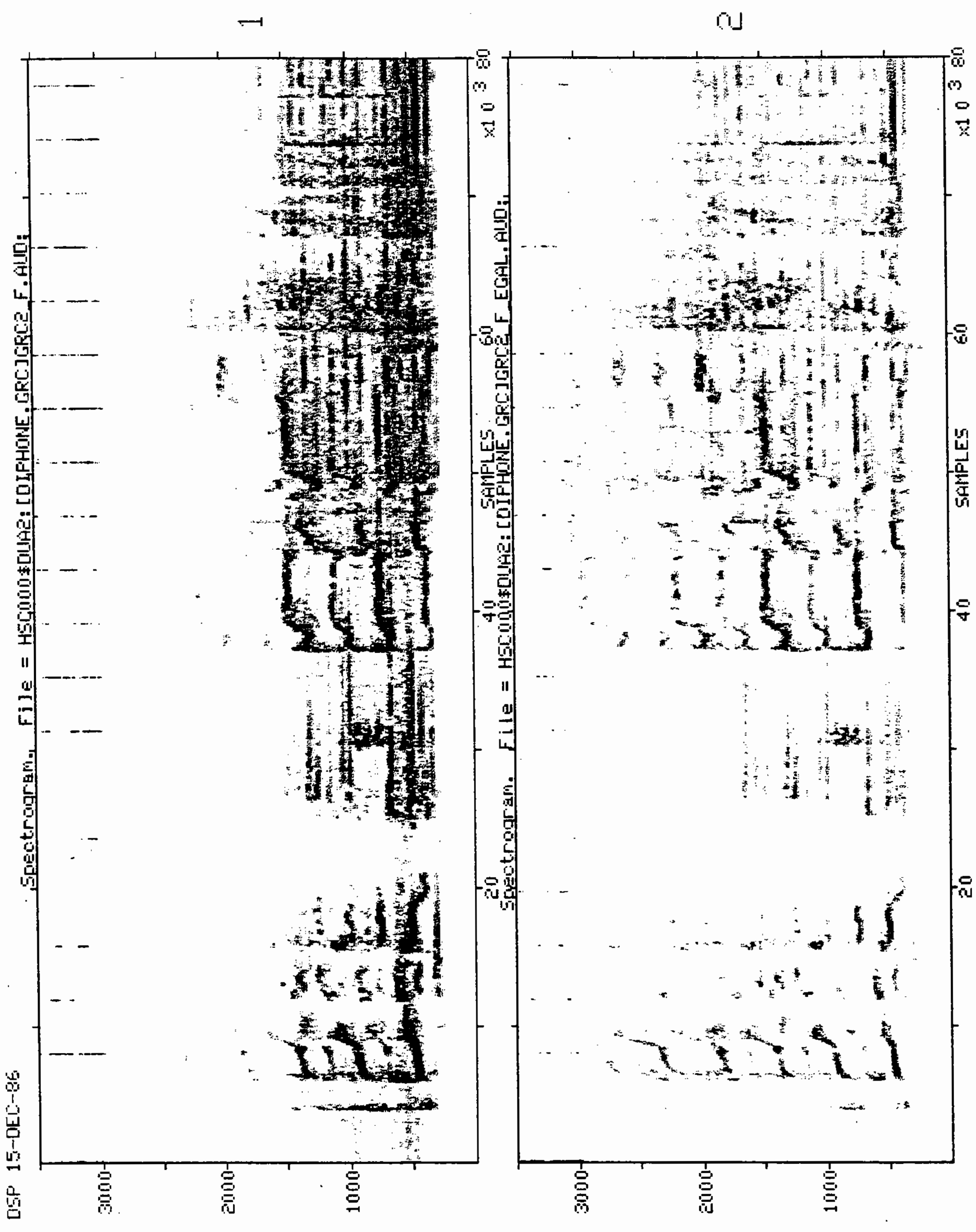


Fig. 16 Spectrograms (narrowband mode) for the intercept signal showing the effect of noise whitening.
Top: Intercept signal before noise whitening
Bottom: Intercept signal after noise whitening

6. Signal Equalization

This technique deals with the equalization of the prefiltered and noise whitened intercept signal toward the prefiltered, and time-aligned reference music signal $r_f^a(n)$. The equalization is desirable not mainly as an end in itself but rather as a preparatory stage before the application of the spectral subtraction process for music cancellation. The aim of such an equalization is to bring the spectrum of the musical component of the intercept signal as close as possible to the spectrum of the reference music signal.

In order to design the equalization filter it is necessary to localize a region in the intercept signal that contains broadband music with no speech. One such segment of duration 0.125 seconds, is composed of a harmonically rich guitar chord.

The filter frequency response was estimated by taking the mean, over 5 frames, of the ratio of the spectra of the filtered, noise-whitened intercept signal $s_{f,nw}(n)$ and of the filtered and time-aligned reference music signal $r_f^a(n)$. For this, the signals were segmented into 256 samples frames, weighted by a 256-point Hanning window, padded with 256 zeros, and transformed into the frequency domain by using a 512-point DFT. The spectral ratio for a single frame is given by

$$H(k) = \frac{|R_f^a(k)|}{|S_{f,nw}(k)|} \qquad \text{for } k = 0 ldots, 511 \ .$$

This spectral ratio is then averaged over 5 frames to form $H'(k)$. The equalization filter frequency response is shown in Fig. 17. We observe that in the frequency band of interest, 280–2800 Hz, the amplitudes $|H'(k)|$ stay approximately within the region delimited by 0 and 20 dB.

The second step of the process consists of the modifying of the equalizer response. First, the range of interest is limited to the band from 280 to 2800 Hz. Outside this band, the filter response was set to the minimum in-band value. Second, the in-band response was smoothed over a 17 band window,

$$|H''(k)| = \frac{1}{17} \sum_{i=-8}^{8} |H'(k+i)| \ .$$

This wide smoothing interval was chosen to average out spectral components which depend on the exact nature of the music in the segment chosen. The frequency response of this modified filter, $H''(k)$, is shown in Fig. 18. In the frequency band between 280 and 900 Hz, the response shows a lowpass frequency rolloff of about 7.4 dB per octave. This lowpass rolloff is known to have some de-reverberating effects.

The equalization process itself consists of the use of the equalization filter to filter the intercept signal. This processing is carried out in the frequency domain by multiplying the spectrum of the intercept signal by the equalization filter response. The spectrum of the intercept signal is formed by segmenting the signal into frames of 256 samples overlapped by 128 samples, multiplying the signal by a Hanning window, padded the result with 256 zeros, and transforming into the frequency

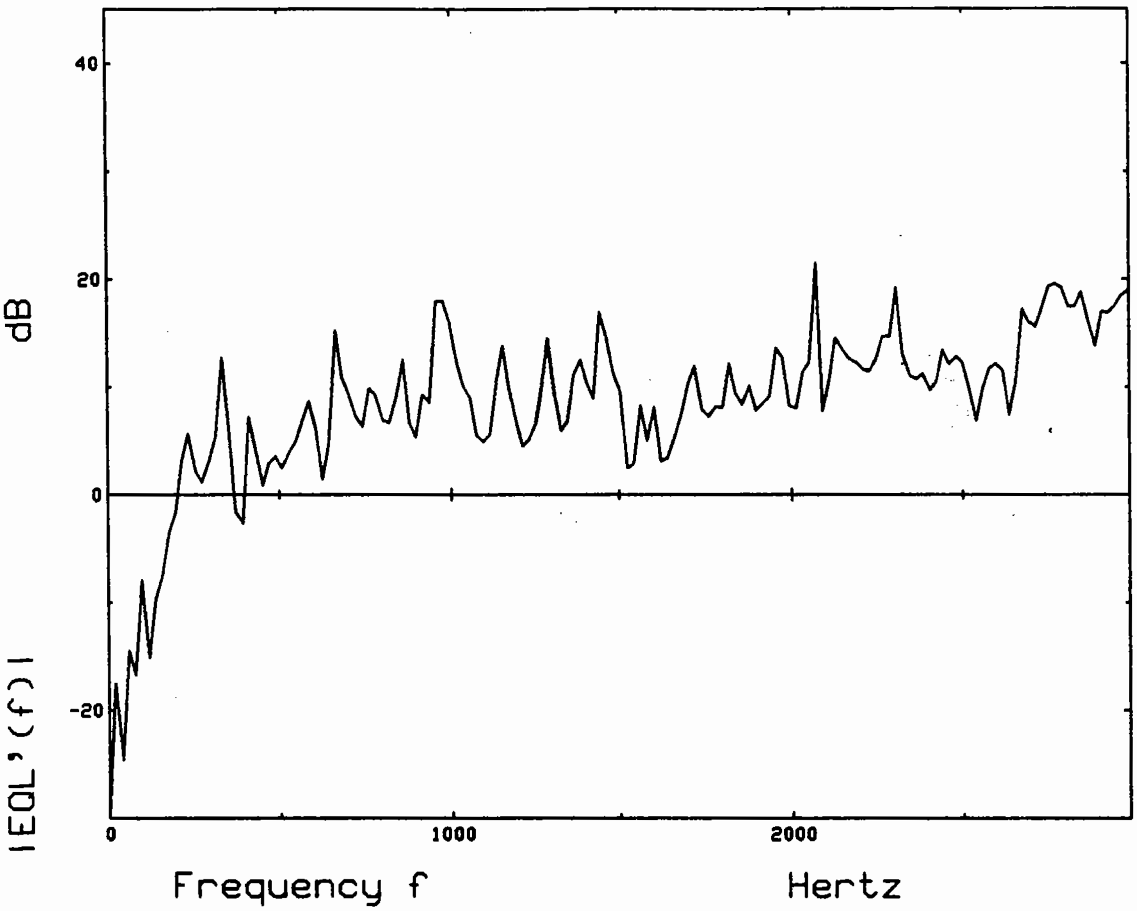


Fig. 17 Frequency response of the equalization filter $H'(k)$.

domain using a 512-point DFT. The final time domain signal is then formed by an inverse DFT with the appropriate overlapping and adding of frames. Note that equalization affects the amplitude alone. The original phase of the intercept signal is preserved.

6.1 Effect of Equalization

Figure 19 shows the spectra of a 512 sample frame of signal that contains music with no speech. As can be seen, the music equalization increases the similarity of the spectra of the intercept signal to the reference music signal. Discrepancies still remain. Some of these can be due to the averaging used to construct the equalizer, or due to the lack of representation of certain frequency components in the segment used to design the equalizer.

Comparing the equalized intercept signal to the original intercept signal, one can hear an increase in intensity. This increase in level can also be seen in Fig. 19. In addition one can hear a high frequency breathing sound and discern that the amplitude of the music components has risen relative

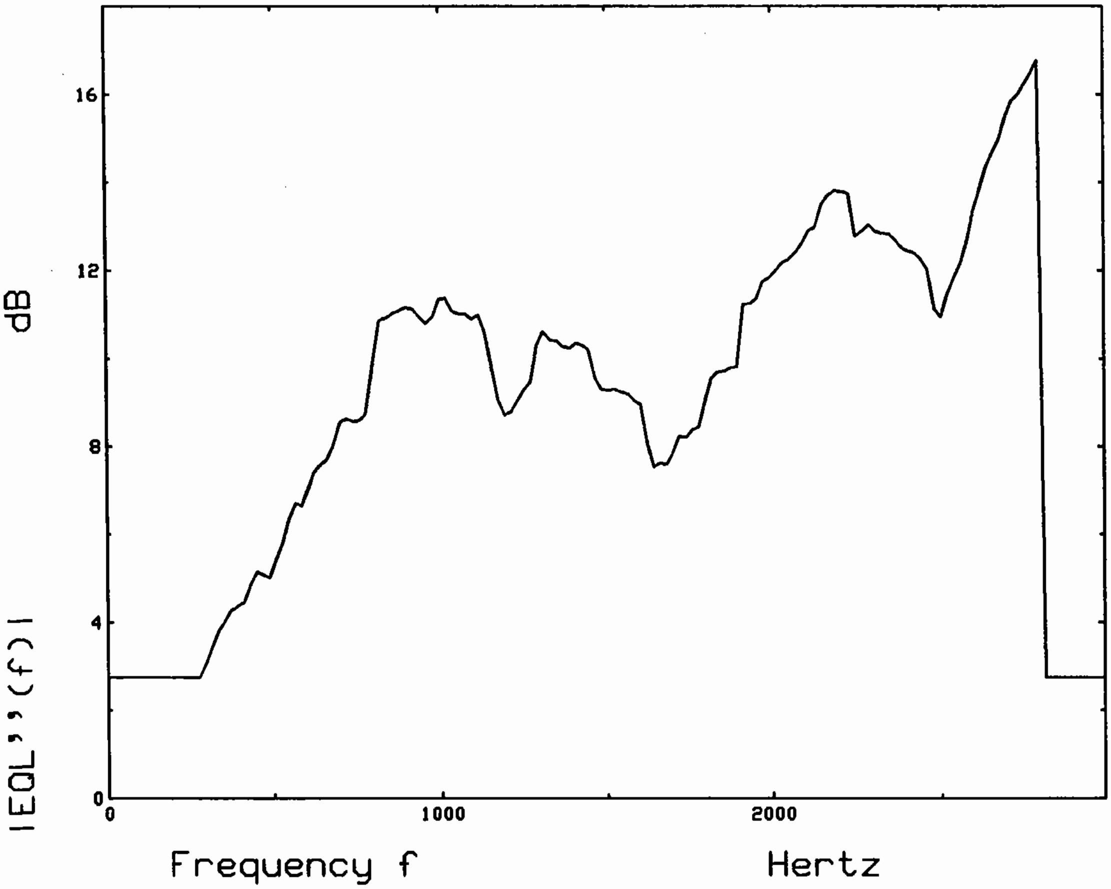


Fig. 18 Frequency response of the modified equalization filter $H''(k)$.

to the speech components. A spectrogram of the resulting equalized intercept signal is shown in Fig. 34.

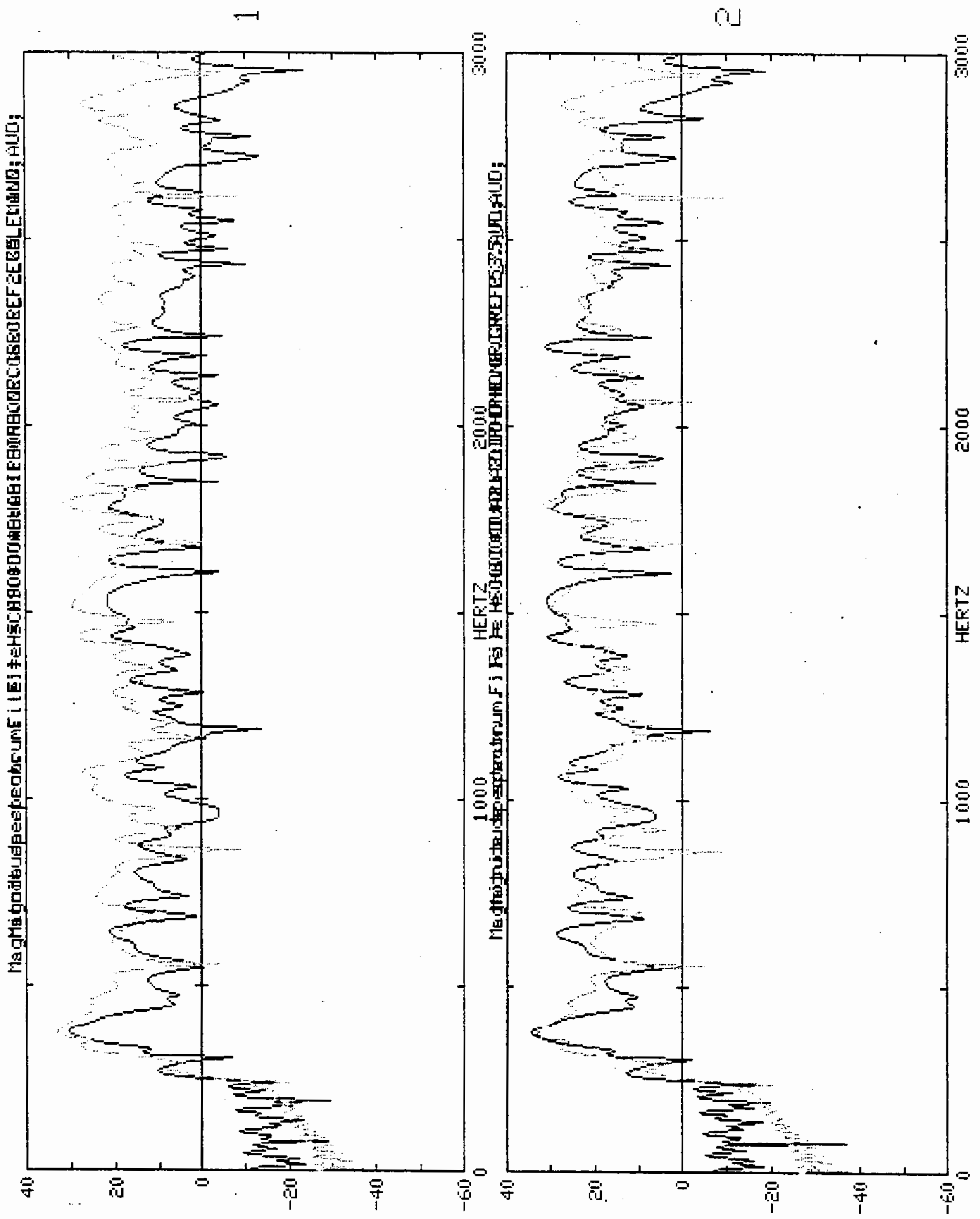


Fig. 19 Spectra of the intercept signal and reference music signal showing the effect of equalization. Dark traces are the intercept signal and light traces are the reference music signal.
Top: Before equalization
Bottom: After equalization

7. Time Domain Cancellation

This section describes the use of adaptive filtering techniques to cancel the music component of the intercept signal. For the purposes of designing the adaptive filter, we can consider two sources of interest: the reference music signal $r_f^a(n)$ which has been filtered and time aligned, and the intercept signal $s_f(n)$ which has been filtered. For the sake of simplified notation, we drop the subscripts indicating the digital filtering operations that were carried out as a preliminary step and the superscript indicating time alignment. Note that this time alignment is relatively coarse and that due to the uncertainty in the anchor point alignment and the variations in speed over a segment, the true time alignment is subject to some variation. If this variation is small enough and occurs slowly enough, the dynamics of the adaptation process will compensate for it to some degree.

The intercept signal contains a music component which suffers from reverberant effects. In addition, in the context of music cancellation both the speech components and the background noise component are interference. The reverberant effects are modelled as a FIR filter acting on the clean music signal. The adaptive filter will try to track the coefficients of this filter and to filter the reference music signal to produce a reverberated signal which matches the music component of the intercept signal. This can then be subtracted form the intercept signal to reduce the level of the music.

The effective reverberation extends over a relatively long time period. The propagation speed of sound corresponds to about 1 ms per foot. The path difference between a direct path and a reflected path can correspond to a large number of ms, and furthermore multiple reflections can extend the reverberant effects to a significant fraction of a second. This means that at a sampling rate of 10,000 samples/second, effective filter lengths must correspond to hundreds, if not thousands of samples. For a filter with M coefficients, the time averages used to update the filter coefficients must extend over intervals significantly larger than M samples. If we violate this condition the filter has enough degrees of freedom to synthesize the other components of the intercept signal and cancel them also. The factor which works agains the use of long time averages is the loss in synchrony between the reference music signal and the music component of the intercept signal due to time alignment changes.

With the above comments as background, we consider two time domain approaches to adaptive signal cancellation [5]. The cancellation system is shown in Fig. 20. Both approaches attempt to minimize the energy of the output signal $y(n)$. If we assume that that the reference music signal is uncorrelated with the speech components and the noise component of the intercept signal, then when the adaptive filter represents the reverberation effects of the room, cancellation of the music components is possible without affecting the speech or noise components.

In the first method used, a block based adaptation of the filter coefficients was used. In the second approach, a stochastic gradient (LMS) technique was used. In this case, the learning process

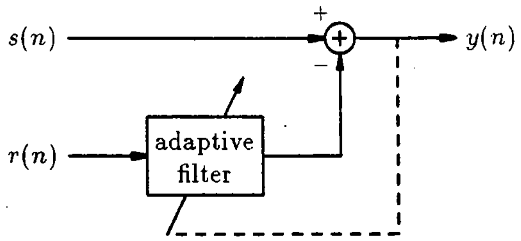


Fig. 20 Basic model for adaptive noise cancelling

was performed on a sample-by-sample basis.

7.1 Block Least-Squares Methods

7.1.1 Conventional Least-Squares Solution

In block least-squares methods one treats successive short time frames over which input data is assumed to be stationary. The covariance method leads to an exact solution to the problem of minimizing the error, in the least-squares sense, over a block of finite length N. It gives the set of coefficients $\{h_0, h_1, ..., h_{M-1}\}$ such that the quantity,

$$\varepsilon = \sum_{n=0}^{N-1} \left[s(n) - \sum_{i=0}^{M-1} h_i r(n-i)\right]^2 ,$$

is minimized. Those coefficients are found as the solution of the following set of M linear equations

$$\sum_{i=0}^{M-1} h_i \sum_{n=1}^{N-1} r(n-i)r(n-j), = \sum_{n=1}^{L} s(n)r(n-j) \quad \text{for } j = 0, 1, \ldots, M-1.$$

The effect of the filter is to remove correlations between the reference signal and the intercept signal. As long as the speech component of the intercept signal is uncorrelated with the reference music signal, it will not be affected. In the block based method, the correlations are in effect estimated by using time averages. If the frame length is too short, residual correlation may be present between the speech component and the reference music signal. The issue of choosing an appropriate frame length will be considered later.

In addition to the fact that this method is computationally intensive, the major limitation of block based algorithms is a maximum number of coefficients that can be used. Large errors may occur when solving the above set of equations with finite precision arithmetic for a large number of coefficients. In order to avoid numerical difficulties, we limit ourselves to solving for at most 30 coefficients.

Before attempting to apply this method to the actual intercept signal, the technique was evaluated on test signals. Such an approach enables us to evaluate some of the factors which affect the performance. In a first experiment, the reference music signal $r(n)$ was added to a known speech signal. For this simple case, the interfering music has not been filtered to simulate reverberant effects. The block based algorithm (12 coefficients) was applied to this composite signal (S+M) in order to suppress the music.

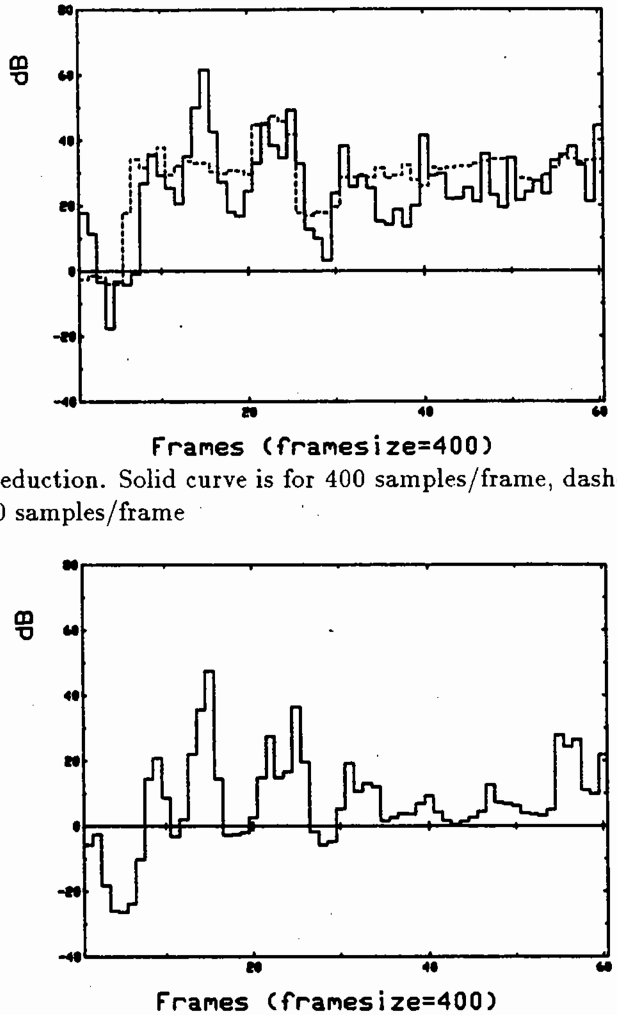
The performance of the adaptive filter is measured in terms of music reduction. We compute the ratio of the music power taken over a segment of the reference signal, to the remaining music power over the same segment in the processed signal. The experiment was run using several values for the frame size. Figure 21 shows some of the results obtained for two particular frame sizes: 400 and 2000 samples. Note that the performance varies significantly over short periods of time depending on whether music or speech dominates in the input signal, more so for the case of the shorter frame size. This causes abrupt changes in noise level throughout the processed signal which affect the ability to comprehend the speech. For larger frame sizes, the contribution of the speech component in the computation of the cross-covariance function tends to be averaged out. Therefore, the cancellation is improved significantly relative to when short frames are used, in those parts of the signal where speech is present. Based on this experiment, the lower bound for the frame size in order to avoid these problems is about 1000 samples. The average music reduction as measured over the entire signal for two different values of the frame size is given in Table 1. For this simple test signal, the cancellation process is very effective.

Frame size (samples)	Average Music Reduction
400	25.0 dB
2000	28.3 dB

Table 1 Average noise reduction for test signal S+M

Somewhat more realistic conditions can be created by passing the test signal S+M through a filter to produce the test signal SF+MF. A 16 pole IIR filter was used. This filter is one that was used to model the spectrum of the background noise alone in the intercept recording. The adaptive filter with 12 coefficients was used to compensate for this (infinite) filter response. As expected the average performance of the adaptive filter is worse than for the previous test signal (see the single pass entry in Table 2). This smaller music reduction is due to the fact that the adaptive filter is too short to effectively model the filter used on the composite signal. The reduction in music level is 11.8 dB for a frame size of 2000 samples.

The experiments just described are appropriate for slowly changing environments (stationary for the duration of the frame) with the music component having undergone filtering with a filter with a short impulse response. For the intercept recording, the environment is not stationary since the



(a) Music reduction. Solid curve is for 400 samples/frame, dashed curve is for 2000 samples/frame

(b) Ratio of music power to speech power

Fig. 21 Results for block based cancellation for test signal S+M

Single Pass		Multiple Pass	
Number of taps	Reduction	Pass number	Reduction
12	11.8 dB	1	11.8 dB
24	18.4 dB	2	14.7 dB
–	–	3	15.2 dB
–	–	4	15.6 dB

Table 2 Average music reduction for test signal SF+MF. For the multiple pass case 12 taps are used for each pass. The frame size is 2000 samples.

reference music signal is not exactly synchronous with the intercept signal and since the equivalent filter response changes with time. The performance of the adaptive filter depends on an appropriate choice for the frame size and on the number of taps used. There are conflicting restrictions on the frame size. The frame size must be long to average out the effect of the speech signal, yet short to allow the signal to be considered to be stationary within that frame.

7.1.2 Multiple Pass Least Squares Method

To avoid numerical problems, the block based schemes must be restricted to solving for a small number of coefficients. This is a major limitation for the present application where a large number of filter coefficients are required to adequately model the reverberation effects in the intercept recording. We describe here another approach to avoid the constraint on the number of coefficients. The scheme applies a sequence of short filters with different offsets. The scheme involves processing the reference signal in multiple passes.

A first step consists of solving for the coefficients $\{h_0, h_1, \ldots, h_{M-1}\}$ of the first group of coefficients. The resulting filter is used to produce an intermediate output signal $y_1(n)$,

$$y_1(n) = s(n) - \sum_{i=0}^{M-1} h_i r(n-i).$$

In the second step, another set of M coefficients is derived in order to minimize the quantity

$$\varepsilon_2 = \sum_{n=1}^{L} \left[y_1(n) - \sum_{i=M}^{2M-1} h_i r(n-i)\right]^2 .$$

Note that these coefficients process the input signal $r(n)$ shifted by M samples. This process is carried out sequentially for additional groups of coefficients. This multiple pass algorithm is sequential in nature and hence gives a different set of coefficients from that given by a single pass scheme with the same total number of coefficients. The interaction between groups of the coefficients is ignored in the multiple pass scheme.

One property of a least squares solution is that the energy in the output signal will be equal to the difference between the energy of the composite signal and the energy of the output of the adaptive filter. This is a consequence of the fact that the resulting output signal is uncorrelated with the reference signal at time lags corresponding to the filter taps. If the composite signal is already uncorrelated with the reference signal, the filter coefficients must be zero in order to produce a zero output signal from the adaptive filter. In short, in a multiple pass scheme, blocks of coefficients will contribute to the output only if the corresponding reference signal components are correlated with the composite signal. The multiple pass scheme will produce a monotonically decreasing overall error,

$$\varepsilon_1 \geq \varepsilon_2 \geq \cdots \geq \varepsilon_k .$$

The multiple pass method was applied to the test signal SF+MF and the results are given in Table 2. Note the suboptimal nature of the sequential multiple pass method. Two passes with 12 coefficients determined in each pass achieves much less reduction than a single pass with 24 coefficients.

7.1.3 Effect of Time Alignment

Differences in time scales between the reference and the signal to be treated affect the process in two ways. First, because the music drifts slightly within a frame as compared to the reference, the computed correlation terms as smeared. This will decrease the amount of cancellation that can be achieved. As a consequence of this effect, the music tends to be better matched in the middle of the frame with the music reduction decreasing towards both ends of the frame. This effect may be very annoying from a subjective point of view. Experiments with test sequences show that with a linear drift of the order of 0.2% those effects are still significant even for short frame size of 400 samples. The second effect of asynchrony is the problem of the time alignment exceeding the span of the adaptive filter. With the music drifting linearly compared to the reference, the offset between the two signals increases from frame to frame until it becomes greater than the number of taps. Beyond this point the adaptive filter becomes much less effective. However, the filter does not become completely ineffective. Even with a misalignment, quasi-periodic components can be still be cancelled if the correlation extends sufficiently to lags which will encompass the filter coefficients.

The effects of alignment can be explored by looking at the performance of a 12 tap adaptive filter as a function of the offset of the reference signal relative to the composite signal. Results for a segment in the composite signal with music alone is shown in Fig. 22. In this case the two signals were perfectly aligned and one can observe a sharp peak in the music reduction with a zero offset value. The shape of the curve reflects the periodic structure of the music signal in the portion studied (the period is about 31 samples). With the frame sizes used, the loss in performance when the reference is shifted one period from the optimal position remains small. Note that the peak music reduction remains essentially unchanged when going from 400 to 1000 samples for the frame size.

The music reduction obtained on the corresponding segment of the intercept signal $s(n)$ is shown on Fig. 23. No speech is present in this segment. The fine structure of the function as obtained using a 400 sample frame, is as expected. However it was found that both the position and the height of the maximum vary greatly from frame to frame and when the frame size is changed. One can also note that the level of music reduction is significantly lower for the longer frame size. This can partially be attributed to the loss of synchrony over the longer frame.

The observed behaviour of the music reduction for the intercept signal suggests that the frame size must be carefully chosen for best results. One can hope that the drift of the optimal offset can be compensated for by increasing the number of passes to give a large overall filter length. The multiple pass method was used to process the full test signal SF+MF. Table 3 shows that the method is very sensitive to the choice of the offset and that the multiple passes cannot fully compensate for the misalignment. A single pass with correct alignment is much better than multiple passes with even a small misalignment.

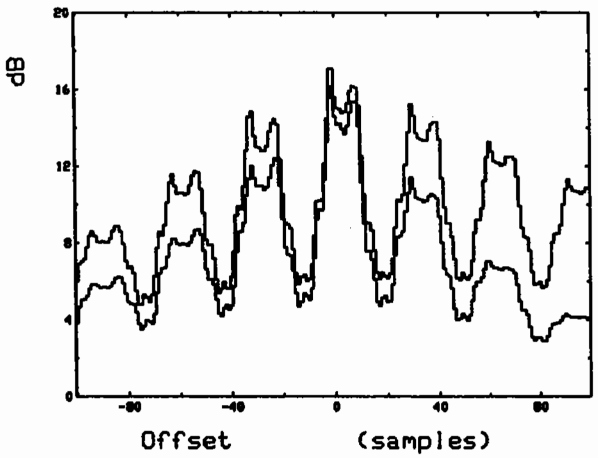


Fig. 22 Music reduction on a segment of test signal SF+MF for a 12 tap filter as a function of the time offset. The upper and lower curves correspond to frame sizes of 400 and 2000 samples respectively.

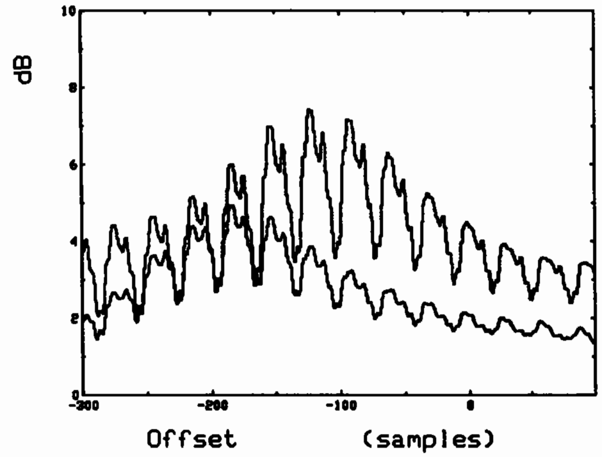


Fig. 23 Music reduction on a segment of the intercept signal for a 12 tap filter as a function of the time offset. The upper and lower curves correspond to frame sizes of 400 and 2000 samples respectively.

7.1.4 Multiple Pass Processing of the Intercept Signal

The multiple pass method was applied to the intercept signal for different offset values. Each pass added 20 filter coefficients. The frame size was 400 samples. The performance of the method was found to be essentially independent of the initial offset value. More music was suppressed at each step up to the third step after which the music power tended to stabilize. But the overall performance remained lower than that of the gradient descent method described in the next section. The varying nature of the signal together with the limitations inherent to our method are among

Step Number	Offset 0	Offset +10	Offset +20
1	13.4 dB	11.4 dB	4.1 dB
2	15.6 dB	13.0 dB	6.9 dB
3	15.8 dB	13.3 dB	8.3 dB
4	–	–	8.6 dB

Table 3 Effect of misalignment for the multiple pass scheme for test signal SF+MF (frame size 2000 samples, 16 coefficients per pass)

the factors which limited the performance of the multiple pass method.

7.2 LMS Adaptive Canceller

The LMS algorithm which uses a gradient descent technique is another approach to the realization of adaptive cancellers. For every pair of input samples, one sample from the distorted signal and one sample from the reference signal, the gradient descent technique updates the filter tap coefficients. In the block algorithm approach, the filter is updated on a block basis. Although the filter realized using the gradient descent is not truly optimal in terms of the minimization of output signal energy, the learning process can be carried out smoothly and continuously. Also, the filter designed using this approach can have a large number of coefficients, whereas a practical implementation of a block based algorithm is limited in the size of the filter that can be used.

In addition, a simplified non-linear model which takes into account low order non-linearities was employed. Let $\mathcal{L}$ and $\mathcal{N}$ denote the time varying responses of the linear filter and the non-linear compensator where

$$\mathcal{L}: \quad y(n) = \sum_{i=0}^{M-1} a_i x(n-i) \ ,$$

$$\mathcal{N}: \quad v(n) = \sum_{j=1}^{3} b_j [u(n)]^j \ .$$

The linear filter uses M coefficients, and the non-linear compensator includes non-linearities up to a cubic term. The non-linear compensator is an instantaneous (memoryless) non-linearity.

The intent of the gradient update scheme is to minimize the energy in the output signal $y(n)$. Let the mean-square value of the output be

$$\varepsilon = E[y^2(n)] \ .$$

The coefficients are updated so as to move in the negative gradient direction. For the non-linear compensator,

$$b_j' = b_j - \beta_j \frac{\partial \varepsilon}{\partial b_j} \ ,$$

where β_j is the step size used for coefficient b_j. Similarly for the linear filter, the coefficients are updated so as to decrease the error at each step,

$$a_i' = a_i - \mu_i \frac{\partial \varepsilon}{\partial a_i} ,$$

where μ_i is the step size used for coefficient a_i.

7.2.1 Canceller Configurations

The linear filter and the non-linear compensator can be combined in a number of different ways. Five configurations were tested as shown in Fig. 24. In these block diagrams, $s(n)$ is the intercept signal and $r(n)$ is the reference music signal.

For the purposes of developing a practical LMS algorithm, the expectation operator in the gradient update scheme is omitted, and the instantaneous value of the squared error is used as an estimate of the mean-square error. The details differ depending on which configuration is being used. Consider configuration A. The update for the non-linear compensator becomes

$$b_j' = b_j + 2\beta_j y(n)[v(n)]^j ,$$

where $v(n)$ is the output of the linear filter,

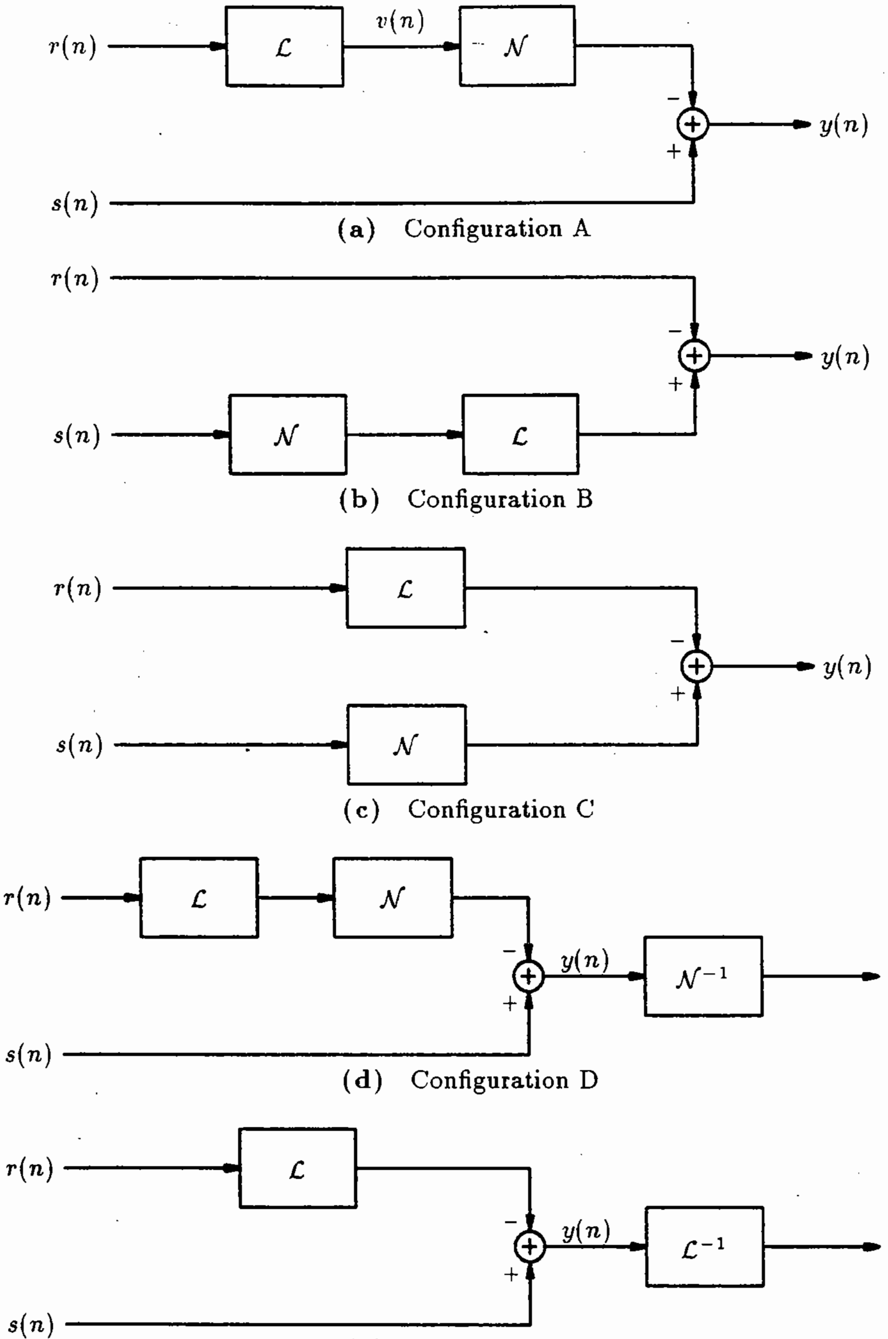
$$v(n) = \sum_{i=0}^{M-1} a_i r(n-i) .$$

For the linear filter, the update becomes

$$\begin{aligned} a_i' &= a_i - \mu_i \frac{\partial \varepsilon}{\partial v(n)} \frac{\partial v(n)}{\partial a_i} \\ &= a_i + 2\mu_i r(n-i) y(n) \sum_{j=1}^{3} j b_j [v(n)]^{j-1} \end{aligned}$$

Note that the updates for the linear filter and the non-linear compensator are interrelated. In the case of configuration A, we determine the coefficients of the non-linear compensator first. Using the updated values of the non-linear compensator, the coefficients for the linear filter are updated.

The different adaptive filtering configurations can be considered in the light of the signal generation model considered in Section 2. In that model, the intercept signal and the reference music signal were assumed to be subject to both linear filtering effects and to non-linear effects. The dominant linear filter corresponds to the room acoustic channel. First, consider the placement of the linear canceller filter. Configurations A, C, D and E subject the reference music to a linear filter to model the room acoustics. The output of the reference music branch before cancellation should be music subject to reverberation effects. Configuration B subjects the intercept signal to a linear filter. However, in this case the linear filter ideally acts as the inverse to the room acoustic channel.



(a) Configuration A

(b) Configuration B

(c) Configuration C

(d) Configuration D

(e) Configuration E

Fig. 24 Canceller configurations

Recall that each of the components in the intercept signal was subject to different room acoustic channels. In this case, since the reference music is serving as the reference signal, the filter will try to model the inverse to the acoustic channel seen by the music component. The output of the intercept branch before cancellation should reveal a signal with decreased reverberation, at least for the music component.

The non-linear compensator can compensate for a memoryless non-linearity. If the major source of the non-linearity in the intercept recording is the recording stage itself (microphone, amplifier and following recorders), then configurations B and C are appropriate. Configuration A applies the non-linearity only in the reference music path. If the music dominates, it will compensate for non-linearities in the recording stage, but not as well as B or C since superposition does not hold for non-linearities.

Configurations D and E are hybrid configurations. Configuration D uses non-linear compensation in the reference music branch and then after cancellation tries to undo the non-linear effects in the output signal. Similarly configuration E uses linear filtering to achieve cancellation, and then applies the inverse room acoustic channel model (derived for the music component) to the uncancelled components. The idea is to try to at least partially dereverberate the speech components.

7.2.2 Adaptive Cancellation

Configuration A is the main focus of the attempts at adaptive cancellation. The different configurations are similar enough that parameter values derived for this configuration are likely to be appropriate for other configurations as well.

Consider first the non-linear compensator $\mathcal{N}$. The squaring and cubing operations can cause overflows. The non-linear compensator design or more specifically the initial coefficient values and the step sizes have to be chosen with care. To start the gradient descent algorithm, the initial values of b_1, b_2, and b_3 were set to 1.0, 0.0, and 0.0 respectively. By trial and error, a combination of step sizes was obtained for the portions of the signals processed. The values of these step sizes were

$$\{\beta_1,\ \beta_2,\ \beta_3\} = \{10^{-15},\ 10^{-18},\ 10^{-27}\}$$

It was found that further increase in any one of these step sizes could drive the filter towards instability.

For the linear filter, the tap coefficients were initialized as follows:

$$a_i = \begin{cases} 1 & i = 0\ , \\ 0 & \text{otherwise}\ . \end{cases}$$

During the initial investigation, the step sizes were equal and time invariant,

$$\mu_i = 10^{-10}\ .$$

As shown in Table 4, the performance of the system in terms of music suppression is a function of the number of coefficients. The music suppression is calculated as the average of the suppressions expressed in dB for 16 ms segments. The segments are chosen from a portion of the signal containing music with no speech present. The suppression for a segment is the ratio of the energy of the

Number Coefficients	Music suppression dB
11	1.43
51	3.54
101	4.71
151	5.40
201	5.91
251	6.31
301	6.64
351	6.97

Table 4 Music suppression as a function of the number of coefficients

adaptively filtered reference music signal to the energy of the intercept signal. If the number of coefficients was increased much beyond 350, the filter became unstable for the given step size.

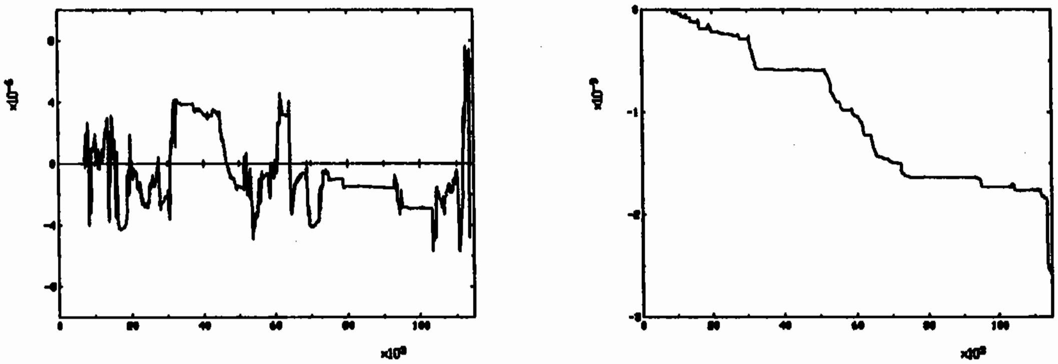
In the next experiment, the number of coefficients was kept fixed at 201 while the step size was varied. The results are shown in Table 5. For these results, the same step size was used for all coefficients and the step sizes for the non-linear compensator were as before. Additional experiments were conducted with step sizes that decayed exponentially as a function of the coefficient index. This strategy gave worse music suppression than having all step sizes equal.

step size	Music suppression dB
10^{-11}	2.09
10^{-10}	5.91
10^{-9}	arithmetic overflow

Table 5 Music suppression as a function of step size for a 201 coefficient adaptive filter

The parameters chosen for music cancellation were a 201 tap adaptive filter with a step size of 10^{-10}. The system based on this design provides noticeable music suppression. Speech that was lost under music is now discernible.

Further investigation of the effect of the non-linear compensator was carried out. Figure 25 shows the variation of the coefficients b_2 and b_3 as a function of time. The value of b_1 remained constant during the adaptation process. It can be noted that the value of b_2 does not seem to be converging, while the value of b_3 ends up at -4.5×10^{-9}. In order to find out the perceptual significance of the non-linear compensator, the non-linear compensator was turned off. Subjectively, there was no difference between the outputs generated with and without the non-linear compensator. Objectively, the difference between the two output signals was 40 dB below the level of the output signal with non-linear compensation.



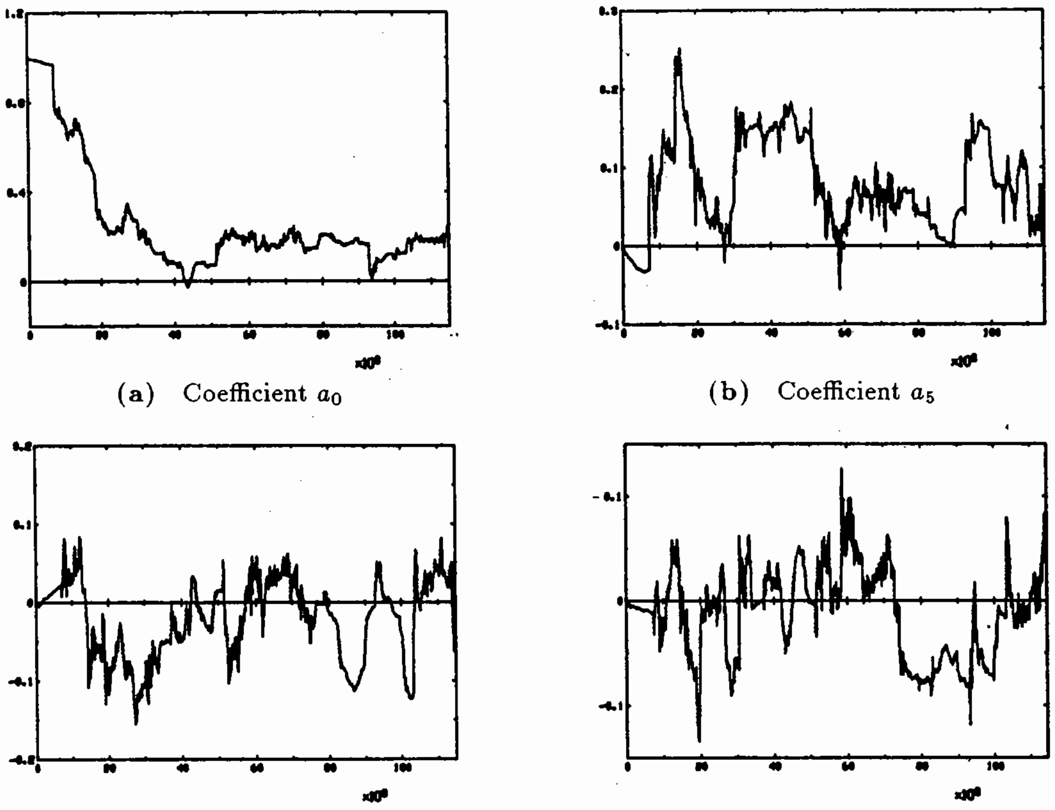
(a) Square term coefficient b_2 (b) Cubic term coefficient b_3

Fig. 25 Time variation of the coefficients of the non-linear compensator

The behaviour of selected tap coefficients of the linear filter is shown in Fig. 26. With the exception of coefficient a_0, which decreases from its initial value of unity, the other coefficients tend to change with no definite signs of convergence. In view of the lack of synchrony between the reference music signal and the intercept signal, variation of the coefficients with time is to be expected. Under ideal conditions with a white reference signal, the impulse response of the filter should maintain a constant shape but with the response drifting back and forth along the filter coefficient axis as the time alignment varies.

Another view of the adaptation process for the filter coefficients is shown in Fig. 27. This is a surface plot of the coefficients over a 20,000 sample segment. The plot is given in two sections, each of duration 10,000 samples. The coefficient values are shown every 100 samples and only the first 101 coefficients are plotted.

Figure 27 shows two different forms of behaviour. In the section labelled section I, the linear filter varies smoothly and in section II, the impulse response changes with more abruptly. The first section consists mainly of a single note of music, while in the second section, the singer's voice appears. Because the signals in both sections I and II were exposed to the same kind of distortion caused by room acoustics and recording devices, one could expect that the linear filter should respond consistently. However, in the first section because of the time correlations present in the sustained note, the optimum canceller filter is not unique. This is most evident from the fact that the first coefficient remains at its initial value. Yet there is nothing inherent that dictates that this value is the best one. Indeed experiments with other starting values show that the first coefficient tends



(a) Coefficient a_0

(b) Coefficient a_5

(c) Coefficient a_{50}

(d) Coefficient a_{200}

Fig. 26 Time variation of the coefficients of the adaptive filter

to keep its initial value in the first section. In the first section, time synchrony is not necessary for good cancellation. When the reference music signal becomes more white, the coefficients quickly change to other values. This is most radically evident in the sudden change in the first coefficient. Note that in the cut between the two parts of the diagram, one can see a smoothness in how the coefficient values change as a function of their index. This is a strong indication that the coefficient values are indeed modelling some filtering action.

Two methods for adapting the step sizes $\{\mu_i\}$ were considered. In the first, the step size was changed according to whether speech energy was present. The rationale is that since speech energy is noise to the adaptation process, the step size should be reduced when speech is present to help average out the effects of the speech. For the initial step size chosen, instability in the adaptation process could be seen when speech was present and the number of coefficients exceeded 300. One strategy tried was to increase the step size over its nominal value determined earlier by a factor of 3 when no speech was present and to reduce it by a factor of 2 when speech was present. This new strategy increases the music suppression to 8.81 dB, while leaving the energy of the speech components almost unchanged. The experiment used manual intervention to determine when speech

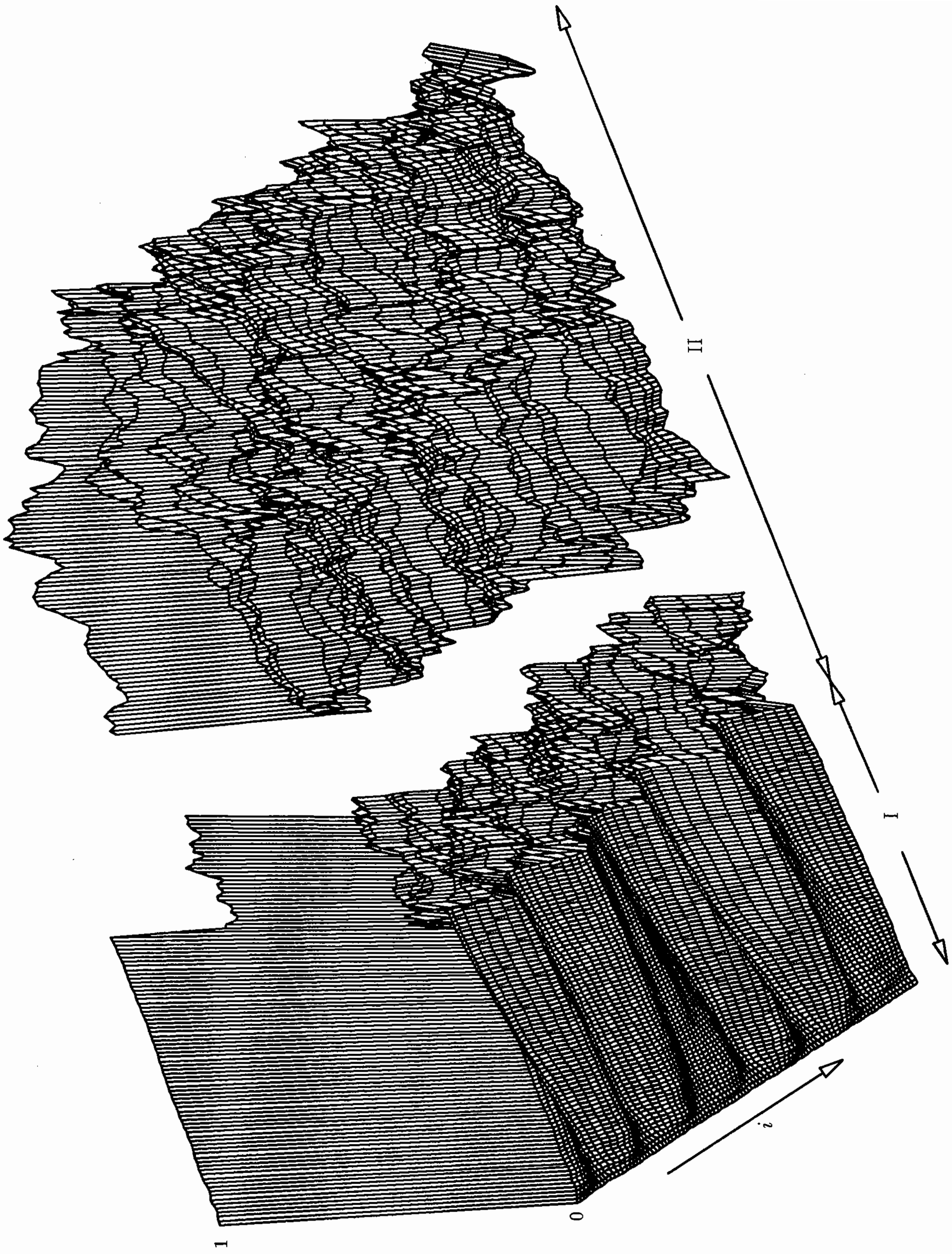


Fig. 27 Three dimensional view of the truncated impulse response of the time varying linear filter. The first 101 coefficients are shown at intervals of 100 samples.

was present. An automatic process to accomplish the change more smoothly is certainly a possibility.

The second method for adapting the step sizes is based on computing an estimate of the optimal step size. The value of the optimal step size is computed by setting the derivative of the mean-square error with respect to the step size to zero. In principle this can be done taking into account the non-linear compensator, but the solution is impractical. Rather, for this method the non-linear compensator is omitted. The optimal step size that will set the error to zero in one step is given by

$$\mu_{\text{opt}} = \frac{1}{2\left(\sum_{i=0}^{N-1} r^2(n-i)\right)} .$$

For practical use, the step size used is this optimal step size times a factor less than unity.

The two methods to adapt the step sizes are very similar in nature. With proper selection of the time varying factors, both approaches were able to suppress the music components in the distorted input.

7.2.3 Configuration B

Recall that the speech and music signals had first been distorted by the room acoustics, possibly differently, and then by non-linearity of various mechanical and electrical devices. The non-linear compensator and the linear filters were placed in the distorted signal path to neutralize the distortions in this order. Several tests were performed with and without the non-linear compensator and with different number of taps for the linear filter. Experimental results showed that regardless of the presence or absence of the non-linear compensator, the performance was unacceptable. With a 201 tap linear filter, the output signal consisted of distorted music and reduced speech energy. When music is present, the adaptive filter does try to cancel the music component. However when music is absent the correlation between the reference music signal and the intercept signal falls to zero and the filter coefficients also go to zero. Given the dynamics of the adaptation process, the filter turns itself off when speech alone is present. The performance of this configuration was unsatisfactory.

7.2.4 Configuration C

This is a combination of configurations A and B. A linear filter with identical characteristics and features as the one used in configuration A was installed on the path of the reference music signal. The non-linear compensator is in the intercept path. For small step sizes for the non-linear compensator, the performance is similar to configuration A. If the coefficient b_1 was allowed to change significantly, the output tended to become very small. The upper path set all coefficients to small values as did the lower path.

7.2.5 Inverse Filtering

Configurations D and E feature the use of inverse filters. It was believed that if the filters could be made to simulate the room acoustics and non-linearity using the gradient descent learning process, the inverse of these filters would be able to undo the corruption and increase the intelligibility of the speech component.

In configuration D, the inverse of the non-linear compensator was used. Let $\mathcal{N}$ and $\mathcal{N}^{-1}$ denote the non-linear compensator and its inverse respectively. The non-linear compensator, $\mathcal{N}$ is defined by

$$z(n) = \sum_{i=1}^{3} b_i[v(n)]^i \ .$$

The inverse compensator $\mathcal{N}^{-1}$ is given by

$$w(n) = \sum_{k} c_k[z(n)]^k \ .$$

With a proper selection of the tap coefficients $\{c_i\}$, the output $w(n)$ of the filter $\mathcal{N}^{-1}$ can be made equal to $v(n)$. In general, an infinite number of terms is required to form the inverse for a finite term non-linear compensator. A truncated version of the inverse filter with 3 coefficients was used. Nevertheless, the resulting system realized with the truncated inverse response was found to be unstable.

Configuration E uses inverse of the linear filter that was used to process the output signal. Let $\mathcal{L}$ and $\mathcal{L}^{-1}$ denote the linear filter and its inverse. With

$$\mathcal{L}: \qquad z(n) = \sum_{i=0}^{N-1} a_i v(n-i)$$

$$\mathcal{L}^{-1}: \qquad v(n) = \frac{1}{a_0}\left(z(n) - \sum_{i=1}^{N-1} a_i v(n-i)\right)$$

This recursive filter $\mathcal{L}^{-1}$ could easily be realized with feedback. Consider three ways to obtain the coefficients for this inverse filter.

1. Update the linear filter $\mathcal{L}$ and use the coefficients for both $\mathcal{L}$ and $\mathcal{L}^{-1}$.
2. Update the linear filter $\mathcal{L}$, generate the output signal without $\mathcal{L}^{-1}$. Realize a time invariant inverse filter using the set of converged tap coefficients.
3. Take the inverse filter into account and update the coefficients of both filters together.

All three ways of obtaining the coefficients were tested. The inverse filter using methods 1 and 2 were very unstable. In order to stabilize the inverse filter using method 1, the maximum step size μ for the coefficients had to be less than 10^{-14}. The resulting system performed poorly in music suppression. This was consistent with the results observed before. The presence of the inverse filter had no effect on the improvement of speech intelligibility, however.

For method 2, a stable inverse filter could not be found. This problem could be attributed to the fact that most of the coefficients a_i were changing with time during the learning process. They did not converge (or tend to converge) to any fixed values. The time-varying filter could not be adequately approximated by a time invariant one.

In method 3, the coefficients were computed by minimizing the squared magnitude of output samples. Stability could be obtained without difficulty. Nevertheless, the resulting system had poor performance in terms of music suppression and improvement in intelligibility. No satisfactory results were obtained.

7.2.6 Separate Gain Factor

There are two dynamics in progress during adaptation. First the filter must adapt to changes in response due to the time alignment drift. However, in addition the filter must try to compensate for the AGC used in the recording process. The filter as set up can compensate for gain changes only by changing all coefficients by a constant factor. As shown in Fig. 28, a separate gain factor was introduced to the system of configuration A (without the non-linear compensator). The idea is to allow the gain G to follow changes due to the AGC while the adaptive filter is allowed to compensate for filter response changes alone.

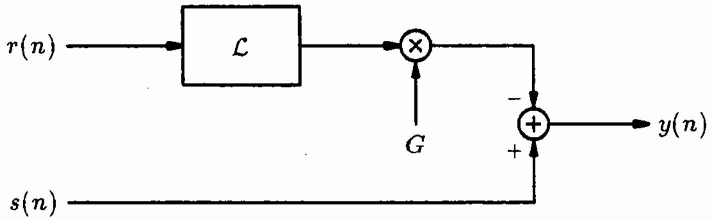


Fig. 28 Decoupling the gain factor from the adaptive linear filter

The gain value was updated using a gradient update. Subjectively, the output was similar to that obtained from configuration A. The absence of the non-linear compensator and the presence of a decoupled gain had no significant effect on the performance of the system.

7.2.7 LMS Adaptive Filter Performance

Amongst the five configurations discussed above, configuration A had the best performance in terms of music suppression. With a proper choice of parameters, the adaptive filter in configuration A could achieve significant suppression of the music component with no stability problems.

The adaptive filtering strategy was applied to cancel the music component in the intercept signal. It has little effect directly on the underlying speech components. Even in sections of the intercept recording with no music interference, the intelligibility of the speech is somewhat hampered by the highly reverberant environment. Some of the adaptive filtering configurations attempted to use inverse filtering in the hope that some of the reverberant effects could be removed. These techniques however suffered from stability problems.

8. Frequency Domain Techniques

8.1 Introduction

Frequency domain techniques operate on the Discrete Fourier Transform of the signals. These techniques have one basic advantage over their time domain counterparts: they can, to some extent, be made to ignore the phase of the signals. This means that time alignment of the signals tends to be less critical for the frequency domain techniques.

Two frequency domain techniques were tested.

1) Comb filtering of intercept signal $s(n)$ to remove the fundamental and its harmonics of the music component, and

2) spectral subtraction of the time-aligned reference music signal from the intercept signal.

8.2 Comb Filtering

The comb filtering technique consists of applying a time-varying comb filter to the intercept signal $s(n)$ in the hope of attenuating those frequencies where there is a concentration of music energy and leaving the superimposed speech undisturbed. These music-intensive frequencies are implicitly assumed to be harmonics of some fundamental frequency f_o so that the spacing of the notches of the comb filter is uniform and equal to f_o.

The application of this method implies two steps:

(i) designing a variable comb filter $H_c(f_o, \omega)$ and

(ii) determining an f_o value for each analysis frame.

8.2.1 Variable Comb Filter Design

Probably the easiest way to design a comb filter is to take a simple lowpass or highpass filter and have it repeat itself τ times in the interval between 0 and $f_s/2$. The prototype filter used is a highpass filter with a notch at low frequencies. In designing an appropriate filter, a first-order IIR structure was selected. For a given order the IIR structure performs better in terms of notch width and passband flatness than an FIR structure. Three highpass filters having different bandwidth notches were designed using the IEEE filter design routines [1]. The frequency response of the prototype filter with the widest notch is shown in Fig. 29. The time domain representation for this filter is

$$y(n) = 0.73008728\ x(n) - 0.73008728\ x(n-1) + 0.46017468\ y(n-1)\ .$$

Starting from the prototype highpass filter, it is easy to generate the desired comb filters. If the frequency domain representation of the prototype filter is $H(e^{j\omega})$ and if the desired comb filter

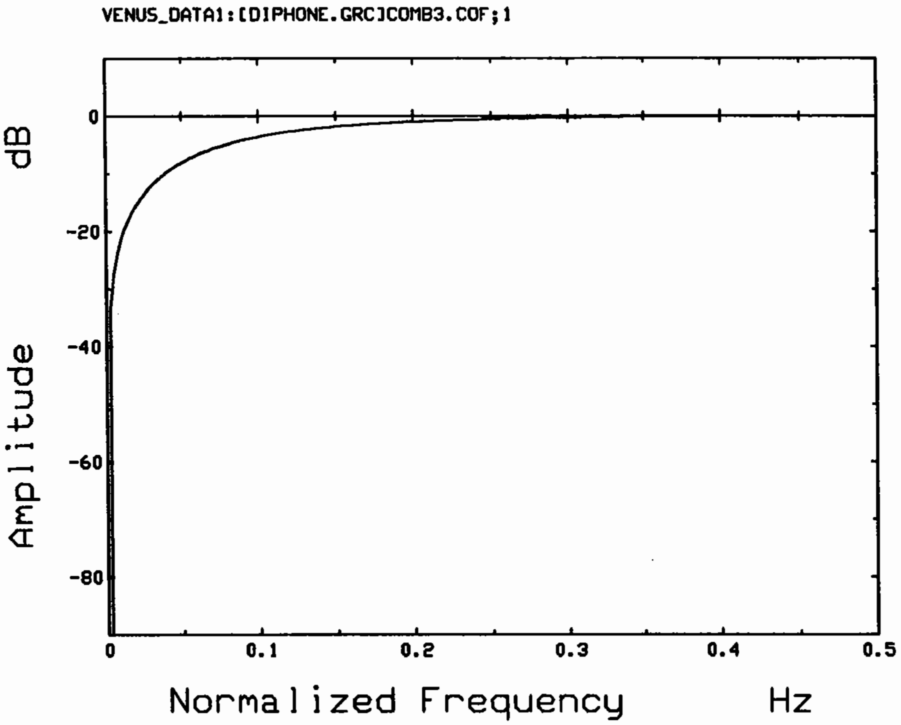


Fig. 29 Frequency response of a prototype notch filter

is to remove every harmonic of f_o then $H_c(e^{j\omega})$, the frequency domain representation of the comb filter, is equal to $H(e^{j\omega\tau})$ where $\tau = f_s/f_o$ and f_s is the sampling frequency. Note that τ should be an integer. In the time domain, the typical comb filter can be denoted by

$$y(n) = a\ x(n) + b\ x(n-\tau) + c\ y(n-\tau)\ ,$$

where a, b, and c are the coefficients of the prototype filter.

Figure 30 shows the frequency representation of a white noise signal (20,000 samples long) filtered by a comb filter with $\tau = 10$. The comb filter used is based on the prototype filter shown in Fig. 29.

8.2.2 Fundamental Frequency Estimation

Several methods were tried in order to estimate the fundamental frequency f_o.

1) the spectral comb multiplication method,

2) the LPC residual correlation method,

3) the spectral peak picking method,

4) the average magnitude difference method, and

5) "manual" tracking.

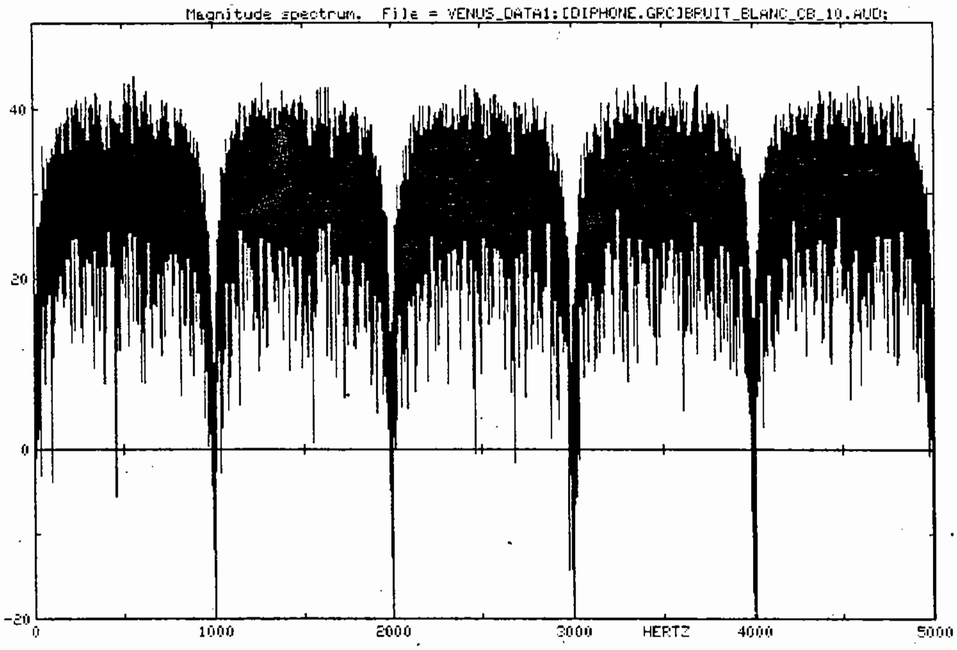


Fig. 30 Frequency domain representation of a white noise signal comb-filtered with $\tau = 10$

Spectral Comb Multiplication Method

In this method, an inverse comb filter is used to select the value of f_o which will maximize (with some constraints) the "energy" which the comb filter will remove from $s(n)$. More specifically, one should find the value of f_o which maximizes

$$\beta(f_o) \int_0^{\pi} P(\omega) \left[R(\omega) \left(1 - H_c(f_o, \omega)\right)\right]^{\alpha} \, d\omega$$

where ${f_o}_{\min} \leq f_o \leq {f_o}_{\max}$. The filter $P(\omega)$ is used to emphasize those frequencies which are perceptually more important. The weighting function $\beta(f_o)$ compensates for the tendency of this integral to be largest for small values of f_o.

In the case where $P(\omega) = 1$, the above integral expression is approximately equivalent to

$$\beta(f_o) \sum_i \left[R(i\tau)\right]^{\alpha} ,$$

where $\tau = f_s / f_o$.

This approximation was used with $\alpha = 1$ and $\beta(f_o) = f_o / {f_o}_{\min}$. The application of this method on the reference signal ($r(n)$) yielded a widely fluctuating f_o estimate pattern which often did not agree with observed values on spectrograms. There are many factors which can account for this behaviour.

1) In order to limit computation time a DFT of 512 points was taken providing a frequency resolution of only 20 Hz.[†] A DFT of 2048 points (resolution of 5 Hz) or more might yield better results.

2) In order to limit computation time, the permissible values for f_o varied from $f_{o_{\min}}$ to $f_{o_{\max}}$ in steps of 5 Hz. Smaller steps (1 Hz or less) should have been used.

3) Setting $\alpha = 1$, $P(\omega) = 1$, and $\beta(f_o) = f_o / f_{o_{\min}}$ is an inappropriate simplification.

LPC Residual Correlation Method

LPC analysis performed on a section of speech helps to remove the short-term correlations in the input signal. The resulting prediction error which has an approximately flat spectral envelope, is seen to be sharply peaked at the beginning of each pitch period. For this reason it is an ideal candidate for fundamental frequency estimation. One may perform an automatic search based on the autocorrelation function of the prediction residual. This function is normally characterized by a series of well defined peaks located at multiples of the pitch period. In the method used here, one simply looks for the location of the maximum of the autocorrelation function in a specified range. This range must be chosen carefully to avoid tracking one of the harmonics of the fundamental.

The method as described above makes direct use of the properties of the prediction residual without any form of preprocessing being needed. By using such a method in the actual context we hope that we will be able at least to reduce the singer's voice in the portions of $r(n)$ where it tends to dominate. One of the major difficulties encountered was in adjusting the search range. Without frequent manual intervention, we could not prevent pitch doubling effects. Moreover the method was found to be much affected by the presence of background music in some segments of the input signal. The music made it impossible to stabilize the measured f_o value around the tracked fundamental in presence of harmonics.

Spectral Peak Picking Method

The spectral peak picking method makes use of a decimated and subsampled version of the lowpass-filtered reference music signal $r(n)$ (after time alignment). Specifically, $r(n)$ was lowpass-filtered at 180 Hz (using a 256 tap FIR filter). The resulting signal was then decimated and subsampled, setting its sampling frequency to 500 Hz, producing $r_{500}(n)$. Such subsampling implies that a 100 sample window in $s(n)$ corresponds to a a 5 sample window in $r_{500}(n)$. Padding these

[†] Note that the DFT was taken on 128 samples of the reference signal padded to the DFT size of 512 samples with zeros.

5 samples with 507 zeros and taking the DFT of the resulting sequence gives a resolution of more than 1 Hz in the spectral domain.

Since the 180 Hz lowpass-filtered $r(n)$ contains at most 3 harmonics, the hope was that by picking the largest non-DC component in the DFT, the frequency of this component would in some sense represent the fundamental frequency.

The results obtained with this peak picking method were mediocre for what are probably some of the following reasons.

1) The presence of multiple "fundamentals" attributable to musical chords made the desired goal of tracking a single fundamental difficult in places.

2) In music, the second and third harmonics are often as intense as the first and, consequently the selected f_o in some regions had a tendency to hop between these values.

3) Having only 5 samples to work with, the reference sequence was not windowed with a tapered window prior to taking its DFT. This favored the appearance of frame discontinuities.

Average Magnitude Difference Method

The Average Magnitude Difference (AMD) Method for f_o estimation is used by the program `f0auto` which is available as part of the INRS-Telecommunications Audio Group Software. The AMD Method works on a center-clipped, lowpass-filtered version of the signal. The clipping levels are established as a certain percentage of the maximum and minimum of the signal over a frame of length N. The result is a signal $s_{cc}(n)$.

Using $s_{cc}(n)$, we can define the AMD function $\Delta(\tau)$.

$$\Delta(\tau) = \sum_{n=0}^{L_c} |s_{cc}(n+\tau) - s_{cc}(n)|$$

where L_c is the length of the correlation window and $(L_c + \tau_{max}) \leq N$.

Using dynamic programming, local to the minima of $\Delta(\tau)$, a pitch estimate is produced along with a "certainty factor" to associate with this estimate. Global dynamic programming on the full length of a given voiced section determines the actual value of the pitch for the current frame.

The AMD method produced the most reliable f_o curve of all the methods discussed so far. The AMD method was applied to $r(n)$, which still has its full complement of low frequencies. It did, however, suffer from occasional doubling of the f_o estimate as well as from some "noisiness" (small fluctuations in the f_o value which are not accompanied by fluctuations of the fundamental in $r(n)$).

Manual Adjustment Method

In order to obtain a smooth and reliable f_o estimate, we took a segment of the f_o estimate produced by the Average Magnitude Difference Method (c.f. previous section) and corrected it manually by superposing it on the spectrogram of the corresponding segment of $r(n)$ and verifying its accuracy.

Manual f_o estimation cannot be applied to the processing of large segments. It was used to estimate the limits of comb filtering as an enhancement method for $s(n)$ given a reliable f_o estimate.

8.2.3 Comparison of the Fundamental Frequency Evaluation Methods

Of all the automatic methods, the Average Magnitude Difference Method produced the most consistent and smooth f_o estimate curve. However, if the goal were to minimize the presence of music in the output signal, it seems likely that the comb multiplication method would be a better way to select the value of f_o.

8.2.4 Comb Filtering Performance

Comb filtering was only marginally successful in removing the reference music signal from $s(n)$. Figure 31 shows 1 Hz bandwidth spectrograms whose frequency ranges are limited to $[300, 800]$ Hz. The frame length is 101 samples. The top spectrogram shows the intercept signal $s(n)$ while the middle spectrogram shows the result of comb filtering the intercept signal. The bottom spectrogram is that of the reference music signal $r(n)$. Note how the comb filtering was quite successful in removing the harmonics around frame 750. On the other hand, it seems to have little effect between frames 800 and 950. Figure 32 shows the frequency domain representation of the intercept signal (top plot) and the comb-filtered intercept signal (bottom plot) for a region where the pitch estimate was constant (110 Hz). The middle plot shows the superposition of the two signals.

In general, the following factors limit the success of our attempts at signal enhancement using comb filtering.

1) Noise generation in the comb filtering process: High levels of low-, mid-,[†] and high-frequency noise are generated in the comb-filtered version of $s(n)$ whenever the value of f_o changes. Figure 33 clearly shows this noise. The top spectrogram is that of the comb-filtered intercept signal. The introduced noise is visible around frames 675 and 810. Note that the noise coincides with f_o discontinuities. The bottom spectrogram shows the first 500 Hz of reference music signal $r(n)$ along with the f_o curve generated by the average magnitude difference method (the light trace line centered around 100 Hz). The multiplicity of fundamentals is also visible in the bottom spectrogram ($f_{o1} \approx 115$ Hz and $f_{o2} \approx 160$ Hz).

   Using a manually corrected f_o estimate significantly reduced the appearance of noise of this type.

2) Wider harmonic bandwidths in $s(n)$: Narrow bandwidth harmonics of f_o in $r(n)$ became "spread out" in $s(n)$. This makes notching them out difficult given the proximity of the neighbouring harmonics.

[†] The presence of mid-frequency noise could not be verified because such noise tends to masked when added to $s(n)$. Its presence, however, is to be strongly suspected given the levels of low- and high-frequency noise which are apparent.

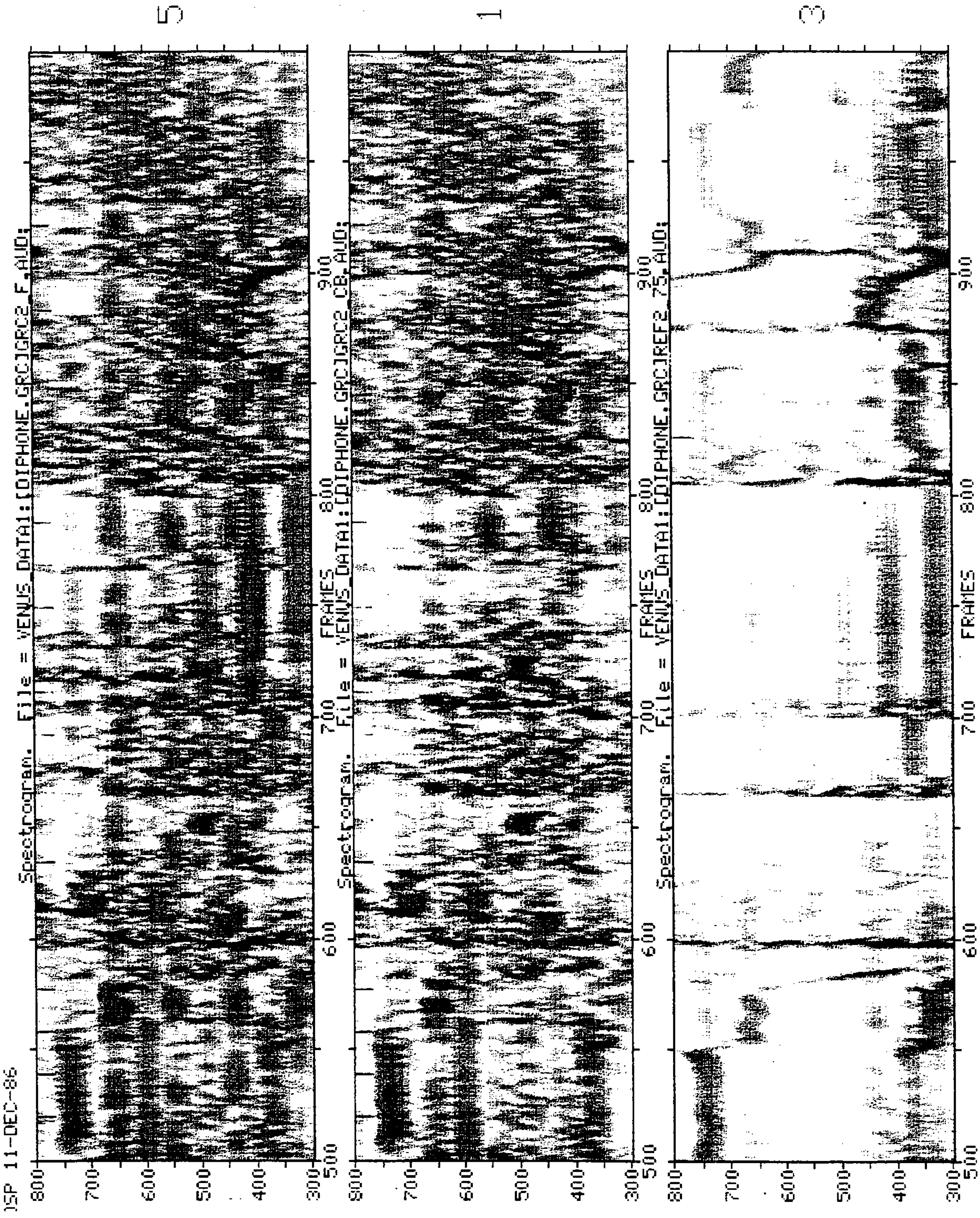


Fig. 31 Spectrograms showing the effect of comb filtering (1 Hz bandwidth spectrogram for frequencies between 300 and 800 Hz).
Top: Intercept signal
Middle: Comb-filtered intercept signal
Bottom: Reference music signal

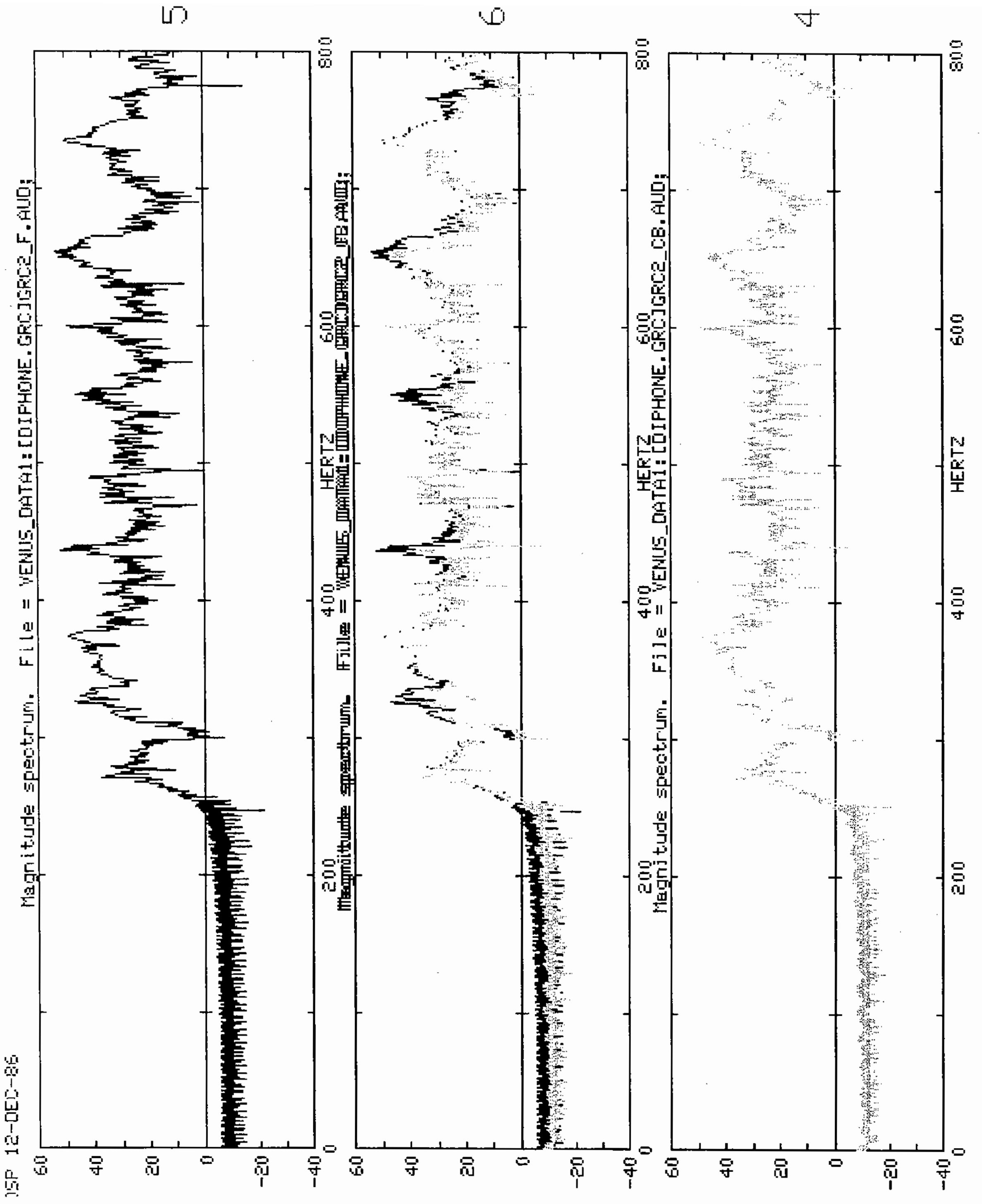


Fig. 32 Spectra showing the effect of comb filtering.
Top: Intercept signal
Middle: Comb-filtered intercept signal superimposed on the intercept signal
Bottom: Comb-filtered intercept signal

3) Integer constraint for f_o: By forcing f_o to be an integer, the resolution is limited to ±0.5Hz which implies that by the i^{th} harmonic, the filter notch is offset by $i/2$ Hz. This offset can be important given the often close spacing of the notches and their narrow bandwidths.

4) Multiplicity of "fundamentals": In musical chords there can be from 2 to 6 simultaneous notes, each with its own fundamental frequency. These frequencies are usually related to the lowest fundamental by $2^{k/12}$, where k is a positive integer (normally greater than 2). The removal of only one of these fundamentals may not suffice to significantly reduce the level of the music.

Problem 1) can probably be solved by using a more sophisticated filter update technique, overlapping of the filtering frames, incorporating preliminary smoothing of the f_o curve, and, possibly, some filtering of the output.

Problem 2) can, in part, be alleviated by selecting a wider bandwidth first order filter as a basis for the comb filters. Of course, this is a compromise because as the bandwidth is enlarged more of the speech in $s(n)$ will be removed along with the harmonics of the music. Our best experimental results coincided with using comb filters based on the filter whose frequency response is shown in Fig. 29. This filter has a relatively large notch bandwidth.

Problem 3) is easily readily bypassed (assuming a precise enough f_o estimate can be generated so that 0.5 Hz becomes significant). A non-integer f_o implies that since τ will be a non-integer, there will be non-integer indices in terms of the form $x(n - \tau)$. Such terms could be evaluated by interpolating $x(n)$ between samples i, and $i + 1$, where $i \leq \tau < i + 1$.

Problem 4), the multiplicity of fundamentals, could be tackled by filtering repeatedly with comb filters having different fundamental frequencies. Preliminary tests with this approach were inconclusive. However, since some of these fundamental frequencies could be relatively close to one another, the spectrum of the repeatedly filtered signal might be mangled beyond recognition.

In summary, we were not particularly successful in enhancing the speech component of the intercept signal using comb filtering. The comb filtering technique has problems in dealing with instrumental music. It however has potential to be more successful in those cases where the singer's voice is the most energetic component of the music spectrum. There, because the singer's fundamental is reasonably well delimited, and occurs at a relatively high frequency, tracking is easier. The fact that the fundamental occurs at a high frequency means also that less of the overall spectrum is removed by comb filtering, leaving the underlying speech less affected.

8.3 Spectral Subtraction

This technique attempts to remove the musical component of the intercept signal by spectral subtraction. At the input to this process, the input signal ($s(n)$) has been prefiltered, inverse filtered to whiten the noise, and equalized in order to match the level of the musical part of the signal with the level of the reference music signal $r(n)$. In spectral subtraction, the prefiltered and time-aligned

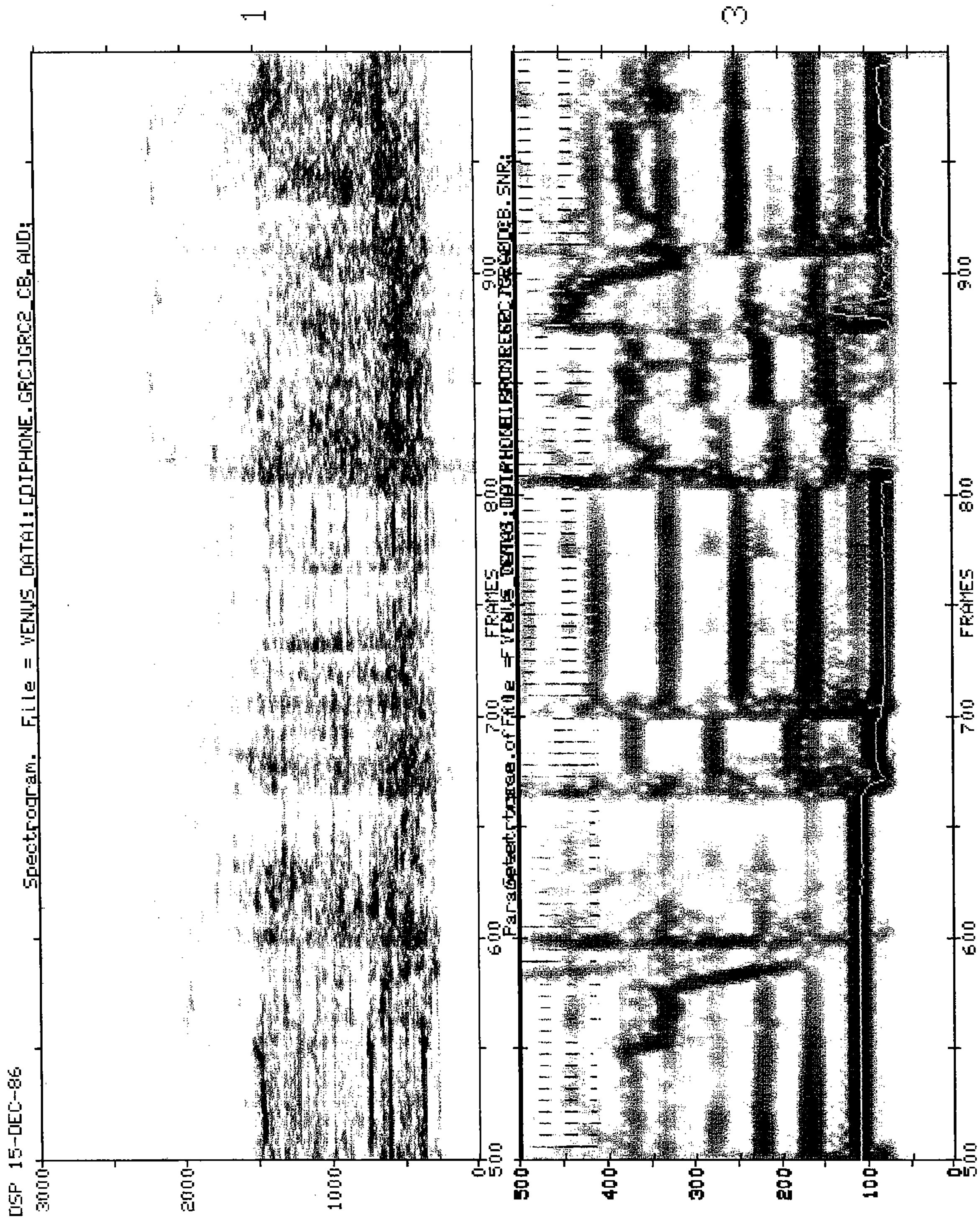


Fig. 33 Spectrograms (narrowband mode) showing the effect of erroneous fundamental frequency estimates.
Top: Comb filtered intercept signal
Bottom: Fundamental frequency estimate (light trace) superimposed on the reference music signal (0 to 500 Hz)

reference signal, $r(n)$, furnishes the spectrum that will be subtracted from $s(n)$. For this, both signals are segmented into frames of 256 samples, with an overlap of 128 samples. The frames of data are then weighted by a 256-point Hanning window, padded with 256 zeros, and transformed to the frequency domain by a 512-point DFT. This produces the spectra $S(k)$ and $R(k)$.

The second step of this process involves spectral subtraction of the magnitudes.

$$|S'(k)| = \begin{cases} \beta\,|R(k)| & \text{if } |S(k)| \leq \alpha|R(k)| \\ |S(k)| - \alpha|R(k)| & \text{otherwise} \ . \end{cases}$$

This operation is carried out frame by frame. The resulting spectrum, $S'(k)$, takes on the phase of the original signal $S(k)$. As far as music suppression is concerned, The best results were obtained with the parameters α and β set to 7 and 0.005, respectively. Experiments were conducted with different values for these two parameters to investigate their effect on artifacts; they are described later.

The resulting spectrum $S'(k)$ for a frame is transformed to the time domain by a 512-point inverse DFT and combined with other frames using overlap and add.

8.3.1 Spectral Subtraction Performance

Figure 34 shows spectrograms of the intercept signal (whitened and equalized) at the top, and the output of the spectral subtraction process at the bottom. The time axis here spans a duration of 3.5 seconds. As can be seen, the harmonics of the singer's voice which show up as thick bands approximately 350 Hz apart in the intercept signal are to a large extent eliminated by the process. The instrumental sounds which show up as thinner bands closer together are partially wiped out by spectral subtraction. The spectrogram of the resulting signal (here there is no speech present) shows short ($\approx$ 40 ms) isolated sounds which are artifacts of the spectral subtraction process. These are in the form of tonal noise.

When listening to the resulting signal, one can hear the speech more clearly. However, it is surrounded by a sea of bubbling sounds. The singer's voice all but disappears in the processed signal and the timbre of the accompanying guitar has been modified.

Attempts were made to mitigate the effects of the tonal noise.

1) Different values for the factor α which determines the fraction of the reference music subtracted were tried. Recall that the reference music signal has been equalized to match the music component of the intercept signal. This means if that a perfect match were present $\alpha = 1$ would be appropriate. However, experiments indicate that overcompensation seems preferable if it is desired to remove most of the music component. A value of 7 gives good music suppression but with accompanying tonal noise. A lower value results in less attenuation of the music but also with less tonal noise. A value of 0.5 gives reasonable amount of music suppression with little or no tonal noise.

2) Another experiment changed the value in β in the spectral subtraction algorithm. This was tried with a view to attenuating the perceptibility of the tonal noise by increasing the

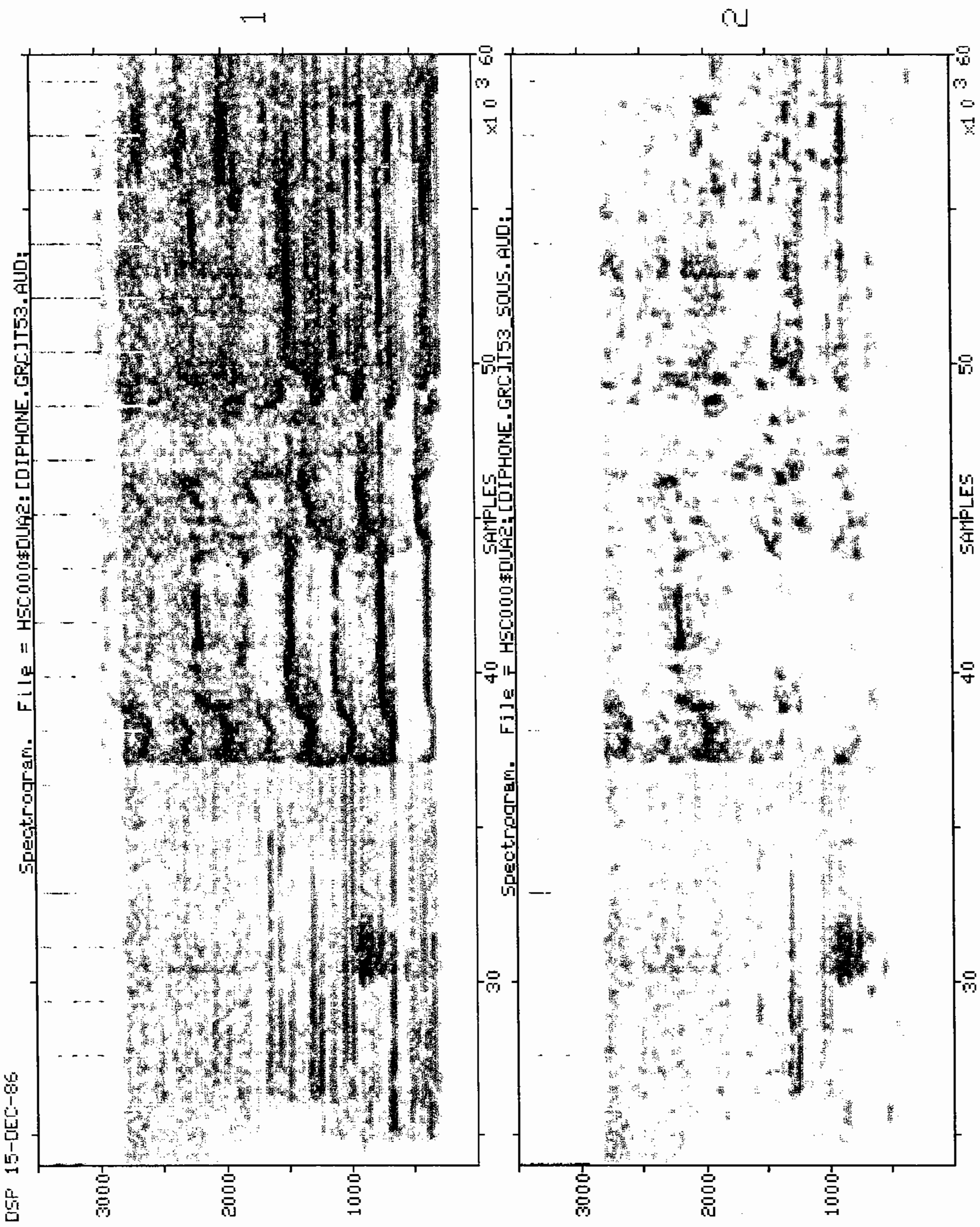


Fig. 34 Spectrograms (narrowband mode) showing the effect of spectral subtraction on the prefiltered intercept signal.
Top: Before spectral subtraction
Bottom: After spectral subtraction

background level. The value $\beta = 0.005$, or equivalently a very low value, gave the best results. An increase of β tends to generate a distinct third signal in the background: a low level version of the reference music signal. This signal coexists with the speech and tonal noise components without significantly masking the tonal noise.

3) Spectral subtraction using a higher frequency resolution was also tried. With a DFT window of 1024 points rather than 512 (256 data points in both cases), the frequency resolution changed from 39 Hz to 19.5 Hz. This modification did not produce any perceptible amelioration of the tonal noises.

4) The bandwidth of the components in the reference signal spectrum was widened before performing the subtraction. For this, the reference signal spectrum, $R(k)$, was replaced by the spectrum:

$$|R'(k)| = \max\left(|R(k-1)|, |R(k)|, |R(k+1)|\right)$$

   This bandwidth widening had the effect of corrupting those frequency bands which were originally occupied mostly by the speech components. In addition, the tonal noise increased.

5) Spectral subtraction with a larger window tends to improve the resulting signal and result in less tonal noise. This involves the use of a larger number of data values in each window. Good results were obtained for 1024 data values. A transform length of 2048 points was used to convert to the frequency domain. It can be speculated that the fact that time alignment is less critical for the larger windows contributes to their superior performance.

6) A last experiment involved equalizing the reference signal, $r(n)$, toward the signal $s(n)$ rather than the other way around, before the spectral subtraction stage. With this it was hoped to reduce the breathy sound introduced by the music equalization stage. The only perceptible difference was a change in overall gain.

8.3.2 Spectral Subtraction: An Evaluation

The net result obtained here after the 3 stages that operated in the frequency domain (namely, the noise inverse-filtering, the music equalization, and the spectral subtraction) is definitely enhanced speech, at least in terms of the intelligibility of the conversation. For the large values of α, tonal noises are introduced. This calls on the listener to ignore the tonal noise. It seems that this can be achieved after a few minutes of listening, because the tonal noises are somewhat unstructured. A better compromise is to use a smaller value of α which gives less music suppression, but also less tonal noise.

In the original intercept signal, the listener tends to follow the music to the detriment of being able to switch concentration to the speech when it is present. With the music suppressed by spectral subtraction, the listener can focus attention of the speech much more easily.

9. Summary and Conclusions

9.1 Processing the Full Length Intercept Signal

The main discussion in this document has centered around the investigation of techniques to enhance the speech intelligibility of a selected segment of the intercept recording. This segment was chosen because it contains a representative set of conditions such as music alone, speech alone, background noise alone and the various combinations of these components.

The next step was to process the entire 25 minute intercept recording using the procedures judged to be the most suitable. This in itself presents problems due to the large amounts of storage required for the digitized signals. The original recording, intermediate signals, and the final output signal all have to be available simultaneously. To facilitate processing, the 25 minute recording was divided into 5 segments. The first and last of these contain no music and therefore were not processed for music suppression. They were however filtered to equalize the spectrum.

The time alignment process for the test segment was obtained with just two anchor points (at each end of the 14.5 second segment). For the longer segments, multiple anchor points had to be determined, lest the time alignment drift substantially within the segment. The anchor points were chosen to be about 50 seconds apart, but their locations were constrained by the availability of suitable acoustic events that could be uniquely identified in both the intercept recording and the reference music recording. The determination of the anchor points was a time consuming task. It was made feasible through the use of our interactive graphics facilities which combine audio playback and visual displays (both time and frequency domain) for digitized signals.

Preliminary results for adaptive cancellation with anchor points only at the ends of the long segments produced an annoying time varying cancellation, with the music level changing in bursts. In the middle of the segment if the time alignment was sufficiently in error, little cancellation was obtained for complex music passages. However sustained notes gave good cancellation. With a larger number of anchor points, these problems do not manifest themselves as severely. Ultimately, it is the lack of good time synchrony which limits the suppression obtainable with the adaptive cancellation technique.

In the longer segments, changes in the gain produced by the AGC were more pronounced than for the test segment. These changes may reduce the effectiveness of spectral subtraction. The adaptive filtering strategy could cope with the gain changes for the most part, although perhaps a resetting of the step sizes would be warranted if the gain changes radically. The inappropriateness of a single step size for the longer segments manifested itself as instabilities in the adaptive filtering. This was avoided by fixing the step sizes at some loss in music suppression capability. There are some signs that the performance suffers from time alignment deficiencies.

The two methods that were applied for music suppression were the LMS adaptive music cancellation and the spectral subtraction technique. It was felt that although the overall intelligibility of the speech after processing is about the same for both methods, the results are somewhat complementary. The adaptive filtering approach has the least effect on the speech but does not achieve as high a level of music suppression. The spectral subtraction method achieves higher levels of music suppression with some local loss of speech content (whenever the speech spectrum significantly overlaps the music spectrum). This means that some portions of the speech are more intelligible in one processed signal than the other. Listening to one and then to the other can enhance overall intelligibility.

One strategy used in hands-free telephony to mitigate the subjective effects of reverberation is to filter the output signal so as to cause a low frequency rolloff. This simple technique was tried, but the results were not judged to be entirely satisfactory. While the subjective effect was that reverberation was less pronounced, speech intelligibility did not seem to improve.

The unique quality of the signal produced by the spectral subtraction method was such that a speed up of the playback may enhance intelligibility. The final output of the spectral subtraction process was produced at the nominal sampling rate, and also oll at a sampling rate 8% higher.

9.2 Summary of the Investigations

Through use of digital signal processing techniques developed at INRS-Telecommunications, we were able to enhance the intelligibility of conversations recorded on an audio tape in the presence of interference and distortion. The distortions included background music and noise (i.e. from the microphone and telephone transmission), as well as room reverberation and imperfections in the original recording process. Removing the background music from the intercept recording was complicated by timing distortions caused by the many mechanical devices involved: the original phonograph, the original tape recorder, our phonograph and tape recorder, each with its own wow and flutter characteristics.

Prior to digital processing, both the intercept (conversation plus music) and the reference (music alone) signals had to be converted to digital format; we employed a more than adequate sampling rate of 10,000 per second, with each sample allocated 16 bits. In the intercept signal, short sections containing only background noise and only music were identified to help model those two sound sources (as transformed by the room acoustics and intercept process).

The successful techniques we employed to enhance the intercept conversation involved filtering, both constant (fixed in time) and adaptive. Fixed filters were designed: 1) to eliminate harmonics of the 60-Hz hum due to power line interference, and 2) to equalize the effects of the noise with respect to the frequency response of the intercept system (mostly due to the microphone and telephone line). For the adaptive filtering, it was necessary to time align the reference and intercept signals to facilitate subtracting a transformed version of the reference from the intercept (thus reducing the amplitude of the interfering music). Frequent manual alignments between the two signals were necessary, using our interactive graphics facilities which combine audio playback and signal display capabilities.

We attempted a noise subtraction technique (due to Boll — designed for eliminating helicopter noise from audio signals); however the output signal, while reduced in noise, suffered from quasi-random tone bursts (tonal noise) which rendered the sound often more disturbing than the original intercept signal. Instead, we used a noise equalization filter which purified the sound, giving the conversation in the intercept more presence than without such processing.

The time-varying filtering was based on adaptive noise cancelling techniques. The adaptive process attempted to estimate the linear and non-linear aspects of the intercept process by passing the reference signal through a adaptive linear filter and a non-linear compensator and subtracting the result from the intercept signal. The characteristics of the filter were obtained by a procedure which minimizes the energy in the final difference signal (i.e. intercept minus processed reference music). By constraining the degrees of freedom in the filter, the filter models the inverse of the intercept process, so that the output of the filter resembles the time-aligned reference music signal, distorted in the same way that the music was originally corrupted at the time of the intercept. While our attempts at non-linear compensation were ineffective, the linear adaptive filter (designed to model the room reverberation effects) provided a good version of how the intercept had linearly transformed the music. After subtracting this modified reference music signal from the intercept signal, the speech became more prominent in the audio signal, and hence easier to understand. The unresolved non-linear distortions and reverberant quality of the speech, however, remain a hindrance to full intelligibility of the conversation. In addition, the full potential of this technique was curtailed by the lack of synchrony between the reference music signal and the intercept signal, especially for the long segments with relatively sparsely spaced anchor points.

The spectral subtraction of the music signal achieved a higher level of music suppression. Experience showed that time synchrony was less of a problem than with time domain cancellation. Spectral subtraction allows for a tradeoff between the degree of music suppression and the level of tonal noise artifacts. A compromise allows for an adequate level of music suppression without introducing a level of artifacts that is unduly annoying. There is some muffling of the underlying speech signal and some sournds tend to be attenuated when there is a large overlap between the spectrum of the music and that of the speech.

9.3 Suggestions for Maximizing the Enhancement Potential

In the process of enhancing the speech component of the intercept signal, we discovered many degradations in the intercept signal which were directly or indirectly attributable to the original recording methodology. Some of these degradations could have been avoided which would have made the enhancement process both easier and more successful.

With regards to maximizing the enhancement potential of a recording, we offer the following specific suggestions.

1) The recording system should be of high quality. This applies to the microphone, the tape recorder, and the quality of the tape itself. It was felt that a major part of the non-linear distortion that was clearly audible was probably due to the microphone. This limits the degree of cancellation possible.

2) It would have been of some use to be able to characterize the impulse response of the room acoustic channel. This can be accomplished simply by recording a sharp noise (or even a clap of the hands).

3) The use of AGC seems to be necessary to accommodate various circumstances. However, the variation in the gain due to AGC can be counterproductive to certain processing techniques. The compromise is to use equipment with a large dynamic range, such that the use of AGC is minimized.

4) Multiple microphones separated in space should be used. In cases where a disturbing noise source (music, fans, etc.) can be identified, a separate pickup recording synchronously on a multiple track recorder would be very useful for noise cancellation.

5) In the case of the situation in the present intercept recording, the availability of a recording of the same music record, played on the same turntable, and recorded through the same intercept procedure would have probably enhanced the potential for music suppression.

9.3.1 Future Research

During the course of the investigation, a number of shortcomings in available techniques were identified. Time did not permit full exploration of all possibilities. We feel significant improvements in cancellation abilities and speech enhancement are possible with techniques that were not available for this project.

1) For the purposes of processing a short segment, manual intervention to determine time alignment is feasible. Further work to automate such a procedure is warranted. The success of the adaptive filtering procedures depends heavily on achieving a good time alignment between the reference signal and the intercept signal. The alignment can in fact be tracked by observing the efficacy of time-shifted versions of the adaptive filter.

2) The adaptive filter in our canceller had to cope with modelling the highly reverberant room acoustic channel and with compensating for the time alignment errors. It is felt that a decoupling of these two tasks may make for a better canceller. Consider a separation of the canceller into two parts: a standard tapped delay line filter and a variable delay element. The variable delay would try to track the time alignment variations, while the filter itself would compensate for the room acoustics. The dynamics of these two processes could then be separately optimized.

3) The adaptive filtering strategies themselves need honing for use in an environment with changing signal energies and signal correlations. Heuristics for changing the step size for

different conditions need to be developed. For this purpose, development of a measure of the whiteness of the reference signal could be useful.

4) The variation in gain due to the effects of AGC probably reduced the effectiveness of the spectral subtraction scheme. The amount to be subtracted was a fixed quantity, when in fact it should have been related to the gain applied to the intercept signal. AGC tracking should be implemented to help compensate for the gain changes. In the case of an intercept with a reference signal, the reference signal can provide clues to an appropriate gain. In addition, the background noise level in the intercept recording can also be used to identify gain changes. A practical scheme could make use of signal levels to track the AGC and undo its effects.

5) Further investigation of the spectral subtraction technique is needed. A bidimensional filter (working in time and frequency) that can identify tonal bursts could be used to eliminate these artifacts and render the output of the spectral subtraction process with more complete music elimination subjectively acceptable. Our spectogram displays show that the tonal noise is clearly identifiable visually. We feel that a processing algorithm can be applied to remove these components.

6) The problem of de-reverberation applies generally to recordings made in unconditioned rooms. Strategies for dereverberation which use homomorphic (cepstral) techniques are good candidates for application [6]. These should have potential in removing the long delay reverberant components to enhance speech intelligibility.

Despite these suggestions for future situations, we are satisfied that we have succeeded in significantly increasing the intelligibility of the conversation on the intercept tape. Taking advantage of our knowledge of how humans perceive speech, we were able to filter the intercept signal in such a way as to reduce the level of the background noise and to subtract out major portions of the interfering music.

References

1. IEEE Acoustics, Speech, and Signal Processing Society, "Programs for Digital Signal Processing", IEEE Press, 1979.
2. G. Oetken, T.W. Parks and H.W. Schussler, "New results in the design of digital interpolators", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 301-309, June 1975.
3. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.
4. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 208–211, April 1979.
5. B. Widrow and S.D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.
6. A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, 1975.