# Technical Note

1978-32

# Homomorphic Pitch Detection

D. B. Paul



15 August 1978

Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-78-C-0002 by

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-78-C-0002 (ARPA Order 2006).





This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

(18) ESD

(19) TR-78-252

(12)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

(6) HOMOMORPHIC PITCH DETECTION.

(10) Douglas B. PAUL
Group 24

(16) YP10



(14) TN-1978-32

(9) TECHNICAL NOTE

(11) 15 AUG 1978

(15) F1962 8-78-C-0002,
W ARPA Order-2006

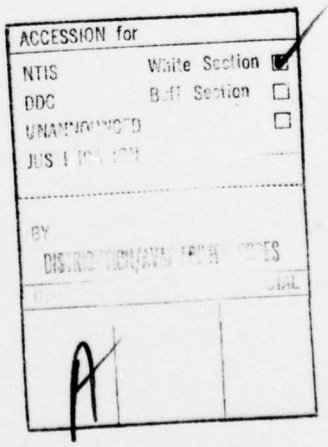
(12) 45p.



LEXINGTON MASSACHUSETTS

## Abstract

This note describes a homomorphic pitch detector which yields good performance on clear speech and moderate robustness to additive broadband noise, narrowband noise, and to degradation by a telephone simulator. It achieves the performance by the use of an adaptive time window, log-spectral windowing, an adaptive voicing threshold, and pitch track smoothing. It has been implemented in a real-time LPC vocoder for testing. Finally, a pilot study of a preprocessor to improve the performance of any coherence seeking pitch detector is presented.

Contents

# I. Introduction

For the past few decades, there has been much interest in speech bandwidth compression systems. The homomorphic vocoder algorithm and its related pitch detection algorithm [5,6], both proposed in the mid-to-late sixties, have undergone little development, primarily due to the complex hardware required to implement the Fourier transforms of the algorithm in real-time. (A real-time capability is required for any practical vocoder and is extremely helpful for research versions.) This obstacle has been removed by recent developments in charge-coupled device (CCD) technology, which promise fast computation of the Fourier transform with relatively simple hardware. The work presented here was undertaken to improve the performance of the homomorphic pitch detection algorithm in anticipation of its eventual implementation using CCD chirp-z transform chips.

First, the human speech production model and the theory of the homomorphic processor are described. Then, some practical difficulties encountered by the basic homomorphic pitch detector are covered in addition to the details of the real-time implementation used here. The final sections summarize and analyze the results of the real-time tests of the algorithm. A pilot study of a method for improving the performance of the homomorphic pitch detector (or any other coherence seeking pitch

detector) is presented in the Appendix.

## II. The Speech Model

The human speech production system consists of an air pressure source (the lungs) feeding through the vocal cords and combined nasal and oral passages. The vocal cords can be caused to vibrate and provide a periodic excitation to the vocal tract. Alternatively, the vocal cords can be abducted to allow airflow into the tract. A constriction higher in the tract will then cause turbulence noise to be generated just downstream of the constriction. The oral and nasal cavities form a variable configuration (straight tube or tube with sidebranch) deformable set of resonators connected to the excitation sources. A simplified version of this model--a periodic pulse or white noise source feeding a time varying filter--is generally used for vocoders derived from speech production models [2].

This simplified model (Figure 1) is adequate most of the time for most speakers. Its description of the voiced excitation breaks down under several situations. At the beginning or end of voicing, the interval between glottal pulses can change very rapidly. Any pitch detector which requires "local stationarity", i.e., stationarity over a small window, is likely to find this zone of speech to be unvoiced. Another troublesome mode of

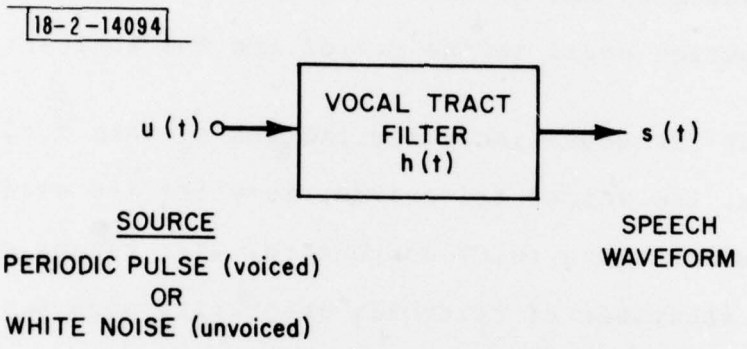



Fig. 1. Speech generation model.

voicing is diplophonia [4], in which alternate pitch periods are more highly correlated than adjacent pitch periods. Not only does this violate local stationarity, but it poses the additional question of how the listener perceives such an excitation--with a period equal to the average of the two adjacent pitch periods or with a period equal to the sum of the two adjacent pitch periods.

Many languages including English contain a class of phonemes, the voiced fricatives, in which the excitation has both periodic and white noise components. The binary source model is clearly incapable of correctly describing such a phoneme. Some African languages include sounds such as clicks which also are not included within the model. The perceptual consequences of such errors in the model vary with the specific error and the language. If the analyzer errs in a forgiving way, many errors will be corrected by the syntactic and semantic constraints of a language.

## III. The Basic Homomorphic Pitch Detector

From a signal viewpoint, the speech generation model is just a periodic or white signal source feeding a filter.

$$s(t)=h(t)*u(t) \tag{1}$$

where $s(t)$=speech signal
$h(t)$=vocal tract filter
$u(t)$=excitation signal--periodic or white
$*$=convolution operator

Time domain convolution results in a frequency domain product:

$$S(f)=H(f)U(f) \tag{2}$$

If one takes the logarithm of the magnitude of both sides,

$$\log|S(f)|=\log|H(f)|+\log|U(f)| \tag{3}$$

the product of the right hand side is transformed into a summation of independent components. $H(f)$ for an adult male speaker uttering a vowel contains, on the average, one formant (resonance) per kilohertz. Therefore, $\log|H(f)|$ is a slowly varying curve. If the utterance is unvoiced, $u(t)$ is white noise and $\log|U(f)|$ is relatively flat. If, on the other hand, the utterance is voiced, $u(t)$ and therefore $\log|U(f)|$ are periodic.

As the typical range of pitches is about 60 to 300 Hz, the frequency domain periodicity is of relatively high frequency. Log|S(f)| is therefore the sum of a slowly varying component and a rapidly varying periodic component.

A final Fourier transform of log|S(f)|

$$c(t) \longleftrightarrow \log|(S(f)| \tag{4}$$

produces the cepstrum (Figure 2) and will generally separate the components along a time scale with the low order components of c(t) representing log|H(f)| and, if voicing is present, a higher order peak at the periodicity of u(t) representing log|U(f)| [6]. This cepstrum is then searched along the zone corresponding to the expected range of pitches for the height and position of the pitch peak. If the peak height is above a threshold, the frame is declared voiced with pitch equal to the position of the peak.

On an ideal speech sound, this homomorphic pitch detector (Figure 3) works quite well. The peak representing the pitch period is sharp and clear when the excitation is voiced and absent when unvoiced. Real speech recorded with high quality equipment in a quiet environment frequently produces a well

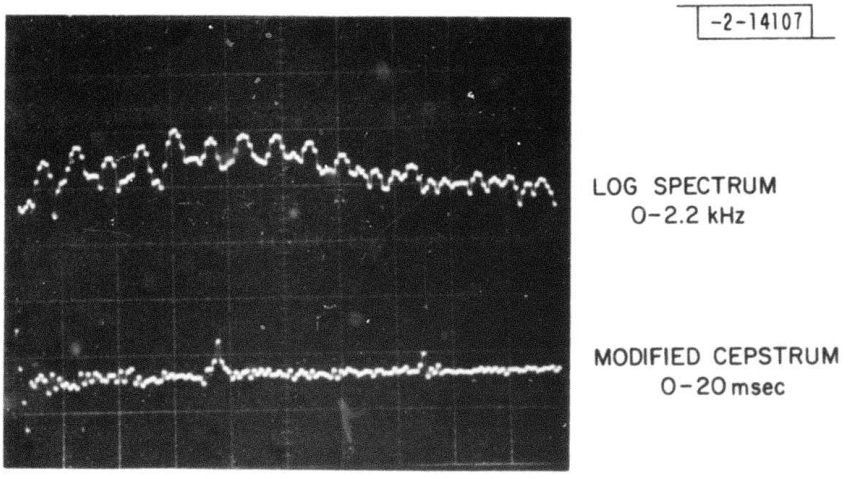



Fig. 2. Waveforms of the basic homomorphic processor

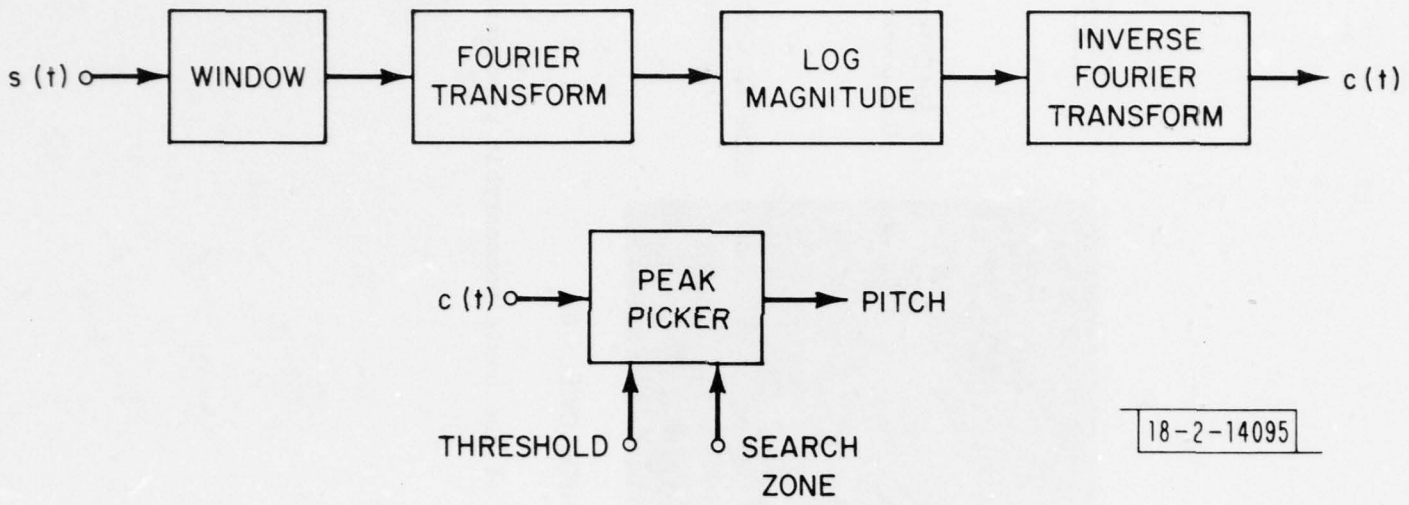



Fig. 3. Basic homomorphic pitch detector.

behaved cepstrum. It also strains the inherent assumptions sufficiently often that the basic homomorphic pitch detector will generally perform fairly poorly over an entire utterance. A practical homomorphic pitch detector therefore requires a number of modifications and additional strategies to give acceptable performance.

## IV. Practical Aspects of Homomorphic Pitch Detection

The first issue which a homomorphic pitch detector must face is the lack of stationarity in the signal. To preserve the local staticnarity that exists, the speech signal must be windowed and analyzed as (hopefully) stationary frames. As some events in (English) speech (flap d and flap t) occur in about 10 mS., a very narrow window is required. The pitch detector, however, requires a high resolution spectrum which suggests that the window be four to five pitch periods long. As typical pitch periods vary from about 2.5 to 20 mS. the above requirements cannot be met. Fortunately, a certain amount of slurring of the excitation is allowable and the optimum size is the minimum width which gives adequate spectral resolution.

Given that a time window is required, which window is the best? The requirements here are not unusual: minimum width main lobe and low sidelobes in the frequency domain with minimum width

in the time domain. The exact choice of window is probably not important as long as one of the standard high quality (i.e., narrow main lobe and minimal sidelobes) windows such as the Hamming window is used.

Observation of high resolution speech spectra will show the spectra of apparently ideal (time domain) speech to be less than ideal. The periodicity of the spectrum may break down above one kilohertz. The voicing periodicity of voiced fricatives may be obliterated above about the same point by the random noise components. Finally, a voiced sound may have as much as 50 db dynamic range in the spectrum with the peak energy concentrated in the region of the first two formants which lie below about 1.5 kHz. These factors suggest that the region up to about 1.5 kHz is the most reliable portion of the spectrum for use in pitch detection.

If a periodic pulse waveform is fed to a homomorphic pitch detector, several phenomena will be observed. First, the height of the pitch period peak in the cepstrum will vary with frequency. This suggests a weighting of the cepstrum such that, over the region of interest, the height of the peak as a function of the period is relatively constant to allow the use of a fixed voicing threshold. This weighting function should be relatively smooth to prevent small pitch errors caused by markedly differing

weights on adjacent samples.

A second observation is that the "floor" from which the pitch period peak rises may not have a constant level. (The cepstrum may be viewed as vocal tract frequency response information in the low order terms and a low level noise floor elsewhere which, if voiced, contains a peak corresponding to the periodicity of the source. The basic problem is discrimination of the peak from the noise and frequency response information.) Not only can this noise floor have a non-zero DC level, but it may be tilted or equivalently have a varying regional DC level (have low frequency terms in the cepstrum). As these deviations from ideal vary, no fixed correction can be used to eliminate them. A good way of processing this noise floor is "local DC removal" or removal of the low frequency terms from the cepstrum. Any threshold operation performed on the pitch period peak now will see just the peak plus noise rather than the peak plus noise plus its local DC level.

A third observation is the presence of cepstral peaks, usually lower in amplitude than the main peak, at multiples of the pitch period. These extraneous peaks, if found by the peak picker instead of the proper peak, will cause a doubling (or tripling) of the pitch period estimate. The amplitude of these peaks is a function of the length of the original time window on

the signal. Too long a window will yield high amplitude extraneous peaks. (For a constant pitch signal, a longer window will yield narrower log-spectral lines. As the log-spectrum approaches a periodic impulse train, so will the cepstrum.) A Hamming window about 4.5 pitch periods long will simultaneously maintain the log-spectral null depth to maintain the desired peak height and suppress the extraneous peaks. Since a fixed window cannot be wide enough for low pitched voices and still avoid doubling on high pitched voices, the window size must be adaptive. The window size required by low pitched speakers is also too wide to preserve short term staticnarity in the speech signal. This loss of stationarity is a compromise which must be made for such speakers but would cause unnecessary degradation of performance for higher pitched speakers.

In general, simple picking of the highest cepstral point within an allowed zone of possible pitch periods and above a certain threshold does not constitute a sufficiently reliable pitch detection and measurement scheme. The desired cepstral peak is subject to amplitude jitter due to the time window vs. time signal phase, staticnarity, and (voiced) signal to (all other) noise ratio. If the sound is voiced, the dominant cepstral peak usually indicates the correct pitch, but the peak amplitude varies. A more accurate scheme than simple peak height

thresholding is required for the voiced-unvoiced decision.

Even with all of the above refinements, a homomorphic pitch detector still makes occasional errors, especially at voicing boundaries. Some form of post processing to correct single errors imbedded in correct pitches is helpful. This smoothing, however, must preserve the voicing boundaries. A linear lowpass operation will not work (especially if an unvoiced frame is represented as a pitch of zero). The nonlinear operation of median smoothing [9] (pick the median of an odd number of the most current estimates) appears to be well suited to the application.

Finally, the pitch detector is implemented in the discrete domain. What sampling rates are required to adequately represent the signals involved? The original signal s(n) can be assumed to be Nyquist rate sampled. S(k), the discrete Fourier transform of w(n)s(n) must have at least as many samples as the time window w(n) to prevent loss of information. log|S(k)|, however, represents a sampled distorted S(f). To lessen the aliasing, log|S(k)| must have a much higher sample rate than that required to represent S(k). The worst case that this sampling rate must meet is the low pitched male speaker. If his fundamental frequency can reach 50 Hz., the sampling must be close enough to adequately represent a comb structure with peaks every 50 Hz. It

is not known how dense a sampling is required, but the sampling used in this implementation (7 Hz) appears to be adequate. It is possible, however, that an even denser sampling would yield improved results.

## V. The Implementation

As the best known judge of a pitch detector is the human ear, development and testing of the pitch detector was carried out by implementing the pitch detector (Figure 4) as part of a real-time vocoder. An existing LDSP [1] implementation of an LPC vocoder was used as the test vehicle. The operator could switch in real time between the homomorphic pitch detector and the original Gold-Rabiner [8] pitch detector, while using the same LPC spectrum analysis and synthesis (20 mS frame interval, 12th order autocorrelation LPC coded to 3.6 kbit). This allowed A-B comparison of the homomorphic pitch detector with a known algorithm.

The original speech signal is lowpassed at 3780 Hz, preemphasized for the LPC and 12 bit sampled at a 132 uS interval (7576 Hz). This sampled signal feeds both the LPC spectral analyzer and the Gold-Rabiner pitch detector. For input to the homomorphic pitch detector, the signal is again lowpassed at a digital frequency of $\pi/2$ (1894 Hz) and downsampled by a factor of

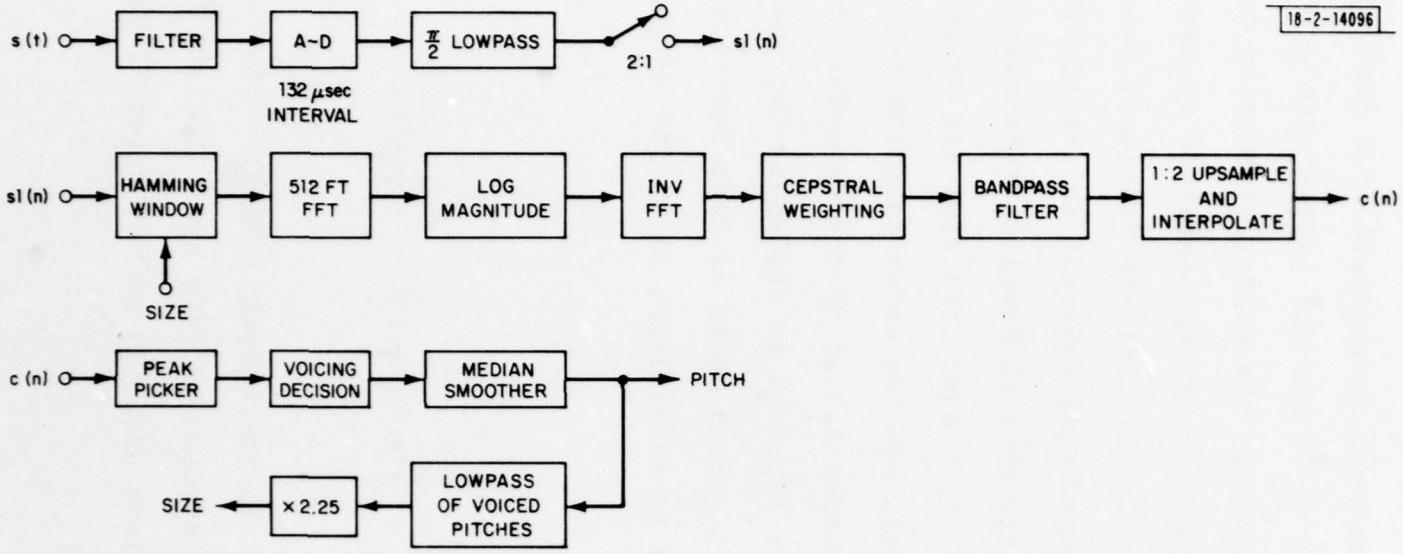



Fig. 4. Homomorphic pitch detector.

2. The downsampled waveform is windowed by a Hamming window of 100 to 250 samples (26 to 66 mS) and padded with sufficient zeros to fill a 512 point buffer. The buffer is Fourier transformed by a real FFT and the logarithm of the magnitude of each frequency point is computed. (Only the positive frequency points need be computed due to the symmetries inherent in the Fourier transform.) This log spectrum is Fourier transformed to produce a downsampled cepstrum. To prevent the spectral window from spreading the high amplitude low order cepstrum into the pitch period zone, the low order 2.3 mS of the cepstrum are zeroed before the cepstrum is filtered to implement the spectral window (Figure 5). The pitch period peaks of this function are now approximately leveled by multiplying it by the weighting function l(n)

$$l(n) = \begin{cases} 1 & n<21 \\ 1+.01(n-21) & 21\leq n\leq 128 \end{cases} \qquad (5)$$

and interpolated back to the original sample rate of 132 uS to create the modified cepstrum (Figure 6). (Periods measured on this cepstrum now relate directly to the original waveform when measured in samples.)

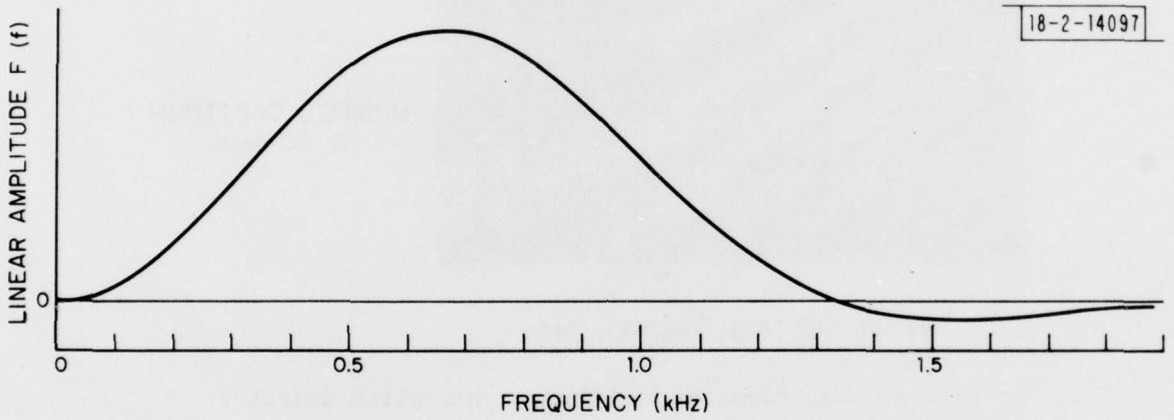



Fig. 5. Log spectral window. (Implemented as a filter on the Cepstrum) Impulse response: f(0) = .52203, f(1) = f(-1) = .23590, f(-2) -.24475, f(3) = f(-3) = .22653, f(4) = f(-4) = -.02637.

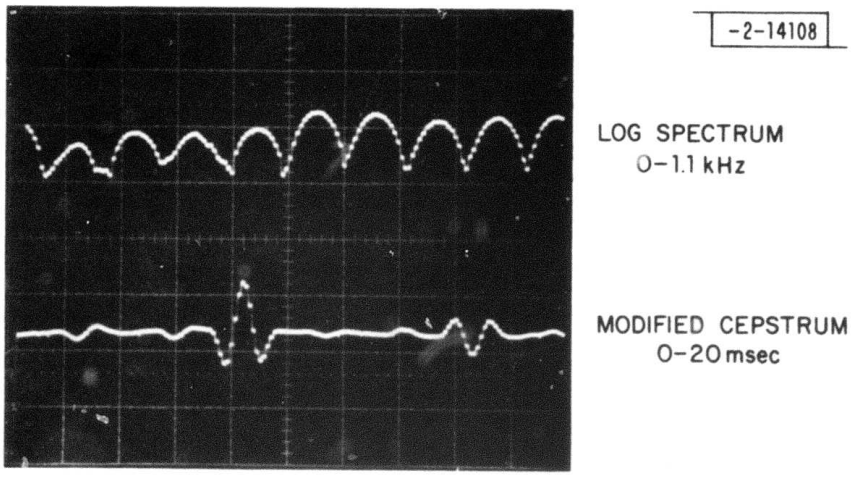



Fig. 6. Waveforms of the homomorphic pitch detector.

A modified peak picker is now applied to the cepstrum. The picker selects the highest point in a window (19 to 152 samples corresponding to pitch periods of 2.5 to 20 mS) and, if the height of the point is above a given threshold and the point is not the lowest order point in the window, chooses this point as the pitch estimate. If the highest point is at the bottom edge of the window, it is assumed to be the side of a higher spectral information peak outside of the window. The threshold (see Table 1) is intentionally low so that any doubtful frames are initially chosen to be voiced. If a pitch estimate is not found, the frame is declared to be unvoiced, which is signaled by a pitch estimate of zero.

As the pitch of voiced speech usually varies smoothly, continuity of the pitch estimates can be used to improve the reliability of the voiced-unvoiced decision. The current frame estimate is given a coincidence score equal to the number of times that its pitch is within 8 samples (1.1 mS) of the pitch of the two adjacent frames. (The pitches of the adjacent frames are the raw outputs of the peak picker rather than an additionally processed estimate to prevent potentially unstable feedback loops.) This score is now used to choose which of three thresholds is to be compared to the current cepstral pitch peak height to decide whether the current frame is voiced. These

thresholds, in arbitrary units, are:

| coincidence score | peak height threshold |
|---|---|
| 0 | 18 |
| 1 | 11 |
| 2 | 8 |
| peak picker | 8 |

Voicing decision thresholds
Table 1

(The units are a function of the implementation choice of the base of the logarithm, the binary points of the logarithm, the spectral window, the gain constant in front of the inverse FFT, and the cepstral weighting function.)

A similar scheme which used the two previous and two following pitch estimates was also tried. Comparison of the current estimate and this set of four yields ten ordered coincidence groups if mirror images are considered identical. Each of these coincidence groups had an associated threshold which was compared with the cepstral peak height of the current frame to make the voiced-unvoiced decision. This scheme yielded only a slight improvement over the simpler scheme and was judged not worth the additional complexity.

The pitch estimates are now fed through a third order median smoother [9]. This operation will remove single errors without shifting the voicing boundaries. The median smoother also frequently corrects erroneous pitch estimates at voice onset and termination and generally yields smoother sounding speech by removing some of the jitter on the pitch track. (Median smoothers will handle unvoiced frames correctly if they are represented by a pitch period of zero.)

The output of the median smoother is used in two ways: it is passed to the synthesizer as the best estimate of the pitch and fed to the adaptive time window size routine. The next window size is computed as follows:

$$\text{size}(n+1) = \begin{cases} \text{size}(n) & p=0 \\ 2.25\,(.9\,\text{size}(n) + .1\,\text{pitch}) & p>0 \end{cases} \qquad (6a)$$

$$\text{limits:} \qquad 26.4 \text{ mS} \leq \text{size} \leq 66.0 \text{ mS} \qquad (6b)$$

(As the window is in the downsampled domain, its true size is 4.5 average pitch periods.) Due to the delays in the voiced-unvoiced decision and the median smoother, the pitch is delayed several frames and therefore the window size is changed several frames late. As the window size need not track the pitch accurately,

this delay appears to cause no degradation of the results.

The above operations are all performed with 16 bit arithmetic. The data for the FFTs are stored in block floating point with right shifts only as required to prevent overflow.

## VI. Results

No clearly defined objective method for the testing of pitch detectors exists. Errors can be of several forms: small pitch errors, gross pitch errors and voicing decision errors. The perceptual significance of each of the errors is a function of the listener, the speaker, the spectral analysis-synthesis algorithm, and where in the speech each error occurs. Therefore, many investigators (including this one) fall back on subjective judgements by trained listeners who can frequently classify the type of error as well as its presence. As the pitch detectors examined here are implemented in a real-time vocoder with real-time displays of the windowed speech, the log-spectrum, the cepstrum, and the pitch track, these observations of performance are based on hours of listening time by several trained listeners who could simultaneously observe the internal workings of the pitch detector as they listened.

On clear speech, the homomorphic pitch detector and the

Gold-Rabiner algorithm perform similarly for male speakers. For female speakers, the homomorphic algorithm makes fewer errors than the Gold-Rabiner algorithm. Both pitch detectors only make occasional errors which are likely to be perceived as glitches in the speech. In silent intervals, the homomorphic pitch detector occasionally finds the pitch of the background 60 Hz power line hum. This, however, is of no perceptual significance as the energy of the synthesized output is too low for the output to be audible. Another characteristic error of the homomorphic pitch detector is occasional "squeaks" caused by spectral envelope information appearing in the pitch zone of the cepstrum being analyzed as a high pitch. The homomorphic algorithm also determines voiced fricatives to be voiced which appears to be a perceptually appropriate decision.

The differences between the pitch detectors become much more obvious when applied to corrupted speech. Both pitch detectors were tested on speech corrupted with additive noise characteristic of the interior of a large jet aircraft [11]. This noise exhibits a broad spectral peak below about 600 Hz. At a signal to noise ratio of about 10 db, the noise degrades the Gold-Rabiner algorithm more than the homomorphic pitch algorithm. The perceptual form of the errors is quite different for the two pitch detectors. The Gold-Rabiner pitch detector jumps in and

cut of voicing at a high enough rate to chop up the speech. (Completely removing the pitch detector and declaring all frames unvoiced would probably be more intelligible.) The homomorphic pitch detector gives fairly good pitch estimates except for occasional zones where it devcices. These zones tend to be of a syllabic duration and are perceived as a devoiced syllable and therefcre appear to do less damage than the chopping tc the intelligibility of the speech. The homomcrphic pitch detector even yields reasonably ccrrect analyses when the signal to noise ratio is so low that the output of the vcccder is unintelligible due to errors in the LPC spectrum analysis.

Comparisons of the twc pitch detectcrs were made with narrowband additive noise. The noise used here is a 100 Hz sine wave. At a signal tc ncise ratio cf abcut 0 db, the homomorphic pitch detector occasionally finds the pitch of the noise during speech silences but is ctherwise unaffected. Under the same conditions, the Gold-Rabiner pitch detector yields badly "chopped up" pitch. At a signal tc noise ratic cf about 10 db the homomorphic pitch detector is essentially unaffected while the Gold-Rabiner pitch detector still yields badly "chopped up" pitch. As the signal tc ncise ratio increases tc about 30 or 40 db, this "choppiness" gradually decreases and vanishes. (The Gold-Rabiner pitch detector should find the pitch of the sine

wave during the silences as it is a direct waveform measurement type of pitch detector. This effect, which is distinct from the "choppiness", ceases above a signal to noise ratio above about 30 or 40 db where the sine wave drops below an energy threshold.)

Comparison of the pitch detectors on telephone degraded speech also indicated differences in the performance of the two pitch detectors. The telephone simulator [10] which was used for the tests has two sets of parameters: "mid" representing a 50th percentile continental US long-distance voice-grade line and a "poor" representing a 90th percentile continental US long-distance voice-grade line. Both settings attempt to simulate the bandpassing, Gaussian and pulse noise, phase distortion, frequency distortion, and nonlinearity of the respective telephone lines. The "mid" telephone line causes essentially no degradation of the homomorphic pitch detector but causes some annoying oscillation in and out of voicing by the Gold-Rabiner pitch detector. On the homomorphic pitch detector, the "poor" telephone line causes some devoicing of syllabic duration and a few "squeaks", neither of which seriously impair intelligibility. The Gold-Rabiner algorithm exhibits severe devoicing and rapid oscillation in and out of voicing which cause severe damage to the intelligibility of the speech. As the telephone simulator high-pass filtered the speech with a cutoff of about 300 Hz,

significant amounts of the information used by both pitch detectors were removed. Versions of the homomorphic pitch detector which used a spectral window which deemphasized the low frequency region exhibited less degradation due to the telephone simulator.

## VII. Discussion

Clear speech allows a pitch detector many design options. Direct waveform processing and measurements are possible. These techniques, which allow the pitch detector to analyze nonstationary voicing almost as effectively as stationary voicing, degrade in the presence of noise. Waveform pitch detectors can no longer accurately locate peaks and zero crossings which may be obscured by additive noise. Distortion for spectral leveling [7], a commonly used preprocessing technique for correlation type pitch detectors, now creates interfering cross terms between the noise and the speech.

Pitch detection in the presence of noise requires the use of the coherence found in the voiced excitation to differentiate the signal from the noise. Pitch detectors which are robust with respect to input speech degradation therefore must yield some of their clear speech performance on nonstationary voicing. The homomorphic pitch detector attempts to explcit this coherence in

several ways to achieve its robustness. Generation of the complex spectrum exploits (and requires) the coherence of the voiced excitation. Taking the magnitude of this complex spectrum maximizes the phase coherence of its periodic line structure. The logarithm, in conjunction with the original time window (which sets the line shape and width), maps the the line structure of the magnitude spectrum into a relatively constant amplitude line structure plus some slowly varying terms. (This log magnitude operation degrades gracefully in the presence of noise by the nulls of the log spectrum becoming "filled in" by the noise.) The second FFT then exploits the coherence of this constant amplitude periodic line structure to generate the cepstral peak which indicates the presence and pitch of voicing.

The majority of the homomorphic pitch detector errors fall into two classes. The pitch is sometimes estimated incorrectly at voicing onset or termination. Nonstationarity in the voicing in both pitch and amplitude tends to concentrate at these points making them obvious trouble spots for a coherent pitch detector. (A possible scheme to improve the pitch detector's tolerance to voicing nonstationarity is outlined in Appendix A.)

The second difficulty is the voiced-unvoiced decision. If the decision is heavily biased toward voicing so that few voiced to unvoiced errors occur, the pitch estimates generally appear

accurate. The voiced-unvoiced decision is based on a pitch continuity dependent threshold placed on the cepstral peak height. What, then, affects this parameter? The height of the peak is the strength of a frequency component in the windowed log spectrum. The strength of this component is a function of the periodic signal-to-noise ratio. (Noise "fills in" the nulls of the comb spectrum and reduces the amplitude of the component.) Varying frequency and amplitude of the voiced excitation reduce the coherence of the voicing and spread and reduce the height of the peaks of the comb spectrum. The vocal tract filter (or audio channel) can have differing delays for the different harmonics of the source which, if the pitch is changing, will spread and lower the cepstral peak. The phase effects of the vocal tract filter and channel are of no consequence since the log spectrum is phase insensitive, except that changing phase shifts in the tract and channel will shift the frequencies of the harmonics and degrade the cepstral peak. Clearly a more accurate measure of voicing signal to noise ratio is desirable.

The spectral window remains the most mysterious part of the pitch detector and a possible target for future development. (The window design is a simultaneous dual domain design problem.) The best windows that were found tended to be quite different from those that one would expect to be optimum. The best so far

appears to have a single broad peak with an upper cutoff at about 1 to 1.2 kHz (Figure 5). Tests of the classic rectangle with smoothed discontinuities give poor results for no apparent reason. Tailoring of the window to specific applications appears to be possible where the signal-to-noise ratio varies over the spectrum. Specific attempts have not been made to design a window for either of the corruptions mentioned earlier, but two windows which give similar performance on clear speech give varying performance on the corrupted speech.

Some of the earlier work on the homomorphic algorithm suggests the possibility of integration of the spectral analysis and the pitch detector [5,6]. This investigation suggests that performance may suffer as a result of such a sharing of processing. Succeeding work has indicated that performance of the spectral estimator over a wide range of pitches requires an adaptive time window which is much smaller (about 3 pitch periods) than the window required for the pitch detector. This would prevent the sharing of the initial FFT as well as any of the succeeding processing.

## VIII. Summary

The homomorphic pitch detector has existed for quite some time in a comparatively undeveloped state due to the difficulties

of real-time implementation. This investigation reveals that, with suitable modification, it is capable of clear speech performance rivaling one of the better pitch detectors in current use. Tests on two forms of corrupted speech indicate the homomorphic algorithm is also more robust with respect to corruption of the speech than is the Gold-Rabiner algorithm, which allows the effective use of a vocoder in a less than ideal environment. As all of the operations required to achieve this performance can likely be implemented in real-time using CCD technology and microprocessors, this pitch detector can probably be built with fairly simple hardware.

# IX. Appendix

The homomorphic pitch detector is one of a class of coherence seeking pitch detectors. These pitch detectors generally use signal processing techniques to process the speech waveform into a form more suitable for pitch detection and estimation. They tend, however, to degrade when the speech deviates form the ideal model in any of several ways. This appendix describes a pilot study of a set of preprocessors for improving the speech generation model error tolerance of coherence seeking pitch detectors.

Basically, the homomorphic pitch detector uses homomorphic techniques for processing the speech signal into a waveform which contains a peak to indicate the presence and pitch of voicing. The problem then becomes detection of the peak's location and a decision based on its height in the face of noise on the waveform.

A large number of factors, as enumerated earlier in this report, lessen the peak height and reduce its discriminability. One of these factors is nonstationarity in the voiced excitation. This nonstationarity can appear in two forms: amplitude and pitch. The amplitude nonstationarity is most severe at voicing boundaries where the perceptual consequences of an error are

minimal. Pitch nonstaticnarity, while frequently severe for a few pitch periods at voicing boundaries, cccurs thrcughout speech. It may be possible tc modify the homomorphic pitch detector to provide a greater tclerance tc these viclaticns of the analysis model.

One means of increasing tolerance tc a particular model error is to postulate the errcr and mcdify the algorithm such that it can function in the presence of the error (even if it fails without the error). A mapping of the speech signal to regenerate a violated assumpticn is one such means. If the original signal is increasing in amplitude, multiply it by a decaying function (such as a decaying expcnential) to restore its amplitude stationarity. If the pitch period of the voiced excitation is changing, time warp the waveform to return it to a ccnstant periodicity signal.

No models exist which give a functicnal descriptiecn of pitch changes. The pitch, however, generally varies smoothly and not toc rapidly. Postulate a periodic signal:

$$f(t)=f(t+p) \qquad \text{(A1)}$$

t=time
p=pericd

Time warp f to create a signal of varying period:

$$g(t) = f(s(t)) \qquad \text{(A2)}$$

s=warp function

$$Pg(t) = (dt)/(ds)\,p \qquad \text{(A3)}$$

Pg=period of g

Assume a linear pitch change:

$$Pg(t) = (1+at)\,p \qquad \text{(A4)}$$

To recover $f(s) = g(t(s))$:

$$ds = (p/Pg)\,dt \qquad \text{(A5)}$$

$$s(t) = \int_0^t 1/(1+at)\,dt \qquad \text{(A6)}$$

$$= (1/a)\log(1+at) \qquad \text{(A7)}$$

$$t(s) = (1/a)\,(e^{as} - 1) \qquad \text{(A8)}$$

$t(s)$ = correcting warp

$$\sim s + (a/2)\,s^2 \qquad \text{(A9)}$$

as small

Thus a linear time warping of a linearly varying pitch signal will approximately map the signal into a constant pitch signal.

Non-real-time simulations of a time warp feeding the homomorphic analyzer have been performed on speech. The time warp was implemented with a 51 section lowpass filter to interpolate the input waveform. The cepstrum (c(n)) as computed by the procedure of Figure 4 was considered the final output from which judgements of the pitch peak enhancement were made. As illustrated in Figures A1 and A2, instances of peak enhancement were found in a short section of speech which was searched.

(The positions of the pitch peaks are skewed as a function of $a$ as t=0, i.e., the point of zero warp, was at the left edge of the Hamming window. In a real implementation, this effect would be corrected so that the point of zero warp would be the center of the window thus removing the dependence of the peak position on $a$.)

A priori, one cannct know whether time warping (or amplitude correction) will enhance the pitch period peak in the cepstrum. An implementation using any of these techniques would therefore be required to run one full pitch detector per preprocessor and base its final decision on the outputs (or intermediate results) of all. As specific attempts will then have been made to correct for deficiencies of the basic pitch detector the voicing decision thresholds of the individual pitch detectors might be raised to lessen the probability of a false voicing detection.

Such a pitch detector might give improved performance in several ways. As most of the clear speech errors of the homomorphic pitch detector occur at voicing boundaries (which are the regions of highest nonstationarity), the technique might prevent some of these errors. In fact, Figures A1 and A2 and most other zones where the technique was found to enhance the pitch peak were at or near voicing boundaries.

The pitch detector might also become more robust. If the speech is degraded by additive noise, the pitch peak height would be reduced. The major perceived error in airborne command post noise is devoicing for occasional sections of approximately syllabic length. If this is a cumulative effect of the noise and nonstationarity in the original speech reducing the peak height, the preprocessor might be able to sufficiently correct the

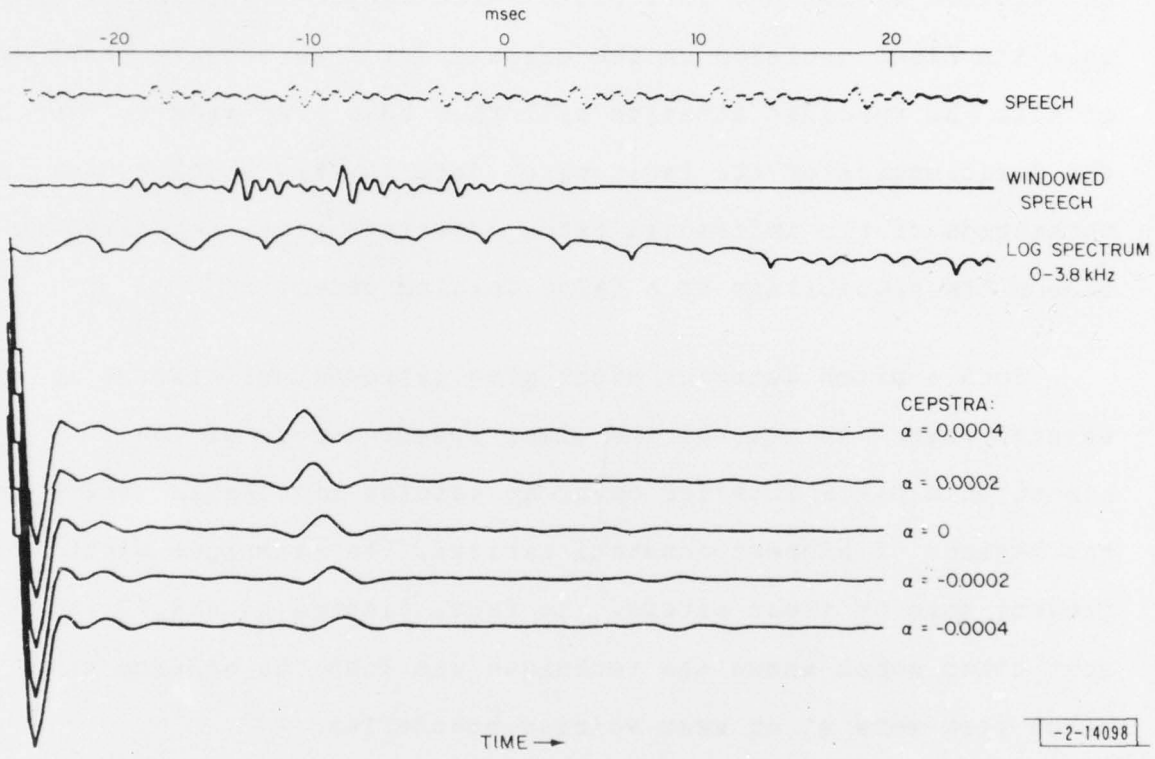



Fig. A-1. Cepstra with time warping.

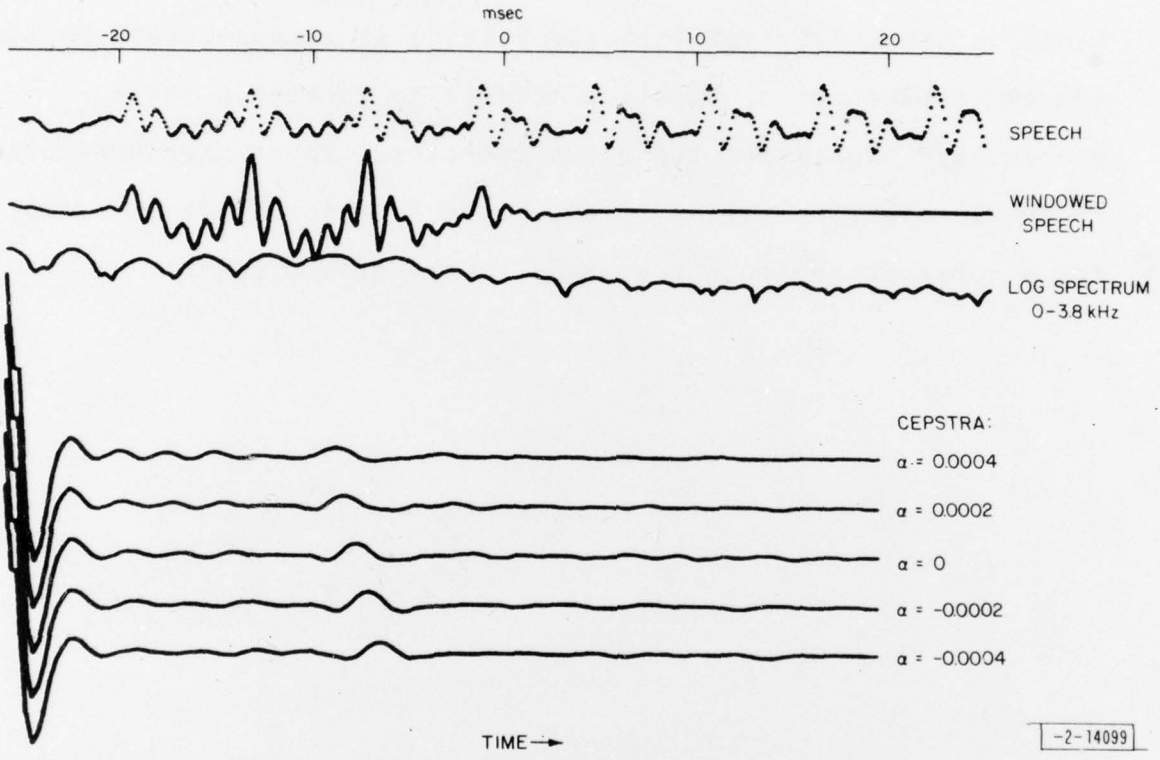



Fig. A-2. Cepstra with time warping.

nonstationarity to allow pitch detection.

Both of the preprocessing techniques postulated here are based on the assumption that the voicing is nonstationary in some way or combination of ways and attempt to provide a "more stationary" signal for the pitch detector. These preprocessors therefore might be useful to any pitch detector which searches for a coherent periodic component in the speech signal.

## X. Bibliography


1. The Lincoln Digital Signal Processor is an advanced version of the LDVT. See P. E. Blankenship et al., "The Lincoln Digital Voice Terminal System," Technical Note 1975-53, Lincoln Laboratory, M.I.T. (25 August 1975), DDC AD-A017569/5; or P. E. Blankenship, "LDVT: High Performance Minicomputer for Real-Time Speech Processing," EASCON'75, pp. 214a-214g.

2. J. L. Flanagan, Speech Analysis Synthesis and Perception (Springer-Verlap, New York, 1972).

3. E. M. Hofstetter et al., "Vocoder Implementations on the Lincoln Digital Voice Terminal," EASCON '75, pp. 32a-32j.

4. P. Lieberman, J. Acoust. Soc. Am. 33, 597-603 (1961).

5. A. M. Noll, J. Acoust. Soc. Am. 41, 293-302 (1964).

6. A. V. Oppenheim and R. W. Schafer, IEEE Trans. Audio Electroacoust. AU-16, 221-226 (1968), DDC AD-678238.

7. L. R. Rabiner, IEEE Trans. Acoust., Speech, and Signal Processing ASSP-25, 24-33 (1977).

8. L. R. Rabiner and B. Gold, Theory and Applications of Digital Signal Processing (Prentice-Hall, Englewood Cliffs, New Jersey, 1975).

9. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, IEEE Trans. Acoust., Speech, and Signal Processing ASSP-23, (1975).

10. S. Seneff, "A Real-Time Digital Telephone Simulation on the Lincoln Digital Voice Terminal," Technical Note 1975-65, Lincoln Laboratory, M.I.T. (30 December 1975), DDC AD-A021409/8.

11. Speech tapes containing noise characteristic of an airborne command post were supplied by the Defense Communications Agency.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>ESD-TR-78-252 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE *(and Subtitle)*<br>Homomorphic Pitch Detection | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical Note<br>6. PERFORMING ORG. REPORT NUMBER<br>Technical Note 1978-32 |
| 7. AUTHOR*(s)*<br>Douglas B. Paul | | 8. CONTRACT OR GRANT NUMBER*(s)*<br>F19628-78-C-0002 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Lincoln Laboratory, M.I.T.<br>P.O. Box 73<br>Lexington, MA 02173 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>Program Element No. 62706E<br>Project Code 8P10<br>ARPA Order 2006 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Defense Advanced Research Projects Agency<br>1400 Wilson Boulevard<br>Arlington, VA 22209 | | 12. REPORT DATE<br>15 August 1978<br>13. NUMBER OF PAGES<br>46 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*<br>Electronic Systems Division<br>Hanscom AFB<br>Bedford, MA 01731 | | 15. SECURITY CLASS. *(of this report)*<br>Unclassified<br>15a. DECLASSIFICATION DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

homomorphic pitch detector
real-time LPC vocoder
broadband noise
narrowband noise

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This note describes a homomorphic pitch detector which yields good performance on clear speech and moderate robustness to additive broadband noise, narrowband noise, and to degradation by a telephone simulator. It achieves the performance by the use of an adaptive time window, log-spectral windowing, an adaptive voicing threshold, and pitch track smoothing. It has been implemented in a real-time LPC vocoder for testing. Finally, a pilot study of a preprocessor to improve the performance of any coherence seeking pitch detector is presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*