MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE VISTA SPEECH ENHANCEMENT SYSTEM FOR AM RADIO BROADCASTING

T.F. QUATIERI
J.T. LYNCH
M.L. MALPASS
R.J. McAULAY
C.J. WEINSTEIN
Group 24

FINAL TECHNICAL REPORT

29 JANUARY 1990

LEXINGTON MASSACHUSETTS

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE VISTA SPEECH ENHANCEMENT SYSTEM FOR AM RADIO BROADCASTING

T.F. QUATIERI
J.T. LYNCH
M.L. MALPASS
R.J. McAULAY
C.J. WEINSTEIN
Group 24

FINAL TECHNICAL REPORT

29 JANUARY 1990

LEXINGTON MASSACHUSETTS

ABSTRACT

A new approach to speech enhancement has been developed for increasing average transmission power in the Voice of America broadcast system. The approach uses a sinusoidal analysis/synthesis framework and integrates phase dispersion, amplitude compression, and spectral shaping to decrease the peak/RMS ratio of the speech waveform so that average transmission power can be increased subject to the peak power limit of the existing transmitters. The processing algorithms adapt dynamically to speech pitch and spectrum, and include a phase dispersion technique adapted from radar signal design which minimizes short-time peakiness of the speech waveform while maintaining the original spectral envelope to minimize perceived distortion. Overall, an advantage of about 3 dB in peak/RMS has been achieved relative to commercial devices with quality which has been judged to be cc hle for the expected conditions of operational broadcast environments. In or ow tests of performance in the real broadcast environment and to investiga .eoffs in degree of peak/RMS reduction versus received speech quality, a real-t: p· has been implemented in the form of a multi-processor based on high-perform- igital signal processing chips. The prototype also provides experimental flexibility through control of the degree of processing (allowing mild, normal, and severe), and allows on-line monitoring of the peak/RMS ratio.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1 Overview

This report describes a research and development effort in audio signal enhancement conducted at MIT Lincoln Laboratory over the period March 1986—March 1989 under the sponsorship of the United States Information Agency/Voice of America (USIA/VOA). The purpose of the effort has been to develop and implement digital signal processing techniques for pre-transmission enhancement of the VOA baseband audio signal. The processing techniques are aimed at reducing the ratio of the peak transmitted speech signal to its RMS level (referred to as peak/RMS ratio), while maintaining speech intelligibility and quality. Peak/RMS reduction increases the average transmission power subject to a peak power limitation at the transmitter, and thus increases robustness of the broadcast system to natural and man-made disturbances.

The signal processing algorithm development effort has focussed on development and test of adaptive signal processing algorithms for spectral shaping, phase dispersion, and amplitude compression. The work has concentrated on reducing the peak/RMS of speech (rather than of music), and has taken advantage of the properties of the speech signal in enhancement algorithm development. The algorithm research and development work has culminated in an enhancement algorithm based on a sinusoidal analysis/synthesis model [1,2] (this model had been developed earlier at Lincoln for speech coding and speech transformation applications), and has achieved significant reduction in peak/RMS ratio with good quality. In a clear listening environment, the intelligibility of the original speech is essentially maintained; while in a noisy listening environment, it is improved when the processed and original speech are compared under a peak constraint. Informal listening comparisons also indicate that the sine-wave preprocessor can achieve about 3 dB more peak/RMS reduction than state-of-the-art commercial units (that were available at the time of the system development) with quality that has been judged to be comparable for the expected conditions of operational VOA broadcast environments.

This algorithm was initally developed and tested in a non-real-time simulation on a general purpose computer facility. During the final year of the project, a prototype multiprocessor system

was developed which implements the algorithm in real time. The real-time system was utilized for additional algorithm development and tuning. A stand-alone version of the prototype real-time processor was delivered to the VOA in March 1989 for future tests on the VOA broadcast system.

The system which has been developed, including the algorithms and implementation, has been named VISTA (Voice Intensification using the Sinusoidal Transformation Algorithm). This name will be used frequently in this report.

1.2 Organization of Report

The organization of this report is as follows. Section 2 provides a summary of the problem being addressed, the goals, the technical approach, and the results achieved in this work. Section 2 is designed to function as a capsule summary of the work, and all items covered in that section are described in more detail in the remainder of the report.

The technical approach to speech enhancement using the sinusoidal analysis/synthesis model is described in detail in Section 3. The sinusoidal analysis/synthesis framework, the algorithm for adaptive phase dispersion using a radar signal design technique, and strategies for amplitude compression and frequency shaping are described. The section concludes with a description of the overall integrated VISTA algorithm, including some representative performance results and algorithm tradeoffs. More detailed results are deferred to Section 5.

The implementation of VISTA in a real-time prototype system using multiple digital signal processing boards is described in Section 4. This section covers: the multiprocessor hardware structure; problems and solutions in moving from a non-real-time, single-processor, floating-point simulation to a real-time, multiprocessor, fixed-point implementation; the multiprocessor software structure and modularization; and the user interface.

Test and evaluation of the enhancement system, which was a continuing process throughout the VISTA development effort, is described in Section 5. Evaluation criteria of two types are described: (1) performance measures which affect transmission range, including peak/RMS ratio and associated transmission power considerations; and (2) speech performance measures including quality,

intelligibility, and performance in noise. Evaluation facilities are described, including a laboratory facility which was developed to allow comparison of VISTA with state-of-the-art commercial audio processors, and arrangements for on-air VOA tests. Speech data bases used for testing are described. Finally, evaluation results are detailed, including comparisons of both the non-real-time and real-time versions of VISTA against state-of-the-art commercial processors, in the dimensions of peak/RMS ratio reduction, speech quality, and speech intelligibility.

Finally, Section 6 summarizes conclusions of this work and outlines areas for potential future work in audio enhancement for broadcast.

2. SUMMARY OF GOALS, TECHNICAL APPROACH, AND RESULTS

2.1 Description of Problem and Goals

The VOA broadcast system includes a number (about 140) of large AM radio transmitters stationed around the world, and operating in the High-Frequency (HF) radio band. The goal of the effort described in this report is to develop techniques for increasing the effective average broadcast power of these transmitters in the presence of natural background noise and man-made interference, including jamming. This is to be accomplished subject to two constraints: (1) each transmitter is subject to a peak power limit; and (2) the home receivers are not under control of the broadcast system designer and, hence, cannot be expected to change. The approach taken in this work is to apply digital signal processing techniques to the speech signal to decrease its peak/RMS ratio prior to transmission and, hence, to increase the average power that can be transmitted under a fixed peak power constraint. The alternative of building bigger transmitters to increase the allowable peak power is prohibitively expensive, so that any reductions made in peak/RMS by digital processing would offer significant economic advantages to VOA.

The problem scenario and desired functions of the speech processor are illustrated in Figure 2-1. The typical input speech waveform is peaky, has a large dynamic range, and has a large peak/RMS (typically 12-14 dB). The goal of the speech processor is to apply phase dispersion, amplitude compression, and spectral shaping in such a way as to produce an enhanced speech waveform which is more dense, has a reduced dynamic range, and has reduced peak/RMS. This is to be accomplished with minimal degradation to the speech quality or intelligibility. To advance the state-of-the-art beyond what was achievable with commercial audio processors, about 7–9 dB reduction in peak/RMS (relative to the original speech) was needed.

2.2 Enhancement Approach

The approach used in processing the speech signal is illustrated in Figure 2-2. A sinusoidal analysis/synthesis framework developed earlier at Lincoln [1,2] is used to decompose the speech into

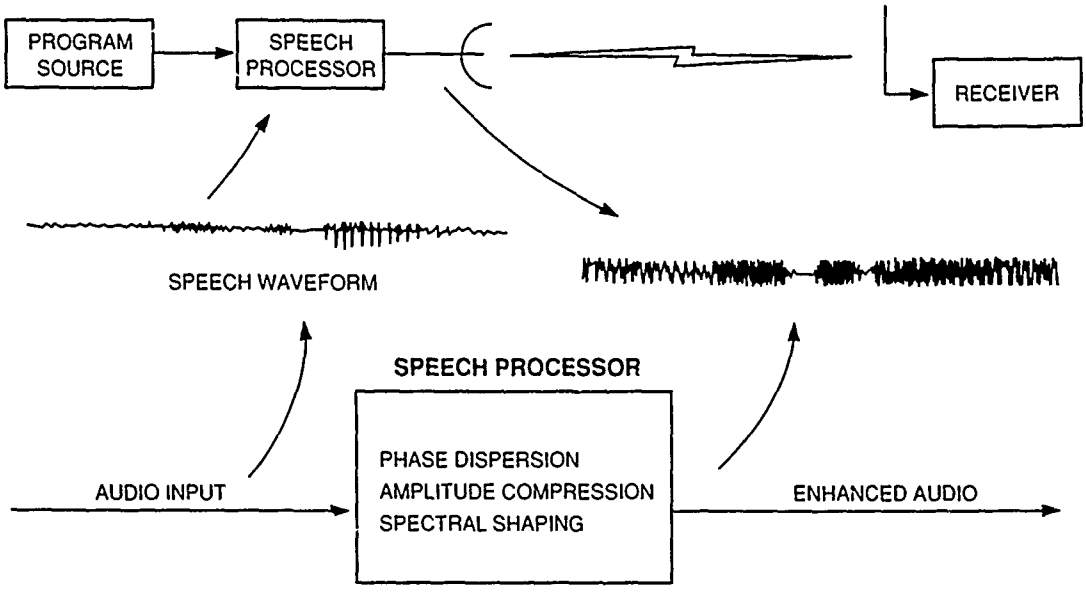


Figure 2-1. Input/output preprocessor waveform characteristics.

a sum of sinusoids with time-varying amplitudes, frequencies, and phases. It was demonstrated in earlier work that, with proper matching and interpolation of these sinusoids over time, a synthetic speech waveform could be produced which was perceptually indistinguishable from the original speech. In this effort, peak/RMS ratio reduction is achieved by an enhancement module which applies phase dispersion, amplitude compression, and spectral shaping, in the frequency domain, to the sinusoidal components of the speech. The approach is distinguished from standard commercial audio enhancement techniques [3] in: (1) the sinusoidal analysis/synthesis framework; (2) the use of adaptive processing; particularly, adaptive phase dispersion to reduce local waveform peakiness; and (3) the coupling of amplitude compression and spectral shaping to phase dispersion. The adaptive phase dispersion technique noted above is also unique in that it applies a technique, due to Key, Fowle, and Haggarty (KFH), originally developed in radar signal design [4] to reduce peak/RMS without modifying the speech spectral envelope.

A more detailed block diagram of the enhancement system is shown in Figure 2-3. Items to be noted are: the exploitation of the properties of the speech signal (pitch, voicing, vocal tract spectral envelope); (2) the signal-dependent adaptive control of the enhancement algorithm; and (3) the tight coupling of spectrum shaping, dynamic range control, and the KFH phase dispersion.

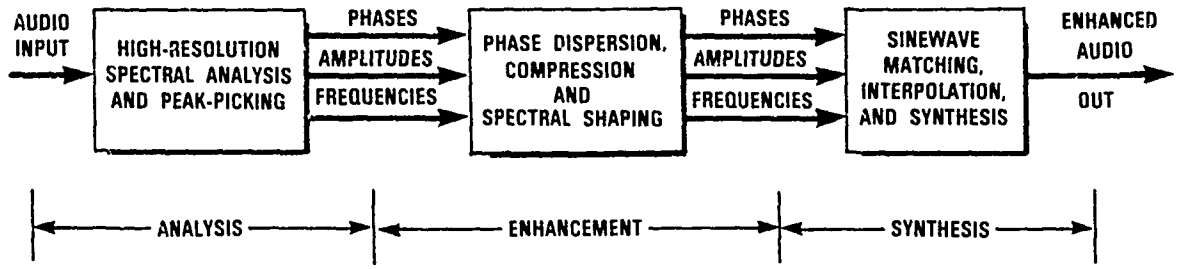


Figure 2-2. Sinusoidal transform enhancement system.

The overall result, as described in more detail below, was an achievement of about 7-9 dB reduction in peak/RMS. Informal evaluations indicate that this represents about 3 dB more reduction over conventional methods. This peak/RMS reduction adds about 3 dB to the average sideband power radiated over the broadcast reception area. (This coverage area is often referred to as the "footprint" in the context of shortwave ionospheric propagation.) The resulting improvement in signal-to-noise ratio (about 3 dB) implies improved listenability over the coverage area.

2.3 System Implementation

The enhancement system, which is referred to as the VISTA (Voice Intensification by the Sinusoidal Transformation Algorithm) system, was initially developed and tested in simulation form on a general purpose, speech research computer facility featuring a SUN3 workstation. When it became clear that substantial improvements in peak/RMS had been achieved with good speech quality, a real-time VISTA prototype using multiple digital signal processor boards was developed, and the real-time software was developed and tested.

The prototype system, which includes seven signal processor boards based on the Analog Devices ADSP2100 processor chip [5], and a microprocessor-based controller board is shown in Figure 2-4. The system is controlled by means of a keyboard and display which allow the user

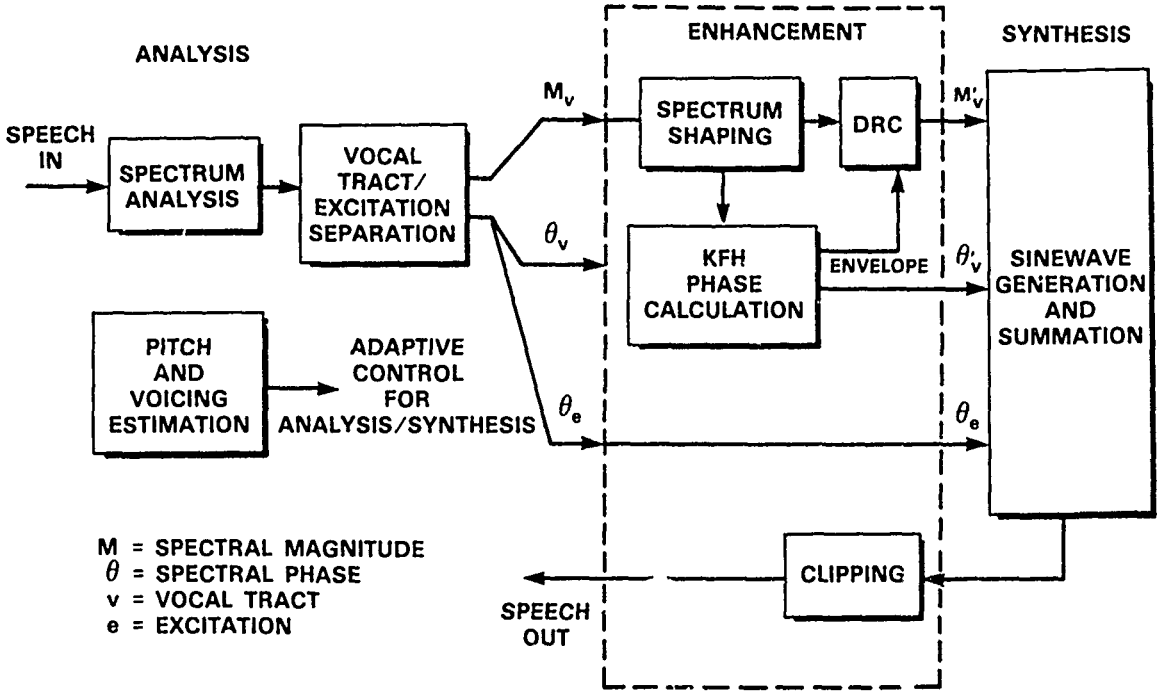


Figure 2-3. Sinusoidal transform system for audio enhancement.

to select a variety of enhancement options including mild, normal, and severe processing. These options produce an increasing degree of peak/RMS reduction, while producing more noticeable quality degradation as the degree of enhancement is increased.

The VISTA prototype was delivered to VOA in March 1989 and is currently operating successfully in the VOA's audio laboratory in Washington, D.C. while plans are being formulated by VOA for live tests over the VOA broadcast system.

2.4 Test and Evaluation Methods and Results

Test and evaluation of the enhancement system described above has been a continuing process throughout the algorithm development. Performance of the developing VISTA system has been compared to state-of-the-art commercial audio enhancement units. A measurement facility was set up to allow calibrated peak/RMS measurements, quality comparisons, and intelligibility testing. In addition, in February and June 1988, tapes processed through the non-real-time enhancement

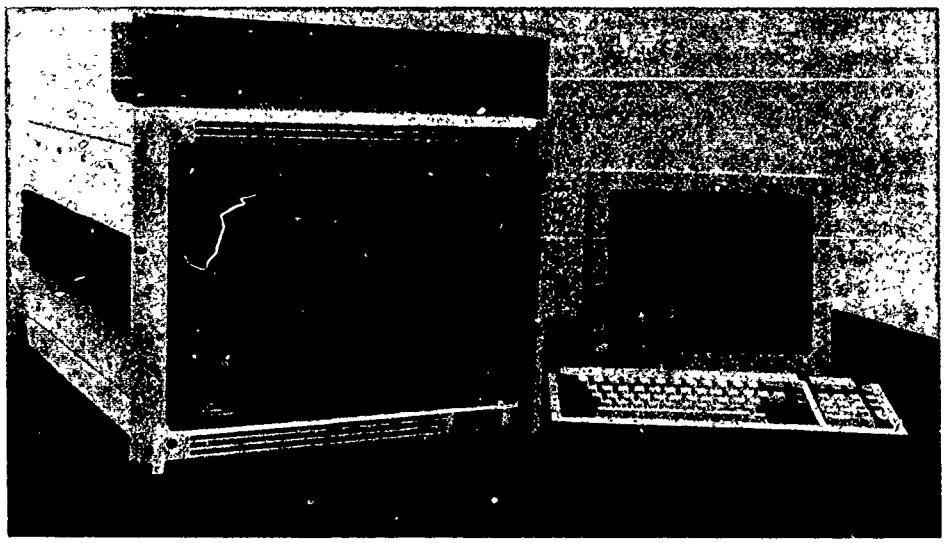


Figure 2-4. VISTA prototype system. VISTA = Voice Intensification by the Sinusoidal Transformation Algorithm.

system were delivered to VOA and transmitted over the VOA channel in live on-air tests Data bases used for testing included: a phonetically-balanced sentence data base; a sentence data base designed specifically for the Sentence Verification Test (SVT); and broadcast program material provided by VOA.

Detailed results of testing, including a selection of intermediate results obtained during the course of the effort, are presented in the body of this report. A representative example of final results obtained with the real-time VISTA system, using a data base of VOA broadcast material, is shown in Figure 2-5. In this case, the output of each processor was AM-modulated and passed through a shortwave radio in the laboratory. The horizontal axis shows peak/RMS ratio reduction relative to the original speech prior to AM-modulation. The vertical axis provides a measure of relative subjective speech quality, on an arbitrary 1-10 scale, at the output of the shortwave radio. The quality judgments were performed using paired comparisons, so that systems at the same level were judged informally to have "equivalent quality". The reduction in quality with greater degrees of enhancement processing (and more peak/RMS reduction) is apparent in the Figure. These

results appear to indicate that, in "NORMAL" enhancement mode, VISTA achieves more than a 3 dB peak/RMS reduction over a state-of-the-art commercial system at comparable quality. The "NORMAL" enhancement mode was judged to be a mode in which the quality degradation would be only slightly noticeable in a real broadcast environment subject to propagation effects and noise. Quality evaluations were, however, subjective and informal. In addition, due to the steepness of the functions in Figure 2-5, quality can change significantly both with speaker and text and with small change in the degree of processing. Consequently, more formal quality testing is needed to provide an objective determination of "equivalent quality" [6]. This testing should be performed in the VOA broadcast environment.

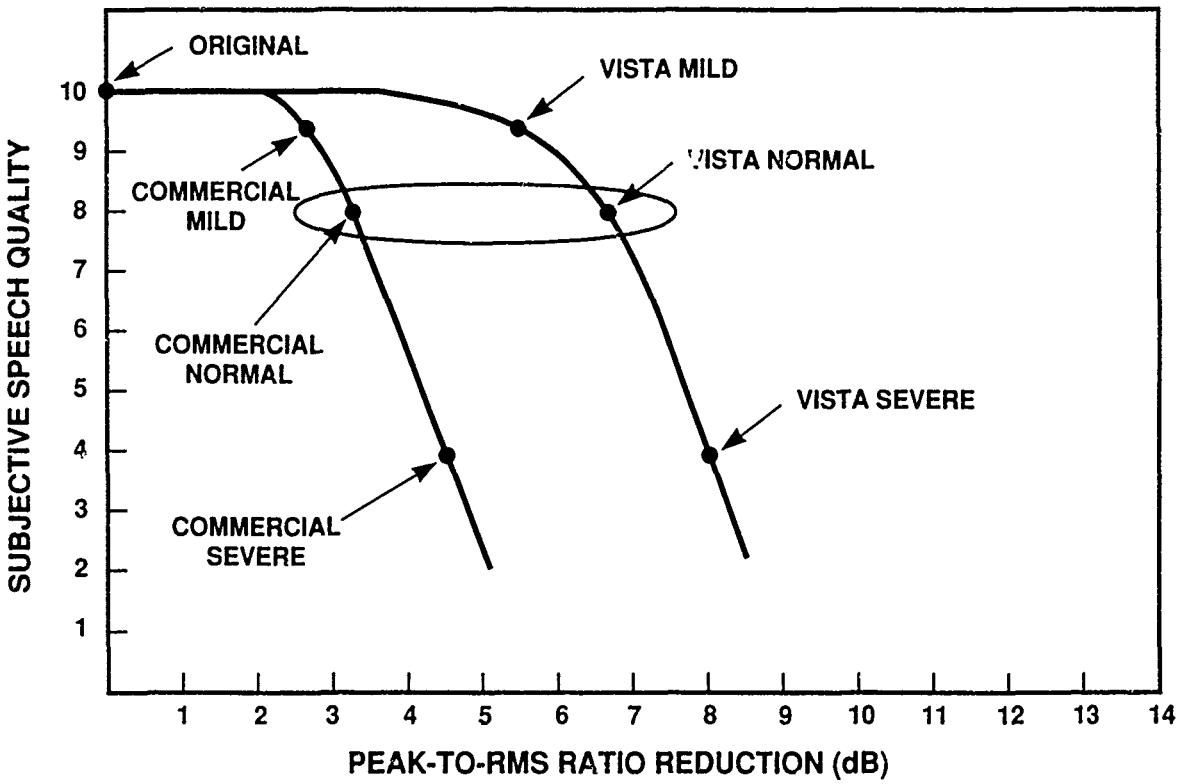


Figure 2-5. Peak-to-RMS/quality tradeoffs.

In summary, an enhancement algorithm and prototype system has been developed which can provide peak/RMS ratio reduction to increase effective broadcast power in the VOA broadcast system, and thus improve listenability of speech in the coverage areas. Live tests on th VOA

channel are planned. Potential future work includes: extensive field testing; more formal quality and intelligibility testing, in the VOA environment, to confirm preliminary evaluations; refinement of the VISTA algorithm based on field test and performance evaluation results; development of enhancement techniques applicable to music as well as speech; and development of a number of VISTA units through technology transfer to industry.

3. SPEECH ENHANCEMENT USING SINUSOIDAL ANALYSIS/SYNTHESIS

At the foundation of the new approach to speech enhancement is an analysis/synthesis system which is based on a sinusoidal representation of speech [1,2]. Enhancement is achieved by manipulating the sine-wave parameters of this analysis/synthesis system to perform phase dispersion, spectral shaping, and amplitude compression as was illustrated in Figure 2-2. The parameters which govern these operations adapt themselves to speech characteristics such as speech spectrum and pitch, achieving a significant peak/RMS ratio reduction.

The primary goal of the sinusoidal processor is to introduce a phase versus frequency characteristic that preserves the spectral magnitude of the speech (since this preserves quality and intelligibility) while producing a waveform that is maximally flat. Since the waveform is already dispersed due to the phase characteristic of the vocal tract, it is necessary to first remove this natural phase dispersion. This is done by constructing a "zero-phase" analysis/synthesis system which produces a very peaky waveform as shown in Figure 3-1. The next step is to add in the phase characteristic that has been designed to optimally flatten the zero-phase waveform. The methods for performing these operations will be developed in the section. It is remarkable that, when amplitude compression is integrated with this dispersion, the spectral information in the resulting processed waveform appears to be embedded primarily within the zero crossings of the modified waveform, rather than the waveform shape. The original intelligibility is essentially maintained without the unacceptable loss in quality suffered by severe clipping.

After a brief background discussion in Section 3.1, Section 3.2 reviews the sinusoidal analysis-synthesis system and describes a "zero-phase" version of this system which removes natural dispersion and provides the basis for the sinusoidal processor. In Section 3.3, a radar signal design solution is used for determining the dispersed speech phase. In Sections 3.4 and 3.5, sine-wave amplitudes are modified for amplitude compression and spectral shaping. Following the sinusoidal processing is a "post-processor", described in Section 3.6, which includes clipping, automatic gain control, and D/A filter compensation. Section 3.7 then integrates all three components and gives

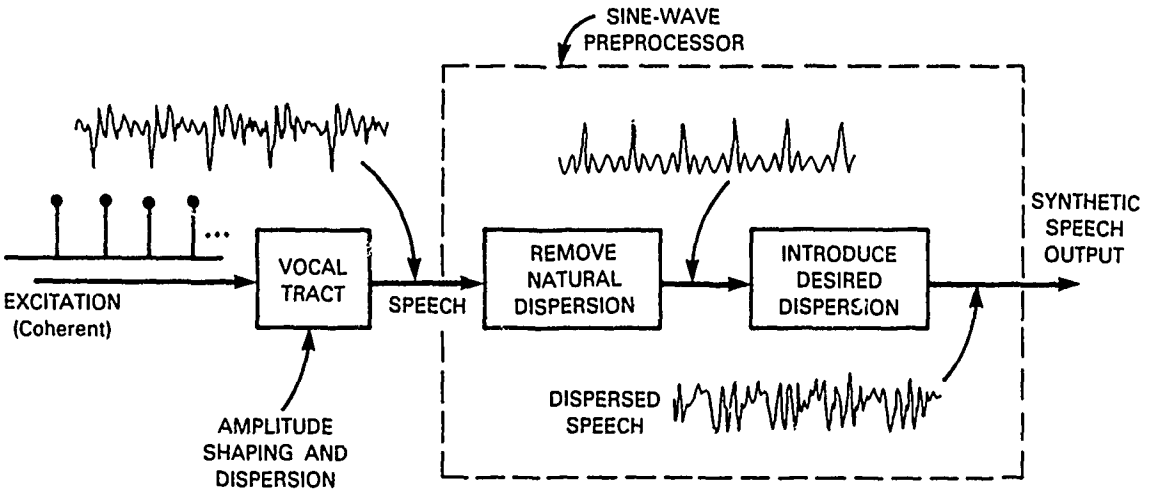


Figure 3-1. Phase dispersion via sinusoidal enhancement.

three enhancement options (mild, normal, and overdriven) which allow the possibility of different degrees of processing for different broadcast ranges and desired quality levels.

3.1 Background and Motivation

The sine-wave analysis/synthesis system, in addition to decomposing the speech waveform into a sum of sine waves, further decomposes each sine wave into its vocal cord excitation and vocal tract contributions, according to the speech production model [1,2]. This frequency-domain representation, therefore, is amenable to a large class of preprocessing operations which adapt to specific speech characteristics. This flexibility contrasts conventional methods of waveform preprocessing for AM radio broadcasting [3] which use primarily time-domain techniques such as clipping and amplitude compression based on time-domain envelope measures. Moreover, conventional methods are often not speech-adaptive. Fixed dispersive networks and fixed preemphasis filters represent two such non-adaptive schemes.

Use of sine-wave-based analysis/synthesis in speech preprocessing was motivated by its earlier

use in speech modification [2,7] and speech coding [8]. The system is capable of modifying the time scale, frequency scale, and pitch of an acoustic waveform by manipulating the durations, frequencies, and phases of the sine-wave components. The sine-wave amplitudes, phases, and frequencies can also be coded and used for efficient speech transmission over narrow-band communication links. In these applications, the speech waveform is decomposed into sine waves, the sine-wave parameter estimates are modified for some desired objective, and a modified waveform is synthesized. In the preprocessing application, the parameters of the sine-wave-based analysis/synthesis system are manipulated to perform dispersion, amplitude compression, and spectral shaping of the speech waveform.

3.2 The Sinusoidal Framework

In this section, the sine-wave analysis/synthesis system is reviewed. A "zero-phase" version of this system is developed for removing natural dispersion in the speech waveform.

3.2.1 Analysis/Synthesis

In the speech production model [9], the speech waveform $s(t)$ is assumed to be the output of passing a vocal cord (glottal) excitation waveform $e(t)$ through a linear system $h(t)$ representing the characteristics of the vocal tract. For simplicity, it is assumed that the glottal pulse shape as well as the vocal tract impulse response is part of the system response $h(t)$. The excitation function can be represented as a periodic pulse train during voiced speech (e.g., the vowel "a"), where the spacing between consecutive pulses is the "pitch" of the speaker,and is represented as a noise-like signal during unvoiced speech (e.g., the fricative "s"). Alternately, the binary voiced/unvoiced excitation model is replaced by a sum of sine waves in the form [1,2]:

$$e(t) = \sum_{k=1}^{L(t)} a_k(t) \cos[\Omega_k(t)] \tag{3.1a}$$

where for the kth sine wave, the excitation phase $\Omega_k(t)$ is the integral of the time-varying "frequency track" $\omega_k(t)$

$$\Omega_k(t) = \int_0^t \omega_k(\sigma)d\sigma + \phi_k \tag{3.1b}$$

where ϕ_k is the fixed phase offset to account for the fact that the sine waves will generally not be in phase. $L(t)$ represents the number of sine waves at time t and $a_k(t)$ is the time-varying amplitude associated with each sine wave. Since the vocal tract impulse response is also time-varying, the vocal tract transfer function (i.e., the Fourier transform of $h(t)$) can be written in terms of its time-varying amplitude $M(\omega;t)$ and phase $\psi(\omega;t)$ components as

$$H(\omega;t) = M(\omega;t)\exp[j\psi(\omega;t)] \tag{3.2a}$$

The system amplitude and phase along each frequency track $\omega_k(t)$ are then given by

$$M_k(t) = M[\omega_k(t);t] \tag{3.2b}$$

and

$$\psi_k(t) = \psi[\omega_k(t);t] \tag{3.2c}$$

Passing the excitation (3.1) through the time-varying vocal tract (3.2) results in the sinusoidal representation for the speech waveform [1]

$$s(t) = \sum_{k=1}^{L(t)} A_k(t)\cos[\theta_k(t)] \tag{3.3a}$$

where

$$A_k(t) = a_k(t)M_k(t) \tag{3.3b}$$

and

$$\theta_k(t) = \Omega_k(t) + \psi_k(t) \tag{3.3c}$$

represent the amplitude and phase of each sine component along the frequency track $\omega_k(t)$. The accuracy of this representation is subject to the caveat that the parameters are slowly varying relative to the duration of the vocal tract system response. Since measurements are made using digitized speech (i.e., with analog-to-digital conversion), sampled-data notation will be used throughout the remainder of this report. In particular, the continuous time variable t is replaced by the integer-valued index $n = 0, 1, 2...$ where $n = t/T_o$ where T_o is the time sampling period. Specifically, $T_o = 100\mu$ sec corresponding to a sampling rate of 10000 samples per second required by a desired 5 kHz speech bandwidth.

Based on the basic sinusoidal representation in (3.1)–(3.3), an analysis/synthesis system has been developed [1,2]. The analysis takes place at a fixed frame interval $Q = 100$ samples corresponding to a 10 ms frame interval. Since the sine-wave parameters are estimated only at the frame boundaries corresponding to the time samples $m = 0, Q\ 2Q$,...henceforth their dependence on time variable t is replaced by their dependance on the frame variable m (i.e., $A_k(t) \rightarrow A_k(m)$). The analysis window duration is set at 2.5 times the average pitch period. This pitch estimate, $\omega_0(m)$, on which this window is determined, is derived using a sinusoidally-based pitch estimation method [10]. Since the pitch estimator itself requires a pitch-adaptive analysis window, the pitch estimation entails two steps: (1) a "coarse" pitch estimate using a fixed 60 ms window and (2) a "refined" pitch estimate using a window derived from the "coarse" pitch. This adaptivity to pitch is necessary to make the system robust over many speakers and speaking conditions. A short-time Fourier transform (STFT) is then computed over this duration with a 1024-point fast Fourier transform (FFT). The excitation frequencies $\omega_k(m)$ are estimated by picking the peaks of the uniformly-spaced (FFT) samples of the short-time Fourier transform magnitude. The sine-wave amplitudes and phases for each analysis frame are then given by the amplitude and phase of the STFT at the measured frequencies.

The first step in synthesis requires association of the frequencies measured on one frame with those obtained on a successive frame. This is accomplished with a nearest-neighbor matching algorithm which incorporates a birth-death process of the component sine waves; i.e., the sine waves are allowed to come and go in time. Amplitude and phase parameters are then interpolated across frame boundaries at the matched frequency sets to recover the original sampling interval.

The amplitude is interpolated linearly and the phase is interpolated with a cubic polynomial. The interpolated amplitude and phase components are then used to form an estimate of the waveform, according to (3.3a), which is essentially perceptually indistinguishable from the original. A block diagram of the baseline sine-wave analysis/synthesis system is illustrated in Figure 3-2.

3.2.2 Zero-Phase Reconstruction

The sinusoidal speech preprocessor depends on the development of a "zero-phase" version of the analysis/synthesis system in Figure 3-2. The essence of the zero-phase reconstruction of speech is elimination of the natural vocal tract system phase during voiced speech segments. The zero-phase waveform consists of coherent sine waves (i.e., in phase) during voiced speech, and is characterized by a symmetric vocal tract impulse response. Use of the natural system phase during unvoiced speech does not change the system's effectiveness in reducing peak/RMS since unvoiced speech contributes little to this measure. Moreover, the preservation of as much of the original waveform as possible helps to preserve the original quality.

The objective then is to estimate and remove the natural phase dispersion during voiced speech. As a first step, to simplify the excitation phase representation in (3.1b), a parameter representing a pitch pulse onset time is introduced [11]. In the context of the sine-wave model, a pitch pulse occurs when all of the sine waves add coherently (i.e., are in phase). Hence, for the *mth* frame, the excitation waveform is modeled as

$$e(n) = \Sigma_{k=1}^{N(m)} a_k(n) \cos \quad [\ (n - n_0(m))\omega_k(m)\ ] \tag{3.4}$$

where $n_0(m)$ is the onset time of the pitch pulse and where the excitation frequency ω_k can be assumed constant over the duration of the analysis window. Comparison of equation (3.4) with equation (3.1b) shows that the excitation phase $\Omega_k(m)$ is linear with respect to frequency. With this representation of the excitation, the excitation phase for the *mth* frame can be obtained through the onset time $n_0(m)$ as

$$\Omega_k(m) = [\ \ (m - n_0(m))\omega_k(m)\ \ ] \tag{3.5a}$$

The amplitude is interpolated linearly and the phase is interpolated with a cubic polynomial. The interpolated amplitude and phase components are then used to form an estimate of the waveform, according to (3.3a), which is essentially perceptually indistinguishable from the original. A block diagram of the baseline sine-wave analysis/synthesis system is illustrated in Figure 3-2.

3.2.2 Zero-Phase Reconstruction

The sinusoidal speech preprocessor depends on the development of a "zero-phase" version of the analysis/synthesis system in Figure 3-2. The essence of the zero-phase reconstruction of speech is elimination of the natural vocal tract system phase during voiced speech segments. The zero-phase waveform consists of coherent sine waves (i.e., in phase) during voiced speech, and is characterized by a symmetric vocal tract impulse response. Use of the natural system phase during unvoiced speech does not change the system's effectiveness in reducing peak/RMS since unvoiced speech contributes little to this measure. Moreover, the preservation of as much of the original waveform as possible helps to preserve the original quality.

The objective then is to estimate and remove the natural phase dispersion during voiced speech. As a first step, to simplify the excitation phase representation in (3.1b), a parameter representing a pitch pulse onset time is introduced [11]. In the context of the sine-wave model, a pitch pulse occurs when all of the sine waves add coherently (i.e., are in phase). Hence, for the *mth* frame, the excitation waveform is modeled as

$$e(n) = \Sigma_{k=1}^{N(m)} a_k(n) \cos \quad [\ (n - n_0(m))\omega_k(m)\ ] \tag{3.4}$$

where $n_0(m)$ is the onset time of the pitch pulse and where the excitation frequency ω_k can be assumed constant over the duration of the analysis window. Comparison of equation (3.4) with equation (3.1b) shows that the excitation phase $\Omega_k(m)$ is linear with respect to frequency. With this representation of the excitation, the excitation phase for the *mth* frame can be obtained through the onset time $n_0(m)$ as

$$\Omega_k(m) = [\ (m - n_0(m))\omega_k(m)\ ] \tag{3.5a}$$

where $b_k(m)$ is a binary weighting function which takes on a value of zero for a voiced track and unity for an unvoiced track

$$b_k(m) = 1 \text{ if } \omega_k(m) \geq \omega_c(m). \tag{3.6b}$$

$$b_k(m) = 0 \text{ if } \omega_k(m) < \omega_c(m). \tag{3.6c}$$

where $\omega_c(m)$ is the voiced/unvoiced frequency cutoff for the mth frame.

In the baseline analysis/synthesis system described in the previous section and illustrated in Figure 3-2, the excitation and vocal tract components were treated as a composite phase. In the zero-phase sine-wave analysis/synthesis system, on the other hand, the excitation and system phase components in (3.3c) and the corresponding amplitude components in (3.3b) are estimated separately. The amplitude estimation is included, as well as the phase estimation, since the separate excitation and vocal tract amplitude components are used later in performing amplitude compression.

The onset time $n_0(m)$, from which the excitation phase is derived, is obtained by interpolating estimates of the pitch period to accumulate onset times from frame to frame. As illustrated in Figure 3-3, whenever the accumulated pitch periods cross a frame boundary, a new onset time is defined for that frame. An estimate of the system phase $\psi(\omega; m)$ at the sine-wave frequencies is then computed by subtracting the estimate of the excitation phase $\Omega_k(m)$ from the measured phase at the spectral peaks. The system amplitude $M(\omega; m)$ is estimated via a smoothing of the high-resolution spectrum similar to that used in the spectral envelope estimation vocoder which exploits a pitch estimate[12]. Following (3.3b) and (3.3c), the excitation amplitude $a_k(m)$ is then estimated by dividing the measured amplitude at the spectral peaks by the estimate of system amplitude $M(\omega; m)$.

In order to determine the voiced/unvoiced track designation required by the phase function (3.6), a frequency cutoff $\omega_c(m)$ must be estimated and a "probability of voicing" measure $V_p(m)$ is derived in the pitch estimation process [10]. $V_p(m)$ falls in the interval [0,1] and gives for each frame the "degree of voicing" (unity being highly voiced). For each frame, the frequency cutoff, $\omega_c(m)$, varies with the voicing probability, $V_p(m)$, as

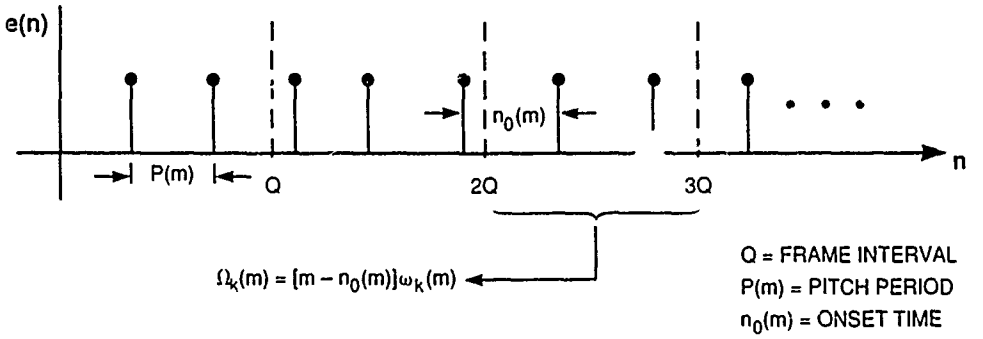


Figure 3-3. Excitation phase estimation via onset time where P denotes pitch period.

$$\omega_c(m) = V_p(m)B \tag{3.7}$$

over a bandwidth B. An example of the transitional properties of the zero-phase system is shown in Figure 3-4 where sine-wave frequency tracks are illustrated as a function of time. Each sine-wave track is designated "voiced" (solid) or "unvoiced" (dashed) according to whether it lies above or below the voicing-dependent frequency cutoff. If a sine-wave frequency track crosses the frequency-cutoff boundary, the track is subdivided into two tracks, one voiced and one unvoiced with linearly rising and falling amplitudes. This process, which will be referred to as "post-matching", will increase the number of tracks and, therefore, the matched frequency set from which the baseline analysis must be updated.

In the synthesis, the composite amplitude estimates of the sine waves is obtained by multiplying vocal tract and vocal cord amplitude contributions and the composite phase estimate is obtained by summing vocal cord and vocal tract phase contributions. As in the baseline analysis/synthesis, the composite amplitude and phase parameters are then interpolated across frame boundaries at the matched frequency sets to recover the original sampling interval. A block diagram of the zero-phase sine-wave analysis/synthesis system is illustrated in Figure 3-5.

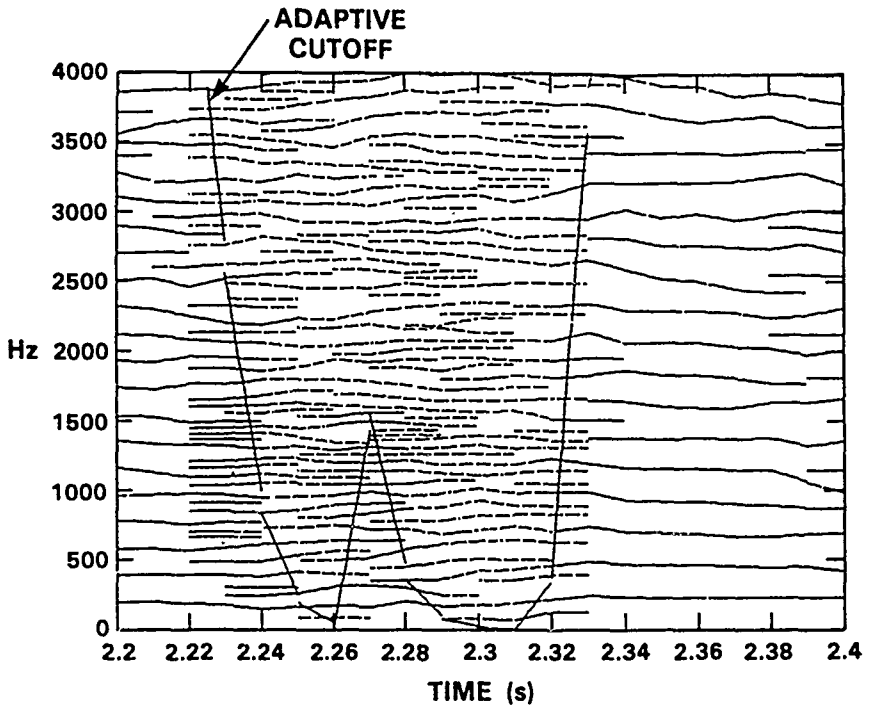


Figure 3-4. Transitional properties of frequency tracks in baseline zero-phase system. Matched frequencies $\omega_k(m)$ are connected with linear interpolation. Broken tracks represent sine-wave births and deaths.

The zero-phase system was subjected to listening tests using an extensive data base ($\approx$ 10's of minutes) and was found to be generally natural and free of artifacts. In testing and refining the new analysis/synthesis system, the voicing probability $V_p(m)$ was first set to zero; since, according to (3.6), when $V_p(m) = 0$, the original system phase is used everywhere, and as would be expected, the reconstruction was essentially perceptually indistinguishable from the original. On the other hand, setting the voicing probability $V_p(m)$ to unity resulted in a "buzzy" reconstruction due to the forced phase coherence during unvoiced speech. When $V_p(m)$ was allowed to take on its measured value, voiced regions almost always consisted of symmetric responses with unvoiced speech largely reproduced with its original naturalness. If $V_p(m)$ was measured high in an unvoiced region (e.g., in voiced/unvoiced transitions), however, a slight "buzz" would occassionally arise. This artifact was avoided by biasing $V_p(m)$ slightly toward the unvoiced decision. This tended to result in the reconstruction of the original speech during transitions; however, this did not effect the effectiveness

of the sinusoidal preprocessor.

3.3 Phase Dispersion

In the radar signal design problem [4,13,14], as in modeling of the voiced speech waveform, the signal is periodic and given as the output of a transmit filter whose input consists of periodic pulses (see Figure 3-6). In this section, one solution to the radar signal design problem is described. The solution is then tailored to the speech dispersion problem and is incorporated into the baseline zero-phase sinusoidal analysis-synthesis system.

3.3.1 The Key, Fowle, Haggarty Solution

The basic unit of the radar waveform is the impulse response, $h(n)$, of the transmit filter illustrated in Figure 3-6. It is expedient to view this response in the time domain as an "FM-like chirp" signal (see Appendix A) with envelope $a(n)$ and phase $\phi(n)$

$$h(n) = a(n)\cos[\phi(n)] \quad 0 \leq n \leq L \tag{3.8a}$$

which has a Fourier transform $H(\omega)$ with magnitude $M(\omega)$ and phase $\psi(\omega)$

$$H(\omega) = M(\omega)\exp[\psi(\omega)] \tag{3.8b}$$

By exploiting the analytic signal representation of $h(n)$, Key, Fowle, and Haggarty [4] have shown that, under a large time-bandwidth product constraint, specifying two of the four amplitude and phase components in (3.8) is sufficient to approximately determine these remaining two components (see Appendix A). How large the time-bandwidth product must be for these relations to hold accurately depends on the shape of the functions $a(n)$ and $M(\omega)$ [4,14].

Ideally for minimum peak/RMS, the time envelope $a(n)$ should be flat over the duration L of the impulse response. With this and the additional constraint that the spectral magnitude is specified (a flat magnitude is usually used in the radar signal design problem), Key, Fowle, and

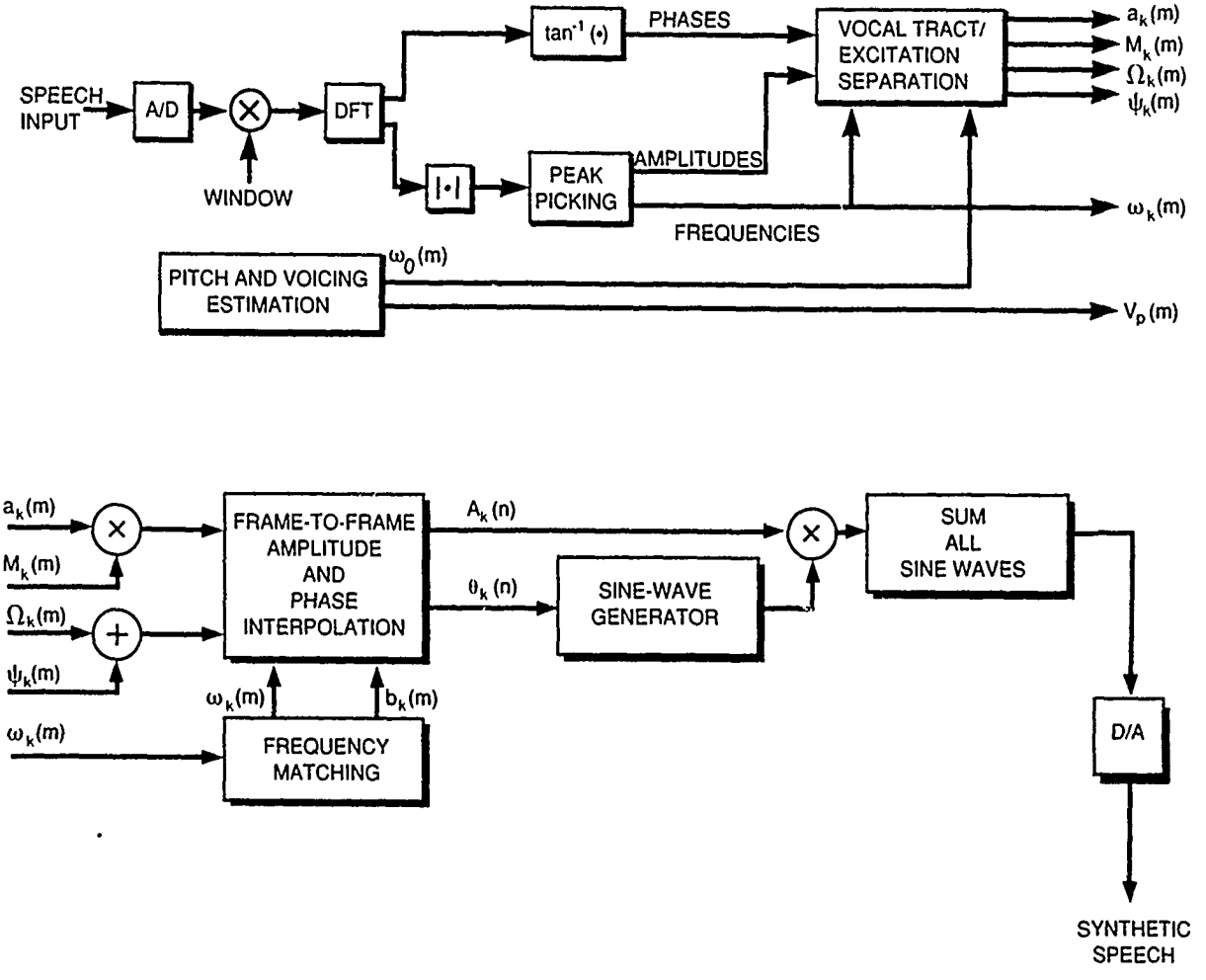


Figure 3-5. Zero-phase sinusoidal analysis/synthesis with vocal tract/excitation separation.

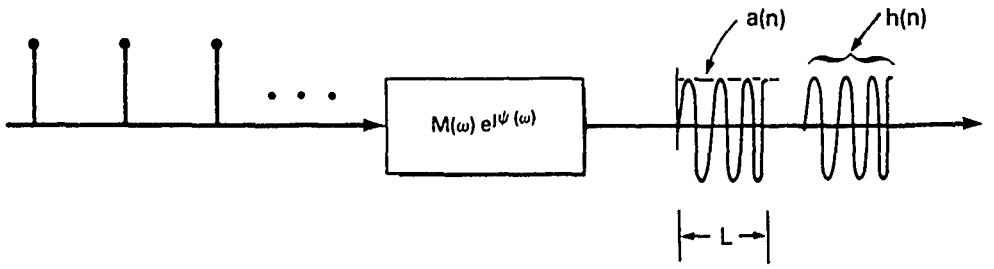


Figure 3-6. Radar signal design for low peak/RMS.

Haggarty's general relation among the envelope and phase components of $h(n)$ and its Fourier transform $H(\omega)$ reduces to an expression for the unknown phase $\psi(\omega)$ as

$$\psi(\omega) = L \int_0^{\omega} \int_0^{\beta} \hat{M}^2(\alpha) d\alpha d\beta \tag{3.9a}$$

where "hat" indicates that the magnitude has been normalized by it's energy, i.e.,

$$\hat{M}^2(\omega) = M^2(\omega) / \int_0^{\pi} M^2(\alpha) d\alpha \tag{3.9b}$$

and where, π represents the signal bandwidth [15] in the discrete time signal representation. The accuracy of the approximation in (3.9) increases with increasing time-bandwidth product. In particular, for this case in which the time envelope is rectangular, and for which the spectral magnitude is smooth and continuous, the time-bandwidth product should be greater than about 20 or 30 to construct a Fourier transform pair with good accuracy using the relation in (3.9) [4,14].

Equation (3.9) shows that the resulting phase $\psi(\omega)$ depends only on the normalized spectral magnitude $\hat{M}_k(\omega)$ and the impulse response duration L. It is shown in Appendix A that the envelope level of the resulting waveform can be determined, with the application of appropriate energy constraints, from the unnormalized spectrum and duration. Specifically it is shown that if the envelope of $h(n)$ is constant over its duration L and zero elsewhere, the envelope constant has the value

$$A = [\frac{1}{2\pi L} \int_0^{\pi} M^2(\omega) d\omega]^{1/2}, 0 \leq n < L \tag{3.10}$$

This amplitude relation will be used in Section 3.4 to develop the frequency-domain-based approach to amplitude compression.

Examples of impulse responses designed with an approximately flat time-domain envelope via the Key, Fowle, Haggarty (KFH) solution are shown in Figure 3-7. In the first two examples in Figure 3-7a and Figure 3-7b, the spectrum is unimodal and the desired impulse response duration is 10 ms. The bandwidth in these examples is 5 kHz and so the time-bandwidth product is 50, which guarantees that the phase-magnitude relation in (3.9) leads to an accurate Fourier transform pair. From Figure 3-7, it can be observed that the resulting chirp responses are dominated by the peak frequency in $M_k(\omega)$. In the third example in Figure 3-7c which illustrates a bimodal spectrum, the two major frequency components of the signal have been mapped to different time slots with about equal duration, reflecting their almost equal intensity levels. This mapping of intensity to frequency duration is a general property of the KFH solution. (This is addressed in more detail in the discussion in Appendix A relating to the group delay properties of the phase $\psi(\omega)$.)

3.3.2 Use of the Key, Fowle, Haggarty Solution for Dispersion in Speech

In the previous section, the phase of the radar transmit filter was designed to satisfy two constraints: a specified spectral magnitude and a flat time-domain envelope over some desired duration. As a consequence, transforming a given response with an arbitrary time-domain envelope to one with a flat envelope requires replacing its phase by that derived from the Key, Fowle, Haggarty (KFH) solution. In the speech dispersion problem, the "filter" is the vocal tract transfer function (3.2). For voiced speech, the waveform is approximately periodic and thus the phase must be extracted from this representation and then replaced with the KFH solution.

Applying the KFH phase to dispersing voiced speech requires the estimation of the spectral magnitude $M(\omega; m)$ of the vocal tract impulse response and the pitch period of the vocal cord excitation $P(m)$. The duration of the synthetic vocal tract impulse response is set close to the pitch period $P(m)$ so that the resulting speech waveform is as "dense" as possible. The analysis

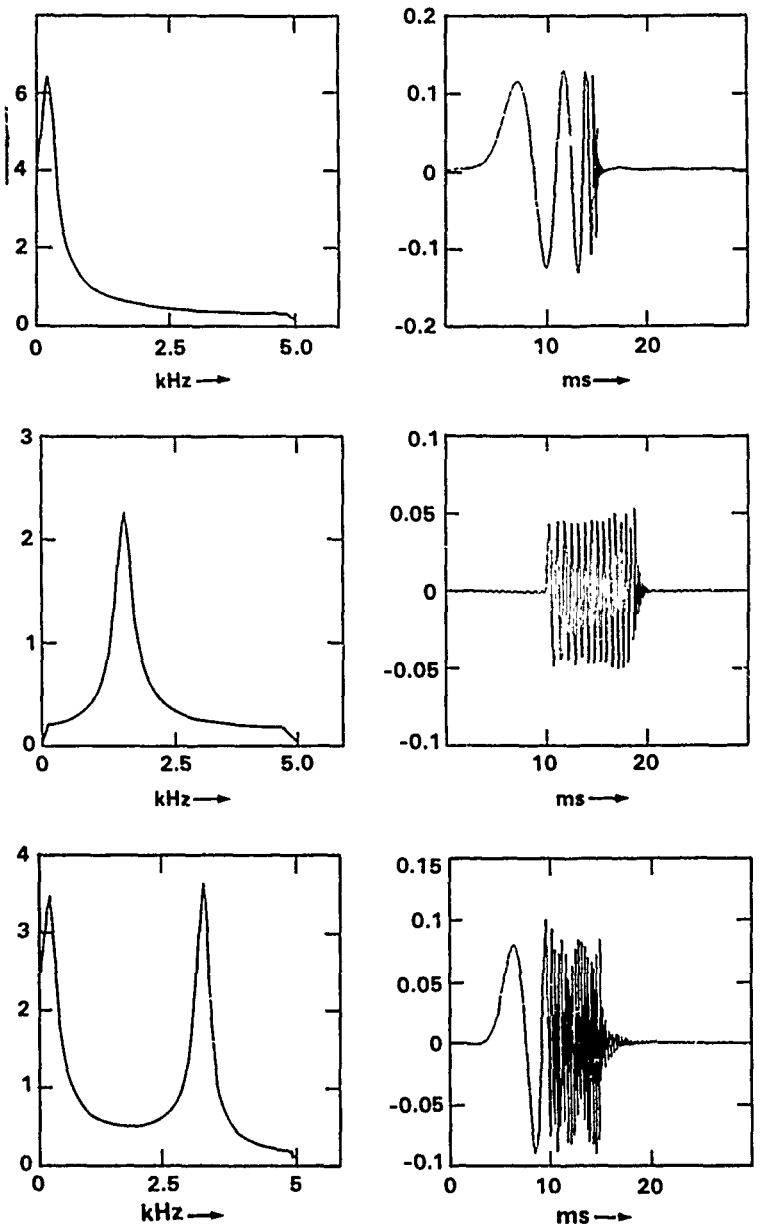


Figure 3-7. KFH responses for various spectra.

component of the zero-phase sine-wave system produces estimates of the spectral and pitch characteristics. The synthetic vocal tract phase derived using the KFH solution, denoted by $\psi_{kfh}(\omega; m)$, is given by

$$\psi_{kfh}(\omega; m) = \mu P(m) \int_0^{\omega} \int_0^{\beta} \hat{M}^2(\alpha; m) d\alpha d\beta \tag{3.11}$$

where "kfh" denotes the KFH phase and where "hat" denotes that the estimated magnitude has been normalized by its energy and where μ, which falls in interval [0,1], is a scale factor to account for a possible desired reduction in the chirp duration less than a pitch period. Since the smallest pitch period assumed encountered in actual speech is about 5 ms and the bandwidth is 5 kHz, there results a minimum time-bandwidth product of about 25 which just satisfies the KFH time-bandwidth constraint of about 20 or 30 [4,14]. (Occasionally, however, very high-pitch speakers with a pitch period as low as 2.5 ms have been observed.) The scaling factor μ is used to reduce the effect of any overlap in successive dispersed impulse responses. This overlap, which can be harmful in maintaining a low peak/RMS ratio, can arise if the time-bandwidth constraint is not adequately satisfied, or if the pitch period estimate is larger than the actual pitch of the speaker.

Applying the KFH phase dispersion solution in the zero-phase synthesis requires that the synthetic vocal tract system phase in (3.11), $\psi_{kfh}(\omega; m)$, which is a continuous function of frequency, be sampled along the sine-wave frequency tracks $\omega_k(m)$

$$\psi_{k,kfh}(m) = \psi_{kfh}[\omega_k(m); m] \tag{3.12}$$

where the subscript "k, kfh" denotes the KFH phase along the kth track. The solution in (3.12) is used only for voiced speech regions where the periodicity assumption holds, whereas in unvoiced regions the original system phase is maintained. Therefore, the KFH phase is assigned only to those tracks designated "voiced". The original system phase is assigned to all tracks designated "unvoiced". Thus the phase assignment for the kth sine wave is given by

$$\theta_k(m) = \Omega_k(m) + b_k(m)\psi_k(m) + [1 - b_k(m)]\psi_{k,kfh}(m) \tag{3.13}$$

where $b_k(m)$, defined in (3.6), takes on a value of zero for a voiced track and unity for an unvoiced track, where $\Omega_k(m)$ is the excitation phase, $\psi_k(m)$ is the original phase and $\psi_{k,kfh}(m)$ is the synthetic phase.

An example of dispersing an artificial speech waveform with fixed pitch and fixed vocal tract spectral envelope is illustrated in Figure 3-8. Estimation of the spectral envelope of the processed and original waveforms in Figure 3-8d used the spectral smoothing technique in [12]. The vocal

tract phase is modified significantly. In Figure 3-8d the magnitude of the dispersed waveform is compared with the original magnitude and the agreement is very close, a property that is important to maintaining intelligibility. It seems remarkable that all of the spectral information can be maintained in such a flat waveform, which suggests that perhaps all the information is coded onto the zero crossings.

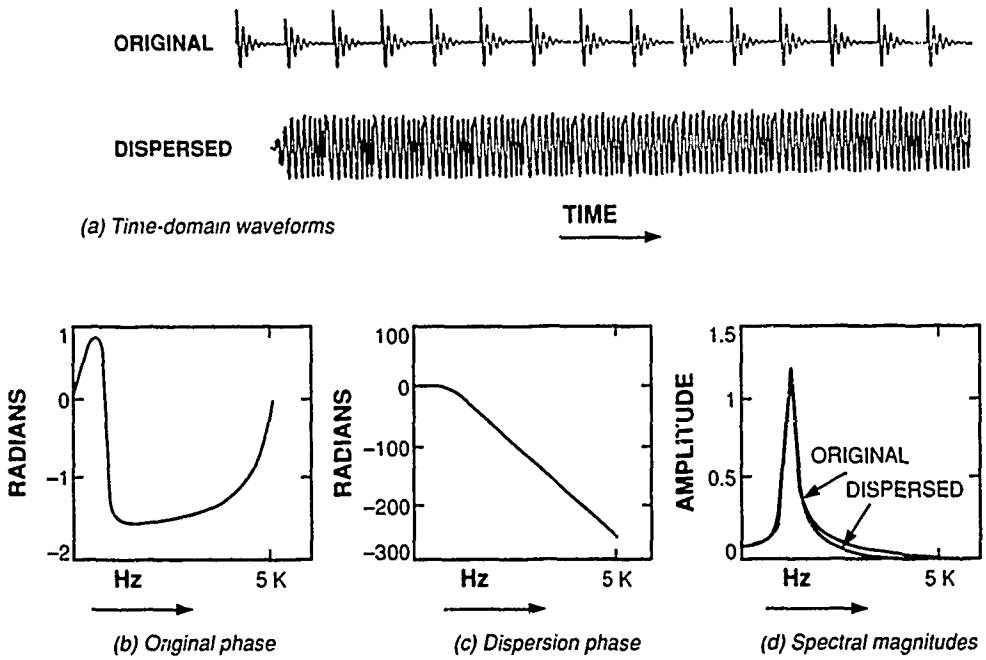


Figure 3-8. Key, Fowle, Haggarty phase dispersion with the sine-wave preprocessor (artificial waveform).

The example Figure 3-8 also illustrates the potential sensitivity of the Key, Fowle, Haggarty phase calculation to small measurement errors in pitch or spectrum. The phase typically traverses from 0 to 300 radians over a bandwidth of 5000 Hz. Consider a small deviation in the pitch period δP added to the actual pitch period. Then it is straightforward to show, that for unity spectral magnitude and a one-sample error in the pitch period, the resulting change in the phase at $\omega = \pi$ is $\pi/2$ — a very large change in the phase over an analysis frame interval. Similar sensitivity can be shown to exist to the spectral magnitude. Such changes in phase can introduce unnatural changes in the sine-wave frequency trajectory. Not only inaccuracies in measurements, but also natural

changes in pitch and spectral magnitude, can cause such frame-to-frame fluctuations.

To reduce large frame-to-frame fluctuations in the KFH phase, both the pitch and the spectral envelope used by the KFH solution are smoothed in time over successive analysis frames. The strategy for adapting the degree of smoothing to speech characteristics is important for maintaining dispersion through rapidly changing speech events. In order that transitions from unvoiced to voiced speech (and vice versa) do not severely bias the averaging process, the degree of smoothing is controlled by the voicing probability, $V_p(m)$ and by spectral and pitch "derivatives" which reflect the rate these parameters are changing in time. Some final smoothing of the phase itself is also performed along frequency tracks designated "voiced" so as to not disturb the original system phase along unvoiced tracks. Under the assumption that speech quality degrades when unnatural changes in phase occur during "steady-state" sounds, the degree of smoothing for all three smoothing operations increases when the spectrum and pitch are slowly varying. Such a design results in little smoothing during speech state transitions or other rapidly-varying events. The smooth phase trajectory is used in (3.13) in place of $\psi_{k,kfh}$(n). The complete scenario for determining a smooth KFH phase is shown in Figure 3-9. The importance of the spectral, pitch, and phase smoothing for preserving speech quality warrants a more thorough description which is given in Appendix B.

An example of a dispersed speech waveform is shown in Figure 3-10. In spite of the above smoothing of spectrum, pitch, and phase, good dispersion was maintained through time-varying speech events. The degree of smoothness was directly related to peak/RMS and loss of quality; the smoother the track, the better the quality, but the higher the peak/RMS.

3.4 Amplitude Compression

KFH phase dispersion reduces the short-time (5-20 ms) waveform fluctuations by "filling in" the waveform over a pitch period. Standard amplitude compression reduces both the short-time and long-time ($\geq$ 20 ms) waveform fluctuations by modifying the waveform envelope. The goal is to flatten the envelope of the waveform during voiced regions for peak/RMS reduction and to modify voiced/unvoiced envelope relations for an increase in the "consonant-to-vowel energy ratio" [16,17,18]. An increase in the unvoiced energy relative to the voiced energy is known to

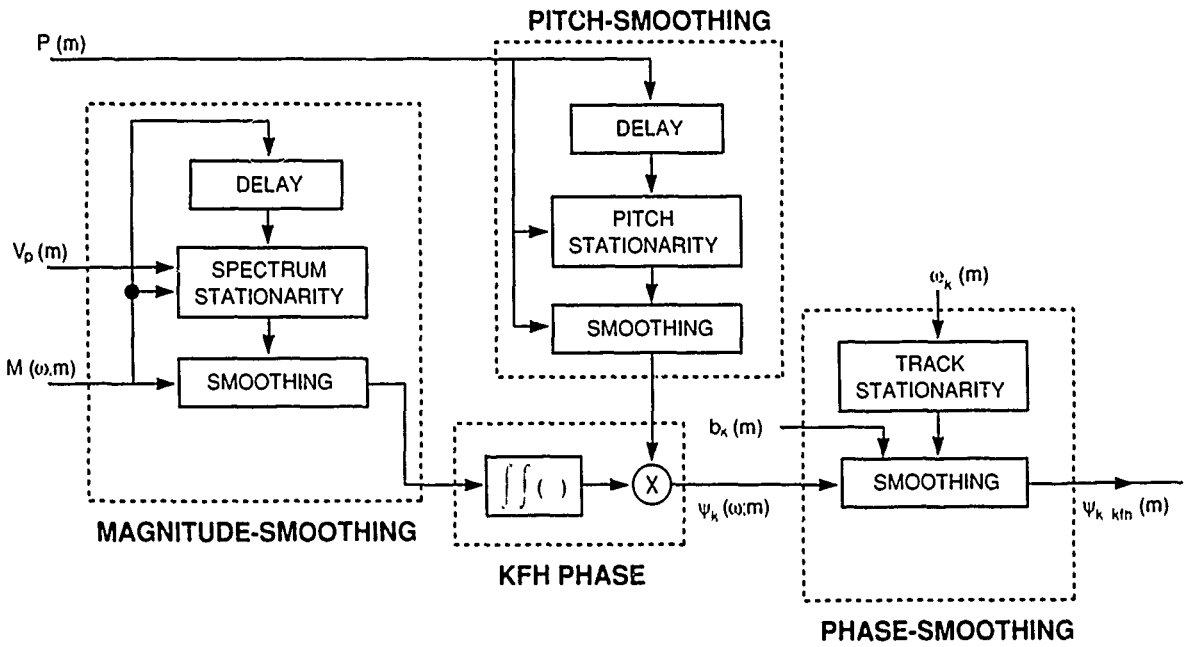


Figure 3-9. Stabilization of KFH phase.

correlate with improved intelligibility [16]. A brief tutorial on the static and dynamic properties of conventional amplitude compression methods, typically used in audio preprocessing, is given in Appendix C.

(a) Original

(b) Dispersed

Figure 3-10. The effect of "optimum" dispersion on a speech waveform.

Conventional amplitude compression methods require a time-domain waveform envelope estimate [3,17,18], unlike the sine-wave enhancement which is based on the KFH solution. This follows from the fact that the vocal tract component of the time-domain envelope of the dispersed waveform is derived directly from the phase calculation in the frequency domain. The excitation component of the envelope can also be obtained in the frequency domain. The composite envelope estimate allows for a simple frequency-domain automatic gain control (AGC) and dynamic range compression (DRC), corresponding to "slow" and "fast" compression dynamics (see Appendix C).

This section begins with a new definition of waveform envelope based on the KFH phase solution and based also on excitation/vocal tract system separation. The envelope estimate is then applied to AGC and finally to DRC. The complete frequency-domain amplitude compression unit is illustrated in Figure 3-11.

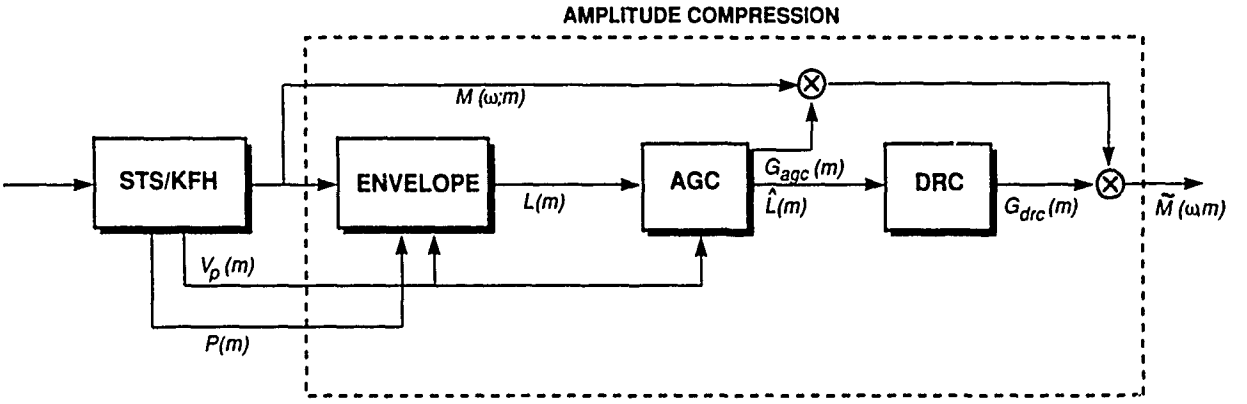


Figure 3-11. Amplitude compression in frequency domain.

3.4.1 Computing the Waveform Envelope in the Frequency Domain

The dispersed waveform, derived using KFH phase, can be thought of as the convolution of a synthetic vocal tract impulse response, with duration roughly a pitch period, and an impulse train with spacing given by the pitch period. From this perspective, the envelope of the resulting waveform can be thought of as the product of the envelopes of the two convolutional components, namely as shown in Figure 3-12.

$$L(m) = L_{sys}(m)L_{exc}(m) \tag{3.14}$$

where the subscripts "sys" and "exc" refer to system and excitation components, respectively. In this section, it is shown that the components in (3.14) and hence the waveform envelope can be computed in the frequency domain.

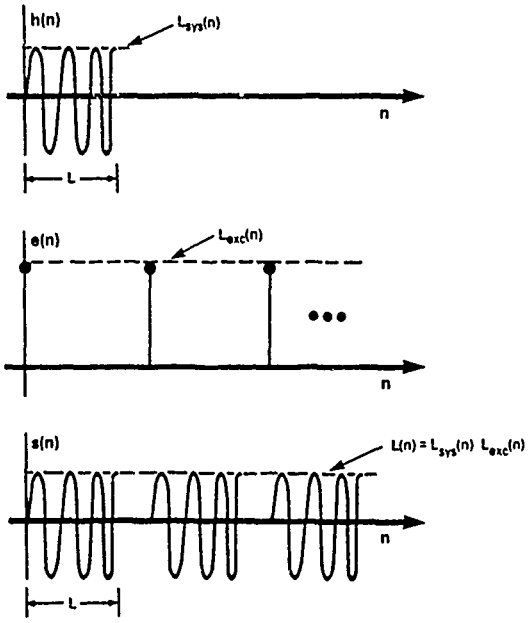


Figure 3-12. Illustration of waveform envelope components ($s(n) = h(n) * e(n)$).

In Section 3.3, the envelope of the radar transmit filter response corresponding to the KFH phase was derived from the spectrum and the chirp duration. In the speech dispersion problem, this envelope varies with the vocal tract spectrum and vocal cord pitch and is given by (sampled at the *mth* frame)

$$L_{sys}(m) = [(1/(2\pi\mu P(m)) \int_0^{\pi} M^2(\omega; m)d\omega)]^{1/2} \tag{3.15}$$

where the scale factor μ is included for a possible desired reduction in duration below a pitch period. For unvoiced speech, the same envelope estimate is used, but the "pitch" is fixed and chosen by dividing the bandwidth (5000 Hz) by an "average" number of peaks (60 peaks) in peak-picking unvoiced spectra; this corresponds to a 12 ms pitch period. In order to obtain a smooth transition from the fixed unvoiced pitch to a time-varying voiced pitch, through voiced to unvoiced transitions

(and vice versa), the pitch-period mapping of Figure 3-13 is used. The cutoff points in the mapping were chosen empirically by ensuring a good envelope fit through speech state transitions.

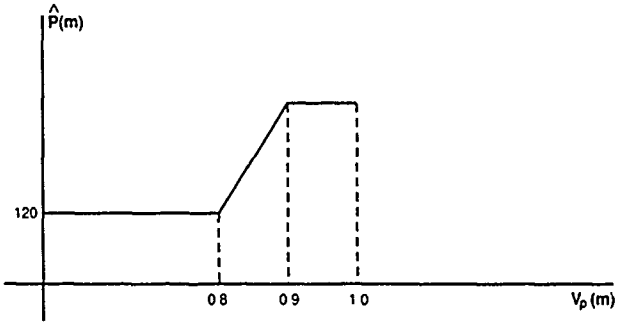


Figure 3-13. Pitch-period mapping for envelope calculation (period is in samples).

The next step is to generate the excitation envelope. Recall from Section 3.2 that the excitation sine-wave amplitudes are obtained by dividing the measured sine-wave amplitudes by the vocal tract system spectral magnitude at the measured frequencies, i.e.,

$$a_k(m) = A_k(m)/M\ [\,\omega_k(m); m\ ] \tag{3.16}$$

Since the spectral magnitude $M(\omega; m)$ is derived from the measured sine-wave amplitudes, then $a_k(m) = 1$ (excluding spurious peaks). Each spectral line has half the height of the corresponding time-domain sine wave (due to the complex representation: $(\exp[j\omega_k(n)] + \exp[-j\omega_k(n)])/2$) and therefore, each sine wave in the time domain has amplitude twice unity. If the sine waves fall within a bandwidth B_w, then the number of sine waves over this bandwidth, for voiced speech, is given roughly by

$$N_{bw}(m) \approx B_w/\ [2\pi/P(m)\ ] \tag{3.17}$$

where $2\pi/P(m)$ is the pitch ($P(m)$ being the pitch period). Since the $N_{bw}(m)$ excitation sine waves are in phase, then it is straightforward to show that the excitation level, i.e., "envelope" in the time domain is given by

$$L_{exc}(m) = 2N_{bw}(m) \tag{3.18}$$

For unvoiced speech a 12 ms pitch period is assumed as before, and the pitch mapping in Figure 3-13 is applied in making speech state transitions.

The composite time-domain waveform envelope is written as the product of the component envelopes (3.14) or from (3.15) and (3.18)

$$L(m) = 2\ [\int_0^{\pi} M^2(\omega;m)/2\pi\mu\ ]^{1/2}\ [B_\omega/\hat{P}(m)^{1/2}\ ] \tag{3.19}$$

where $\hat{P}(m)$ is the pitch period derived from the mapping of Figure 3-13. An example of the waveform envelope estimate is shown in Figure 3-14. The waveform shape is closely tracked even through speech transitions. However, since the phase dispersion and envelope estimates are derived under a large time-bandwidth product assumption [4,14] and since the pitch period estimate used in (3.19) may at times be larger than the actual pitch, the resulting impulse response envelope is not always flat and, as demonstrated, peaks in the resulting waveform can rise above the estimated envelope.

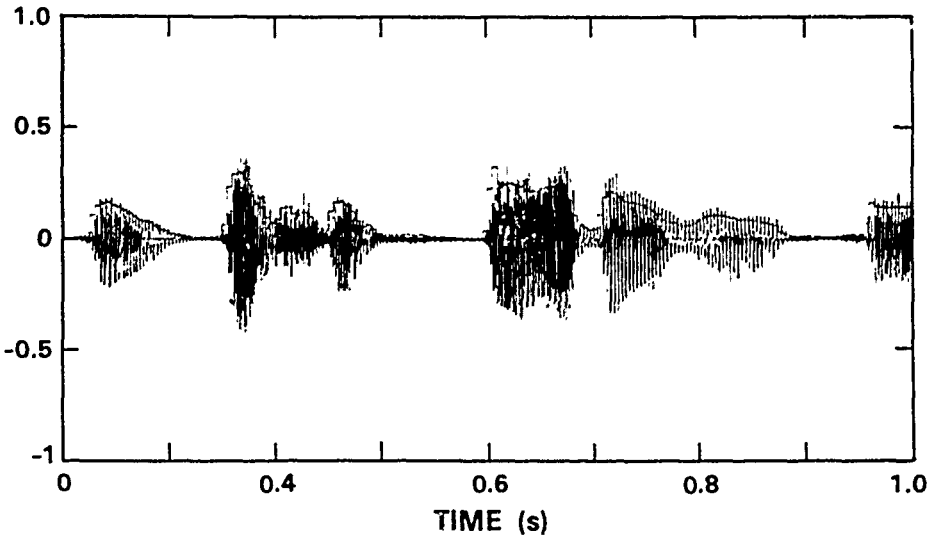


Figure 3-14. Example of waveform envelope estimation.

3.4.2 Automatic Gain Control (AGC)

In automatic gain control (AGC), a slowly time-varying gain is applied to compress long-time fluctuations in the waveform envelope as occur, for example, with different speakers or speaking conditions. The goal is to modify the waveform level so that the DRC which follows AGC is given a meaningful 0 dB reference level. This should be accomplished without a noticeable change in the speech quality. A problem typical of AGC is that its dynamic component (see Appendix C) often causes an increase of low-level sounds such as noise during silence, or in what is traditionally referred to as "pumping" [17,18]. Although the release time is on the order of hundreds of milliseconds, the gain derived from the (static) input/output envelope characteristic (referred to as IOEC in Appendix C) will boost low-level signals as it passes from the compression through the expansion region of the IOEC. If the release time is set too long to attempt to avoid this problem, then the AGC will be ineffective during long stretches of voiced speech.

The frequency-domain AGC used by the STS enhancement avoids this problem by exploiting the voicing probability $V_p(m)$ derived in the pitch extraction unit. Specifically, the release and attack parameters, α_r and α_a, respectively, which govern the release and attack times of the AGC dynamics (Appendix C) are a function of $V_p(m)$:

$$if\ [V_p(m) \leq .9] \quad \alpha_r = 1.0 \ \ \alpha_a = 0.0 \tag{3.20a}$$

$$if\ [V_p(m) \geq .9] \quad \alpha_r = .9 \ \ \alpha_a = 0.2 \tag{3.20b}$$

from which an average envelope measure is derived. A gain $G_{agc}(m)$ is obtained by dividing this average envelope into the desired envelope level (the 0 dB reference level). This forces the average envelope of the dispersed waveform to be at the desired 0 dB reference level. The gain , $G_{agc}(m)$, is then applied to the original envelope measure

$$\hat{L}(m) = G_{agc}(m)L(m) \tag{3.21}$$

which is to be used by the DRC. The voicing-dependent release parameter which takes on unity in (3.20) (an infinite release time) trys to prevent the gain from changing during transitions and

unvoiced speech, thus attempting to maintain voiced/unvoiced energy relations and to prevent "pumping".

3.4.3 Dynamic Range Compression (DRC)

In DRC, the envelope $\hat{L}(m)$ in (3.21) is first smoothed according to dynamics which are governed by attack and release times faster than those used by the AGC. A gain, $G_{drc}(m)$, is then computed from the IOEC similar to that in Figure C-1 of Appendix C. This gain is combined with the AGC gain and the composite gain is applied to the vocal tract spectrum $M(\omega; m)$

$$\tilde{M}(\omega; m) = G_{drc}(m) G_{agc}(m) M(\omega; m) \tag{3.22}$$

as illustrated in Figure 3-11. This has the effect of modifying the time-domain waveform envelope even though the modification takes place in the frequency domain.

3.5 Spectral Shaping

Both adaptive and fixed spectral shaping are applied to the vocal tract spectral magnitude estimate. The adaptive shaping "enhances" the spectrum and consists of two components; the first filter, $H_p(\omega; m)$, preemphasizes the spectrum and the second filter, $H_s(\omega; m)$, sharpens spectral energy concentrations (i.e., the formants). These operations give the processed speech a more "crisp", and "clean" quality, and can improve intelligibility. The fixed shaping filter, $H_r(\omega)$, is applied to compensate low-pass receiver characteristics. The composite spectral shaping filter, denoted $H_c(\omega; m)$, is given by

$$H_c(\omega; m) = H_p(\omega; m) H_s(\omega; m) H_r(\omega) \tag{3.23}$$

where any combination of components can be applied. Typically, in the clear (without radio transmission), only the adaptive components are applied; otherwise, all three components are used.

The composite shaper (3.23), as illustrated in Figure 3-15, is applied prior to dispersion since the KFH phase calculation (3.11) requires the modified spectrum given by

$$\tilde{M}(\omega;m) = H_c(\omega;m)M(\omega;m) \tag{3.24}$$

The spectrum $\tilde{M}(\omega;m)$ is smoothed by the operations of Section 3.3.2 prior to the KFH phase calculation. It is also modified by the amplitude compress unit of the previous section and used in the final waveform reconstruction. This integration of spectral shaping with dispersion and with amplitude compression contrasts conventional methods which perform these operations independently.

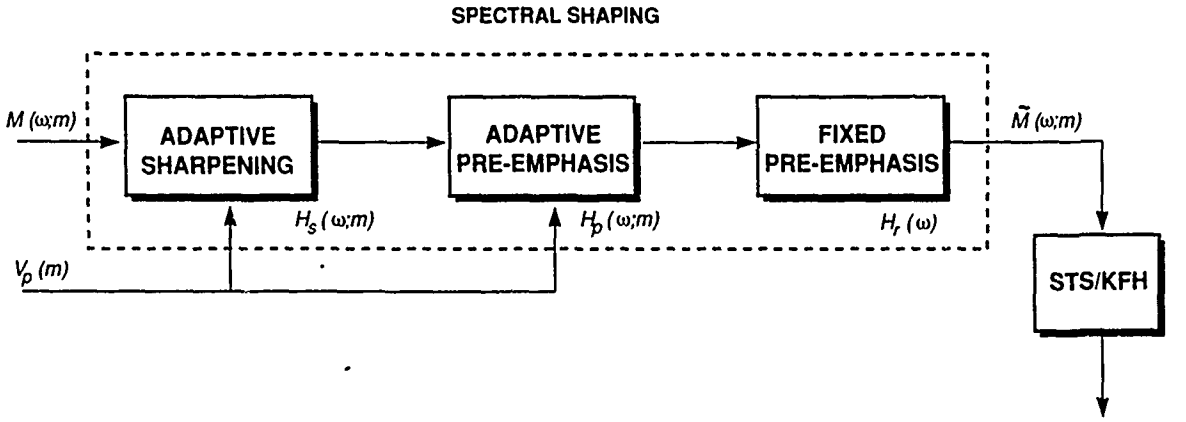


Figure 3-15. The integrated spectral shaper.

3.5.1 Adaptive Shaping

The preemphasis stage of the adaptive spectral shaping compensates for the natural roll-off in the speech spectrum during voiced speech. The idea is to increase the signal-to-noise ratio of speech in the high end of the bandwidth so that high frequencies will be just as "intelligible" as low frequencies in the presence of background noise. Given the natural roll-off of the voiced speech spectrum, a filter was designed by Griffths [19] to maximize the speech "articulation index". This design was later extended by Niederjohn, et al. [20], who determined an "optimally" intelligible fixed preemphasis of 6 dB/octave starting at 1100 Hz. This fixed filter raises the high frequencies for both voiced and unvoiced speech as well as during pauses. Our experience with this filter is that, although improving intelligibility in noise, the resulting speech can be "tinny" and "noisy" due to boosting fricatives and low-level noise during pauses.

An alternate scheme is to adapt the preemphasis so that it is applied only during voiced speech. In unvoiced speech a gain is not applied; the amplitude compression unit increases the intensity of these regions. The degree of preemphasis is a function of the voicing probabilty $V_p(m)$.

$$H_p(\omega; m) = P[\,\omega; V_p(m)\,] \tag{3.25}$$

where the preemphasis filter $P\ [\omega; V_p(m)\ ]$ and its functional dependence on $V_p(m)$ is illustrated in Figure 3-16. The maximum gain at the high-frequency end of the band (5000 Hz), $1 + G_o$, is reached when $V_p(m) = 1$.

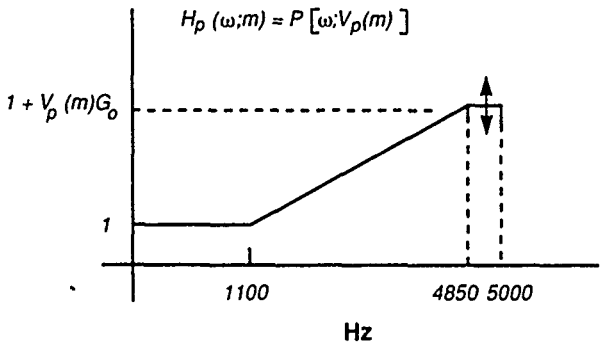


Figure 3-16. Adaptive preemphasis characteristics.

The purpose of the second adaptive filter is to sharpen the formants which can be slightly smeared in the zero-phase sine-wave analysis/synthesis process. The adaptive filter is derived (Figure 3-17) from the vocal tract spectral envelope $M(\omega; m)$ by first removing any spectral tilt, denoted $T(\omega; m)$, and then raising the resulting spectrum to a fractional power β:

$$H_s(\omega; m) = [M(\omega; m) - T(\omega; m)\ ]^{\beta * V_p(m)} \tag{3.26}$$

and where the scaling of the power β by $V_p(m)$ ensures that the spectral modification takes place only during voicing. (Excessive sharpening during unvoiced speech can generate a tonal sound.) The tilt is derived by using the first two coefficients of the cepstral representation of $M(\omega; m)$ [15,21].

The specific parameter choices for (3.25) and (3.26) are described in Section 3.7 and are a function of the desired degree of processing.

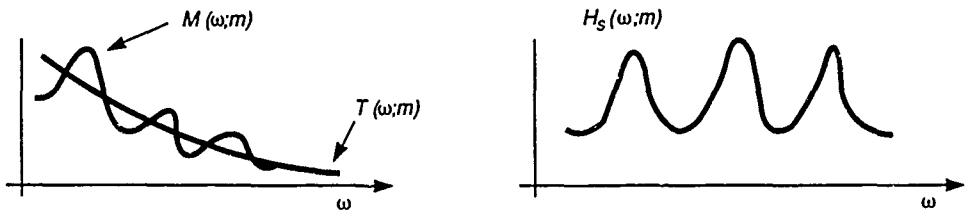


Figure 3-17. Adaptive sharpener characteristics.

3.5.2 Fixed Shaping

The purpose of the fixed (non-adaptive) spectral shaping $H_r(\omega)$ is to compensate for the low-pass filter built into the typical shortwave home receiver and to introduce a high-pass filter at DC to protect the AM transmitter from low-frequency components. The high-pass filter has the additional benefit of adding a "crisper" sound to the processed speech. The high-pass filter is down by 6 dB around 100 Hz and the preemphasis is typically up by 8 dB or more at 5 kHz. Various degrees of fixed receiver compensation are provided in Section 3.7.

3.6 Post-Processing

The operations which follow the STS synthesis are: (1) clipping, (2) AGC, and (3) amplitude and phase compensation of the D/A filter. The post-processing unit is illustrated in Figure 3-18. Each component will be briefly described.

3.6.1 Clipping

The purpose of the clipper is to provide a peak constraint to protect the transmitter, to remove spurious peaks due to, for example, overlap of the synthetic vocal tract chirp response at pitch-period boundaries, and to further reduce the peak/RMS ratio. With respect to the last objective, one of the unique features of the sine-wave processor is that via the KFH phase calculation, spectral information appears to be mapped essentially to zero crossings in the time-domain waveform. This is

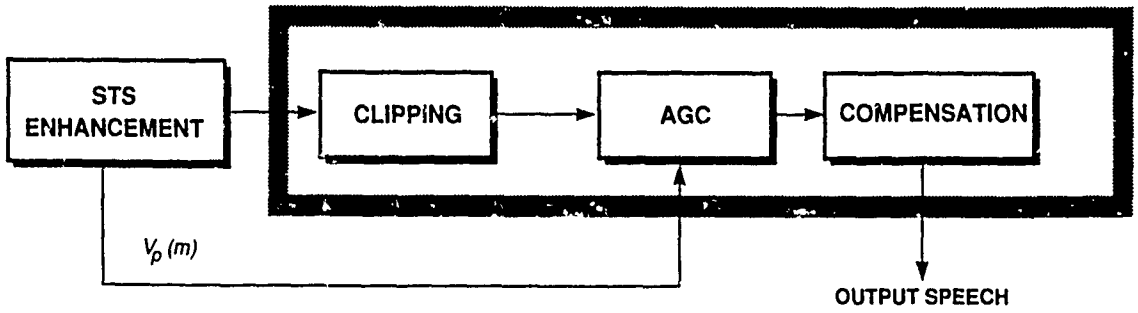


Figure 3-18. Post-processing.

because the relative spectral intensities are mapped to the relative durations of spectral components of the chirp over a pitch period. This property seems to allow considerably deeper thresholding by the clipper than can be applied to the original waveform. This is particularly important in the "severe" processing mode described in Section 3.7 which achieves surprisingly good quality for a low clipper threshold. Various levels of the clipper are determined by enhancement needs and are described in Section 3.7.

Since the clipper is simulated (and implemented) digitally, the input to the clipper must first be upsampled and low-pass filtered to extend the frequency range (from 5 kHz to 20 kHz) so that the nonlinear clipping operation will not cause aliasing. Low-pass FIR filters were designed to roll off at 4500 Hz in order to allow adequate passband and stopband ripple. After clipping the upsampled waveform, the resulting signal is downsampled and low-pass filtered.

3.6.2 Output AGC

A slow AGC is applied to reduce any residual waveform envelope fluctuations. In order to avoid problems typical of conventional AGC's (e.g., "pumping" alluded to earlier), a voicing-dependent release time (illustrated in Figure 3-18) was used. This approach is similar to that used in the voicing-dependent (frequency-domain) AGC described in Section 3.4.2. In the output AGC, in contrast to the frequency-domain AGC, the gain acts on a sample-by-sample basis and is derived

from an IOEC with mild compression and expansion characteristics. A decrease in the peak/RMS of roughly .5 dB was achieved.

3.6.3 D/A Filter Compensation

In order to convert the digital waveform back to an analog representation, the processed waveform is passed through a D/A converter and thus a final low-pass filter is required. This low-pass filter introduces both amplitude and phase distortion and can result in as much as a 3 dB increase in the peak/RMS. Hence, a digital compensator was designed and applied prior to the D/A filter. The design of this compensator is described in Appendix D.

Figure 3-19a shows the impulse response of the D/A filter prior to applying the compensation. Figure 3-19b shows the response after passing an input impulse through the digital compensator. The resulting reponse is symmetric with linear phase shown in Figure 3-19c. With this compensation, the 3 dB peak/RMS increase (prior to compensation) was removed.

3.7 The Integrated System: System Parameters and Tradeoffs

In this section, the components of the previous sections are integrated to form the VISTA (Voice Intensification using the Sinusoidal Transformation Algorithm) preprocessor. The non-real-time computer simulation of the system is described. Parameter settings are made to create different processing modes for different operating conditions and desired peak/RMS ratios and quality levels.

3.7.1 The System

The complete audio preprocessor is illustrated in Figure 3-20. In the analysis stage, the short-time Fourier transform (STFT) is computed with a 1024-point FFT at a fixed frame rate of 10 ms and with a window duration 2.5 times an average pitch period. The pitch extraction requires both a "coarse" and "refined" pitch estimate as described in Section 3.2.1. A voicing probability is also obtained in the pitch extraction unit. The refined pitch estimate is used in the separation of the excitation and system phase functions, as well as in the separation of the excitation and

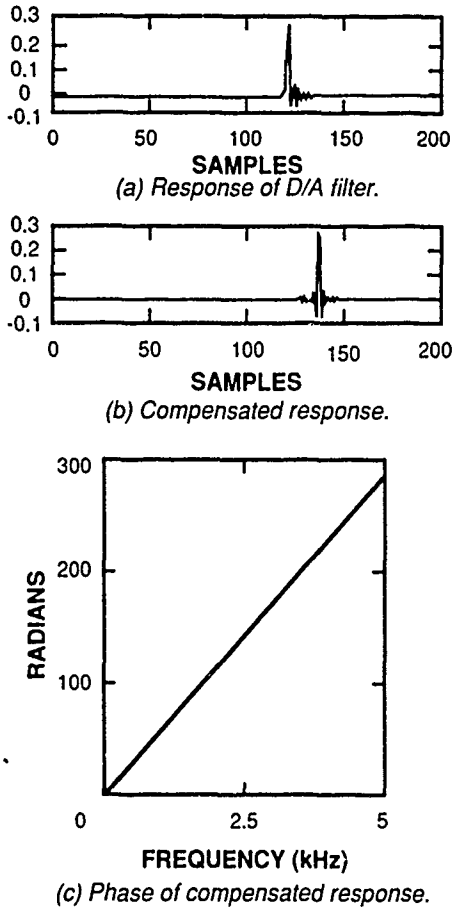


Figure 3-19. Effect of D/A compensation.

system amplitude components. The matching unit includes frequency matching, post-matching which updates the matched frequencies $\omega_k(m)$ according to a voicing-dependent frequency cutoff, and a "voiced" and "unvoiced" frequency-track assignment $b_k(m)$ which takes on a value of zero for a voiced track and unity for an unvoiced track.

In the enhancement unit the three operations of dispersion – including smoothing (Figure 3-9), spectral shaping (Figure 3-15), and amplitiude compression (Figure 3-11) work synogistically to create a waveform with a low peak/RMS. The synthetic vocal tract phase is used along all frequency tracks designated "voiced" and the original system phase is maintained along all tracks designated "unvoiced".

$$\tilde{\psi}_k(m) = b_k(m)\psi_k(m) + [\,1 - b_k(m)\,]\hat{\psi}_{k,kfh}(m) \tag{3.27}$$

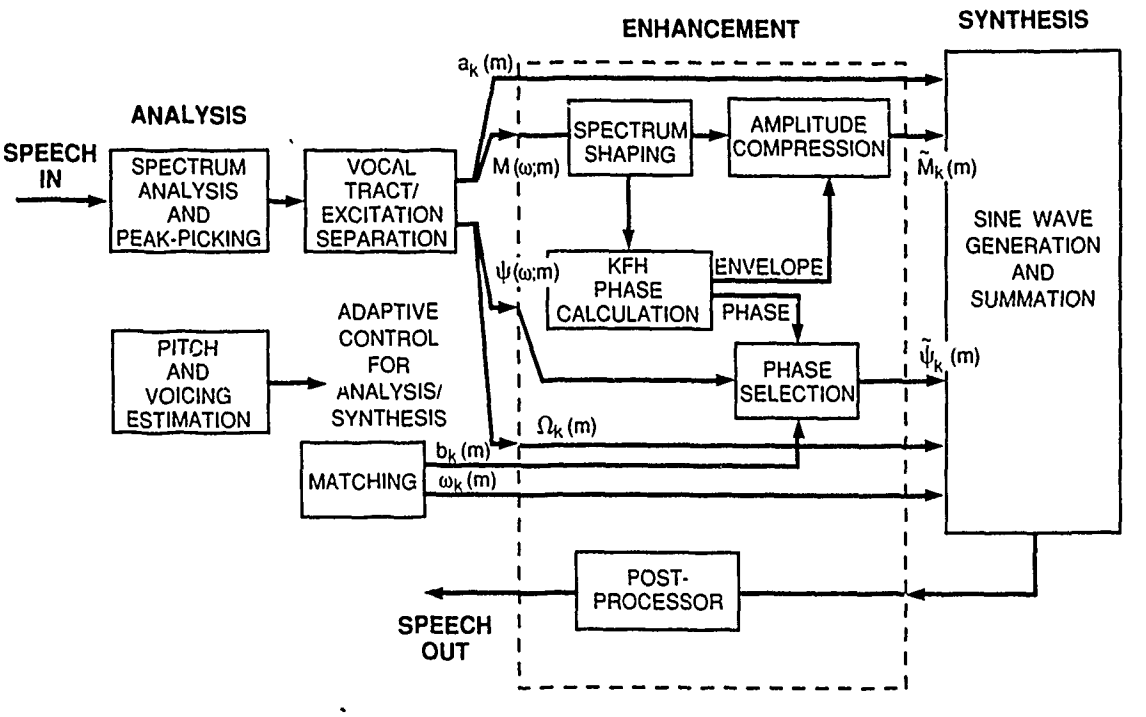


Figure 3-20. Sinusoidal Transform System for audio enhancement.

where the notation "hat" denotes the smooth KFH phase. The spectral shaper $H_c(\omega;m)$ and amplitude compression unit (with gain $G_{drc}(m)G_{agc}(m)$) modify the vocal track spectral magnitude on each frame to yield

$$\tilde{M}(\omega;m) = G_{drc}(m)G_{agc}(m)H_c(\omega;m)M(\omega;m) \tag{3.28}$$

which is sampled at the sine-wave frequencies to obtain the "compressed" vocal track amplitude

$$\tilde{M}_k(m) = \tilde{M}(\omega_k(m);m) \tag{3.29}$$

In the synthesis stage (see Figure 3-5b) sine waves are generated and amplitude modulated by the compressed sinewave amplitude components. Specifically, the excitation and system amplitudes at the matched sine-wave frequencies are multiplied and then interpolated linearly across each frame to form the amplitude modulation at the original sampling interval n .

$$\tilde{A}_k(n) = \tilde{M}_k(n)a_k(n) \tag{3.30a}$$

The excitation phase is added to the system phase at frame boundaries and this sum is interpolated using a cubic interpolating polynomial [1,2]. (Computation more amenable to real-time implementation can be performed via a quadratic interpolating function as discussed in Section 4.) The resulting phase along sine-wave frequency tracks at the original sampling interval n is given by

$$\tilde{\theta}_k(n) = \Omega_k(n) + \tilde{\psi}_{k,kfh}(n) \tag{3.30b}$$

where $\Omega_k(n)$ is the excitation phase. The processed waveform is then generated as

$$\tilde{s}(n) = \sum_{k=1}^{L(n)} \tilde{A}_k(n)\cos[\tilde{\theta}_k(n)] \tag{3.30c}$$

After sine-wave synthesis, post-processing (Figure 3-18) is performed. This includes clipping, a (voicing-dependent) AGC, and D/A filter compensation.

3.7.2 The Non-Real-Time Simulation

The integrated preprocessor was simulated on a floating point SUN3 computer with a floating point processor and tested on a number of Lincoln Laboratory and Voice of America data bases. Due to the requirement of a 1024-point FFT in the analysis and the requirement of up to 140 sine waves (up to 70 spectral peaks were allowed) in the synthesis, the simulation ran at roughly 300 times real time. The non-real time simulation evolved through a number of iterations to achieve consistency with the real-time implementation structure. The non-real time simulation was also extended and generalized to emulate fixed point and time limitations of the real-time system such as the analysis frame interval (12 ms), phase interpolation (quadratic in contrast to a cubic interpolator), phase quantization, and cosine table lookup for sine-wave generation. These issues will be further discussed in Section 4.

Figure 3-21 illustrates an example of processing a waveform with VISTA prior to post-processing and obtained using the non-real-time simulation. Two important changes in the waveform have

taken place. The peakiness with respect to a pitch period has been reduced via adaptive dispersion and the short-time and long-time envelope functuations have been reduced by amplitude compression. The two waveforms have been peak normalized so that since the processed one has a lower peak/RMS, it will sound louder than the original. There is of course some loss in quality which is typical of this kind of processing. These quality-peak/RMS tradeoffs will be further discussed in the next section and in Section 5.

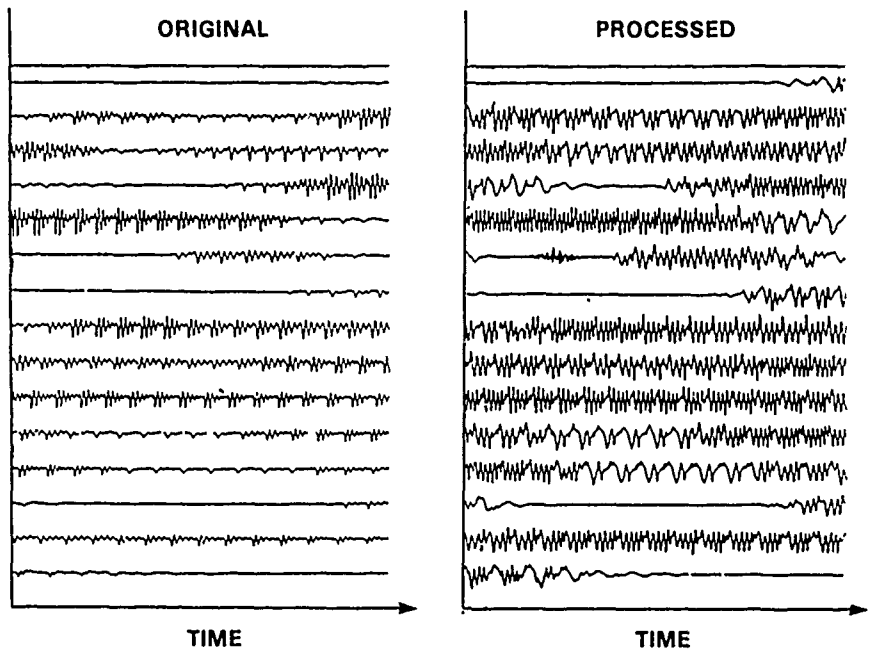


Figure 3-21. Comparison of original waveform and processed speech with combined dispersion and DRC.

3.7.3 Degrees of Processing

The diversity of the STS parameters allow for the chosing of different degrees of processing. In particular, parameter settings were determined for "mild", "normal", and "severe" processing modes. The mild mode uses extreme smoothing of enhancement parameters to achieve very high quality speech with moderate peak/RMS enhancement. In the normal mode, a balance between quality and peak/RMS is achieved. Finally, the severe mode achieves minimum peak/RMS at the

expense of speech quality but without compromising speech intelligibilty. In addition, the system allows for three different degrees of preemphasis for receiver compensation ("none", "medium", and "heavy"). Figure 3-22 show the resulting nine possible processing modes. A description of when to use the different modes in a practical setting is given in Section 4.4 which describes the user interface to VISTA.

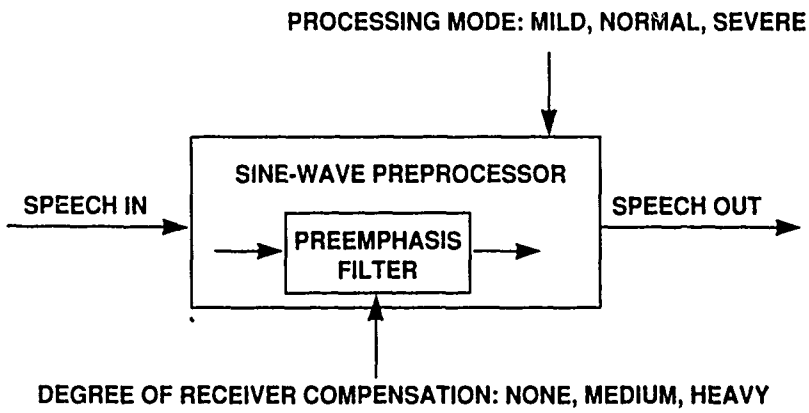


Figure 3-22. Enhancement options.

Table 3-1 shows a representative subset of parameter values used in different operations and their values in achieving the three desired processing levels, all for the medium receiver compensation. The parameter sets are subdivided into four groups: (1) dispersion, (2) frequency-domain AGC and DRC, (3) spectral shaping, and (4) output AGC and clipping. In traversing from mild to severe processing, dispersion is increased by relaxing spectral smoothing and relaxing smoothing of phase along frequency tracks. The "chirp" response duration (corresponding to the Key-Fowle-Haggarty phase calculation) is increased to do more "filling in" of a pitch period. The release and attack times for the frequency-domain AGC and DRC are reduced to make the amplitude compression more responsive. The IOEC for frequency-domain DRC is also changed according to the processing mode (see Figure 3-23). The pre-emphasis gain G_o in Figure 3-16 and the compression factor β in (3.26) are increased to achieve more spectral enhancement. The output "voicing

dependent" AGC release time is relaxed during voicing and the clipper threshold is lowered. The IOEC for (voicing-dependent) output AGC does not change with mode (Figure 3-24). Preliminary settings were made in the non-real-time system and later refined in the real-time implementation described in Section 5.

TABLE 3-1.

Parameter Selections for Ennancement Options (Medium Preemphasis)

OPERATION	*MILD*	*NORMAL*	*SEVERE*
DISPERSION			
Magnitude Smoothing (α)	.5	.7	.8
Phase Smoothing (β)	.999	.95	.999
Chirp Duration (μ)	.5	.8	.97
AGC and DRC			
Release Time (α_r)	.98/.4	.95/.2	.9/.1
Attack Time (α_a)	.0/.2	.0/.2	.0/.1
SPECTRAL SHAPING			
Adaptive Gain (G_0)	.1	.2	.3
Sharpening Factor (β)	.23	.23	.25
OUTPUT AGC AND CLIPPER			
Release Time (α_r)	.995	.997	.99
Threshold (w/r signal max)	.73	.67	.4

The options for the receiver compensation $H_r(\omega)$ are shown in Figure 3-25. These options are independent of the processing mode (i.e., "mild", "normal", or "severe") and were included per recommendation of the VOA. The "none" mode exists for listening to the processor directly from its output (i.e., without radio transmission). "Medium" mode was set so that the processed speech, when listening in the laboratory, sounds "sufficiently crisp" through the Lincoln Kenwood short-wave radio and "comparable in crispness" to a state-of-the-art commercial processor [21] when its preemphasis is also set in a medium mode (see Section 5). When combined with adaptive preemphasis (3.25), the resulting composite preemphasis $H_p(\omega; m)H_r(\omega)$ in normal operating mode and for steady-state voiced speech, at 5 kHz, is about 8 dB above baseband. Finally "heavy" preemphasis was chosen for operation in a more severe real-world operating environment. When combined with

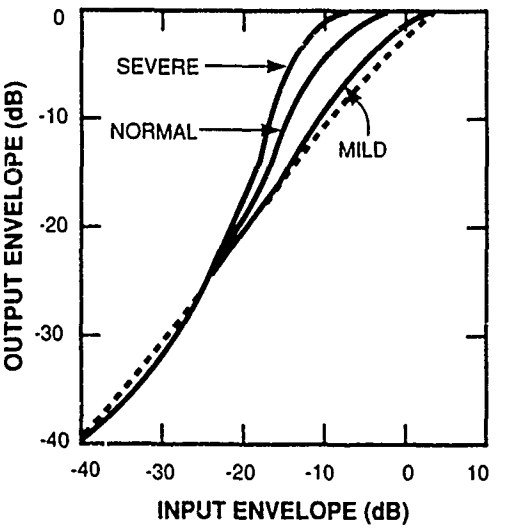


Figure 3-23. Input/output envelope characteristics (IOEC) for frequency-domain DRC.

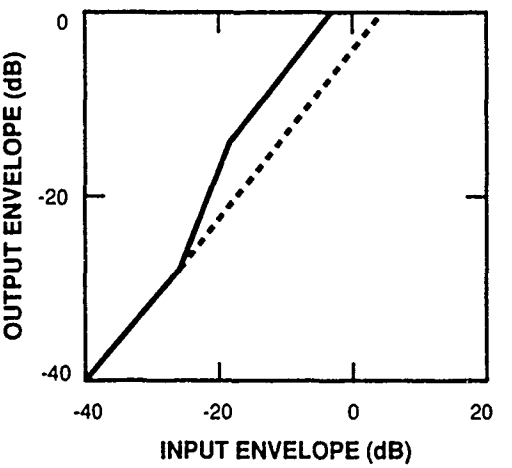


Figure 3-24. Input/output envelope characteristics (IOEC) for output time-domain AGC.

adaptive preemphasis, the resulting preemphasis in normal operating mode and for steady-state voiced speech, at 5 kHz, is about 14 dB above baseband. All three receiver compensation modes use a high-pass filter, recommended by Birth [18], which rolls down by about 6 dB/octave starting at about 300 Hz. This was included to protect the transmitter from DC. It has the additional advantage of improving "brightness" of quality. Only one composite preemphasis with a maximum boost of about 6 dB at 5 kHz was incorporated in the non-real-time system in the early stages of system development; while in the real-time system all three options with the highpass filter were incorporated and refined (see Section 5).

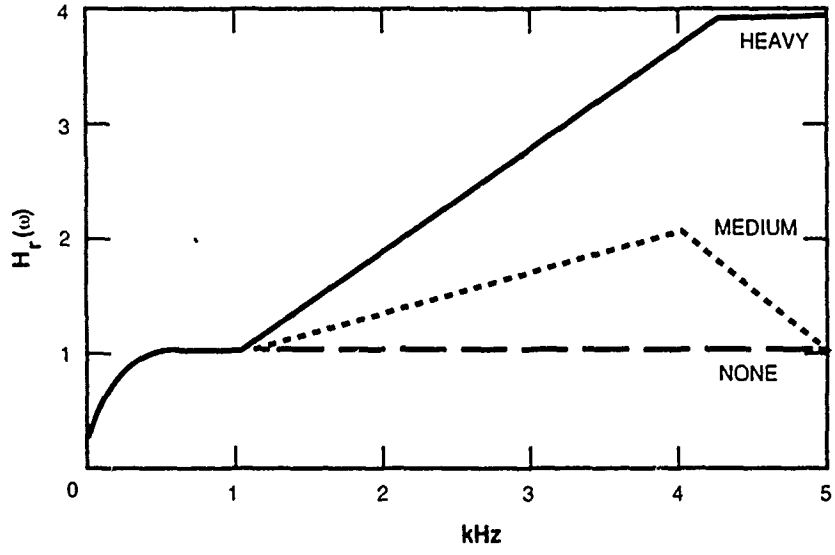


Figure 3-25. Receiver compensation options.

The specific peak/RMS for the nine processing states are illustrated in Table 3-2 (see Section 5.2.4 for a description of the data bases). The VISTA parameters were refined for each state so that for a particular processing mode, the peak/RMS remains roughly intact for different choices of receiver compensation. The peak/RMS decreases roughly to 2 dB in going from mild to normal processing mode, and decreases again by roughly 2 dB in going from normal to severe processing. The peak/RMS values in Table 3-2 should be compared against a peak/RMS of roughly 14.5 dB for unprocessed speech. The measurements were made using the *histogram method* described in Section 5.1.4. These peak/RMS reductions correspond to an increase in loudness (under a peak-

power constraint) and some quality loss. Further discussion of these and other measurements, the associated quality levels, and comparative performance with state-of-the-art commercial devices is given in Section 5.

TABLE 3-2.

Average Peak/RMS (Real-Time) for Processing Mode and Preemphasis Options

PREEMPHASIS	*PROCESSING MODE*		
	Mild	Normal	Severe
None	9.0 dB	6.8 dB	5.0 dB
Medium	8.7 dB	6.8 dB	5.1 dB
Heavy	8.9 dB	6.8 dB	5.0 dB

Note:
- Averaged two data bases (Lincoln and NOVA)
- Unprocessed average peak/RMS = 14.5 dB

3.8 Summary

In this section, a new frequency-domain approach to speech enhancement (VISTA) was presented which is based on a sinusoidal representation of speech and which uses a radar signal design solution. Significant reduction of peak/RMS and increase in waveform loudness (under a peak-power constraint) were obtained. Various processing options were defined and some peak/RMS measurements were illustrated. The next sections investigate the real-time implementation of VISTA and give further measurements of peak/RMS, quality, and intelligibility.

4. REAL-TIME VISTA PROTOTYPE SYSTEM

In this section, the implementation of VISTA in a real-time prototype system using multiple digital signal processing boards is described. The problems and solutions in moving from a non-real-time, single processor, floating-point simulation to a real time, multi-processor, fixed-point implementation is described. The user interface is also discussed.

4.1 Multi-Processor Hardware Structure

The real-time VISTA system operates in a serial array of seven processors designed around the ADSP2100 (Analog Devices) microchip [5]. The ADSP2100 was chosen primarily for its ease of programming and speed of execution. The ADSP2100 is a 16-bit fixed-point unpipelined processor with a 125 ns cycle time, an arithmetic/logic unit, a multiplier/accumulator, a barrel shifter, no overhead loop control, and a powerful multifunction capability. The ADSP2100-based multiprocessor architecture supports high-speed interprocessor communication together with analog I/O and DSP peripheral I/O. Additional resources include 16K 24-bit words of DSP program instruction memory, 16K 24-bit words of program data memory, 8K 16-bit words of DSP data memory, and a set of special control/status registers. Adjacent processors communicate directly via an interprocessor dual-port memory, a block of 1024 data memory locations which is shared by the two processors. Access to memory shared with the processor to the left involves transfer over a VMEbus and requires two cycles for a read or write which is invisible to the program itself. Analog I/O is available on boards #2 and #7 only, and the sampling clocks on these boards are independent. Options such as sampling interval and preemphasis and deemphasis are programmable.

In the development stage of the real-time VISTA system, the multi-processor array was connected via a VMEbus and bus repeater to a SUN workstation with a UNIX operating system. This configuration is shown in Figure 4-1. In addition to the software tools provided by Analog Devices, a multi-processor debugging package developed at Lincoln Laboratory was available. This provided the means for downloading the processors, setting breakpoints, and reading and writing both program and data memory during real-time execution. The debugger made it possible to perform many A:B listening tests to determine the final setting of the system parameters.

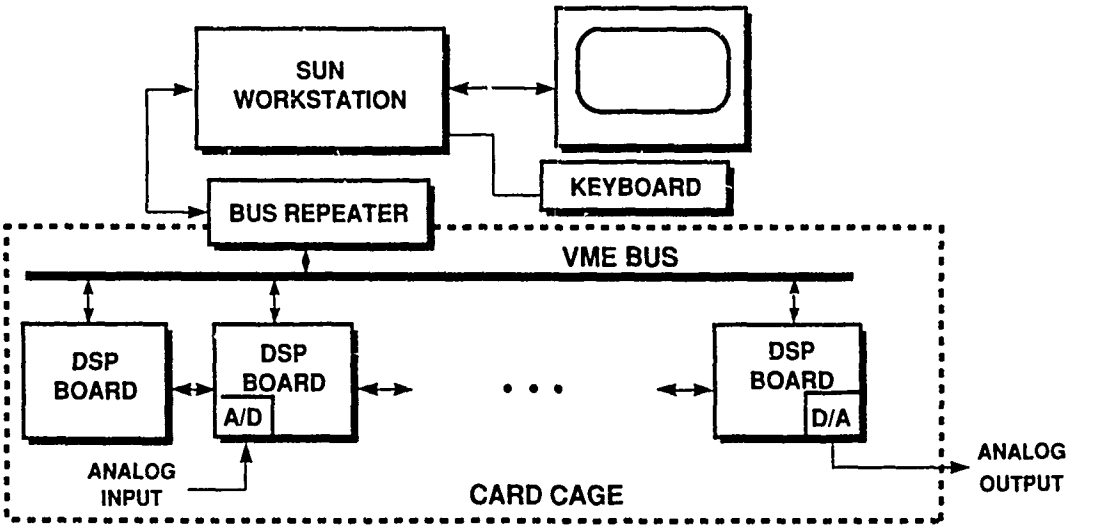


Figure 4-1. VISTA system development facility.

4.2 Implementation Problems and Solutions

There were two major issues to be addressed in converting the VISTA system from non-real-time to real-time. The first was going from floating point to fixed point arithmetic without compromising performance. The second was streamlining certain non-real-time modules to keep the number of required processors reasonable.

4.2.1 Floating- to Fixed-Point Conversion

Even when 16 bits are sufficient to accommodate the range of final results, intermediate calculations may overflow or underflow this limit. A standard technique for maintaining accuracy during fixed point calculations is the use of block normalized data with a corresponding block exponent. For example, before each stage of the 1024-point FFT, the data is shifted to prevent overflow and avoid unnecessary underflow, and the block exponent is updated. An exponent detector, as part of the barrel shifter in the ADSP2100, greatly eases the burden of preserving accuracy using this method. Where appropriate, logarithmic arithmetic has been used, and computations have been done in double and triple precision, where critical. Multiprecision is often trivial to achieve as the

multiplier/accumulator of the ADSP2100 supports a 40-bit product. Computing the phases would be cumbersome even in floating point arithmetic, and execution time would exceed a reasonable limit. Phases are determined via a table lookup, and the resulting quantization has produced no perceptible loss in performance of the system. A piecewise quadratic sine-wave synthesis has replaced the cubic sine-wave synthesis used in the non-real-time system because it obtains an accurate polynomial fit to the phase trajectory and is much more straightforward to implement.

4.2.2 Algorithmic Streamlining

A number of the non-real-time modules were modified for implementation in real time. Increasing the frame interval from 10 ms to 12 ms provided a more comfortable cushion in the processor dedicated to performing the 1024-point FFT. The "coarse" pitch extractor whose sole function is to establish the adaptive window for the 1024-point FFT needs only 1000 Hz of the available 5000 Hz bandwidth, so the input data was passed through a downsampling filter and required only a 256-point FFT to determine the spectral information. The resolution of the candidates for the computationally-intensive, "refined" pitch extractor was increased from .5 Hz (non-real-time) to 1.22 Hz (real-time) to reduce execution time significantly.

Another issue which arose in the conversion to real time involved the method of choosing peaks and the limit on the number of sine waves after frequency matching. In choosing the peaks in the non-real-time system, the largest 70 peaks were first found and then they were thresholded so that no peak below about 50 dB of the maximum peak was allowed. In the real-time system all peaks were found over the band and above this same threshold. After matching and "post-matching" (Section 3.2.2) in the real-time system, a maximum of 120 frequencies were enforced, while a limit of 140 was imposed in the non-real-time system after matching (corresponding to 70 sought peaks), and any number of additional frequencies were allowed after "post-matching". Due to this limit on the peak number after matching in the real-time system, the allowable frequency band for peak picking was reduced from 5000 Hz to 4500 Hz. This reduced the possibility of exceeding the 120 limit and, thus, the abrupt termination of a sine wave which can produce jitter and glitches in the waveform construction. The implication of reducing the band will be discussed in Section 5.

Additional real-time implementation issues included partitioning of the modules for a multiprocessor system, efficient interprocessor data flow, and the accommodation of separate clocks for the A/D and D/A. These issues will be discussed in Section 4.3.

4.2.3 Comparing the Real-Time and Non-Real-Time Systems

Throughout the development of the real-time system, comparisons were made with the non-real-time system. Some of the real-time enhancement subroutines were tested with the exact digital data used in the non-real-time system so that differences in numerical accuracy could be observed and assessed. The quantized and streamlined algorithms of the real-time system were incorporated into the non-real-time system for experimentation. In particular, phase quantization, piecewise quadratic synthesis, a 12 ms frame interval, downsampled input to the coarse pitch extractor, and less resolution in the refined pitch extractor were all simulated in the non-real-time system to verify integrity of performance. Finally, many hours were spent listening to the audio output of both systems as a prelude to fine tuning the parameters of the real-time VISTA system.

4.3 Software Structure and Modularization

The configuration of the seven processors and their functions are shown in Figure 4-2. The results of the partially processed data travel from boards #2 through #7, one frame at a time, while board #1 serves as a satellite processor for board #2. Each processor introduces an additional frame delay with the exception of #6 which operates in parallel with #5. Since the analog I/O of boards #2 and #7 are on separate clocks, *Analyzer A* in board #2 functions as the master timer for the entire system. At the 12 ms frame boundary, a flag in dual-port memory is set for board #3 which is idling in anticipation of this event. Similarly, boards #4 through #7 receive indication of the frame boundary from their respective left neighbors. The *Postprocessor* in board #7 must provide feedback through the system of the number of samples it has played out during this frame, which may vary by ±-1, so that during the next frame time, the sine-wave synthesizer will accommodate clock drift by generating the correct number of output samples. The *Postprocessor* tolerates this jitter by including a very small cushion in its double-buffered output. At every frame boundary,

all or part of the necessary data from each board resides in interprocessor memory for its right neighbor. In the case where the amount of data to be passed exceeds 1024 words, a second batch of data is written to right memory when the first transaction has been completed. This additional data exchange between a single processor and its neighbors was designed carefully to interleave transactions to maximize efficiency.

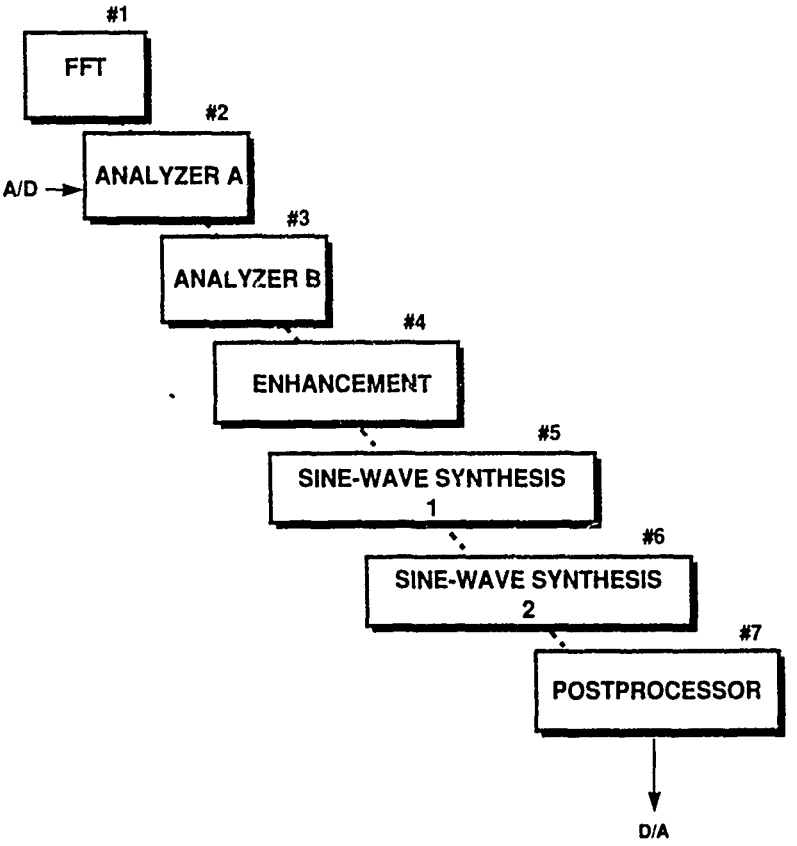


Figure 4-2. Processor functions.

The various modules of the system are partitioned among the seven ADSP2100 processors as follows:

1. Board #1 contains a 1024-point FFT which is executed upon demand by *Analyzer A*. This computation takes about 75% of the available 12 ms.

2. Board #2 consists of *Analyzer A* which controls the frame timing of the entire system and includes the A/D input, a downsampling filter, a 256-point FFT, a coarse pitch extractor to determine the window for the 1024-point FFT, a peak-picker on the results of the previous 1024-point FFT, and the first portion of the refined harmonic pitch extractor. The maximum measured computation time is 77.5% of real time.

3. Board #3 accommodates *Analyzer B* which includes the last portion of the refined pitch extractor, a spectral envelope estimator (SEE), a generator of phases at the SEE peaks, a spectral enhancer, and an excitation amplitude generator. The measured computation time is 84.2% of real time.

4. Board #4 contains the *Enhancement* module which includes a pitch smoother, an excitation phase generator, spectral preemphasis, an envelope smoother, a Key, Fowle, Haggarty phase generator, voicing-dependent AGC, dynamic range control, a system phase generator, a peak-matcher, a post peak-matcher, a Key, Fowle, Haggarty phase smoother, and a parameter generator for the sine-wave synthesis. The measured computation time is 87.1% of real-time.

5. Boards #5 and #6 contain *sine-wave synthesis* 1 and 2, each of which is capable of reconstructing 60 frequency tracks. The resulting waveforms are summed to produce the final waveform to be passed on to board #7. The measured computation time for each of these processors is 97.7% and 98.5% of real time, respectively.

6. Board #7 accommodates the *Postprocessor* which includes the D/A output, the upsampling/downsampling filter for the clipper, the compensation filter, a voicing-dependent AGC, three peak/RMS ratio calculations for intervals of $\approx$ 1.6 seconds, $\approx$ 2.5 seconds, and $\approx$ 3.6 minutes (see Sections 4.4 and 5.1), and an option to play out the raw input speech which has been peak-normalized to match the processed speech (see Section 4.4). The maximum measured computation time is 96.7% of real time.

4.4 The User Interface

4.4.1 Overview

A photograph of the entire VISTA system was shown in Figure 2-4. The photo shows (1) the enclosure for the special purpose digital hardware, (2) the terminal, and (3) the keyboard. The

power switch and reset button are located on the lower right side of the large enclosure. The front panel shown in Figure 4-3 has the following elements:

1. power indicator
2. digital VU meter to monitor input level to A/D
3. input attenuator (between input and A/D)
4. overflow indicator (at input to A/D)
5. processor/bypass switch

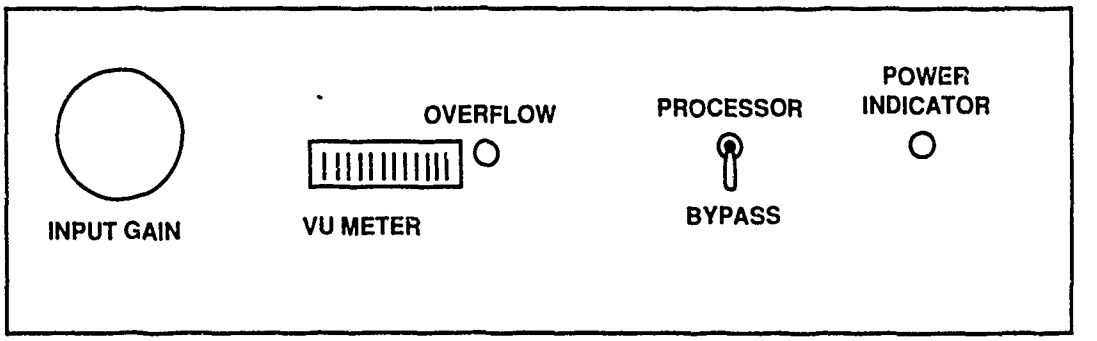


Figure 4-3. Front panel VISTA processor.

Turning the power switch on initiates a process which downloads machine code from non-volatile memory on the EPROM board to the volatile memory on the ADSP boards. When all the boards are loaded the real-time program will be automatically started. The enhancement processor will then be in the normal enhancement mode without preemphasis — the modes recommended for studio listening. The processor/bypass switch on the front panel should be set to processor.

The processor modes are controlled by typing commands from the terminal. The software has been programmed to respond to the minimum sequence of input keystrokes which uniquely specifies a legal command. Typing an illegal command causes the unit to type a message suggesting that the user type "HEL" for "help" to get a list of the legal commands. In general, a carriage return

is not required following a command. Pushing the "RESET" button near the on/off switch causes the system to reload its programs and come up in the default mode.

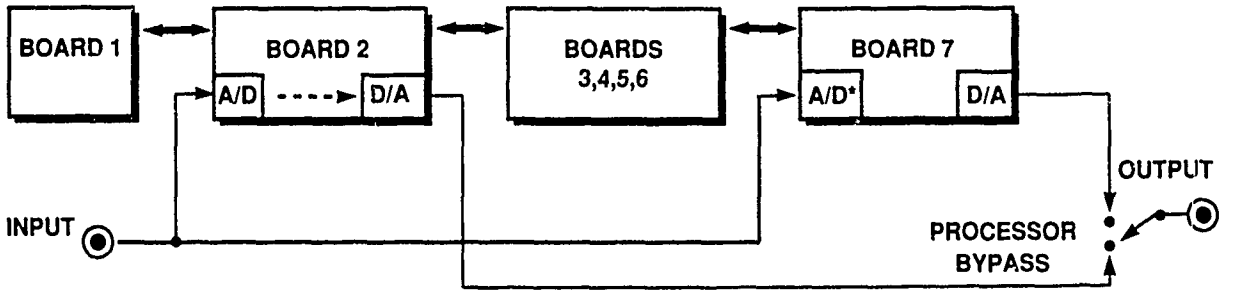
The enhancement processor has two features for bypassing the VISTA processor. The first feature is controlled by a switch on the front panel labeled "Processor/Bypass". In the "PROCESSOR" mode, the input A/D on board 2 provides speech to the rest of the boards while the D/A on board 7 drives the output, as shown in Figure 4-4. In the "BYPASS" switch position, the system output is driven by the D/A on board 2. The digital input to this D/A is taken directly from the A/D on the same board. This feature allows the user to test the analog circuits on board 2 and ensure that overload in the A/D is not occurring.

The second feature is "bypass with AGC". This mode is controlled by a terminal command, "R" for "RAW". In this mode, the VISTA system is bypassed and in its place a very long release time AGC is performed. This speech will sound the same as the input speech except it will be gain-normalized to have the same peak level as the processed speech. The feature allows loudness comparison of processed speech with unprocessed speech. This feature is achieved by having the input come from an A/D on board 7 as shown in Figure 4-4. A digital AGC algorithm on board 7 normalizes the speech which is then output by the D/A on the same board.

4.4.2 Enhancement Modes

Four enhancement modes are provided. Modes are switched via the terminal by typing just enough letters to define a unique command word. The system will respond by echoing the full command name. Typing unnecessary characters and/or a carriage return is not advised as this will generally result in an "unrecognized command" message. In particular, typing extra characters after the command "MO" for entering the monitor mode will pop the user right back out of it. The command word, the mode name, mode description, and intended use are given in Table 4-1.

The four enhancement modes provided allow the user to match the degree of enhancement to the anticipated broadcast environment. In the "NORMAL" mode, the enhancement unit uses parameters which allow the performance to strike a balance between decreased peak/RMS ratio and maintenance of speech quality. The normal mode is intended for most broadcast conditions. In the



*A/D INPUT ON BOARD 7 IS UNDER SOFTWARE CONTROL (See "Raw" Command)

Figure 4-4. Diagram showing A/D and D/A connections.

"MILD" mode, higher quality is achieved by using less aggressive DRC and AGC algorithms and by using less phase dispersion. The "MILD" mode is intended for broadcasting dramatic or political material where speaker effect and emphasis is important. In the "SEVERE" mode, the parameters are modified to give close to maximal decrease in peak/RMS ratio. The severe mode is intended for maximum punch-through for noisy and long-range broadcasting where the information content is judged to be most critical. Long-term listening in this mode could be fatiguing to listeners and, therefore, it should be restricted to short programs. The "NATURAL" mode is provided for testing and comparison purposes. All of the enhancement algorithms are bypassed in this configuration. The speech is, however, processed by the sine-wave analysis/synthesis.

Three modes of preemphasis are provided. It is very important that the preemphasis mode match the intended use or the speech will sound either slightly "tinny" or low-passed. Modes are switched at the terminal by typing a unique command string. The commands, the mode name, the description of the mode, and the intended uses are given in Table 4-2.

TABLE 4-1.

Enhancement Modes

Command	*Mode Name*	*Description*	*Intended Use*
NA	NATURAL	By-passes the enhancement algorithms but uses the sine-wave analysis/synthesis algorithms.	Comparison tests.
MI	MILD	Uses extensive smoothing of enhancement parameters to achieve very high quality speech with less peak/RMS enhancement.	Radio plays or political speeches where speaker voice modulation is important.
NOR	NORMAL (default mode)	Balance between quality and peak/RMS ratio has been achieved.	Most radio broadcasts.
SE	SEVERE	Maximal enhancement in peak/RMS ratio. Speech quality is compromised but intelligibility is not.	Short broadcasts with high information content where maximum radio range is desired.

4.4.3 Monitoring and Verifying Performance

As shown in Figure 4-5, typing the command word "MO" (for monitor) will cause the terminal to display the peak/RMS ratio for three time intervals:

1. 1.6 seconds (16384 samples)
2. 26.2 seconds (16 x 16384 samples)
3. 3 minutes, 29.7 seconds (128 x 16384 samples)

The peak/RMS ratio is given in dB to two decimal places, although variations of less than ± 0.2 dB are deemed insignificant. In order to change the mode of the processor or to simply terminate the peak/RMS readout, type any character. The procedure for measuring peak/RMS is given in Section 5.1.4, and is referred to as the *interval method.*

TABLE 4-2.

Preemphasis Modes

Command	*Mode Name*	*Description*	*Intended Use*
NON	No Preemphasis (default mode)	Optional preemphasis is not activated.	Studio listening.
ME	Medium Preemphasis	Preemphasis is designed for typical HF receiver.	Most broadcasts.
HEA	Heavy Preemphasis	Preemphasis is designed for very narrow band receivers.	Broadcasts where a very "bright" sound is desired.

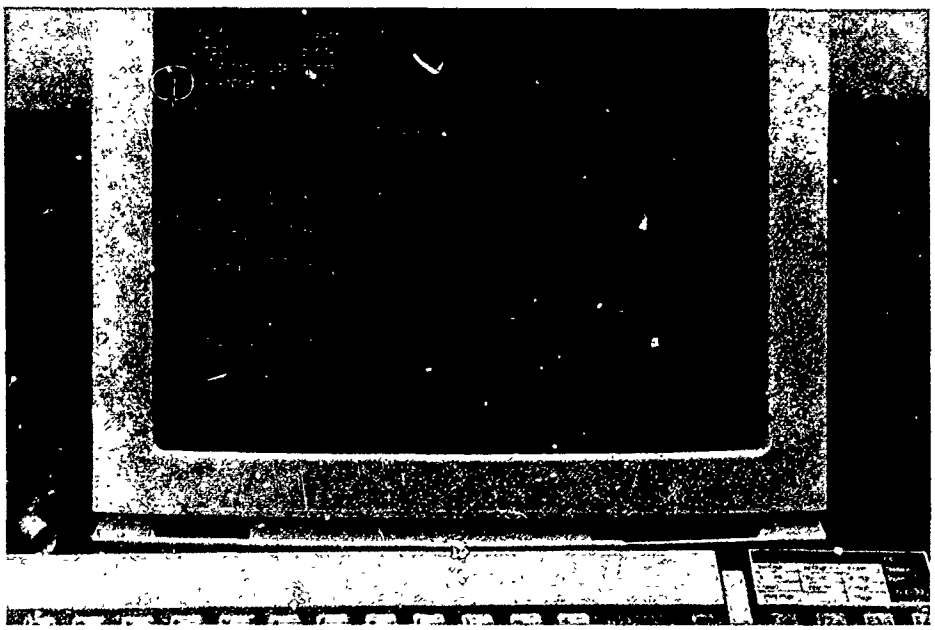


Figure 4-5. Monitor displaying peak/RMS ratios..

Two other modes are available for test purposes. The first test mode replaces the digitized audio input with an internally-generated pulse train with a 6 ms period. This mode is activated by typing "PU" for "pulse". In any of the three enhancement modes, the output signal will be a periodic chirp signal. This is an optimally-dispersed signal with a uniform envelope and a very low peak/RMS ratio (5.6 dB for normal mode). This test mode verifies that all of the digital processor

boards, as well as the output D/A circuits, are working. To return to the mode of processing speech, simply type "SP" for "speech".

The second test mode is useful for understanding the operation of the system. This mode is called the zero-phase mode and can be activated by typing "Z" for "zero". In this mode the optimal KFH dispersing phase has not been applied to the speech, but only a linear excitation phase as described in Section 3.2.2. The natural phase of the speech has been removed thus yielding a "zero-phase" system. The resultant speech waveform is symmetric about each pitch-pulse. When compared to unprocessed speech with the same peak levels, the zero-phase speech will sound very quiet.

Another feature of the user interface allows comparisons to be made. The user sets up one desired mode then types "SA" for "save". The system responds by asking the user to specify "A" or "B". After this selection, the user can select another mode and save it labeled as "B". Thereafter, until different designations are made, the user can type "A" or "B" to restore the system to the predefined modes. Thus, rapid switching for listening tests can be accomplished.

If for any reason the user interface software gets into a non-functional mode, push the "RESET" button next to the on/off switch on the lower right side of the large enclosure. We have programmed a test mode which allows us to check many features of the system. Thus it is possible, although very unlikely, that the user might accidentally type some command which places the unit in such a mode.

All of the system commands have been further summarized in a user interface manual [22].

5. VISTA ALGORITHM TEST AND EVALUATION

As part of the algorithm development process, it was necessary to establish a variety of test facilities and several data bases in order to make objective quantitative measurements of system performance. These facilities and data bases were used to assess the state-of-the-art of commercial processors, to provide insight into problem areas with the Lincoln Laboratory algorithms as they were being developed, and to provide a reliable and consistent way to evaluate the final system. An explanation of the concepts underlying the evaluation of the speech enhancement algorithms, and a description of the facilities and data bases used to perform the evaluations follow. Finally, both non-real- and real-time test results will be given, and comparisons with commercial devices will be made.

5.1 Evaluation Criteria

5.1.1 Choice of Criteria

In order to establish appropriate evaluation criteria, we first need to establish the purpose of the speech communication system and the environment in which it will be used. For the VOA speech enhancement application, it is instructive to consider two distinct listening environments. The first audience is listeners of HF radios located near enough to the transmitter so that background noise (natural or man-made) is negligible. For these listeners the naturalness of the speech or its "quality" is of prime importance. The second audience is listeners located far from the transmitter who receive low signal levels and therefore hear the speech with background noise. For these listeners, intelligibility-in-noise is of prime importance.

Intelligibility and quality, therefore, are the two primary speech characteristics to be measured. Throughout the development of VISTA a number of different procedures, both informal and formal, were used in evaluations. This section first describes our subjective measures of intelligibility and quality and other more formal measures. The section ends with a discussion of peak/RMS which is an objective measure of loudness and, thus, can also be used as a measure of intelligibility-in-noise. The implications of lowering peak/RMS on transmitter power are also described.

5.1.2 Intelligibility Testing

In performing speech preprocessing, the intelligibility of the processed speech should ideally equal that of the original speech when listening in the clear. When listening in a noise environment, the intelligibility should ideally improve when compared with the original peak-normalized speech. To measure the intelligibility one could use single words, single sentences, or longer passages and paragraphs. Single words, which can be as short as 300 to 500 ms, can fail to exercise (or excite) some of the characteristics of speech enhancement algorithms. Thus, well known tests such as the Diagnostic Rhyme Test (DRT) [23] may not be the most appropriate test. The DRT may also not be appropriate for testing in a noise background. The DRT is based on discrimination of sounds such as "b" vs "d" in, for example, the words "bad" vs "dad". In particular, the listener is presented one of the two words and asked to choose one word from a pair. Since the speech preprocessors often have a "settling time" from initial operation, testing subtle differences in initial consonants may not be fair. On the other hand, the use of very long passages usually requires the listener to comprehend the passage. Because the difficulty of passages can vary and because the experience, education, and intelligence of listeners can vary, the use of long passages generally gives test scores with very large variance. Sentence tests, therefore, appear to be most appropriate.

The Sentence Verification Test (SVT) is one such test developed by Pisoni [24]. In the SVT we make measurements at several signal-to-noise ratios and therefore, can calibrate system effectiveness in terms of increased immunity to noise. Since obtaining immunity to noise is a major goal of the system, the SVT test results can be taken as a measure of the degree of the success in achieving the major goal. The results of SVT are an average of the score of many human listeners on an intelligibity task. Because the test requires the use of many human subjects, it is expensive to run and difficult to schedule. The SVT is therefore best used to validate system performance rather than an on-going tool useful for algorithm development.

The SVT evaluation is based on listeners responding "true" or "false" to a short sentence such as "snow is black". In addition, the listeners are asked to transcribe what they hear. Three measures have been used to generate results:

1. accuracy of the true/false response

2. reaction time of the correct true/false responses

3. accuracy of the transcription in terms of percent words correct.

We have found the transcription accuracy to provide test results with the least variance and hence, we use this measure in reporting results. We believe, however, that the tests (1) and (2) above may also be sensitive to degradation of the speech and might be useful for our purposes.

The tests were conducted at the Speech Research Laboratory at Indiana University under the direction of Dr. David Pisoni. Professor Pisoni served as a consultant on the VISTA project and has intensive experience in measurement of the quality and intelligibility of speech. Each data point, corresponding to a specific system under test and a defined signal-to-noise ratio, requires the use of several subjects to reduce the variance of the test scores. Twelve subjects per point were used, 3 systems, and times 3 signal-to-noise ratios, bringing the total number of subjects to 108.

5.1.3 Quality Testing

Speech quality is a matter of human judgement. The judgments can be made in a variety of ways (e.g., blind – where the listeners don't know the identity of the system being compared; double blind – where, in addition, the test evaluation personnel doesn't know the system identities), can use a variety of scales from pseudo-absolute (e.g., good, fair, bad) to strictly comparative (A better than B, B better than C, etc.), and can involve a single listener or a large panel of listeners. The more formal the procedure, the more costly and time-consuming and hence inappropriate for on-going system development. Hence, during development we have opted to have the system developers and others at Lincoln Laboratory assess system quality. With experience, such individuals develop extreme sensitivity to the nuances of speech which makes them highly valuable for the continual assessment of speech quality. Evaluation of the final product is another matter. Because of the subjective nature of the quality assessment, a formal quality assessment is appropriate.

Our informal speech quality measure vs degree of enhancement can be effectively illustrated with the aid of a graph of the form shown in Figure 5-1. This figure sketches speech quality as a function of the reduction in peak/RMS ratio. The quality level of 10 is assigned to the natural

speech and 1 is assigned to the bandlimited clipped speech. Typical quality decreases as peak/RMS is reduced. Two systems with "equivalent quality" through informal pairwise comparisons will fall roughly on a horizontal line, as illustrated. It should be emphasized, however, that this quality measure is subjective. Different subjects will perceive quality differently and so may not agree on the relative quality level. A particular artifact may not be as objectionable to one subject as to another. Consequently, quality judgement of one speech processor may not be consistent among subjects. Inconsistencies in quality assessment may also arise among speakers being judged.

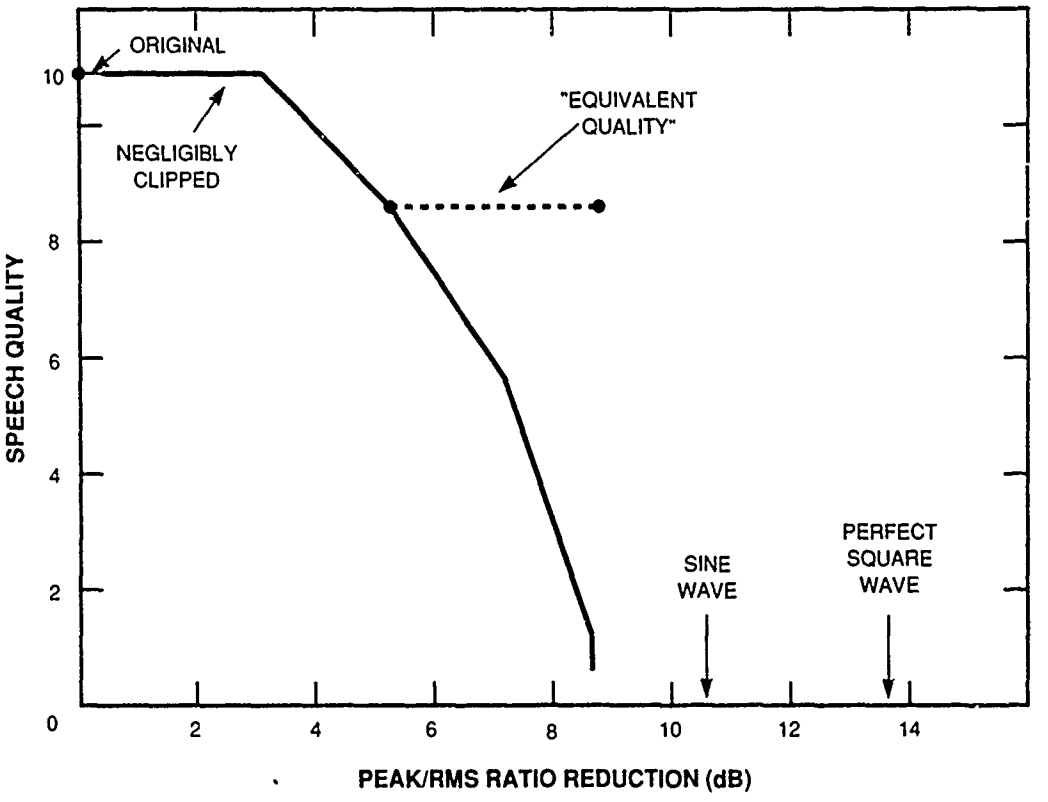


Figure 5-1. Quality vs peak/RMS ratio.

To clarify the informal quality comparison process used in the ongoing development, the procedure will be described in detail. The equipment to be compared were located in a conference room which was the same size as a regular two-person office (9 x 18 ft). The room had no sound absorbers nor sound insulation. The ambient noise level was average for an office environment.

The two systems to be compared were adjusted to give approximately the same quality as judged by the system developers (four listeners). The output gain of one system was adjusted to a comfortably loud listening level and the gain of the second system was adjusted so that in the opinion of the listeners it was at the same loudness as the first system (see next section for a discussion on loudness). This judgement was made to within ±1 dB. We then fine-tuned the settings of the two systems until the quality of the two systems were comparable. Two speech processors simply do not sound the same, and they both sound different from natural speech. The four listeners did not always agree on the nature or the degree of the distortions or artifacts in each of the speech systems. To one listener, a system might sound slightly muffled on a particular speech segment but another listener may hear it as less high-end frequency boost. Even when the listeners could agree on the nature or degree of a particular distortion they would not agree on the importance of that particular distortion relative to some other distortion. We also found that different talkers and loudspeakers or head phones affected our judgments. What we did find, however, is that for large quality differences we could agree on an overall judgment that A sounded "better" than B. After we reached a consensus on speech quality, we then recorded a passage and measured the peak/RMS for each system.

Quality assessment has been made formal in a test called the Diagnostic Acceptability Measure (DAM) [6]. The test is formal in the sense that the listeners are asked to rate various aspects of their perception of the speech (e.g., harsh, nasal, low-pass, buzzy, etc.). Many listeners are used and the various scores are averaged. The listeners are also trained on a variety of speech types and can calibrate them.

This procedure is expensive and time-consuming to use, and is difficult to use for interactive refinement of the system. In addition, formal testing of quality and acceptability for the VOA application should properly be performed in the VOA broadcast environment, where the listening criteria may be different than for the DAM. As a consequence, we did not use the DAM test during our system development process, but did use it once on one configuration of the real-time system, just prior to shipment to VOA.

5.1.4 Peak/RMS, Loudness, and Transmitter Power Considerations

We assume the louder a signal in noise, the greater its intelligibility. The measurement of the loudness of speech can be very complex if one chooses to take into account the multiple perceptual phenomena which play a part. For example, the perceived difference in loudness of two signals depends on the absolute levels of the signals. Perceived loudness also depends on the spectral and temporal character of the signals. For reasons of expediency, we have opted for two simple measures. The first loudness measure is purely objective, it depends on the average power in speech. We assume that if two signals are normalized to have the same peak value, as is required by an AM transmitter, then the signal with the larger RMS value will be louder. Thus, the ratio of the RMS value to the peak value is the objective measure. For historical reasons, this is expressed in decibels as the peak/RMS ratio, thus 7 dB peak/RMS is louder than a 13 dB peak/RMS signal. How this ratio is actually measured will be discussed shortly. Our second loudness measure is totally subjective. A listener is asked to adjust the level of two signals so that they are perceived to be the same loudness. This level is set to be comfortably loud. The signals are then peak-normalized (using an oscilloscope). The difference of the attenuator settings (we used a stepped attenuator with 1 dB increments) is then the loudness difference. The precision of the measurement is $\pm.5$ dB and the accuracy for a single listener could be as poor as ± 1 dB. By using many listeners, the accuracy is increased.

There are several important issues in measuring the peak/RMS ratio of speech. The first issue is how to deal with pauses or gaps in the speech. If these pauses are included in the measurement of the RMS level, then the measure will depend on the characteristics of the speaker thus diminishing the ability to make sensitive distinctions between enhancement algorithms. For any digital system, the problem is avoided by including a sample in the RMS measure only if it exceeds some threshold. We set this threshold at 1/100 (-40 dB) of the nominal peak value of the signal. While this is an arbitrary criterion, it solves the practical problem of making reproducible and meaningful measurements which are less dependent on the test signal characteristics. The second issue is how to define the peak mathematically in a way consistant with how the peak level is defined operationally at an AM broadcast station. At an AM broadcast station, the peak of the modulating signal is

defined by a clipper which prevents phase reversals in the transmitter carrier. If the operator sets the input speech level too low so that it is never clipped, then the efficiency of speech transmission is reduced. If the operator sets the input speech level too high, then the speech is clipped frequently and will sound distorted. When the distortion caused by the clipper is just barely perceptible, then we say that the peak level of the speech is the clip level. This criterion, however, involves human perception rather than a strictly objective mathematical criterion and is therefore not useful for ongoing testing.

We therefore have developed an objective mathematical approach as an approximation to the operational approach. The mathematical approach we have developed is based on a histogram of the absolute value of the speech samples. We refer to this technique as the *histogram method.* The expected amplitude range is divided into 128 equal intervals or bins. Each incoming sample will cause one of the bin counters to be incremented depending on the sample absolute value. The histogram for an entire speech file (which may be several minutes) is then normalized by the total sample count and then integrated (starting with the smallest value) to produce a distribution function which is monotonically increasing from 0.0 to 1.0. The 99.99 percentile of the distribution function is defined to be the "peak" of the speech waveform; that is, 0.01% of the samples will be above the "peak". These samples correspond to those clipped portions of the waveform which occur infrequently enough to be perceptibly insignificant. Again, we have an arbitrary constant embedded in the definition of the measurement. This means that in order to compare peak/RMS ratios between various laboratories, it is important that these constants be explicitly stated. Through this procedure we have gained immunity to artifacts of the speech waveform which would obscure small but important differences in the enhancement processors being compared.

There is, however, a simpler and faster technique for estimating the peak/RMS ratio which has proved valuable in providing a performance indicator for the real-time digital system. This technique is called the *interval method.* Samples of incoming speech are segmented into groups (or "intervals") of 16384 (2^{14}) samples which exceed the threshold of -40 dB below the normal peak (thus, the interval is at least 1.6 seconds long). The second largest peak is found for each interval, which is used to compute the peak/RMS ratio in dB for that interval. Then, the peak/RMS ratio in dB is averaged for 16 intervals (at least 26.2 seconds). Then, 8 of these averages are further

averaged for at least 3 minutes (see Section 4.4.3).

The interval measure of peak/RMS ratio is usually within 0.1 dB of the histogram measure for speech processed with the VISTA. For unprocessed speech or speech from the commercial audio processors, the interval technique can differ by 1/2 to 1 dB from the histogram technique, due to the possible presence of spurious peaks. The VISTA generates speech with very well-defined peak values so that almost any method of determining a "peak" will yield almost identical results. Thus, the simpler interval method is perfectly adequate for monitoring the system performance and was used in the real-time VISTA system. All comparative measurements were made with the histogram method in non-real time.

Another measure has been used by others, similar to the peak/RMS ratio, and is sometimes confused with it. In this other method, a sine wave is used to modulate a transmitter. If its peak is set at the maximum without causing phase reversals in the transmit waveform, then we say that we have 100% modulation. A peak-normalized speech signal will have a lower average modulation level which can be expressed as a ratio (M) to the 100% modulation level of the sine wave. The modulation level is proportional to the RMS level of the speech. Thus, we have $M = \frac{RMS\ sine}{RMS\ speech} = \frac{pk/RMS\ speech}{pk/RMS\ sine}$ since the peaks are the same, or $M = 20\log(pk/RMSspeech) - 3dB$. Thus, if processed speech has a peak/RMS of 7 dB, the modulation ratio M is 4 dB.

This change in modulation ratio will have an effect on transmitted power. In particular, lowering the peak/RMS implies increasing the transmit power and therefore the fuel bill. Details on transmitter fuel and broadcast power are discussed in Appendix E. There it is shown that a doubling of speech power (+ 3 dB) must be supported by a 9% increase in the average transmitter power.

5.2 Facilities

5.2.1 Algorithm Development Computer Facility

The enhancement algorithm was first developed in the C programming language and ran in non-real-time on a VAX-11/780 computer or a SUN3 computer. The VAX is time-shared among

many researchers in the speech technology group; in addition, for the kind of processing involved, it was about a factor of two slower than the SUN (assuming single user on each machine). On the SUN it took about 10 minutes to process a two-second speech file; 60 such sentences (one of our data bases) would therefore take about 10 hours – an overnight run). Before allocating a SUN full time to the project, the same experiment might have taken three or four days.

The development of the real-time software was largely done on a different SUN computer which was connected via a VME-bus to a special purpose digital processor designed and built by Lincoln Laboratory; this will be described in the following section. This development system includes debugging software which permits inspection of registers in a hardware simulation, the placement of conditional traps, and a myriad of other features which ease software development.

Some of the real-time software – those modules dealing with subtle and critical parts of the enhancement algorithm – were first tested in isolation in the following manner. The non-real-time C program was used to process about one second of speech. The non-real-time input and output data for each analysis speech frame was then written to a disk file. The "input" data was converted to fixed-point hex format, and then, by use of the debugger software, it was processed by the real-time module under test. The module output was written to a disk file which was subsequently converted to floating point by another C program. Finally, yet another C program compared the output of the real-time program with the output of the non-real-time program. Because of differences in the processors (16 bit vs 32 bit or 16 bit with exponent vs 32 bit), the results were not identical. However, often we found ways to improve the accuracy of the real-time software or found programming bugs which less extensive testing had failed to reveal. Incorporation of each new module into the rather larger real-time software was facilitated by this extensive module testing.

5.2.2 Commercial Equipment Measurement Facility

Three commercial audio processors were purchased and tested to establish a baseline standard against which the enhancement algorithms could be compared. A block diagram of the facility is shown in Figure 5-2 and a photograph of it is shown in Figure 5-3. This facility allowed rapid

switching between processors to aid listening tests, allowed independent input level adjustment to each processor, and allowed each processor output to be peak normalized. In addition, an HF modulator (HP-8656A) and an HF receiver (Kenwood R2000) were options to allow the listener to hear how the output of each processor sounded through a typical bandlimited shortwave radio. Noise could also be added at the output to further simulate the listening condition when tuned to a distant transmitter. A "home-made" clipper was also introduced. The clipper was a symmetric device which clamped the instantaneous positive and negative signals which exceeded the identical positive or negative threshold. The threshold could be set at 1V, 100mV, or 20mV. Since the clipped signal had very wide bandwidth, it was filtered to 5 kHz to match the other processors. The characteristics of the filter (ORBAN 8-channel Equalizer Model 672A) were adjusted to minimize the peak/RMS ratio of the output.

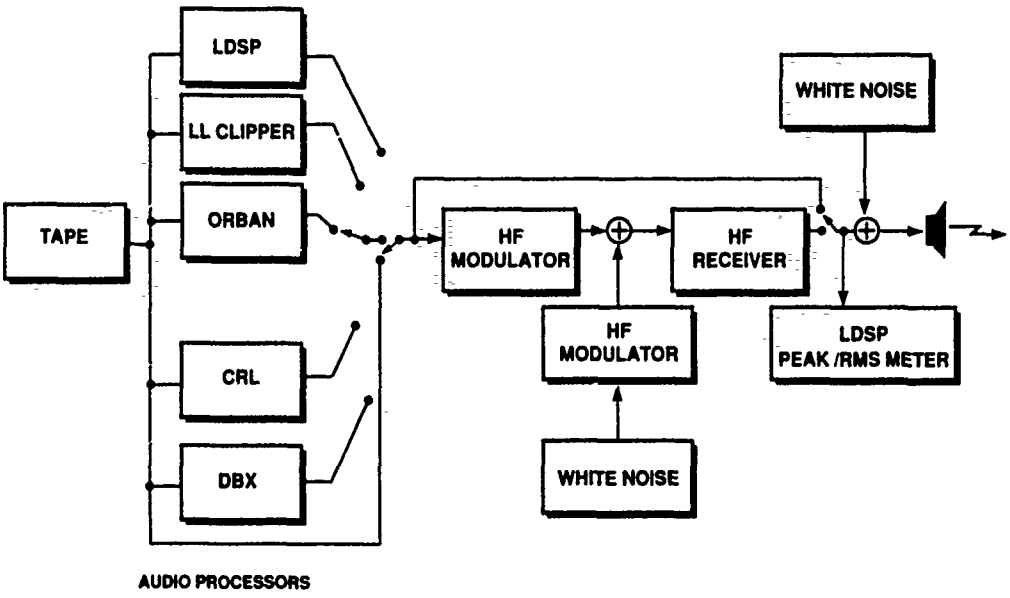


Figure 5-2. Audio evaluation facility block diagram.

In performing the planned measurements, it was necessary to account for the variability of different talkers and different speech sounds to ensure the generality of the results. Therefore, an automated facility was developed to store, reproduce, and analyze a large number of speech files so that any processor could be rapidly tested to generate reproducible results. In the early stages of system development, the data handling system used a PDP-11 and a high speed, special

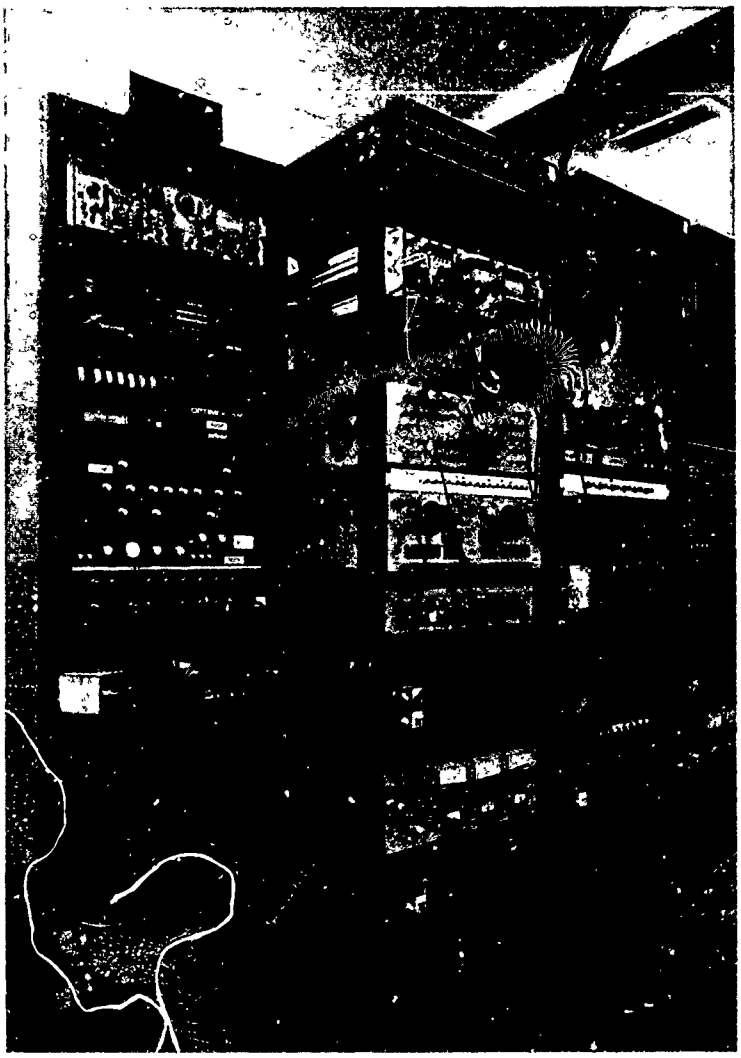


Figure 5-3. Photograph of audio evaluation facility.

purpose digital signal processor designed and built at Lincoln Laboratory (LDSP) [25]. In one testing scenario, speech files, corresponding to about two seconds of speech, were stored on a disk. The host transfers each such file to an outboard memory on the LDSP in non-real-time. After obtaining the complete file, the LDSP outputs the speech in real-time with its D/A converter (16-bit samples, 10,000 samples/s) to the analog commercial audio processors under test. Simultaneously the processed speech is passed through an A/D and stored in its outboard memory. As each sample is stored, a peak-searching algorithm is run and the sample is squared (16-bit result) and added to a double precision (32-bit) accumulator. When the writing and reading of the real-time speech file are

completed, the peak/RMS ratio (in dB) referred to earlier as the histogram method, is computed and, it and the new processed speech file are transferred back to the host in non-real-time. The procedure is then repeated for each speech file used in the study. The processed speech file can later be analyzed to debug the experimental setup and to better understand or evaluate processor algorithms.

In the later stages of development and testing, a second facility was established to compare the real-time VISTA processor with one of the commercial processors. This second facility used a MASSCOMP computer to sample the output speech so that the peak/RMS ratio for each processor could be measured in non-real-time using the histogram method. In particular, the digitally processed speech was first returned to analog form and then resampled by the MASSCOMP. As shown in Figure 5-2, the output signals of each system could also be peak normalized. This procedure insured that any problems with the output D/A and filters on the real-time system would be identified by the peak/RMS measurement. Since the speech was already bandlimited to 5 kHz, the input filters on the MASSCOMP A/D were removed so that any distortion they might process would not contaminate the result.

5.2.3 VOA Radio Channel Test Setup

The output of an HF receiver can be significantly degraded relative to the broadcast studio signal. This degradation can be due to the studeo and transmission equipment, the characteristics of the ionospheric transmission channel, the special properties of the noise and interference of the channel, and the characteristics of the radio receiver itself. The enhancement processing does not attempt to mitigate the effects of the transmission channel. Thus, it was not necessary to have frequent access to a test facility which simulated the channel. It was important, however, to understand whether the channel had a deleterious effect on the enhanced signal, greater than on the unprocessed signal. We, therefore, conducted a transmission test into a dummy transmitter load and a live on-the-air transmission test. The results of these tests will be described in Section 5.3.7.3.

The characteristic of the radio receiver and the presence of noise can have some influence on the enhancement algorithm. The receivers have narrow bandwidths; i.e., about 6 dB bandwidth of 2.7 kHz in narrowband mode and about 6 dB bandwidth of 6 kHz in wideband mode. This restriction can be compensated somewhat by preemphasis in the processor as described in Section 3.5. In addition, noise at the receiver can mask certain artifacts of the enhancement processing. We, therefore, included noise in some listening tests to determine the extent of the masking and to demonstrate the utility of peak/RMS ratio enhancement in making the speech more intelligible in noise. In order to test these effects, we exploited the HF receiver which is fed by the HF modulator. We also added an audio noise source (white uniform) prior to the receiver, as illustrated in Figure 5-2.

5.2.4 Data Bases

Several different data bases were used in the evaluation process. Our first data base was the single male-speaker sentence "autumn leaves turn yellow" which was extracted from a standard test tape (the "S1" tape) used in evaluating vocoders. We refer to the second data base as the Lincoln Data Base or the phoneme-specific-sentence (PSS) data base. It consisted of a total of 60 utterances, representing 10 speakers each speaking 6 different sentences. The sentences were designed by Huggins and Nickerson [26] to emphasize specific phonetic features. The six sentences with descriptions of the phonetic characteristics are as follows (each sentence is denoted by a three-letter identifier):

1. bln: "The little blankets lay around on the floor."
   - general non-diagnostic, more rapid, unstressed and reduced syllables
2. fth: "His vicious father has seizures."
   - emphasizes voiced and unvoiced fricatives
3. nan: "Nanny may know my meaning."
   - emphasizes nasals and nasalized vowels
4. roy: "Why were you away a year, Roy"?

- has only vowels and glides and no abrupt changes in level

5. swm: "The trouble with swimming is that you can drown."

   - general non-diagnostic, more rapid unstressed and reduced syllables

6. tea: "Which tea party did Baker go to"?

   - has all the stops and affricatives except /j/

The data base was obtained by having 5 males and 5 females read the sentences in a quiet room. All the speakers were American-born native speakers of English and did not have any marked regional accents. After the sentences were digitized, the files were all normalized to have the same peak value and then truncated to have about 50 ms of silence before and after the utterance. The third data base consisted of 5 minutes of a VOA science notebook, #2540, Super NOVA and often referred to as NOVA. This was spoken by a trained male announcer and its general level was very tightly controlled (manually, we assume). We often used the first 60 seconds of the NOVA tape to make peak/RMS measurements. These later two data bases were used extensively in testing the non-real- and real-time systems during the final stages of the project.

In addition to these data bases, we had available about another hour of VOA broadcast material which we used near the completion of the project. These broadcast materials included several speakers and were used for quality assessment and peak/RMS measurement of the real-time system. In particular, 10 extractions were made (roughly 10 minutes) consisting of 5 male and 5 female speakers. Two male speakers were taken from what appears to be telephone interviews. One male passage is the same VOA passage (NOVA) used earlier.

Finally, data bases supplied by Dynastat, Inc. were used in DRT and DAM evaluations in testing one configuration of the real-time system. A data base of true/false sentences was supplied by Pisoni for SVT evaluations on one configuration of the non-real-time system.

5.3 Evaluation Tests and Results During VISTA Development

5.3.1 Preliminary Evaluation of Commercial Devices

Three commercial audio processors were chosen for experimentation among the ten or so readily available on the market. In addition, the "home-made" clipper described above and a real-time simulation of a single-channel dynamic range compression (DRC) [17,18] algorithm, on the LDSP signal processing computer [25], were tested.

The parameters of the commercial processors and the input signal level were adjusted to minimize the peak/RMS ratio of the processed signal of the particular device under test. As a result, the units were driven at higher levels than recommended by the manufacturers at some loss in the speech quality. In standard use the processors would exhibit higher quality, but would produce a signal which would be peakier, and thereby result in peak/RMS ratios that would be higher than the results reported here. In these preliminary evaluations, the reason for overdriving the commercial processors was to determine the limit of their achievable peak/RMS ratio so that a comparative reference could be established for subsequent digital signal processor enhancement algorithms. To establish a reference, one of the processors was also set up to yield very low distortion by driving at recommended input levels.

Although the clipper was designed without regard for speech quality, it was tested to establish a meaningful peak/RMS goal for any digital signal processing algorithm. The simplest form of a dynamic range control algorithm (see Appendix C for a review of DRC) was also tested in order to gain some understanding of the critical issues in the design of such algorithms. This simple algorithm performs DRC on the full-band speech signal (i.e., operates "single-channel") in contrast to the multi-band (2 to 5 channels) technique used in many commercial processors. In addition, this simple algorithm did not use preemphasis, did not use any distortion-cancelling, nor did it use any phase dispersion. The DRC algorithm [17,18] was essentially a limiting amplifier (if the envelope is within 20 dB of the peak level, the gain is adjusted to produce an envelope at that peak level) with a 2 ms release time and an instantaneous attack time. The 2 ms release time is very fast (the recommended release time to ensure high quality is 40 to 60 ms) and was so chosen to

maximize the performance of the algorithm while giving acceptable speech quality. The quality of this DRC algorithm is less than that of any of the commerical processors (which are close to each other in quality) but significantly better than the quality of the bandpass-filtered clipper.

In the first experiment, the peak/RMS ratio was calculated for all of the unprocessed utterances in the PSS data base. These were found to vary from 12.6 to 20.0 dB (mean = 15.6 dB; st. dev. = 2.1 dB). Thus, speech of sentence length (about 2 sec) is not long enough to produce reliable and statistically significant estimates of peak/RMS. Averaging over 60 sentences (about 2 minutes of speech) reduces the standard deviation of the measured peak/RMS by $\sqrt{60}$ or to less than 0.3 dB for the unprocessed speech and less than 0.15 dB for the processed speech. The principal result of the study is shown in Table 5-1 which gives the average peak/RMS ratio for each processor and for the unprocessed speech using the PSS data base. The average reduction (relative to unprocessed speech) obtained by the three commercial processors is 6.1 dB (taking the peak/RMS ratio from 15.6 dB to 9.5 dB). The 0.8 dB difference range covered by the three commercial devices is relatively small. The bandlimited clipper gave a peak/RMS reduction of 10.3 dB (taking the peak/RMS ratio from 15.6 to 5.3 dB). The clipped speech is thus perceived to be very loud, however, as mentioned earlier, very distorted, although still intelligible. The simple DRC algorithm had a peak/RMS reduction comparable to the best commercial processor (it measured 0.2 dB better, an insignificant difference); however, the quality produced by the simple DRC algorithm was generally much less than that with the commercial processors.

These results were obtained by overdriving each of the processors beyond the levels recommended by the manufacturer. This was done to minimize the peak/RMS ratio of the output at the expense of speech quality. Since the processors were severely overdriven, the resulting quality was far from the original. To understand this trade-off, one of the processors (C) was driven at a level which gave almost no distortion. In this condition, it gave an average peak/RMS ratio of 12.4 dB. This value is a 2.8 dB reduction over the unprocessed speech but is 3.5 dB less reduction than the value achieved by overdriving the same processor.

Table 5-2 shows the results of comparing the peak/RMS ratio of the various sentences for processor C, the clipper, and the unprocessed waveform. This data is averaged over the 10 talkers.

TABLE 5-1.

Average Peak/RMS Ratio for Each Processor (in dB)

Processor	*Peak/RMS*	*Reduction in Peak/RMS*
No Processor	15.6	—
Commercial Processor A	10.1	5.5
Commercial Processor B	9.0	6.6
Commercial Processor C	9.3	6.3
Single Channel DRC Algorithm	8.8	6.8
Bandlimited Clipper	5.3	10.3

Prior to processing, the difference between the maximum and minimum peak/RMS ratios is 2.1 dB, while after processing by processor C the difference has been reduced to 1.2 dB. After processing, the two peakiest sentences are "fth" and "tea" which are characterized by fricatives and stops, respectively, and thus would be expected to be peaky. The two least-peaky sentences after processing are "nan" and "roy" which are both entirely voiced and have no stops or frication, and thus would be expected to have smooth envelopes and thus low peak/RMS ratios. That these expectations are not met prior to processing is perhaps due to speakers modulating the intensity of their voices. This intensity modulation can dominate the above more subtle effects.

We also notice that the peak/RMS ratio for the clipper for each individual sentence is 3.95 $\pm$ 0.15 dB below that for processor C, suggesting that the variability in the measured peak/RMS ratio is inherent in the sentences. The other two commercial processors show a similar trend.

Table 5-3 shows the peak/RMS ratio for each talker for unprocessed speech and for speech processed by the same two devices shown in Table 5-2. There is a 3.5 dB difference in peak/RMS between the peakiest and least-peaky talker before processing, but this is reduced to 1.7 dB after processing by processor C or to 0.7 dB after processing by the clipper. Again, it is noted that the difference between processor C and the clipper is within 0.5 dB of a constant for each individual talker, suggesting that the variability in the measured peak/RMS ratio is inherent in the talker.

The next result comes from computing the average peak/RMS ratio for male and female talkers.

TABLE 5-2.

Average Peak/RMS Ratio for Each Sentence for Selected Processors (in dB) – PSS

Sentence	*No Processor*	*Processor C*	*Clipper*
bln	15.3	9.0	5.2
fth	15.8	9.9	5.9
nan	14.7	9.0	4.9
roy	15.9	8.7	4.7
swm	16.8	9.3	5.4
tea	15.1	9.7	5.6
Note: Averaged over talkers			

For unprocessed speech the difference in the averaged peak/RMS ratio is only 0.25 dB, and only 0.07 dB after processing by processor C. Thus gender differences appear to be insignificant with respect to peak/RMS ratio.

The principal conclusion drawn from this study is that it is possible to reduce the peak/RMS

TABLE 5-3.

Average Peak/RMS Ratio for Each Talker for Selected Processors (in dB) – PSS

Talker No.	*No Processor*	*Processor C*	*Clipper*
1	14.3	9.0	5.0
2	14.9	8.8	5.1
3	15.3	9.5	5.2
4	14.3	9.2	5.5
5	15.3	9.0	5.1
6	17.0	10.5	5.3
7	16.2	9.4	5.6
8	15.1	9.1	5.1
9	16.7	9.2	5.7
10	16.2	9.3	5.2
max-min	3.5	1.7	0.7
Note: Averaged over sentences			

ratio of 5 kHz bandlimited speech by about 6 dB using commerical processors, and by about 9.5 dB using a bandlimited clipper but in both cases at the expense of quality. When one of the commercial processors was driven at a level to give very high quality, it reduced the peak/RMS ratio by about 3 dB, i.e., about 3 dB less reduction than when the processor was overdriven to produce maximum reduction in peak/RMS ratio. With more sophisticated processing we should hope to improve the quality without changing the 6 dB peak/RMS reduction. These improvements would represent about 3 dB of improvement over the current state of the art. It is clear from the result of the simple DRC algorithm that brute-force techniques are not adequate.

It is further concluded that there can be significant variation in peak/RMS ratio measured over talkers or speech material. When research on algorithm development attempts to achieve the last dB of improvement, and measurements with an accuracy of 0.2 dB or so are necessary, then it would be necessary to take the variation due to talkers and to speech material into account in evaluating the algorithms by averaging over several minutes of speech material.

5.3.2 Comparison of Non-Real-Time VISTA System with Commercial Devices

5.3.2.1 Peak/RMS versus Quality Results

Throughout the development of the VISTA system we have evaluated its peak/RMS performance relative to one commercial processor representative of the state-of-the-art. The processor developed by ORBAN, Inc., was chosen since its performance was representative of the state-of-the-art (we did not observe a great variation in the performance of the three processors we purchased for our laboratory tests), and also because the VOA was concurrently using the ORBAN processor for some of its own tests. The commercial processor was setup in these tests to have a quality "equivalent" to the VISTA system for different processing modes, as described in Section 5.1.3. As with VISTA (see Section 3.7.3), three modes of operation for the commercial system were setup ("mild," "normal," "severe") with 3 independent preemphasis settings ("none", "medium", and "heavy"). These tests were initially made in the "normal" mode without preemphasis on the Lincoln data base (phoneme-specific-sentences), but were also conducted on the Pisoni and VOA data

bases throughout development. The history of these tests using the non-real-time VISTA is shown in Table 5-4. Peak/RMS measurements were made with the histogram technique.

TABLE 5-4.

History of Peak/RMS Measurements (Normal/No Preemphasis)

Quarter	*Dates*	*Peak/RMS* lower than commercial	*Quality* relative to commercial	*Data Base*
4	Dec86 - Feb87	1.6 dB	"equivalent"	1 male, 1 female, 6 sentences each
5	Mar87 - May87	2.0 dB	slightly lower	PSS
6	Jun87 - Aug87	2.8 dB	slightly lower	PSS
		2.9 dB	slightly lower	Pisoni
8	Dec87 - Feb88	3.0 dB	"equivalent"	PSS/NOVA
*PSS = Lincoln Data Base – 60 phoneme-specific sentences				

The peak/RMS – quality tradeoffs for the last results in Table 5-4 (NOVA only) for all three modes (no preemphasis) are shown in Figure 5-4. As expected, as we drive the processors harder (going from mild to normal to severe mode) peak/RMS decreases at the expense of quality. The encircled region represents the normal operating mode for the two devices. This level of quality is acceptable in the sense that in a real operating environment (with background noise and channel degradation) the quality is roughly that of the original. For this NOVA data base, in normal mode, VISTA does about 3 dB better than the representative commercial system for quality informally judged to be "equivalent"; this maps to a factor of two improvement in the broadcast area.

5.3.2.2 SVT Results

The Sentence Verification Test (SVT) was made under three conditions: unprocessed, the commercial processor, and the VISTA system non-real-time simulation as it existed in July 1987. Both were in normal operating mode without preemphasis. This version of the VISTA included Key-Fowle-Haggarty (KFH) dispersion, dynamic range compression, and clipping but did not contain

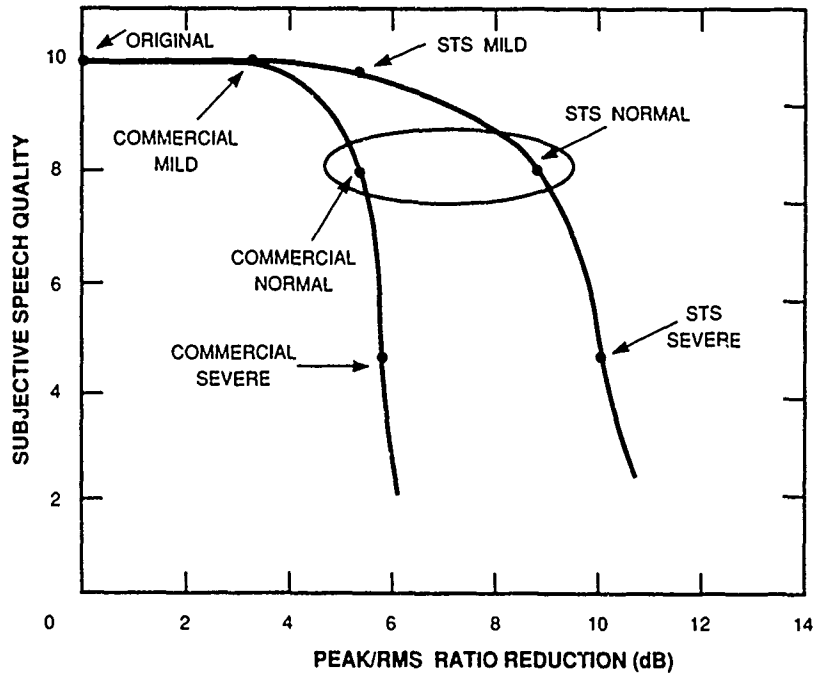


Figure 5-4. Peak/RMS-quality tradeoffs.

many more subtle refinements which have improved both the quality and the peak/RMS ratio enhancement. Noise was added to the speech signal just prior to the earphones. A noise level of 0 dB means that the RMS level of the noise is approximately the same as the RMS level of the unprocessed speech.

Figure 5-5 shows the transcription score from SVT for the three conditions as a function of signal-to-noise ratio. The curves shown are cumulative Gaussian distribution curves whose parameters are picked to best match the data points. Where the curves cross a transcription score of 50%, the commercial system has a 7 dB advantage over the original speech, while the VISTA system has an additional 3 dB advantage over the commercial system.

5.3.2.3 Broadcast Station Tests

Two tests were made at the VOA transmitter complex at Greenville, N.C. The first, on February 16 and 17, 1988, used a high power transmitter operating into a dummy load. The second,

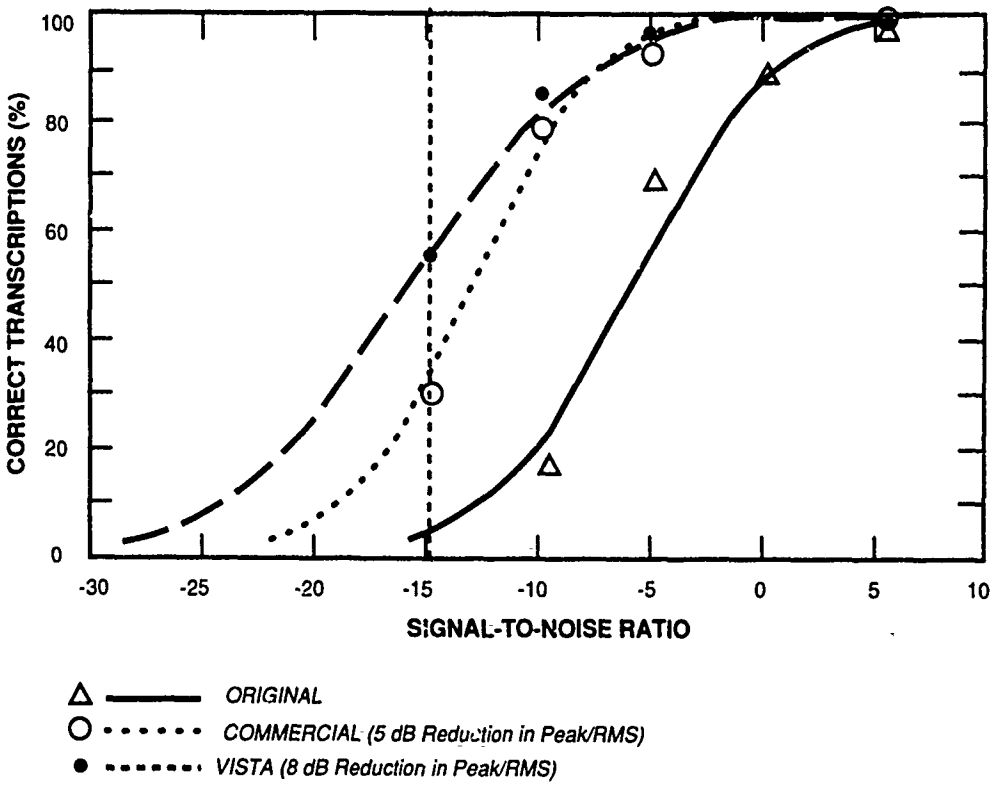


Figure 5-5. Transcription performance.

made on June 9, 1988, was a live on-the-air test and involved several listening sites.

The dummy load test was made to determine if the transmitter site equipment might introduce distortion, which if not compensated for, could reduce the effectiveness of the VISTA enhancement processing. Audio recordings of the output of an HF receiver were made to establish an upper bound on the distortion caused by the transmitter equipment. The experimental transmission was into a dummy load which is a resistor with a high power rating, rather than over the air. Analog recordings of the unprocessed, VISTA processed, and VOA-UREI (a VOA enhancement unit) processed speech were made. At Lincoln Laboratory the first 60 seconds of the processed recordings (which had been though the HF transmitter and receiver) were digitized, along with the VISTA processed speech which had earlier been recorded at Lincoln. The resulting peak/RMS ratios were:

VISTA processed (original digital version)	7.1 dB
VISTA processed (from tape)	8.5 dB
VISTA processed (from tape to transmitter-receiver)	8.3 dB
VOA processed (unprocessed from tape to the UREI processor to transmitter-receiver)	9.6 dB

The live on-air test, conducted on June 9, 1988, compared a version of the VISTA enhancement algorithm with the commercial processor and with the VOA-UREI. The commercial processor was a prototype of a new "HF" design by ORBAN, Inc., at that time not commercially available, and was run in two modes – "normal" and "severe" processing. This was the first opportunity during the course of the Lincoln program with VOA to test the Lincoln processing in an actual broadcast situation. The test was very informative and helpful both in getting an initial evaluation of the algorithm and in providing a basis for design of future tests.

The material played over the air was the 5-minute NOVA program used in previous tests. For the VISTA test, a recorded version of the program was processed though the non-real-time simulation of VISTA in "normal" mode, and the resulting output recording was shipped to VOA and re-recorded onto a master tape which also contained the unprocessed program. Thus, the VISTA-processed material went through two extra tape-recording steps prior to transmission. (Previous results had indicated about a 0.7 dB increase in peak/RMS for each recording. It is also likely that the final post-transmission recording would have greater effect on the phase-control-dependent VISTA system than on the other processors.) In addition to the tape recording losses, it was known (as discussed in Section 3.6.3) that a (correctable) distortion due to imperfect compensation for the D/A system in the non-real-time simulation was causing another 1.0 dB increase in peak/RMS of the VISTA material.

Informal listening to the broadcast through the Kenwood shortwave receiver (at Lincoln) on June 9 produced the following subjective observations. The VISTA and commercial (Normal) were at about the same quality and loudness levels. The commercial (Severe) was slightly louder, but noticeably poorer in quality. The VOA system was not quite as loud, and was comparable in

quality. For the material recorded at the modulation monitor at Greenville on June 9, the following peak/RMS ratios were measured:

VOA Processor	10.7 dB
Commercial (normal)	10.1 dB
VISTA (normal)	9.96 dB
Commercial (severe)	8.4 dB

These results, together with listening done on June 9, generally confirmed the informal observations noted above. It is believed that the VISTA system has a potential 2.5 to 3.0 dB advantage in peak/RMS over representative commercial systems at comparable quality, but that advantage was not realized in the on-air test due to losses caused by the multiple tape recordings (estimated at up to $3 \times 0.7 = 2.1$ dB) and to uncompensated dispersion (estimated at 1.0 dB) in the D/A filters. The tape recordings will not be present in an actual broadcast using a real-time VISTA system. However, more work was needed to reduce the residual D/A filter distortion. This compensation was successfully performed at a later date (see Section 3.6.3 and Appendix D).

Other lessons learned from the June 9 test include: (1) the VISTA algorithm should be tested in "severe" processing mode as well as in "normal" mode, (2) the preemphasis used with the non-real time VISTA should be increased to match the commercial preemphasis, which will yield a crisper sound after HF transmission (incorporated in final system); and (3) recordings from the HF receiver should be made with the receiver in wideband mode, which appears to increase intelligibility at a penalty in perceived background noise level.

5.4 Tests and Results with the Real-Time VISTA System

In Section 3.7.3, some results were reported which used a 60-sentence (phoneme-specific) data base and 60 seconds from a VOA broadcast. The measurements there were done in real-time on the last (7th) multi-processor ADSP board. In the testing in this section, the real-time VISTA system is compared with the commercial processor used above. As in all comparative studies, the peak/RMS measure was made in non-real time using the histogram-based method described in

Section 5.1.2. These experiments were performed with medium preemphasis and quality ratings were determined through the short-wave radio.

In our laboratory tests, two data bases were used. First the 60-sentence phoneme-specific (PSS) data base. Second, 10 passages (roughly 10 minutes) from VOA broadcast material was used, consisting of 5 male and 5 female speakers – as described earlier. Two male speakers were taken from what appears to be telephone interviews. One male passage is the same VOA passage (NOVA) used earlier.

Figure 5-6 shows the peak/RMS – quality tradeoffs for the PSS data base for all three operating modes. The peak/RMS was computed separately for each sentence, and then an average was computed. Table 5-5 gives the average peak/RMS for the sentences for mild, normal, and severe processing modes. Figure 5-7 shows the peak/RMS – quality tradeoffs for the VOA data base. As with the PSS data base, we obtained about 3 dB peak/RMS with VISTA in normal mode. Although not shown here, the peak/RMS levels for the telephone interview speech were less striking, e.g., 2.1 dB improvement over the commercial device for the normal case. This may bias an average measure for the typical broadcast speakers. Figure 5-8 gives the same results but for the NOVA VOA passage only. Note in Figure 5-8 that the real-time results for the NOVA passage are consistent (excluding an absolute peak/RMS level) with those for the non-real-time results of Figure 5-4. In both cases, about a 3 dB peak/RMS advantage was obtained with VISTA in normal mode. Finally, Table 5-6 shows the peak/RMS for the male and female speakers from the 10-speaker VOA data base. Note that the VISTA favours females, while the commercial system (slightly) favours males, although these differences may not prove to be significant.

TABLE 5-5.

Sentence Data Base – Peak/RMS Results

PROCESSOR	*MILD*	*NORMAL*	*SEVERE*
Original	14.1	14.1	14.1
VISTA	7.2	6.2	5.17
Commercial	12.0	10.34	9.99

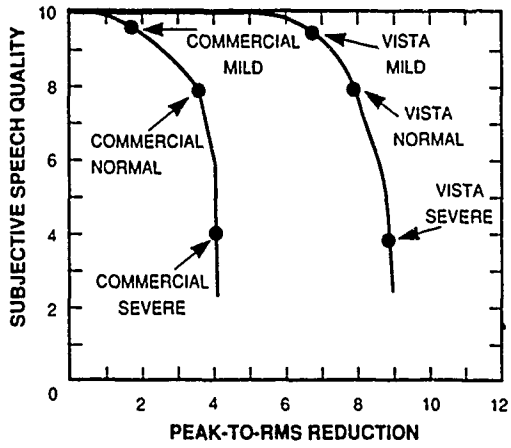


Figure 5-6. Peak/RMS vs. quality tradeoffs – PSS data base.

TABLE 5-6.

Peak/RMS for Male (M) and Female (F) Speakers from the 10 Sentence VOA Data Base

PROCESSOR	*MILD*		*NORMAL*		*SEVERE*	
	M	F	M	F	M	F
VISTA	8.8	7.97	8.05	7.17	7.18	6.42
Commercial	11.48	11.79	10.73	10.88	10.26	10.62

Finally, the Dynastat test tape was run through the processors just prior to shipment of VISTA to VOA. A DRT and DAM evaluation was performed in the normal/medium-preemphasis mode. These tests were run only for a single setting of the VISTA and ORBAN parameters, because our emphasis for formal testing was to deliver the system to VOA for test in their broadcast environment. Test sentences from Dynastat were run from analog tape into the VISTA and the commercial processor and the output of the short-wave radio was recorded directly onto analog tape. Consequently, the histogram method (non-real-time) peak/RMS measurements were not made. Nevertheless, the similarity of DAM sentences to PSS leads us to project a greater than 4 dB advantage in peak/RMS with the VISTA system (see Figure 5-6 and Table 5-5) for the PSS data. The processed DRTs were 93, 91, and 95 for the unprocessed, VISTA and commercial, respectively; and for the DAMs: 59, 50, and 57 for the unprocessed, VISTA and commercial, respectively. The results were somewhat surprising and indicated that the quality of the processors' output is

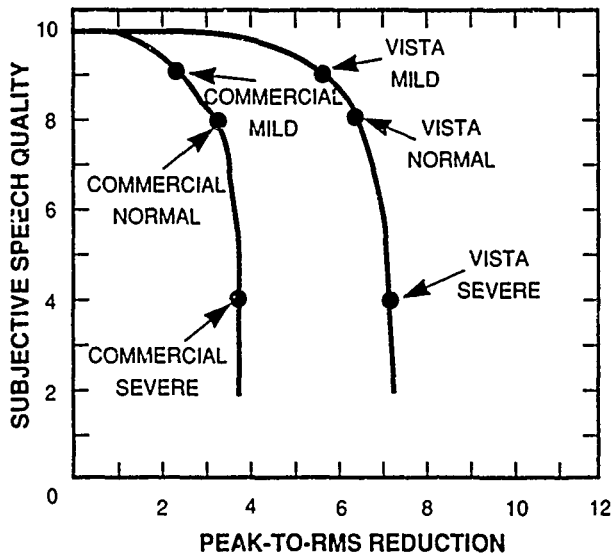


Figure 5-7. Peak/RMS vs. quality tradeoffs. VOA data base – 10 speakers.

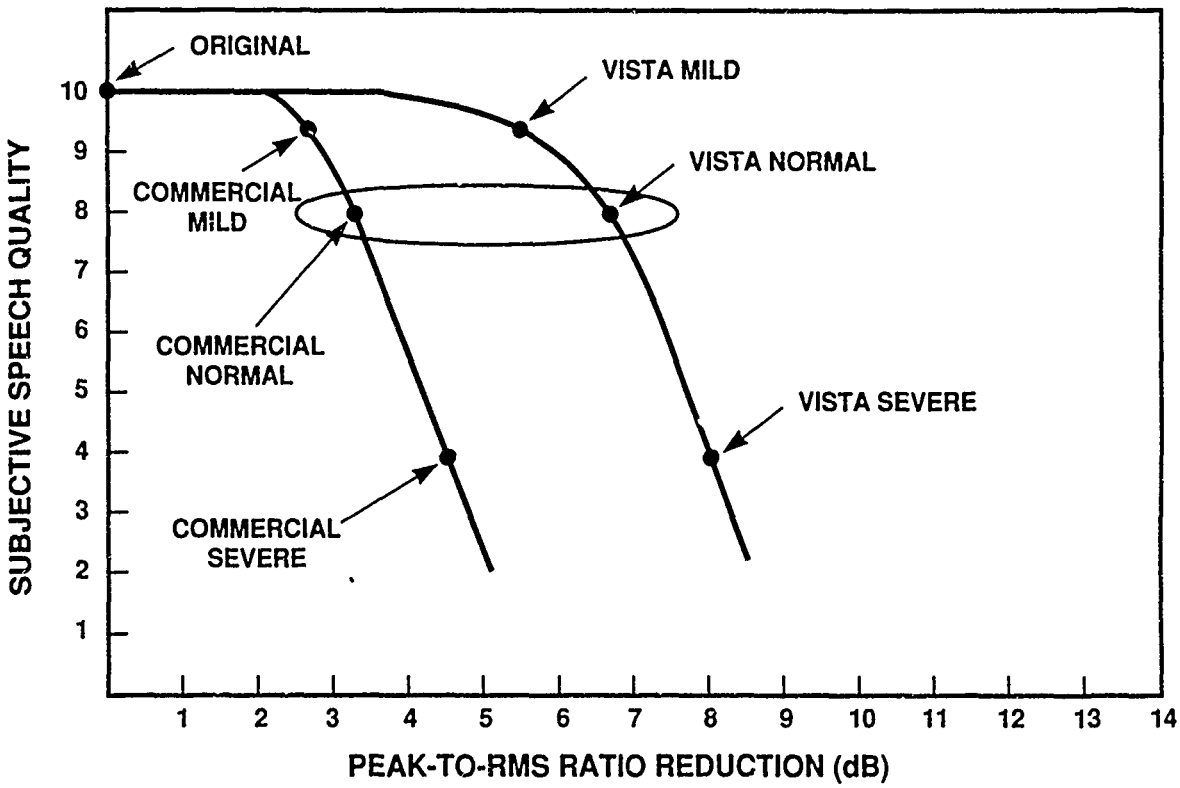


Figure 5-8. Peak/RMS vs. quality tradeoffs. VOA data base – NOVA.

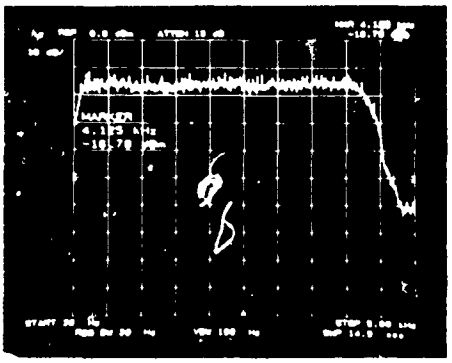
sensitive to system parameter settings, as illustrated by the steepness of the quality-peak/RMS functions. Due to the steepness of these curves, it is important that tests in the VOA environment be conducted for a range of parameter settings of the VISTA processor. In addition, the results and other observations raised a number of issues for future consideration, as discussed in the next section.

5.5 Issues for Future Testing

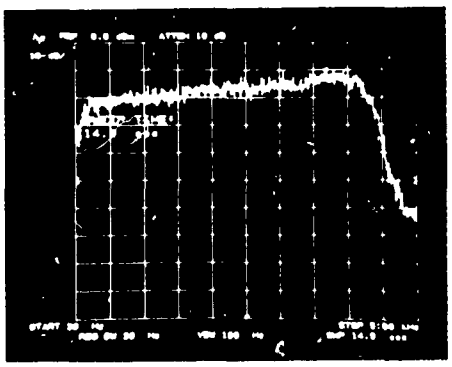
The results of the previous section have helped in planning future tests and development. The difference in DRT and DAM test results may be partially explained by the loss of about 500-800 Hz in the high-frequency end by VISTA due to the real-time constraints. The overall impulse-train response of VISTA in normal mode for different preemphasis settings is shown in Figure 5-9. The spectral response shows a high-frequency rolloff starting about 4.2 kHz. This bandwidth is in contrast to about 5 kHz for the commercial device. (The radio receiver has a bandwidth of roughly 6 kHz in wideband mode.) Methods to improve bandwidth should therefore be explored in the future. Any additional testing should require equal bandwidths for both processors.

Figure 5-9 also illustrates the nature of the preemphasis curves and the filter at DC that was described in Section 3.5.2, both of which can severely influence quality, as well as intelligibility. It was observed that the "harshness" of preemphasis is strongly dependent on the speaker's characteristics as well as the degree of preemphasis. (This was true of the commercial device as well as of VISTA.) The time constraints of the project did not allow a careful study of preemphasis. It is hypothesized that a more sophisticated *adaptive preemphasis,* which accounts for the speaker's spectral tilt, will help remove the speaker dependence and occasional excessive harshness of quality. Likewise, the high-pass filter at DC also appears to strongly influence quality and intelligibility and is speaker and text dependent. Both functions, designed for receiver compensation, need further refinement and understanding.

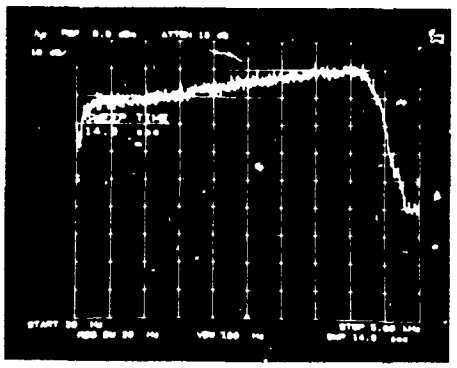
Another important issue involves how, more generally, the many parameter settings of VISTA (and likewise, the commercial devices) generate the steep quality/peak-to-RMS tradeoff curves. This steepness implies a need to carefully control settings. It also appears to correspond to changes



(a) Normal/None



(b) Normal/Medium



(c) Normal/Heavy

Figure 5-3. VISTA spectral response to internal pulse train.

in qua h different speakers and text. Future testing should therefore require formal quality ratings, w. .1 . ·rs, for many settings, to obtain a finer sampling of the quality-peak/RMS curves for both processors and thus a better measure of "equivalent" quality.

In terms of intelligibility testing, the SVT was not performed on the real-time system. The Pisoni data base was however, processed and delivered and is awaiting evaluation.

Quality and peak/RMS may not be the ultimate performance of a system. Although the quality of two systems may differ, each has its own "style". The VISTA is often described as "authoritative" - a quality perhaps desirable in a real operating environment. Thus further testing should account for any advantages due to "style" in a real VOA operating enviroment.

6. CONCLUSIONS AND FUTURE WORK

In this report, a new frequency-domain approach to speech preprocessing for reducing peak/RMS was presented which is based on a sinusoidal representation of speech. The phase dispersion and amplitude compression via sine-wave modification is controlled by a radar signal design technique. Significant reduction of peak/RMS and increase in waveform loudness (under a peak-power constraint) were obtained. The application to AM radio broadcasting for the Voice of America was described. For this application, various processing options were defined and some peak-to-RMS and quality measurements were illustrated. A real-time prototype system was based on seven ADSP2100 boards was described. The prototype system was recently shipped to VOA for on-site testing.

Some items for future work were outlined in Section 5.5. These involved adaptive preemphasis high-pass filtering and parameter refinements. Another area of improvement is phase dispersion. Methods other than that of Key, Fowle, and Haggarty dispersion for reducing peak/RMS ratio have appeared in the literature [28,29.30]. Although these methods can also achieve large reduction in peak/RMS, they are either computationally infeasible or do not have the flexibility required in the speech context. Nevertheless, this is not to say that there may not exist better (and in particular, more robust) ways of doing phase dispersion. The sensitivity to spectral magnitude and pitch of the Key, Fowle, and Haggarty approach warrants further study in adaptive phase dispersion. Other potential future work includes: refinement of the VISTA algorithm based on field test results; and development of enhancement techniques applicable to music.

Another future item is motivated by the desire to reduce spectral congestion and to eliminate the wasteful carrier in AM modulation which led to the increasing use of single-sideband, suppressed-carrier modulation (SSB-SC or simply SSB) in point to point communication [31]. Although an SSB receiver is more complicated than its AM counterpart, this is regarded as a small price to pay to gain its advantages. However, its use is known to cause troubles in the transmitter because the desire to use modulating signals that are compressed in amplitude can lead to undesirable peaks in the SSB modulation envelope [31]. In particular, the phase relations of the signal

can be undesirably altered. These concerns about SSB transmission of preprocessed speech signals leads to some important potential research efforts in the understanding of the usefulness of the VISTA preprocessor in SSB.

ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of the following Lincoln Laboratory personnel: Gerald O'Leary, Joseph Tierney, Albert Huntoon, Thomas Parks, and Donald Chapman for major contributions to the prototype processor design and development, and for valuable technical discussions; Peter Blankenship for valuable technical discussions; and Walter Morrow for his initial suggestion that digital speech processing should be applicable in the Voice of America broadcast environment, and for his support and encouragement.

We would also like to acknowledge the contributions and support of the following Voice of America personnel: Donald Messer for his initiative in getting this project started, and for his support and valuable technical discussions through its duration; John Birch for his support and for sharing his expertise in audio processing for broadcast; and Gerald Berman for his support and encouragement in the latter stages of the project.

APPENDIX A
THE KEY, FOWLE, HAGGARTY SOLUTION

It is expedient to view the radar transmit filter impulse response, $h(n)$ (Figure 3-6), in terms of its analytic signal representation[13]. Let $\hat{h}(n)$ be the Hilbert transform of $h(n)$. Then the analytic signal counterpart $r(n)$ of $h(n)$ is given by

$$r(n) = h(n) + \hat{j}h(n) \tag{A.1}$$

where " j " denotes imaginary component of the complex signal $r(n)$. The *envelope* of $r(n)$ is given by

$$a(n) = |r(n)| \tag{A.2}$$

and the *phase* is given by

$$\phi(n) = arctan[r(n)] \tag{A.3}$$

Thus the analytic signal can be written as

$$r(n) = a(n)exp[j\phi(n)] \tag{A.4}$$

which has an inverse transform written as

$$R(\omega) = M(\omega)exp[\psi(\omega)] \tag{A.5}$$

where π represents the bandwidth (5k Hz) in the discrete-time representation [15]. The real part of the analytic signal $a(n)\exp[j\phi(n)]$ is the desired time-domain signal $h(n)$:

$$h(n) = Re[a(n)exp(j\phi(n))] = a(n)\cos[\phi(n)] \tag{A.6}$$

The radar signal design problem can be stated as follows: Given a time-domain envelope $a(n)$ and a frequency-domain spectral magnitude $M(\omega)$ find the phase $\psi(n)$ in time and the phase $\theta(\omega)$ in frequency such that the following Fourier transform relation is satisfied:

$$a(n)\exp[j\phi(n)] = F[M(\omega)\exp[j\psi(\omega)] \tag{A.7}$$

where "F" denotes Fourier Transform. Key, Fowle, and Haggarty [4] have shown that, under a large time-bandwidth product constraint, (A.7) can be manipulated to form the relation:

$$|\ddot{\psi}(\omega)| \;=\; c\,|M(\omega)|^2 /\; a(n_\omega) \tag{A.8}$$

where

$$n_\omega = -\frac{1}{2\pi}\dot{\psi}(\omega) \tag{A.9}$$

That is, the time parameter is a function of the phase derivative and c is a constant of proportionality. The time parameter n_ω is called the *group delay* and gives the "time" at which the frequency ω occurs.

If the time-domain envelope is set to a constant over a desired duration L (typical for a radar response) and if (A.8) is integrated twice, the desired phase function in frequency results:

$$|\ddot{\psi}(\omega)| \;=\; c\,|M(\omega)|^2 \tag{A.10a}$$

so that

$$\psi(\omega) = c\int_0^\omega \int_0^\beta M^2(\alpha)d\alpha d\beta \tag{A.10b}$$

The constant of proportionality can be obtained by noting that the group delay at π must equal the desired signal duration, $\dot{L}$; i.e.,

$$n_\pi = L = -\frac{1}{2\pi}\dot{\psi}(\pi) \tag{A.11}$$

that is, since the phase derivative is monotonically increasing, the largest frequency must map to the largest non-zero point in time. From (A.11) the constant, c, can be shown to be proportional to the duration and inversely proportional to the signal energy, i.e.,

$$c = L/\int_0^{\pi} M^2(\alpha)d\alpha \tag{A.12a}$$

so that

$$\psi(\omega) = L\int_0^{\omega}\int_0^{\beta} \hat{M}^2(\alpha)d\alpha d\beta \tag{A.12b}$$

where "hat" denotes the energy-normalized spectrum. In this solution, then, the phase adapts to the response duration and the spectrum.

If instead of specifying a flat time-domain envelope in the general solution (A.8), the phase is made quadratic, then from (A.8), $\ddot{\psi}(\omega) = constant$. With this quadratic constraint, (A.8) becomes:

$$a(\mathrm{n}_\omega) = c|M(\omega)|^2 \tag{A.13}$$

so that $a(n)$ takes on the shape of the spectral envelope. Use of the quadratic phase then is not appropriate for speech since the vocal tract spectral envelope is not flat over the speech bandwidth. Furthermore, changing the constant c for additional quadratic dispersion simply time-expands the filter output while keeping its time envelope intact. These observations have helped explain the limitations of quadratic phase filtering of the speech waveform.

With the KFH solution, the envelope level of the chirp response can be computed. As a consequence of the Hilbert transform relation between the real and imaginary parts of the analytic signal (A.1), the analytic signal $r(n)$ can be shown to have a Fourier transform

$$R(\omega) = M(\omega)exp[\theta(\omega)] \quad for \;\; 0 \leq \omega \leq \pi \tag{A.14}$$

and

$$R(\omega) = 0 \quad for \;\; -\pi \leq \omega \leq 0 \tag{A.15}$$

With $a(n) = A$ for $0 \leq n < L$, it is therefore straightforward to show that the envelope level can be obtained from the spectral energy and the duration L. Consider Parseval's relation:

$$\sum_{0}^{L} a^2(n) \;=\; 1/2\pi \int_0^{\pi} M^2(\omega) d\omega \tag{A.16}$$

Then the constant envelope level A can be written as

$$A = [\; 1/(2\pi L)(\int_0^{\pi} M^2(\omega))]^{1/2} \quad \text{for} 0 \leq n < L \tag{A.17}$$

In the context of speech preprocessing, this level estimate is useful in frequency-domain amplitude compression.

APPENDIX B
ACHIEVING A SMOOTH KFH PHASE

This appendix describes the smoothing of the spectrum and the pitch period prior to the KFH phase calculation, and describes the smoothing of phase along frequency tracks which follows the KFH calculation. The entire smoothing scenario, which was illustrated in Figure 3-9, indicates that all smoothing operations depend on measures of "stationarity" for controlling the degree of smoothing.

Smoothing Spectral Magnitude

In spectral smoothing, the spectrum is divided into a base band (0 to 2500 Hz) and a high band (2500 to 5000 Hz)and the spectrum in each band is smoothed independently in time with different smoothing dynamics. The smoothing dynamics are controlled by a "spectral derivative" in each band and this measure reflects the entire time history of spectral change. The smoothing technique is described in terms of a generic band.

The "energy" in each band is given by

$$E_m = [\int_{B_w} M(\omega; m)^2]^{1/2} \qquad \text{(B.1)}$$

where m denotes the analysis frame number and B_w is either the base band or high band. The spectral derivative between two consecutive frames $m-1$ and m is defined by

$$D(m) = \int_{B_w} [M(\omega; m)/E_m - M(\omega; m-1)/E_{m-1}]^2 \qquad \text{(B.2)}$$

which can be shown to lie in the range [0,1]. This measure is then raised to a power given by the voicing probability $V_p(m)$ (derived via the pitch extractor [10]) to obtain a measure of spectral change

$$\delta(m) = [D(m)]^{V_p(m)} \qquad \text{(B.3)}$$

This ensures that during unvoiced speech when $V_p(m) = 0$, the derivative is effectively set at unity thus prohibiting smoothing of the magnitude.

APPENDIX B
ACHIEVING A SMOOTH KFH PHASE

This appendix describes the smoothing of the spectrum and the pitch period prior to the KFH phase calculation, and describes the smoothing of phase along frequency tracks which follows the KFH calculation. The entire smoothing scenario, which was illustrated in Figure 3-9, indicates that all smoothing operations depend on measures of "stationarity" for controlling the degree of smoothing.

Smoothing Spectral Magnitude

In spectral smoothing, the spectrum is divided into a base band (0 to 2500 Hz) and a high band (2500 to 5000 Hz)and the spectrum in each band is smoothed independently in time with different smoothing dynamics. The smoothing dynamics are controlled by a "spectral derivative" in each band and this measure reflects the entire time history of spectral change. The smoothing technique is described in terms of a generic band.

The "energy" in each band is given by

$$E_m = [\int_{B_w} M(\omega;m)^2]^{1/2} \tag{B.1}$$

where m denotes the analysis frame number and B_w is either the base band or high band. The spectral derivative between two consecutive frames $m-1$ and m is defined by

$$D(m) = \int_{B_w} [M(\omega;m)/E_m - M(\omega;m-1)/E_{m-1}]^2 \tag{B.2}$$

which can be shown to lie in the range [0,1]. This measure is then raised to a power given by the voicing probability $V_p(m)$ (derived via the pitch extractor [10]) to obtain a measure of spectral change

$$\delta(m) = [D(m)]^{V_p(m)} \tag{B.3}$$

This ensures that during unvoiced speech when $V_p(m) = 0$, the derivative is effectively set at unity thus prohibiting smoothing of the magnitude.

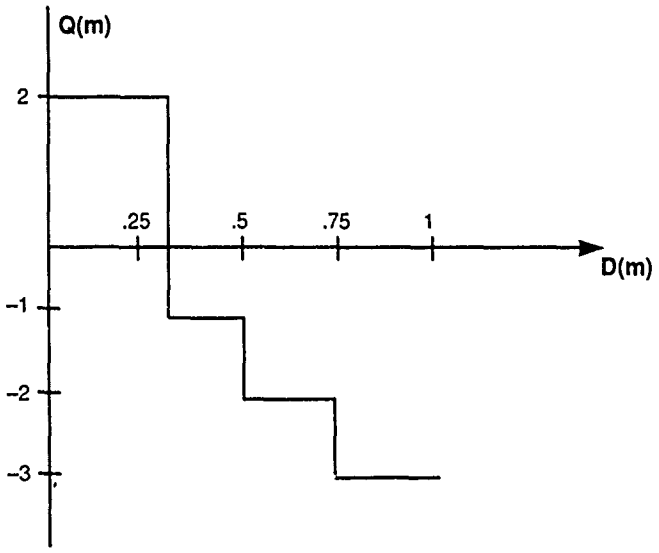


Figure B-1. Quantization levels for spectral derivative.

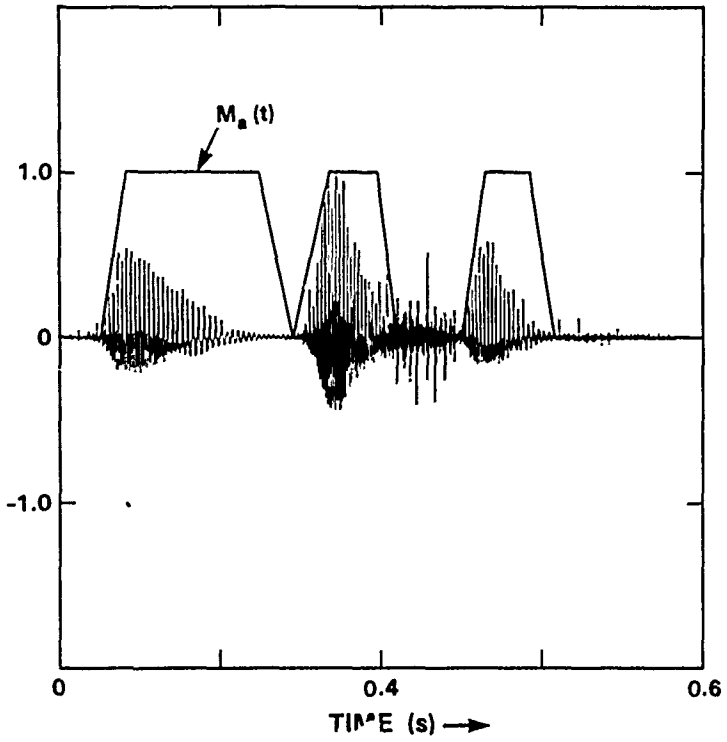


Figure B-2. Illustration of accumulator $M_a(m)$, giving the "degree of stationarity".

and unvoiced spectra which can degrade the performance of the KFH dispersion due to its dependence on the smooth magnitude. An example of these properties is shown in Figure B-2 where the base-band accumulator $M_a(m)$ is superimposed on the dispersed time-domain speech waveform.

Smoothing Pitch

The pitch-smoothing scenario uses the same approach as above; smooth heavily when the pitch is not expected to change significantly since the ear is most sensitive to perturbations in these regions. As with the smoothing of spectral magnitude, a measure of "pitch stationarity" is employed. Whenever a change in pitch period $P(m)$ over two consecutive frames is less than one eighth the current pitch-period estimate [10], a "pitch-period change accumulator", $P_a(m)$, is updated by one (as the accumulator grows, the smoothing increases). If a change in pitch period over two consecutive frames is greater than one-eighth of the current pitch-period estimate, the accumulator is reset to zero (the pitch period is changing very fast and no smoothing is performed).

$P_a(m)$ is then used to compute a smoothing parameter by raising a constant, β, to the power $P_a(m)$

$$\alpha(m) = 1 - \beta^{P_a(m)} \tag{B.7}$$

which is then used in a first-order recursive smoothing loop

$$P'(m) = \alpha(m)P'(m-1) + [1 - \alpha(m)]P(m) \tag{B.8}$$

where as $\alpha(m)$ approaches unity less smoothing is performed.

Smoothing the Phase Along Frequency Tracks

The Key, Fowle, Haggarty phase computation is performed on the smooth spectral magnitude. This phase function (which is a function of frequency) is then smoothed in time along frequency tracks designated "voiced". Smoothing occurs at the frame rate and the degree of smoothing along a particular track, k, depends on the length of the frequency track, denoted $T_k(m)$, (defined as the

number of frames a frequency trajectory exists). The smoothing parameter for the *kth* track is given by

$$\alpha_k(m) = 1 - \beta^{T_k(m)} \tag{B.9}$$

The parameter $\alpha_k(m)$ is used in a first order recursive loop. The result of the phase smoothing is a set of smooth phase trajectories along the "voiced" sine-wave frequency tracks, evaluated at analysis frame boundaries. The smooth phase at frame boundaries will later be added to the excitation phase and interpolated across frames in the synthesis.

APPENDIX C
STATIC AND DYNAMIC PROPERTIES OF AMPLITUDE COMPRESSION

Typically, in amplitude compression, a time-varying gain is applied to reduce envelope variations. This gain is derived from a desired input/output envelope characteristic (IOEC) such as the one illustrated in Figure C-1. This curve is characterized by a compression region which maps a range of input envelopes to a single value, a region in which no modification occurs (to avoid boosting silence or pauses which can cause "breathiness"), and an expansion region which boosts the input envelope. To apply this curve, an input envelope needs to be measured and this conventionally is performed in the time domain as, for example, by waveform rectification or waveform peak-tracking methods. Before applying the curve in Figure C-1, the instantaneous (i.e., raw) envelope estimate is smoothed in time to avoid abrupt changes. These smoothing characteristics are referred to as the *dynamic* component of the amplitude compression, in contrast to the *static* component of Figure C-1. When the time constants which control the dynamics are long (e.g., > 100 ms) , the amplitude compression is referred to as automatic gain control (AGC) and when they are short (e.g., < 100 ms) the compression is referred to as dynamic range compression (DRC). AGC accounts for volume fluctuations (e.g., going from one speaker to another), and DRC accounts for natural short-time fluctuations over a single speaker.

Dynamic Characteristics

Amplitude compression requires an envelope for each time sample (for each analysis frame in the sine-wave based system), denoted by $L(n)$. This envelope measure, $L(n)$, is smoothed according to certain "attack" and "release" dynamics to form the smooth envelope $\hat{L}(n)$. The smoothing dynamics for the release state are given by the recursion

$$\hat{L}(n) = \alpha_r \hat{L}(n-1) + (1-\alpha_r)L(n) \tag{C.1a}$$

and are invoked when the envelope is falling below its average, i.e.,

$$if\ L(n) < \hat{L}(n-1) \tag{C.1b}$$

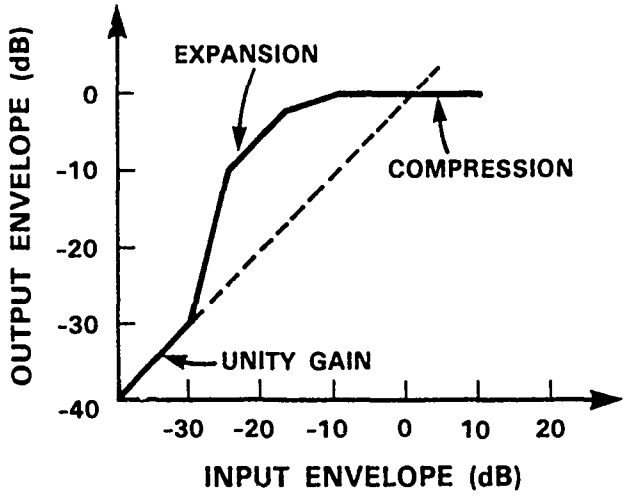


Figure C-1. Typical input/output envelope characteristics (IOEC) for amplitude compression.

The smoothing dynamics for the attack state are given by the recursion

$$\hat{L}(n) = \alpha_a \hat{L}(n-1) + (1-\alpha_a)L(n) \tag{C.1c}$$

and are invoked when the envelope is rising above its average, i.e.,

$$if\ L(n) \geq \hat{L}(n-1)\ . \tag{C.1d}$$

It is important that the α_r and α_a parameters be choosen to avoid "breathiness", "pumping", and other artifacts typical of amplitude compression [17,18]. For example, an excessively slow attack time may give the speech a "muddy" or "bassy" characteristic. Typically the attack time is much faster (almost instantaneous) than the release time.

Static Characteristics

In the static operation, the smooth envelope, $\hat{L}(n)$, is converted to decibels (dB) and applied to the static IOEC to obtain a gain. The gain is then applied to the waveform to perform amplitude compression. The 0 dB reference level, which is a key element in forming the IOEC, is obtained

by making an estimate of the largest envelope of the waveform. Generally this is easy to estimate since an AGC is applied prior to the DRC.

With the 0 dB reference level, denoted by L_0, the gain can be derived. First the input level is computed in dB by

$$L_{in} = 20 * \log[\hat{L}(n)/L_0] \tag{C.2}$$

The gain for an output level L_{out} (in dB and obtained from the IOEC) is then given by

$$gain = 10^{(L_{out}-L_{in})/20} \tag{C.3}$$

Finally, the compressed waveform is obtained by applying the gain to the original waveform. (In the sine-wave processor this modification is made in the frequency domain.)

APPENDIX D
COMPENSATOR DESIGN

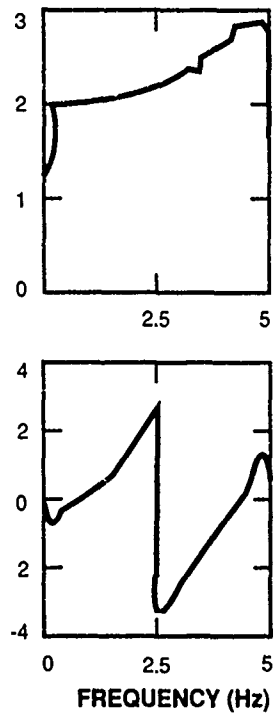
In designing the D/A compensating filter, the response of the D/A filter was measured by exciting the filter with a series of impulses spaced 40 ms apart to avoid response overlap. Fifteen of these responses $h_k(n)$ were averaged (in time) to obtain an estimate $h(n)$

$$h(n) = \sum_{i=1}^{15} h_i(n) \tag{D.1}$$

with an effective length of roughly 200 samples. This estimate (which is noncausal) was alligned with the time origin and then windowed with a 200-point hamming window to effect a smoothing of the spectral magnitude and phase:

$$\hat{h}(n) = w(n)h(n) \tag{D.2}$$

The resulting smooth spectral amplitude and phase of $\hat{h}(n)$ are shown in Figure D-1. In order to avoid excessive preemphasis of the high frequencies (given the unreliablity of the filter estimate near 5 kHz), the filter magnitude was saturated above 4500 Hz as illustrated in Figure D-1.



(a) Spectral Magnitude

(b) Phase

Figure D-1. Characteristics of D/A compensation filter.

APPENDIX E
TRANSMITTER POWER CONSIDERATIONS

The goal of the VOA enhancement project is to increase the RMS level of the speech modulation signal relative to its peak in order to increase the average sideband transmitter power and, hence, improve the robustness of the VOA transmission to natural or man-made noise. This strategy allows the system to be enhanced without upgrading the system transmitters; however, since the average power of the transmission is increased, then so is the fuel bill. This section presents details of calculations quantifying the tradeoff between peak/RMS reduction and increase in total transmit power.

First assume the unprocessed modulation signal is characterized as follows:

$$\text{MAX} \mid s(t) \mid = 1 \tag{E.1}$$

$$[s(t)]_{\text{AVE}} = \lim_{T\to\infty} \frac{1}{T}\int_0^T s(t)dt = 0 \tag{E.2}$$

and

$$([\mid s(t) \mid^2]_{\text{AVE}})^{1/2} = \text{RMS}[s(t)] = [\lim_{T\to\infty}\int_0^T \mid s(t) \mid^2 dt]^{1/2} = k \tag{E.3}$$

so that

$$\text{peak/RMS}[s(t)] = k^{-1} \tag{E.4}$$

The processed signal has an increased RMS level $C \geq k$, and has a peak/RMS ratio C^{-1}. Finally, the transmitted signal is

$$x(t) = [1 + s(t)]\cos\omega_c t \quad . \tag{E.5}$$

Note that the peak constraint on the speech signal guarantees that the transmit signal will not have a phase reversal due to $1 + s(t)$ becoming negative.

We now calculate the average power in $x(t)$ as follows:

$$P = \lim_{T\to\infty} \frac{1}{T} \int_0^T x^2(t)dt \tag{E.6a}$$

$$P = \lim_{T\to\infty} \frac{1}{T} \int_0^T [\cos^2 \omega_c t + 2s(t)\cos^2 \omega_c t + s^2(t)\cos^2 \omega_c t]dt \tag{E.6b}$$

We then use the identity $\cos^2\theta = \frac{1}{2} + \frac{1}{2}\cos 2\theta$ to expand the integrand to six terms. Assuming that $s(t)$ is continuous and slowly varying relative to $\cos^2 \omega_c t$ ($S(t)$ has a 5 kHz bandwidth while $\frac{\omega_c}{2\pi}$ is in the 5-30 MHz range) the integral of the three terms involving $\cos 2\omega_c t$ all vanish because $\int_0^{2\pi} \cos\theta d\theta = 0$. To be more precise $\int_0^T \cos 2\omega_c t\ dt \leq \frac{1}{2\omega_c}$, thus $\lim_{T\to\infty} \frac{1}{T} \cos 2\omega_c t\ dt \leq \lim_{T\to\infty} \frac{1}{T}\frac{1}{2\omega_c} = 0$. We now have for the average power $P = \lim_{T\to\infty} \frac{1}{T}\int_0^T (\frac{1}{2} + \frac{S(t)}{2} + \frac{S^2(t)}{2})dt = \frac{1}{2} + \frac{1}{2}k^2$, which comes from the second term vanishing because $s(t)$ has zero mean, and the third term is the mean square value of $s(t)$. The expression for the average power in the processed signal has the same form. The RMS level of the processed signal is denoted C, and thus,

$$P_p = \frac{1}{2} + \frac{1}{2}C^2 \quad . \tag{E.7}$$

The power ratio R of the average power for the processed signal to the average power of the unprocessed signal is

$$R = P_p/P = (\frac{1 + C^2}{1 + k^2}) \quad . \tag{E.8}$$

To interpret this expression let us assume that the processor increases the RMS level of the speech by 3 dB, i.e., $C = \sqrt{2}k$. The power ratio is then $R = (\frac{1+2k^2}{1+k^2})$, which is plotted in Figure E-1 If $k = 0.316$ (corresponding to a peak/RMS ratio of 10 dB), then $C = 0.446$ (peak/RMS = 7 dB) then the ratio is 1.090 or 0.37 dB. Thus, a doubling of the speech power (+3 dB) must be supported by a 9% increase in the average transmitter power.

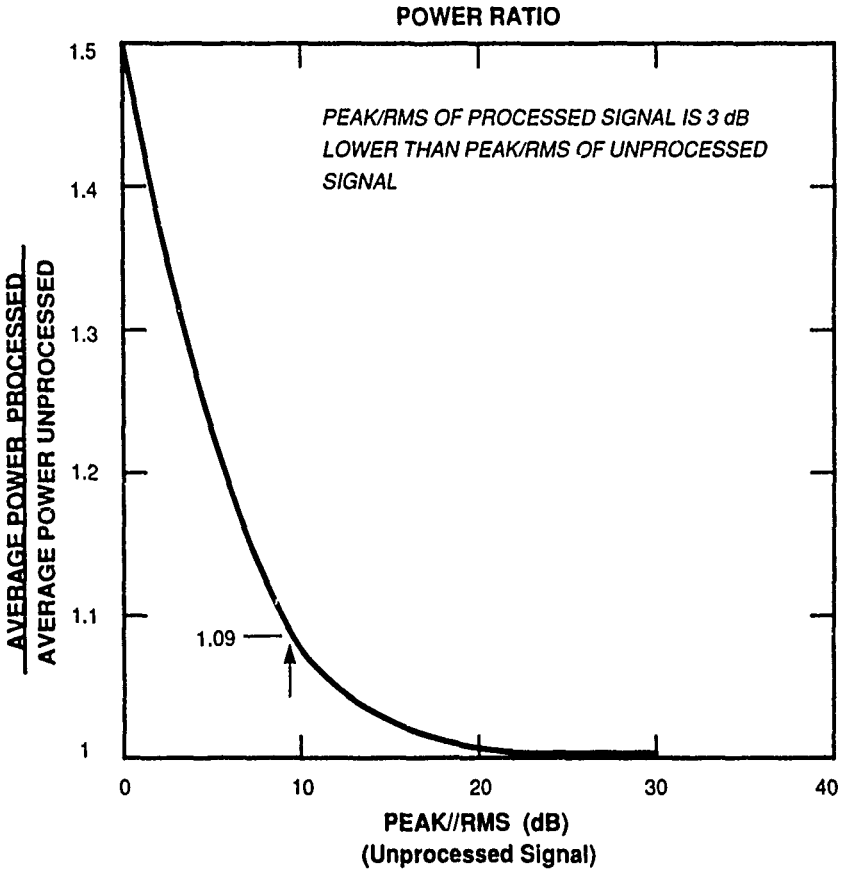


Figure E-1. Peak/RMS.

REFERENCES

1. R.J. McAulay and T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-34, No. 4, pp. 744–754 (August 1986).

2. T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-34 No. 6, pp. 1449–1464 (December 1986).

3. Modulation-Processing Techniques for Sound Broadcasting, Tech. 3243-E, Technical Center of the European Broadcasting Union, Bruxelles, Belgium (July 1985).

4. E.L. Key, E.N. Fowle, and R.D. Haggarty, "A Method of Pulse Compression Employing Nonlinear Frequency Modulation," Technical Report 207, Lincoln Laboratory, MIT. (13 August 1959), DDC 312903.

5. A.H. Huntoon, "LINC21C 12-Bit Audio Data Acquisition Board," Lincoln Manual LM-160, Lincoln Laboratory, MIT (26 May 1989).

6. S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements, *Objective Measures of Speech Quality* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

7. T.F. Quatieri and R.J. McAulay, "Mixed-Phase Deconvolution of Speech Based on a Sine-Wave Model," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., Dallas, Texas, 6–9 April 1987, Vol. 2, pp. 649–652.

8. R.J. McAulay and T.F. Quatieri, "Multirate Sinusoidal Transform Coding at Rates from 2.4 kbps to 8 kbps," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., Dallas, Texas, 6–9 April 1987, Vol. 3, pp. 1645–1648.

9. L. Rabiner and R. Schafer, *Digital Processing of Speech* (Prentice Hall, Inc., Englewood Cliffs, NJ, 1978).

10. R.J. McAulay and T.F. Quatieri, "Pitch Estimation Based on a Sinusoidal Speech Model," (to be published).

11. R.J. McAulay and T.F. Quatieri, "Phase Modeling and Its Application to Sinusoidal Transform Coding," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., Tokyo, Japan, 8–11 April 1986, pp. 1713–1715.

12. D. Paul, "The Spectral Envelope Estimation Vocoder," IEEE Trans. Acoustics, Speech and Signal Process., Vol. ASSP-29, No. 4, pp. 786–794 (August 1981), DTIC AD-A107063.

13. C.E. Cook and M. Bernfeld, *Radar Signals* (Academic Press, 1967).

14. E.N. Fowle, "The Design of FM Pulse-Compression Signals," IEEE Trans. Information Theory, No. 10, pp. 61-67 (January 1967).

15. A. Oppenheim and R. Schafer, *Digital Signal Processing* (Prentice Hall, Inc., Englewood Cliffs, NJ, 1968).

16. M.H.L. Hecker, *A Study of the Relationship Between Consonant-Vowel Ratios and Speaker Intelligibility,* Stanford University, PhD Thesis, 1974.

17. B. A. Blesser, "Audio Dynamic Range Compression for Minimum Perceived Distortion," IEEE Trans. Audio Acoust., Vol. AU-17, No. 1 (March 1969).

18. J. Birch, "Evaluation of Clipping and Limiting Amplifiers," Project Report No. 636, Audio Branch, Engineering Division IBS/ET, United States Information Agency (September 1974).

19. J.D. Griffiths, "Optimum Linear Filter for Speech Transmission," J. Acoust. Soc. Am., Vol. 43, No. 1 (1968).

20. R.J. Niederjohn and J.H. Grotelueschen, "The Enhancement of Speech Intelligibility in Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-24, No. 4 (August 1976).

21. Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," Technical Report 753, Lincoln Laboratory, MIT (10 November 1986), DTIC AD-A176068.

22. "User's Guide for the Lincoln Laboratory Real-Time VISTA System," private communication (February 1989).

11. R.J. McAulay and T.F. Quatieri, "Phase Modeling and Its Application to Sinusoidal Transform Coding," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., Tokyo, Japan, 8-11 April 1986, pp. 1713-1715.

12. D. Paul, "The Spectral Envelope Estimation Vocoder," IEEE Trans. Acoustics, Speech and Signal Process., Vol. ASSP-29, No. 4, pp. 786-794 (August 1981), DTIC AD-A107063.

13. C.E. Cook and M. Bernfeld, *Radar Signals* (Academic Press, 1967).

14. E.N. Fowle, "The Design of FM Pulse-Compression Signals," IEEE Trans. Information Theory, No. 10, pp. 61-67 (January 1967).

15. A. Oppenheim and R. Schafer, *Digital Signal Processing* (Prentice Hall, Inc., Englewood Cliffs, NJ, 1968).

16. M.H.L. Hecker, *A Study of the Relationship Between Consonant-Vowel Ratios and Speaker Intelligibility,* Stanford University, PhD Thesis, 1974.

17. B. A. Blesser, "Audio Dynamic Range Compression for Minimum Perceived Distortion," IEEE Trans. Audio Acoust., Vol. AU-17, No. 1 (March 1969).

18. J. Birch, "Evaluation of Clipping and Limiting Amplifiers," Project Report No. 636, Audio Branch, Engineering Division IBS/ET, United States Information Agency (September 1974).

19. J.D. Griffiths, "Optimum Linear Filter for Speech Transmission," J. Acoust. Soc. Am., Vol. 43, No. 1 (1968).

20. R.J. Niederjohn and J.H. Grotelueschen, "The Enhancement of Speech Intelligibility in Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression," IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-24, No. 4 (August 1976).

21. Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," Technical Report 753, Lincoln Laboratory, MIT (10 November 1986), DTIC AD-A176068.

22. "User's Guide for the Lincoln Laboratory Real-Time VISTA System," private communication (February 1989).