Tuning Time-Frequency methods for the detection of metered HF speech

Douglas J. Nelson[1] and Lawrence H. Smith

R523
U.S. Dept. of Defense
Ft. Meade, MD 20755, USA

June 13, 2002

ABSTRACT

Speech is metered if the stresses occur at a nearly regular rate. Metered speech is common in poetry, and it can occur naturally in speech, if the speaker is spelling a word or reciting words or numbers from a list. In radio communications, the CQ request, call sign and other codes are frequently metered. In tactical communications and air traffic control, location, heading and identification codes may be metered. Moreover metering may be expected to survive even in HF communications, which are corrupted by noise, interference and mistuning. For this environment, speech recognition and conventional machine-based methods are not effective. We describe Time-Frequency methods which have been adapted successfully to the problem of mitigation of HF signal conditions and detection of metered speech. These methods are based on modeled time and frequency correlation properties of nearly harmonic functions. We derive these properties and demonstrate a performance gain over conventional correlation and spectral methods. Finally, in addressing the problem of HF single sideband (SSB) communications, the problems of carrier mistuning, interfering signals, such as manual Morse, and fast automatic gain control (AGC) must be addressed. We demonstrate simple methods which may be used to blindly mitigate mistuning and narrowband interference, and effectively invert the fast automatic gain function.

1. INTRODUCTION

We address the problem of processing HF signals for the purpose of detecting the presence of metered speech. In so doing, we both address the challenges of HF environment and demonstrate new procedures, which are effective in extracting prosodic speech features from HF signals. We then demonstrate a maximum likelihood approach, which is effective in detecting the presence of speech utterances, which are spoken in a metered style.

There are many cases in speech where it is necessary to recognize events which represent a style which is different from the normal style of the conversation or monolog. The classic example of this is the use of the rhymed couplet in Shakespearian plays to cue the entrance of actors or the end of a scene. In normal speech, speaking style can dramatically change when dictating numbers, such as telephone numbers, or when dictating a list of words. In radio communications, the CQ and MAYDAY calls, call signs and other codes sound noticeably different from the normal conversational speech on these channels. In each case, these events sound different because the speaker wishes to convey structured information to the listener, and the speaker wants to insure that the exact message is received by the listener, with no errors. If we consider the problem of a speaker dictating or spelling a list of words of similar length to a listener who must transcribe the words, there are definite limitations placed on the speaker. To transmit the entire list efficiently, the speaker must speak at reasonably fast rate, but not not faster than the listener can write. In addition, if the dictated information is structured, one might well expect the speaker to adopt a style in which the cadence of the syllables becomes somewhat regular. We call such a style of speech metered speech. It occurs naturally in poetry, where the poet intentionally chooses words for each line, such that, in reading the poem, the cadence is natural, and the reader naturally falls into a metered style. Metered speech may also occur naturally, if the speaker is spelling a list of words, or dictating a call sign, flight number, position and bearing, etc.

We pose the problem of detecting a dictated list transmitted over a tactical radio communication channel and propose a detection method based an expected metering of the speech during the list transmission. For the problem

[1] Corresponding author: waveland@erols.com, www.wavelandplantation.com

to represent a realistic tactical situation, we assume a single sideband (SSB) HF transmission, which has all the normal anomalies, including a mistuned carrier, narrowband interference, channel noise and a receiver with a fast automatic gain control (AGC). We demonstrate methods, which are effective in mitigating these signal conditions and detecting occurrences of metered cadences in speech, which may result from recitation of structured lists or multiple repetitions of words or short phrases. The method is based on detection of the harmonic structure resulting from the quasi-periodic amplitude modulation of metered speech.

This paper is structured as follows. In the next section, we briefly discuss the paradigms of the speech signal and metered speech. We then present a brief section, in which we describe the problems of HF channels and the process of simulating HF data. In section 4, we outline processes which mitigate the conditions of the HF channel. In sections 5, we outline the construction of a spectrum which supresses harmonic energy, except at the fundamental, and in section 6, we describe a filter process, which greatly improves isolation of the harmonic fundamental. In the final section, we adapt the methods to the problem of detecting metered speech in an HF environment.

2. THE SIGNAL PARADIGM

Speech is an acoustic signal which is produced as the vocal tract is excited by two excitation functions. These are voicing, in which a series of quasi-stationary pulses are produced at the glottis at the back of the vocal tract and frication, which results from turbulence in the front of the vocal tract, caused by air passing through constrictions at the lips, teeth, or vellum. Information is encoded as the vocal tract changes configuration, resulting in a dynamically changing resonant structure. In speech, data is transmitted as a sequence of words or symbols, each of which consists of one or more syllables. These syllables, which are basic units of speech, may themselves be represented as ordered triplets (CVC) of phonemes, in which the middle phoneme is a vowel, and the leading and trailing phonemes are consonants, either, or both, of which may the the null consonant. While consonants may be be voiced or unvoiced, vowels are always voiced. In speech which has has not been modified, the majority of the signal energy is contained in the vowel, and vocalic stress may generally be observed as an increase in the energy of the stressed vowel. This is the primary factor used in the metered speech detection algorithm we present.

In detecting metered speech, we operate under the assumption that there is a measurable cadence in metered speech. In this cadence, the syllabic stresses are quasi periodic, or the words in the recited list are spoken in groups of similar time duration. If this paradigm is valid, we would expect to see periodicity, representing the stress/group rate, in the energy envelope of the speech signal. This is not an unreasonable assumption, since we have been conditioned to group sequences of numbers into short groups, the length of which depend on the context. For example, an eleven digit telephone number is typically spoken as four groups. As a telephone number, the number 1nnnnnnnnnn is generally grouped as (1)(nnn)(nnn)(nnnn). A 9 digit social security number is grouped as (nnn)(nn)(nnnn). In each case, it is natural to adjust the time between each group so that the group rate is more-or-less constant. In spelling words, or reciting mixed numbers and digits, such as a call sign, each letter other than "W", and each digit, other than "zero" and "seven" consists of one syllable and requires approximately the same amount of time to enunciate, resulting in a natural cadence, whose fundamental is the rate at which the letters or digits are spoken. In military communications, the phonetic alphabet is represented mostly by two syllable words, the first syllable of which is stressed, so we should expect to observe a metered cadence in spelled words, call signs and coordinate locations.

Assuming the model to be correct, what are the observables? We may expect the metered structure to be observable in the envelope of the speech signal and the spectrum of the signal envelope. Since the expected signal envelope is quasi periodic, its expected spectrum is a harmonic structure, whose fundamental is the measured group rate, which we may reasonably expect to be in the approximate range of 0.5 to 2.0 Hz.

3. HF SIGNAL CONDITIONS AND HF SIGNAL SIMULATION

It is our goal to detect the occurrences of metered speech in tactical communications. In particular, we consider the case of an HF communication channel, which adds an additional layer of complexity to the problem. The HF band 3 - 30 MHz is typically very dense, with the possibility of long transmission distances, resulting from ionospheric ducting. The most common modulation is single sideband AM modulation (SSB-AM). Most carrier recovery processes exploit signal symmetry and are based on a phase doubling or similar process [1]. In such processes, the time-domain signal is squared or raised to a higher power. In the frequency domain, this process is a convolution, or iterated convolution of the spectrum. If there is no residual sideband energy in the SSB-AM signal, carrier recovery is nearly impossible.

Because signal strength can vary considerably, most HF receivers are equipped with a fast automatic gain control (AGC), which adapts to changing signal power in a fraction of a second and maintains a nearly constant output signal power, even if that output is pure noise. This poses no problem for human listeners, but renders ineffective any machine-based process which is dependent on signal amplitude. Finally, there are frequently interfering signals, such as co-channel speech and manual Morse. We make the simplifying assumption that there is no interfering speech and that the total bandwidth of interfering signals which are stronger than the signal of interest is small. We make no other assumptions. In particular, we assume a fast AGC and that the carrier may be mistuned by as much as 200 Hz.

A reasonable simulation of a received HF signal may be obtained from a clean speech signal by the following process:

1. Add noise and interference to desired SNR and SIR levels.
2. Complete the noisy analytic signal using the Hilbert transform to estimate the imaginary part of the signal.
3. Frequency translate to simulate carrier offset.
4. Filter signal with a bandpass filter with passband 300 Hz - 3 KHz.
5. Simulate fast AGC.
6. The simulated HF signal is the real part of the previous step.

In simulating the AGC, we may use a leaky integrator of the form

$$\mathrm{G}_{f\alpha}(t) = \alpha \mathrm{G}_{f\alpha}(t - \Delta t) + (1 - \alpha)|f(t)|, \tag{1}$$

where $f(t)$ is the synthesized HF signal without the fast AGC. The signal, with the AGC may then be estimated as

$$f_{AGC}(t) = max(\zeta, f(t)/max(\epsilon, \mathrm{G}_{f\alpha}(t))), \tag{2}$$

where ϵ is some small threshold whose purpose is to bound the denominator away from zero, and ζ is the threshold at which the resulting signal is clipped. Manual Morse interference may be synthesized by adding gated (intermittent) sine waves at step 1.

4. SPEECH DETECTION, INTERFERENCE MITIGATION AND FAST AGC

We address each of the HS signal problems, starting with the detection of the presence of speech, the mitigation of interference, and the mitigation of the fast receiver AGC. We assume that the interference is primarily narrowband and not dense in the speech channel. We assume that the adaptation time of the AGC is on the order of 1/10 second. We further assume an additive noise model, with sparse random impulsive noise bursts. The noise need not be Gaussian or white. Impulsive noise may be dense in the spectrum, but of short duration, and therefore only corrupts a sparse set of analysis frames. An interference. environment fitting this model may consist of a few manual Morse signals with intermittent shot noise.

To detect the presence of speech, we use a variation of the NP algorithm to estimate the SNR of the signal [2]. In the NP algorithm, an instantaneous SNR is estimated as the ratio of the average energy of spectral components of spectral components which have a high probability of being speech related and the average energy of components which are probably not speech related. Assuming a narrowband spectral representation, most of the speech energy is concentrated in pitch harmonics, which occupy approximately 25% to 50% of the nominal 3 kHz channel bandwidth. While the narrowband interference may have greater peak power than the speech signal, we assume that the total aggregate bandwidth of the narrowband interference is less than 10% of the channel. With these assumptions, we may estimate the instantaneous SNR of the speech, with the contributions of the interference and shot noise removed. To do this, we estimate the narrowband spectrogram [3]

$$F(\omega, T) = \left| \int_{-\infty}^{\infty} f(t) w(T - t) e^{-i\omega t} dt \right|^2, \tag{3}$$

where we follow the convention that lower case letters represent the time domain signal and upper case letters represent the spectrogram.

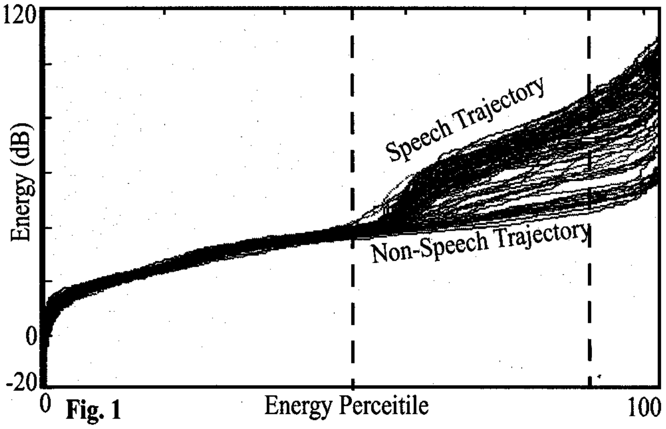

Figure 1. Sorted spectra of clean speech from the SWITCHBOARD database, showing different trajectories for non-speech spectra and speech spectra. Lower traces represent the absense of speech and the upper traces represent speech. The two vertical dashed lines represent the 50^{th} and 90^{th} percentiles, respectively.

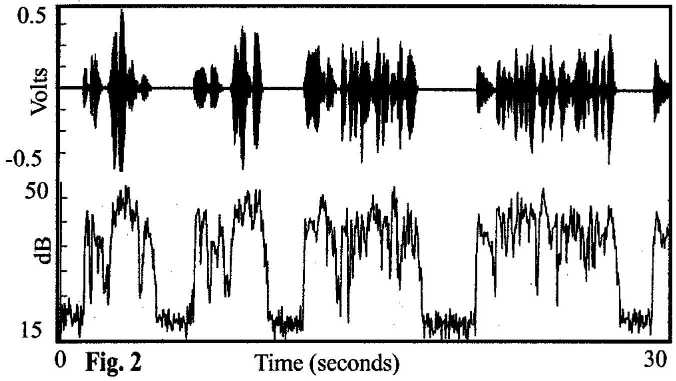

Figure 2. A sample of clean speech representing a read list. Bottom trace is the instantaneous SNR representing the ratio of the 90^{th} and 50^{th} percentile energies expresssed in dB.

For each T_0, we reorder the spectrum in ascending values of $F(\omega, T_0)$, and index the reordering by percentile. This can be done as a permutation of frequency $\omega = \omega_\alpha$, $0 < \alpha < 100$, where $F(\omega_{\alpha_1}, T_0) \geq F(\omega_{\alpha_0}, T_0)$, if $\alpha_0 < \alpha_1$. Under our assumptions, the energy in the 90^{th} energy percentile is dominiated by the speech signal, if it is present, and by noise, if there is no speech. Similarly, energy in the 0 to 50^{th} percentile range represents noise. In the reordered spectrum, the components representing the percentiles 50 and 90 are in fixed locations and may be easily extracted. By letting $\omega_\alpha(T)$ be the value of ω for which the α^{th} energy percentile is achieved at time T. We may therefore estimate the SNR of speech in the HF channel as

$$\mathrm{SNR}_f(T) = 10 \log_{10} \frac{F(\omega_{90}(T), T)}{F(\omega_{50}(T), T)}. \tag{4}$$

Eq. (4) is an estimate of the SNR of the speech, with the narrowband interference removed. If there is no speech present, $|\mathrm{SNR}_f(T)|$ may be expected to be small. Since it is based on a power ratio, $\mathrm{SNR}_f(T)$ is essentially independent of the AGC function, if the AGC function does not change too rapidly. Fig. (1) is a representation (dB)

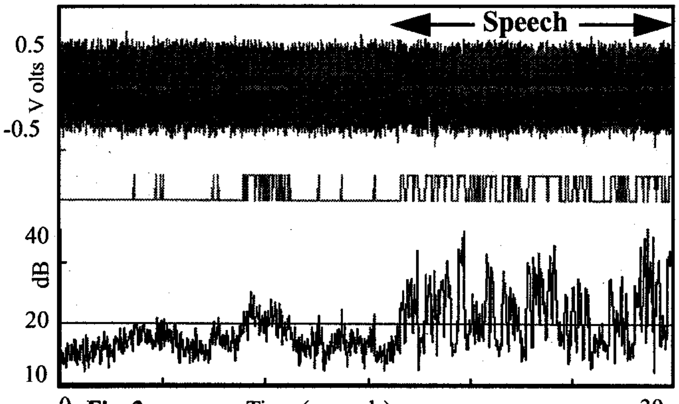

Fig. 3

Figure 3. Thirty seconds of noisy signal with a very fast AGC. Top trace: the signal voltage, Bottom trace: The instantaneous SNR (dB), Middle trace: Voicing indicator.

of the sorted spectra, $\tilde{F}(\Omega,T)$, of a noisy signal, where the spectra were computed from speech sampled at 8 kHz, windowed with a 513 sample Hanning window, zero-filled to 1024 points and Fourier transformed. Only the 513 frequency components representing DC to the Nyquist frequency (4 kHz) were retained. Of note is the the fact that there are essentially two trajectories on the display. The lower trajectory is approximately linearly increasing as a function of α and represents the portion of the signal where speech is not present. The traces in the upper trajectory are approximately 10 to 40 dB higher than those in the lower trajectory, for values of α larger than approximately 55. The upper trajectory represents the presence of speech. The 50^{th} and 90^{th} percentiles are represented by vertical dashed lines. Fig. (2,3) represent, respectively, $\mathrm{SNR}_f(T)$ for a clean signal from the SWITCHBOARD database [4] and a signal with very fast AGC, and non-white noise, consistent with the simulation outline of the previous section. In both cases, speech has been properly indicated, but with some errors in the case of the noisy signal. The middle trace in Fig. (3) represents the binary (speech/not speech) decision, based on a 20 dB threshold applied to $\mathrm{SNR}_f(T)$.

5. CORRELATION SPECTRUM

We next address the issues of carrier mistuning, detection of voiced speech, and estimation of the instantaneous energy of the voiced component of speech. We base this process on the correlation spectrum proposed by Nelson and Tempkin for unambiguous estimation of ω_0, the excitation fundamental of voiced speech* [6]. In our application, we adapt the method to mitigate a mistuned carrier and estimate a time-varying energy envelope approximating the instantaneous energy in the excitation function. In addition, we propose the use of a "thumbtack" highpass filter (THPF) to remove the bias from the spectrogram and autocorrelation surfaces. Use of this filter simplifies the estimation of ω_0 in the NT algorithm, and removes most of the dependence on secondary tests, which were needed in the NT algorithm.

In the correlation spectrum, the normal Fourier spectrum and a spectral representation estimated directly from the time-lag autocorrelation function (TLACF)

$$\mathrm{r}_f(\tau,T) = \int_{-\infty}^{\infty} f^*(t-\frac{\tau}{2})f(t+\frac{\tau}{2})w(T-t)dt \tag{5}$$

are combined to produce an harmonic-free spectral representation.

If, as is the case for speech, the excitation function is a quasi-periodic pulse train, then both the Fourier spectrum and the TLACF have "harmonic" structures, in which energy is concentrated at exact integer multiples of a

*We use the notation ω_0 for the excitation fundamental rather than the more common F_0 to maintain the convention that the signal is represented by variations of the letter "f", while frequency and time are represented, respectively, by variations of "ω" and "t".

fundamental

$$\begin{aligned} \omega_k(T) &= (k+1)\omega_0(T) \quad , k = 0, 1, \ldots \\ \tau_k(T) &= (k+1)\tau_0(T) \quad , k = 0, 1, \ldots, \end{aligned} \tag{6}$$

where $\omega_0(T)$ is the excitation fundamental frequency, and $\tau_0(T)$ is the excitation fundamental period at time T.

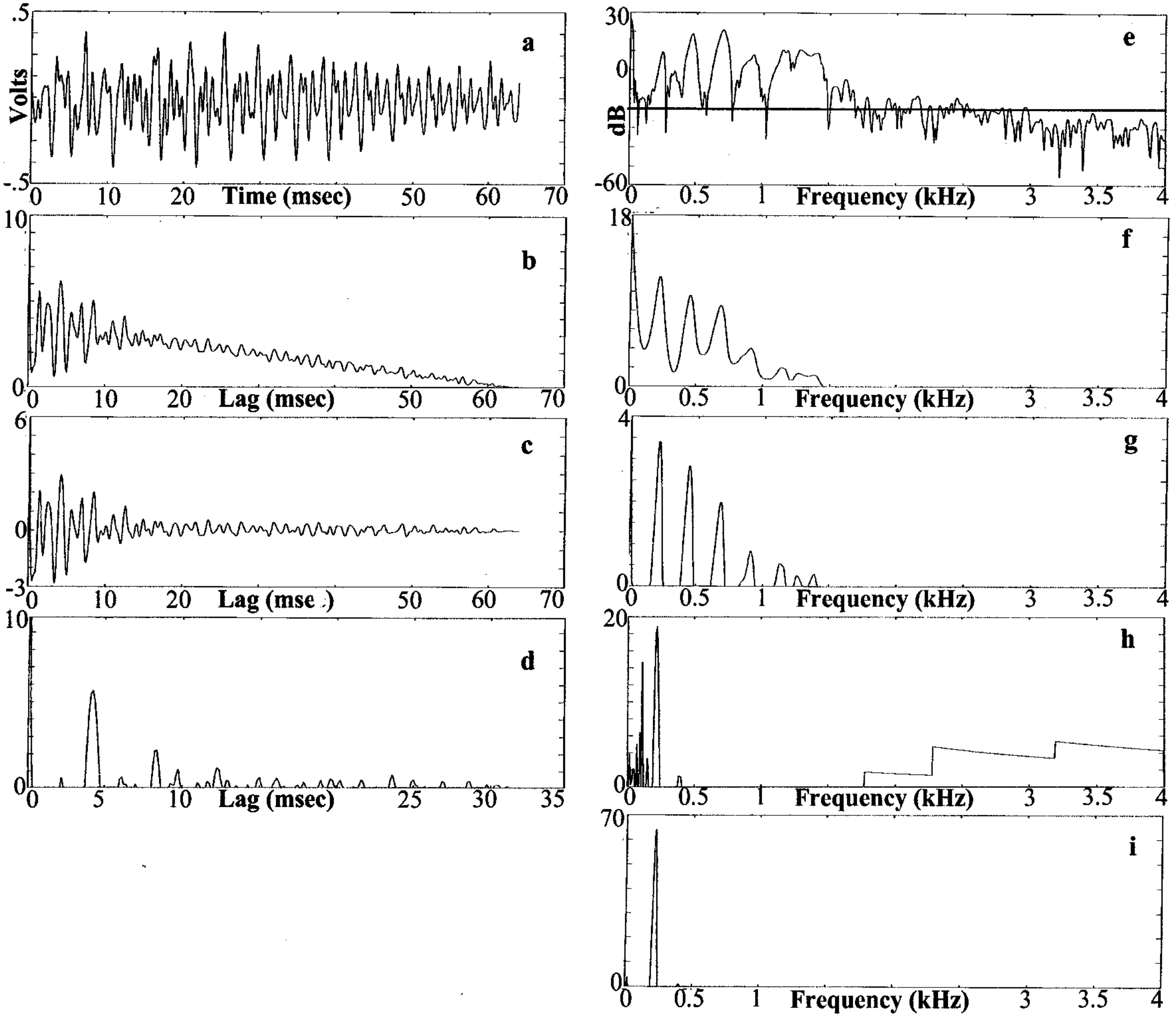

Figure 4. Correlation spectrum of a noisy signal with a 0.1 volt DC offset and a 25 Hz carrier offset **a**: Time domain signal; **b**: Time-lag autocorrelation; **c**: Highpass filtered time-lag autocorrelation; **d**: Half-wave rectified time-lag autocorrelation function calculated as the highpass filtered inverse FFT of log power spectrum of Fig. (4e) **e**: Signal power-spectrum (dB) **f**: Frequency-lag autocorrelation function; **g**: Highpass filtered and half-wave rectified Frequency-lag autocorrelation function; **h**: Reciprocally ordered TLACF from Fig. (4d); **i**: correlation spectrum (product of representations in Fig. (2g) and Fig. (2h))

We note that the receiver may easily mistune the carrier by as much as 200 Hz, resulting in a translated received Fourier spectrum, but only a phase rotation of the autocorrelation coefficients. For the mistuned signal, the Foruier spectral and TLACF "harmonics" satisfy

$$\begin{aligned} \Omega_k(T) &= (k+1)\omega_0(T) + \omega_c \quad , k = 0, 1, \ldots \\ \tau_k(T) &= (k+1)\tau_0(T) \quad\quad\quad , k = 0, 1, \ldots, \end{aligned} \tag{7}$$

where ω_c is the mistuned carrier offset, and $\Omega_k(T)$ is the received frequency of the k^{th} harmonic at time T.

We must translate the "harmonic" bulges in the mistuned signal spectrum to their correct locations, we must reconstruct a frequency fundamental, and we must isolate that fundamental. This is done in the correlation spectrum. To compute the correlation spectrum, we apply a few simple tricks, as depicted in Fig. (4). The first of these is to represent the spectrum as the frequency-lag autocorrelation (FLACF) of the Fourier spectrum

$$\mathrm{R}_f(\omega,T)=\int_0^{\nu} F^*(\omega-\frac{\zeta}{2},T)F(\omega+\frac{\zeta}{2},T)d\zeta, \tag{8}$$

where "*" represents conjugation, and we have assumed for generality that $F(\omega,T)$ may be complex. From the TLACF, we may estimate a frequency spectrum by simply re-indexing the time-lag autocorrelation function

$$F_{\mathrm{r}}(\omega,T)=\frac{1}{\omega}\mathrm{r}(\frac{1}{\omega},T) \qquad ,\omega>0. \tag{9}$$

Finally, the correlation spectrum is computed as the product

$$\mathrm{P}_f(\omega,T)=\mathrm{R}_f(\omega,T)F_{\mathrm{r}}(\omega,T). \tag{10}$$

If we assume that $F(\omega,T)$ is the spectrogram, the units of the magnitude of $\mathrm{P}_f(\omega,T)$ are $[\mathrm{POWER}]^3$, and as a distribution, it is expected to have significant energy only where both the TLACF and FLACF functions both have energy (*i.e.* at the excitation fundamental.) We may use $\mathrm{P}_f^{1/3}(\omega_0(T),T)$ as an estimate of the excitation energy, up to a multiplicative constant. This is only an estimate, since it is dependent on the time-varying formant structure of speech, and it is conditional on removal of the AGC. However, we may assume that it still reflects the gross characteristics of excitation energy function.

5.1. Rolloff and a "thumbtack" filter

The correlation spectrum is critically dependent on the spectral and autocorrelation representations used in its estimation. This factor led Nelson and Tempkin to introduce secondary indicator functions to reduce estimation errors [5, 6]. The primary source of these errors is that the harmonic bulges have only relatively more energy than the components around them. The energy bias may be quite large, making it difficult to identify the fundamental, ω_0. The effects of a small bias are depicted in Fig. (4b, 4c, 4f). We would like a representation in which the only energy resides in a small neighborhood of the fundamental and each of the harmonics. To approximate this, we model the distributions as representations of the desired form with a slowly-varying, additive bias. For example, for the time-lag autocorrelation function

$$\mathrm{r}_f(\tau,T)=\rho_f(\tau,T)+B(\tau,T), \tag{11}$$

where $\rho_f(\tau,T)\geq 0$, and is assumed to be zero for values of τ which are not near some harmonic $\tau_k(T)$.

If we assume the bias to be slowly varying as a function of τ, $\rho_f(\tau,T_0)$ may be approximated by applying to $\mathrm{r}_f(\tau,T_0)$ a highpass filter, which preserves energy peaks and their signs. Since we may assume that the TLACF surface is real, such a highpass filter will result in positive bulges in the filtered surface. Therefore, we need only half wave rectify the filtered surface to obtain an estimate of the surface $\rho_f(\tau,T)$ with the desired characteristics. We must, therefore, only demonstrate an highpass filter with the desired characteristics. A "thumbtack" finite impulse response (FIR) filter, with the following impulse response is one possible such filter.

$$\Psi_\nu(\zeta,0)=\delta(\zeta)-\frac{2}{\nu}(1+cos(2\pi\frac{\zeta}{\nu})) \qquad ,\zeta\in(-\nu/2,\nu/2), \tag{12}$$

where $\delta(\zeta)$ is the Dirac delta function. Eq. (12) represents a zero-phase FIR filter, which is the difference between the original signal and a lowpass representation of the signal, which is created by convolving the signal with a Hanning window. We have written the impulse response as a function of two variables, since we apply it as a spatial filter to surfaces. For the TLACF, the filtered surface is computed, for each value of T_0 as

$$\mathrm{r}_{f\Psi}(\tau,T_0)=\Psi_\nu(\tau,0)*\mathrm{r}_f(\tau,T_0). \tag{13}$$

The filter Eq. (12) preserves local peaks, while tending to remove any slowly varying (in τ) bias. The function $\rho_f(\tau,T_0)$ of Eq. (11) may be approximated, by half-wave rectification as

$$\rho_f(\tau,T_0)=\begin{cases}\mathrm{r}_{f\Psi}(\tau,T_0) & ,\mathrm{r}_{f\Psi}(\tau,T_0)>0\\ 0 & ,\mathrm{r}_{f\Psi}(\tau,T_0)\leq 0\end{cases} \tag{14}$$

An application of this filter to the TLACF is depicted in Fig. (5). It should be noted that, while we have only discussed the application of the highpass filter to the TLACF, it applies equally well to the other surfaces we have discussed. It should be noted that not all highpass filter have the desired properties. For instance, a differentiating filter results in bulges large in absolute value if the surface is changing rapidly in the cross-time variable, and is zero at local maxima.

6. DETECTION OF METERED SPEECH

We now use the tools developed in the previous sections to address the problem of detection of metered speech. As mentioned in Section 2, for metered speech, we expect t a somewhat uniform syllable stress rate, represented by a bulge in the spectrum of the energy envelope, $|f(t)|^2$, of the time domain signal. This bulge we would expect in the frequency range 0.25 - 2.0 Hz. The strategy we follow is to extract, at each time T_0, a vector of features. On a marked training set, we determine the maximum likelihood model and then make decisions on the test data, based on the training model. To winnow the features to a manageable set, we determine how well each feature correlates with the data marks and retain only those features which correlate best. Several features were tested, including harmonic energy in several narrow bands, and $\mathrm{SNR}_f(T)$ of Eq. (4).

The three surviving features were $\mathrm{SNR}_f(T)$, the harmonic energy in the 0.25 - 1.0 Hz band and the harmonic energy in the 2.0 - 4.0 Hz band, computed as follows. We assume that the data have been filtered to approximately 3.0 kHz, sampled at 8 kHz and mu-law encoded. The data were then processed as follows:

1. Segment the data into frames of length 513 samples, with a 263 sample overlap
2. Multiply each frame by a Hanning window
3. Zero-filled to 1024 samples
4. Compute squared magnitude of FFT of each frame
5. Retain only 513 spectral components, representing DC to 4 kHz

Computed as outlined, the time and frequency quantization of $F(\omega, T)$ are $\Delta T = 1/32$ sec and $\Delta\omega = 1/128$ kHz, respectively. The function $\mathrm{SNR}_f(T)$ was estimated from the sorted spectra by Eq. (4). The harmonic energy was estimated for each frame by peak picking the correlation spectrum in a neighborhood of the expected excitation fundamental

$$\tilde{\mathrm{P}}_f(T_0) = \mathrm{P}_f^{1/3}(\omega_0(T_0), T_0) \approx \max_{\omega} \mathrm{P}_f^{1/3}(\omega, T_0) \qquad , |\omega - \mathrm{E}(\omega_0)| < \epsilon, \tag{15}$$

where $\mathrm{E}(\omega_0)$ is the expected value of ω_0, and ϵ is the expected standard deviation of ω_0, estimated over a large set of speakers. In computing the correlation spectrum, the thumbtack filter of the previous section, was applied to both the TLACF and the FLACF.

To compute the in-band harmonic energies, steps 1 - 5 above were applied to $\tilde{\mathrm{P}}_f(T_0)$, using an overlap of 481 in step 1 to produce the surface, which we will simply denote by $P(\omega, T)$. With these parameters, the surface has a time quantization of 1 second and a frequency quantization of approximately 0.031 Hz. The energies in bands 0.25 - 1.0 Hz and 2.0 - 4.0 Hz were estimated by averaging, for each T_0, the surface energy within the respective band at that time.

The algorithm and the various components of it have been applied to several speech files containing metered speech. Typical performance is depicted in Fig. (5).

REFERENCES

1. Bell Telephone Laboratories Technical Staff,*Transmission Systems for Communications,* Bell Telephone Laboratories, Fifth Edition, 1982.
2. D. Nelson and J. Pencak, "Pitch based methods for speech analysis," in *Proc SPIE Adv Sig Proc Conf* , San Diego, CA. vol. 2303, 1994.
3. D. Gabor,"Theory of Communication,"Proc. of the Inst. of Elect. Eng., vol. 93, no 26, pp. 429-457, 1946

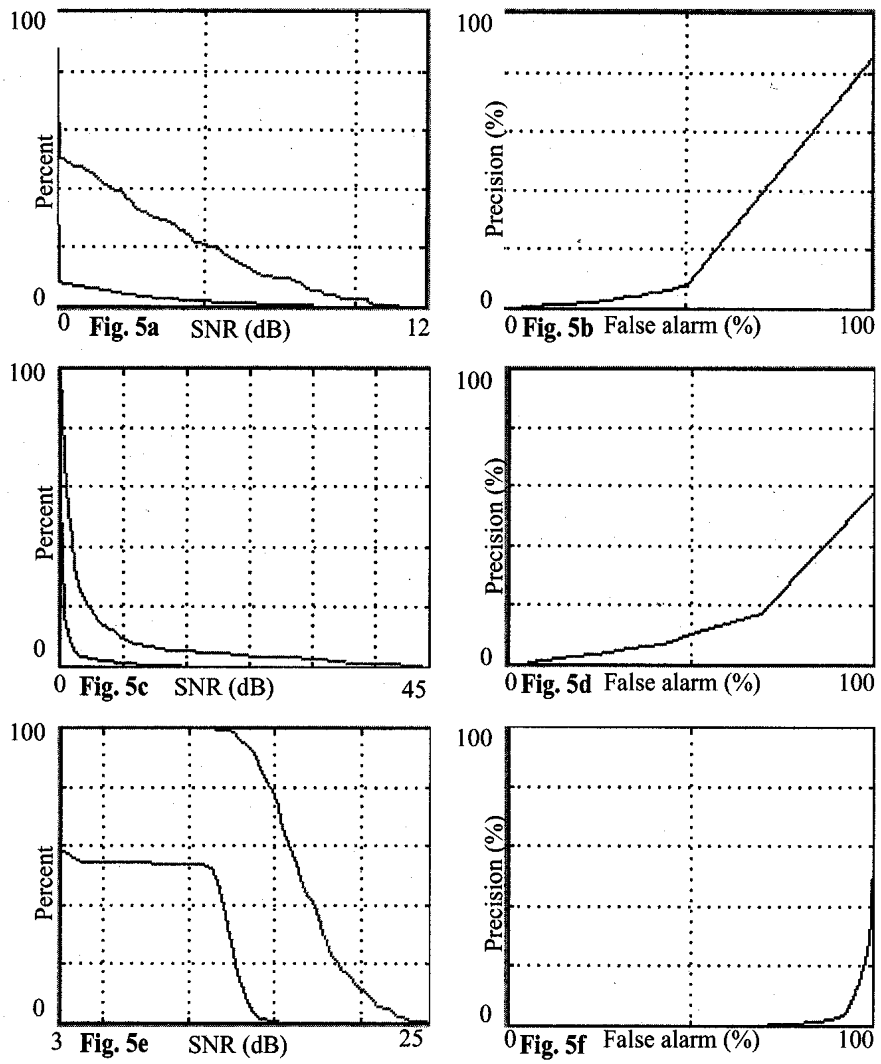

Figure 5. Representative precision/recall tests **a**: Precision/recall using all three features outlined in the text; **b**: Precision as a function of false alarm rate for all three features; **c**: Precision/recall for voicing indicator alone; **d**: Precision as a function of false alarm rate for voicing indicator alone; **e**: Precision/recall for SNR alone; **f**: Precision as a function of false alarm rate for SNR alone.

4. Switchboard-2 Phase 1, The Switchboard Corpus of Recorded Telephone Conversations (available through the Linguistic Data Consortium, University of Pennsylvania), NIST, 1996.
5. J. A. Tempkin and D. J. Nelson, "A Spectral Phase Algorithm for Detecting and Estimating Pitch," in *Proc. of the SPIE Adv. Sig. Proc. Conf.*, August, 2001.
6. D. J. Nelson and J. A. Tempkin, "Signal exploitation by time and frequency correlation," in *Proc. of the SPIE Adv. Sig. Proc. Conf.*, August, 2001.