SIGNAL RECONSTRUCTION FROM CONCENTRATED STFT PEAKS

D.J. Nelson

U.S. Dept. of Defense
Ft. Meade, MD 20755-6514

ABSTRACT

We propose a new method for adaptively removing noise and interference from a signal. In this method, isolation of the signal and interference components is facilitated by a concentration process applied to the short time Fourier transform (STFT). Unwanted components are removed from the concentrated surface, and a clean signal is estimated by integration. The concentrated STFT is a linear representation, free of cross terms and having the property that signal and interference components are easily recognized because their distributions are more concentrated in time and frequency. We demonstrate the advantages of the proposed method over conventional methods.

1. INTRODUCTION

A frequently recurring problem in communications is the need to remove noise and interference from a signal. There have been many approaches suggested to address this problem. For stationary narrowband interference, a simple notch filter may be effective. Adaptive notch filters have been suggested to remove narrowband non-stationary interference (c.f. [1, 2]). For removal of broadband noise from speech, Wiener filter techniques, such as spectral subtraction are frequently used (c.f. [3, 4]).

In adaptive filter implementations, the time-varying instantaneous frequency (IF) of the interfering signal is estimated from a TF distribution, such as a Wigner distribution (c.f. [5, 1]). The interfering signal is then translated to a constant frequency and removed with a notch filter, and the resulting clean signal is frequency translated back to baseband. In this operation, the interference bandwidth and IF are estimated and the IF must be accurately tracked. For a single narrowband interfering signal and a wideband signal of interest, this may not be a problem, but isolating and tracking the IF in the presence of multiple narrowband components and cross-terms may be difficult.

In Wiener filtering or spectral subtraction, components which are dominated by noise are removed from the spectrogram, and the clean signal is estimated by a pseudo inversion process (c.f. [3, 4]). In general, a true inverse of the modified spectrogram does not exist, so the solution must be chosen to minimize an error criterion.

We propose a new signal reconstruction method. This method is based on a recently proposed linear TF paradigm [6]. In the linear TF distribution paradigm there are two assumptions: (1) such distributions are linear representations of the signal; (2) the value of the signal at each time is distributed in frequency, in the sense that the value of the signal at each time may be estimated as the integral of the surface with respect to frequency. This second condition is a linear time marginal (LTM). Linear distributions are significantly different from energy distributions. Energy distributions are not linear. In addition, for energy distributions, the time marginal condition is the requirement that integration of the TF surface with respect to frequency results in the energy density of the signal. Because of linearity and the LTM, we may remove unwanted components from linear TF surfaces and recover a signal whose TF representation is the modified surface.

We present a method for estimating signal components with narrow FM bandwidth directly from a wideband STFT representation. In the proposed method, the STFT is modified, concentrating the components along curves functionally representing the instantaneous frequencies of individual signal components. The concentration process may be seen to preserve linearity and the LTM property. From the concentrated STFT, the actual values of an individual signal component may be estimated by evaluating the concentrated STFT along a curve functionally representing the instantaneous frequency of that component.

We apply our methods to estimate a clean speech signal from an interference environment. In applying these methods, speech formants are reduced to narrowband components by the concentration process. A clean signal may then be estimated as the sum of relatively few concentrated components.

2. THE SIGNAL REPRESENTATION, IF AND STFT

We assume an analytic multi-component signal of the form $s(t) = a(t)e^{i\omega t}$, $a(t) \geq 0$. We further assume that the signal may be decomposed as the sum of signal components

$$s(t) = \sum_k s_k(t) \quad (1)$$

$$s_k(t) = a_k(t)e^{i\phi_k(t)} \quad , a_k(t) \geq 0, \phi_k(t) \in \Re \quad (2)$$

where noise and interference may be represented as one or more of the signal components. This decomposition is not assumed to be unique.

2.1. Instantaneous Frequency

While it is not essential, it is useful to consider the individual signal components to be narrowband in the sense that their instantaneous frequencies are slowly varying as

a function of time. For the signal Eq. (1), the respective instantaneous frequencies of $s(t)$ and $s_k(t)$ are

$$\omega_s(t) = \frac{d}{dt}\arg\{s(t)\} = \frac{d}{dt}\phi(t) \quad (3)$$

$$\omega_{s_k}(t) = \frac{d}{dt}\arg\{s_k(t)\} = \frac{d}{dt}\phi_k(t) \quad (4)$$

2.2. Fourier Transform and STFT

The Fourier transform of a signal, $s(t)$, is defined by

$$S(\omega) = \int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt. \quad (5)$$

The complex-valued STFT [7] has the representation

$$\mathbf{S}_{kh}(\omega, T) = \int_{-\infty}^{\infty} s_k(t+T)h(-t)e^{-i\omega t}dt. \quad (6)$$

$$\mathbf{S}_h(\omega, T) = \int_{-\infty}^{\infty} s(t+T)h(-t)e^{-i\omega t}dt \quad (7)$$

$$= \sum_k \mathbf{S}_{kh}(\omega, T), \quad (8)$$

where Eq. (8) follows from Eq. (1). In the following discussion, we assume for simplicity that the window, $h(-t)$, and its frequency response, $H(\omega)$, are both real and symmetric and that both have mean zero and small variance.

For a signal component, $s_k(t)$, of the form Eq. (2) to be narrowband, both $a_k(t)$ and $\phi_k(t)$ must change slowly as a function of time. In this case, $\phi_k(t)$ has a power series expansion $\phi_k(t+T) \approx \phi_k(T) + \omega_{s_k}(T)t$, and the STFT has the representation

$$\mathbf{S}_{kh}(\omega, T) \approx a_k(T)\int_{-\infty}^{\infty} e^{i\phi_k(t+T)}h(-t)e^{-i\omega t}dt \quad (9)$$

$$\approx a_k(T)e^{i\phi_k(T)}\int_{-\infty}^{\infty} e^{i\omega_{s_k}(T)t}h(-t)e^{-i\omega t}dt \quad (10)$$

$$= s_k(T)H(\omega - \omega_{s_k}(T)). \quad (11)$$

2.3. The CIF representation

To distinguish the signal instantaneous frequency from the derivative of the STFT phase with respect to time, we will use the notation [8, 9]

$$\mathbf{CIF}_{\mathrm{S}_h}(\omega, T) = \frac{\partial}{\partial T}\arg\{\mathbf{S}_h(\omega, T)\} \quad (12)$$

$$\approx \frac{1}{\epsilon}\arg\{\mathbf{S}_h(\omega, T+\frac{\epsilon}{2})\mathbf{S}_h^*(\omega, T-\frac{\epsilon}{2})\}, \quad (13)$$

where ϵ is assumed to be small.

3. ESTIMATING SIGNAL COMPONENTS FROM THE STFT

We will now demonstrate that signal components may be estimated directly from the STFT. To do this, we need to define the concepts of separability and linear time marginals.

3.1. Separability

We define a TF surface, $\mathbf{S}(\omega, T) = \sum \mathbf{S_k}(\omega, T)$, to be separable (with respect to a given decomposition) at a point, (ω_0, T_0), if $\mathbf{S}_k(\omega_0, T_0) \neq 0$ for at most one value of k. A surface is approximately separable at (ω_0, T_0) if for some k

$$|\mathbf{S}_k(\omega_0, T_0)| \gg |\mathbf{S}_l(\omega_0, T_0)|, k \neq l. \quad (14)$$

Separability is a local condition which is either true or not true at each point on the surface. At a separable or approximate separable point, (ω_0, T_0), the component, $s_k(t)$, whose surface magnitude, $|\mathbf{S}_k(\omega_0, T_0)|$, is largest is defined to be the dominant component. If $s_k(t)$ is the dominant component, at a separable point, (ω_0, T_0), then $\mathbf{S}(\omega_0, T_0) = \mathbf{S}_k(\omega_0, T_0)$. Assuming Eq. 1,2, separability is the exact condition under which one signal component can be isolated, and estimated from the surface of the composite signal.

Since the STFT is linear, separability is the condition under which the local contribution of a signal component can be isolated on the STFT surface. By Eq. (11), for coordinates, (ω, T), near a separable point, (ω_0, T_0), the dominant component, s_k, satisfies [9]

$$\mathbf{CIF}_{\mathrm{S}_h}(\omega, T) \approx \omega_{s_k}(T_0). \quad (15)$$

3.2. The Linear Marginals

For energy distributions, the marginal conditions are the requirement that the integrals of the surface with respect to time and frequency respectively result in the power spectral and the signal energy density functions respectively. We define the linear marginals as

$$\mathfrak{s}(T) = \int_{-\infty}^{\infty} \mathbf{S}(\omega, T)d\omega \quad (16)$$

$$\mathfrak{S}(\omega) = \int_{-\infty}^{\infty} \mathbf{S}(\omega, T)e^{-i\omega T}dT \quad (17)$$

For the STFT, the linear time marginal is

$$\mathfrak{s}_h(T) = \int_{-\infty}^{\infty} \mathbf{S}_h(\omega, T)e^{i\omega t}d\omega\Big|_{t=0} \quad (18)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(-t)s(t+T)e^{-i\omega t}dte^{i\omega t}d\omega\Big|_{t=0} \quad (19)$$

$$= 2\pi h(0)s(T) \quad (20)$$

3.3. Estimating a "Narrowband" Component

We now combine the LTM and separability properties with the representation Eq. (11) to estimate the value of a signal component from the STFT. By Eq. (11) and Eq. (20), we may estimate the value of a narrowband signal component at a point T_0 as

$$s_k(T_0) \approx \int_{\omega_{s_k}(T_0)-\epsilon}^{\omega_{s_k}(T_0)+\epsilon} \mathbf{S}_h(\omega, T_0)d\omega, \quad (21)$$

where the magnitude of the error of Eq. (21) is approximately

$$2\left|\int_{\epsilon}^{\infty} H(\omega)d\omega\right| \quad (22)$$

and we have assumed that $\omega_{s_k}(T_0)$ is known. If we evaluate Eq. (21) for each T_0, the resulting signal is effectively the result of applying an adaptive bandpass filter to the signal, $s(t)$, where the center frequency of the filter at any time is the instantaneous frequency of the component, $s_k(t)$, and the bandwidth of the filter is 2ϵ. The adaptive filter, Eq. (21), can be made to adapt to the instantaneous signal bandwidth by making ϵ time dependent.

4. CONCENTRATING THE STFT

The representation, Eq. (11), may be related to the uncertainty principle for the STFT. Assuming a real and symmetric window, $h(-t)$, the time uncertainty is essentially the standard deviation of $h(-t)$. The frequency uncertainty is the bandwidth of the windowing function. In addition, there is the uncertainty in the signal, representing the combined AM and FM signal modulation.

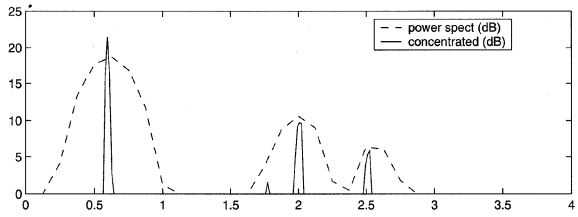

Fig. 1. The spectrum of 3 msec of voiced speech. Dashed line: power spectrum (dB); Solid line: magnitude of concentrated spectrum (dB)

We will now use the CIF to concentrate the STFT. In this concentration process, we effectively remove much of the STFT and signal uncertainty. For a single component with slowly varying amplitude and instantaneous frequency, this concentration process will result in a surface whose energy is narrowly concentrated along a curve functionally representing the IF of the signal component. The concentration process may create a narrowband representations of signal components whose spectral bandwidths are broad.

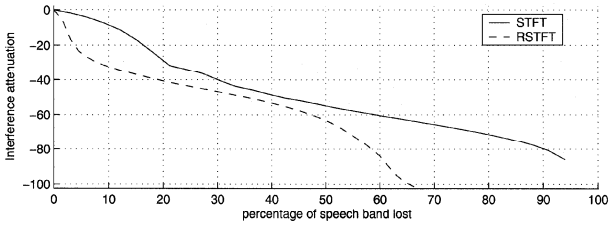

Fig. 2. A comparison (Solid line: STFT; Dashed line: weighted concentrated STFT) of the attenuation of the sinusoidal FM interference of Fig. 3 as a function of percentage of the TF surface lost.

First compute $\mathbf{CIF}_{\mathrm{S}_h}(\omega, T)$. By Eq. (15), $\mathbf{CIF}_{\mathrm{S}_h}(\omega_0, T_0)$ represents an estimate of the IF of the dominant signal component at separable point, (ω_0, T_0). We reassign the STFT component, $\mathrm{S}_h(\omega_0, T_0)$ to the coordinate, $(\mathbf{CIF}_{\mathrm{S}_h}(\omega_0, T_0), T_0)$. In this process, the values of all STFT components mapped to (Ω_0, T_0) are accumulated and assigned to that coordinate [1]. We denote by $\mathrm{S}_{\mathrm{CIF}}(\omega, T)$ the concentrated STFT in which each surface component has been assigned to its new frequency. We have dropped the dependence on h from the notation, since concentration effectively mitigates the effect of the windowing function, as may be seen in Fig. 1. Since we have assumed that $h(t)$ and $H(\omega)$ are both real with mean zero and small variance, we may assume that the concentrated STFT preserves phase and timing of the dominant signal component at each separable point. In the concentration process, at each time, T_0, most of contribution of each signal component may be concentrated to a very narrow frequency band.

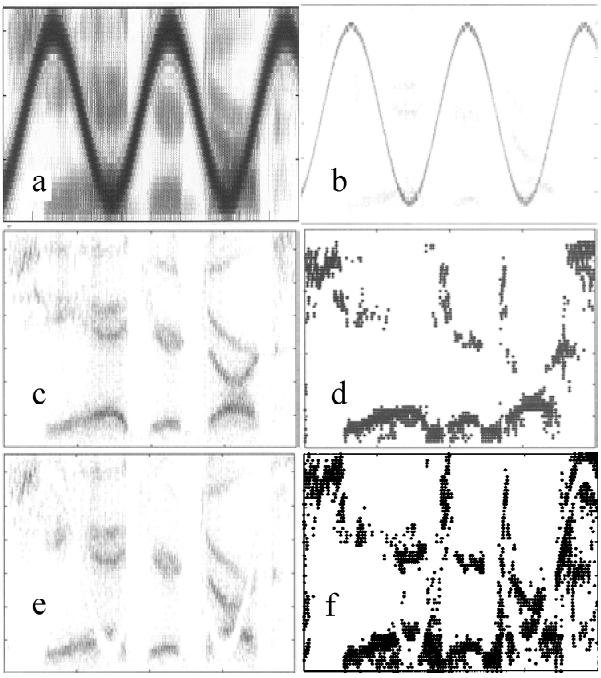

Fig. 3. TF representations of signal before removing interference. Displays represent one second of data computed with a 3 msec Hanning window. a: Spectrogram of speech and interference (dB); b: Concentrated STFT (dB) of speech and interference; c: Concentrated STFT of clean speech (dB); d: Locus of strongest peak of Fig.3c; e: Concentrated STFT (dB) of speech with concentrated interference removed; e: Locus of the strongest peak of concentrated STFT with interference removed

We now apply frequency smoothing to the concentrated STFT at each time

$$\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T) = \mathrm{S}_{\mathrm{CIF}}(\omega, T) \circ_\omega W(\omega), \tag{23}$$

where $\circ_\omega$ represents convolution with respect to ω and $W(\omega)$ is a smoothing window. We may estimate the value at time,

[1] For a similar process applied to the spectrogram, c.f [9]-[11].

T_0, of individual signal components by evaluating the "spectrum", $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T_0)$, at values of ω for which the magnitude of $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T_0)$ is relatively maximized.

We may estimate the analytic signal at time, T_0, as the sum of the complex-valued "peaks" of $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T_0)$. A real signal is then computed as the real part of the resulting sum. Narrowband interference may be removed by excluding interference peaks from the sum, and since noise has random phase, noise contributes little to the peak values.

In the concentration process wideband signal components may be reduced to a narrowband representation. The bandwidth compression of signal components may be considerable. In Fig. 2, we represent the signal energy (dB)as a function of bandwidth of the sinusoidal interference represented in Fig. 3. Displayed are the distributions for the un-modified STFT and the concentrated STFT. The bandwidth reduction of the component is significant. At the 20 dB energy threshold, the STFT bandwidth is 16% of the total spectrum. For the concentrated STFT, the corresponding bandwidth is 3%.

5. SIGNAL ESTIMATION AND REMOVAL OF INTERFERENCE AND NOISE

In testing the method, We selected files from the TIMIT database [12]. In each case, $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T)$ was computed using a 3 msec Hanning window. For each T_0, the 5 components with the largest peak magnitude were selected and summed to approximate the analytic signal. An audio signal was constructed as the real part of the analytic signal. The signal constructed in this manner was nearly distortion free.

Next, as indicated in Fig. 3, a sinusoidally FM modulated signal was added to the speech, and the surface, $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T)$, was computed as before. The 5 strongest peak components at each time were computed as before. Since the interference was much stronger than the speech signal, the interference component was removed by discarding the strongest peak component, and the clean signal was estimated as the sum of the remaining 4 components. The resulting signal was nearly clean speech, with some distortion noted when speech components could not be distinguished from the interference component on the TF surface.

As a final test, a signal representing only the single strongest peak value of $\widetilde{\mathrm{S}}_{\mathrm{CIF}}(\omega, T_0)$ at each time was computed from the clean signal. This single component was completely intelligible and undistorted. The fact that it resulted in clean speech is quite amazing. For voiced speech, this component generally represented the first formant, and for unvoiced speech, this component generally represented the frequency at which frication "energy" was concentrated, as represented in Fig. 3.

6. CONCLUSIONS

We have presented a new complex-valued linear TF representation in which multi-component signals, such as speech may be represented as a sparse sum of narrowband components. We have demonstrated that this representation may be easily estimated from the STFT by a simple concentration process. This representation may be used to remove non-stationary interference and in data compression.

7. REFERENCES

[1] M. Amin, C. Wang, and A. Lindsay, "Optimum Interference excision in Spread Spectrum Communications Using Open-Loop Adaptive Filters", in IEEE Trans. Sig. Proc., vol. 47, no. 7, pp. , July 1999.

[2] D. Rich, "Cochannel FM Interference Suppression Using Adaptive Notch Filters", in IEEE Trans. Comm, vol

[3] S.F. Boll, "A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech", in IEEE Conf. on Acoust. Speech and Signal Proc., vol. 4, pp. 200-203, April, 1975.

[4] G. Whipple, "Low Residual Noise Speech Enhancement Utilizing Time-Frequency Filtering", in IEEE Conf. on Acoust. Speech and Signal Proc., vol. 1, pp. 5-8, 1994.

[5] B. Boashash and M. Mesbah, "Signal Enhancement By Time-Frequency Peak Filtering", in IEEE Transactions On Signal Processing, vol. 52, no. 4, pp. 929-937, Apr. 2004.

[6] D. J. Nelson, D.C. Smith and R.C. Masenten, "Linear distribution of signals," to appear in SPIE , Denver, 2004.

[7] D. Gabor, "Theory of communication,", in J. IEE, vol. 93, 429-457, 1946.

[8] D. Nelson, "Special Purpose Correlation Functions", internal report, 1989, (republished in part in: "Special Purpose Correlation Functions for Improved Parameter estimation and Signal Detection," Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, Minneapolis, pp. 73-76, April,1993).

[9] D. J. Nelson, "Cross-spectral methods for processing speech", in J. Acoust. Soc. Am., vol.110, num. 5, pt. 1, pp. 2575-2592, Nov. 2001.

[10] K. Kodera, R. Gendrin and C. de Villedary, "Analysis of time-varying signals with small BT values," in IEEE Trans. Acoust. Speech Signal Proc.,Vol.26,pp.64-76, 1978.

[11] F. Auger and P. Flandrin, "Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method," in IEEE Trans. Sig. Proc., vol. 43, no. 5, pp. 1069-1089, May, 1995.

[12] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST, 1990.