Suppression of Acoustic Noise in Speech Using Spectral Subtraction

STEVEN F. BOLL, MEMBER, IEEE

Abstract—A stand-alone noise suppression algorithm is presented for reducing the spectral effects of acoustically added noise in speech. Effective performance of digital speech processors operating in practical environments may require suppression of noise from the digital waveform. Spectral subtraction offers a computationally efficient, processor-independent approach to effective digital speech analysis. The method, requiring about the same computation as high-speed convolution, suppresses stationary noise from speech by subtracting the spectral noise bias calculated during nonspeech activity. Secondary procedures are then applied to attenuate the residual noise left after subtraction. Since the algorithm resynthesizes a speech waveform, it can be used as a preprocessor to narrow-band voice communications systems, speech recognition systems, or speaker authentication systems.

I. Introduction

BACKGROUND noise acoustically added to speech can degrade the performance of digital voice processors used for applications such as speech compression, recognition, and authentication [1], [2]. Digital voice systems will be used in a variety of environments, and their performance must be maintained at a level near that measured using noise-free input speech. To ensure continued reliability, the effects of background noise can be reduced by using noise-cancelling microphones, internal modification of the voice processor algorithms to explicitly compensate for signal contamination, or preprocessor noise reduction.

Noise-cancelling microphones, although essential for extremely high noise environments such as the helicopter cockpit, offer little or no noise reduction above 1 kHz [3] (see Fig. 5). Techniques available for voice processor modification to account for noise contamination are being developed [4], [5]. But due to the time, effort, and money spent on the design and implementation of these voice processors [6]–[8], there is a reluctance to internally modify these systems.

Preprocessor noise reduction [12], [21] offers the advantage that noise stripping is done on the waveform itself with the output being either digital or analog speech. Thus, existing voice processors tuned to clean speech can continue to be used unmodified. Also, since the output is speech, the noise stripping becomes independent of any specific subsequent speech processor implementation (it could be connected to a CCD channel vocoder or a digital LPC vocoder).

The objectives of this effort were to develop a noise suppression technique, implement a computationally efficient algorithm, and test its performance in actual noise environments. The approach used was to estimate the magnitude frequency spectrum of the underlying clean speech by subtracting the noise magnitude spectrum from the noisy speech spectrum. This estimator requires an estimate of the current noise spectrum. Rather than obtain this noise estimate from a second microphone source [9], [10], it is approximated using the average noise magnitude measured during nonspeech activity. Using this approach, the spectral approximation error is then defined, and secondary methods for reducing it are described.

The noise suppressor is implemented using about the same amount of computation as required in a high-speech convolution. It is tested on speech recorded in a helicopter environment. Its performance is measured using the Diagnostic Rhyme Test (DRT) [11] and is demonstrated using isometric plots of short-time spectra.

The paper is divided into sections which develop the spectral estimator, describe the algorithm implementation, and demonstrate the algorithm performance.

II. Subtractive Noise Suppression Analysis

A. Introduction

This section describes the noise-suppressed spectral estimator. The estimator is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. Spectral information required to describe the noise spectrum is obtained from the signal measured during nonspeech activity. After developing the spectral estimator, the spectral error is computed and four methods for reducing it are presented.

The following assumptions were used in developing the analysis. The background noise is acoustically or digitally added to the speech. The background noise environment remains locally stationary to the degree that its spectral magnitude expected value just prior to speech activity equals its expected value during speech activity. If the environment changes to a new stationary state, there exists enough time (about 300 ms) to estimate a new background noise spectral magnitude expected value before speech activity commences. For the slowly varying nonstationary noise environment, the algorithm requires a speech activity detector to signal the

Manuscript received June 1, 1978; revised September 12, 1978. This research was supported by the Information Processing Branch of the Defense Advanced Research Projects Agency, monitored by the Naval Research Laboratory under Contract N00173-77-C-0041.

The author is with the Department of Computer Science, University of Utah, Salt Lake City, UT 84112.

0096-3518/79/0400-0113$00.75

program that speech has ceased and a new noise bias can be estimated. Finally, it is assumed that significant noise reduction is possible by removing the effect of noise from the magnitude spectrum only.

Speech, suitably low-pass filtered and digitized, is analyzed by windowing data from half-overlapped input data buffers. The magnitude spectra of the windowed data are calculated and the spectral noise bias calculated during nonspeech activity is subtracted off. Resulting negative amplitudes are then zeroed out. Secondary residual noise suppression is then applied. A time waveform is recalculated from the modified magnitude. This waveform is then overlap added to the previous data to generate the output speech.

B. Additive Noise Model

Assume that a windowed noise signal $n(k)$ has been added to a windowed speech signal $s(k)$, with their sum denoted by $x(k)$. Then

$$x(k) = s(k) + n(k).$$

Taking the Fourier transform gives

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega})$$

where

$$x(k) \longleftrightarrow X(e^{j\omega})$$

$$X(e^{j\omega}) = \sum_{k=0}^{L-1} x(k)e^{-j\omega k}$$

$$x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega k}\, d\omega.$$

C. Spectral Subtraction Estimator

The spectral subtraction filter $H(e^{j\omega})$ is calculated by replacing the noise spectrum $N(e^{j\omega})$ with spectra which can be readily measured. The magnitude $|N(e^{j\omega})|$ of $N(e^{j\omega})$ is replaced by its average value $\mu(e^{j\omega})$ taken during nonspeech activity, and the phase $\theta_N(e^{j\omega})$ of $N(e^{j\omega})$ is replaced by the phase $\theta_x(e^{j\omega})$ of $X(e^{j\omega})$. These substitutions result in the spectral subtraction estimator $\hat{S}(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = [|X(e^{j\omega})| - \mu(e^{j\omega})]\, e^{j\theta_x(e^{j\omega})}$$

or

$$\hat{S}(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$$

with

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|}$$

$$\mu(e^{j\omega}) = E\{|N(e^{j\omega})|\}.$$

D. Spectral Error

The spectral error $\epsilon(e^{j\omega})$ resulting from this estimator is given by

$$\epsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega})\, e^{j\theta_x}.$$

A number of simple modifications are available to reduce the auditory effects of this spectral error. These include: 1) magnitude averaging; 2) half-wave rectification; 3) residual noise reduction; and 4) additional signal attenuation during nonspeech activity.

E. Magnitude Averaging

Since the spectral error equals the difference between the noise spectrum N and its mean μ, local averaging of spectral magnitudes can be used to reduce the error. Replacing $|X(e^{j\omega})|$ with $\overline{|X(e^{j\omega})|}$ where

$$\overline{|X(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |X_i(e^{j\omega})|$$

$X_i(e^{j\omega}) = i$th time-windowed transform of $x(k)$

gives

$$S_A(e^{j\omega}) = [\overline{|X(e^{j\omega})|} - \mu(e^{j\omega})]\, e^{j\theta_x(e^{j\omega})}.$$

The rationale behind averaging is that the spectral error becomes approximately

$$\epsilon(e^{j\omega}) = S_A(e^{j\omega}) - S(e^{j\omega}) \cong \overline{|N|} - \mu$$

where

$$\overline{|N(e^{j\omega})|} = \frac{1}{M} \sum_{i=0}^{M-1} |N_i(e^{j\omega})|.$$

Thus, the sample mean of $|N(e^{j\omega})|$ will converge to $\mu(e^{j\omega})$ as a longer average is taken.

The obvious problem with this modification is that the speech is nonstationary, and therefore only limited time averaging is allowed. DRT results show that averaging over more than three half-overlapped windows with a total time duration of 38.4 ms will decrease intelligibility. Spectral examples and DRT scores with and without averaging are given in the "Results" section. Based upon these results, it appears that averaging coupled with half rectification offers some improvement. The major disadvantage of averaging is the risk of some temporal smearing of short transitory sounds.

F. Half-Wave Rectification

For each frequency ω where the noisy signal spectrum magnitude $|X(e^{j\omega})|$ is less than the average noise spectrum magnitude $\mu(e^{j\omega})$, the output is set to zero. This modification can be simply implemented by half-wave rectifying $H(e^{j\omega})$. The estimator then becomes

$$\hat{S}(e^{j\omega}) = H_R(e^{j\omega})X(e^{j\omega})$$

where

$$H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2}.$$

The input-output relationship between $X(e^{j\omega})$ and $\hat{S}(e^{j\omega})$ at each frequency ω is shown in Fig. 1.

Thus, the effect of half-wave rectification is to bias down the magnitude spectrum at each frequency ω by the noise bias determined at that frequency. The bias value can, of course,

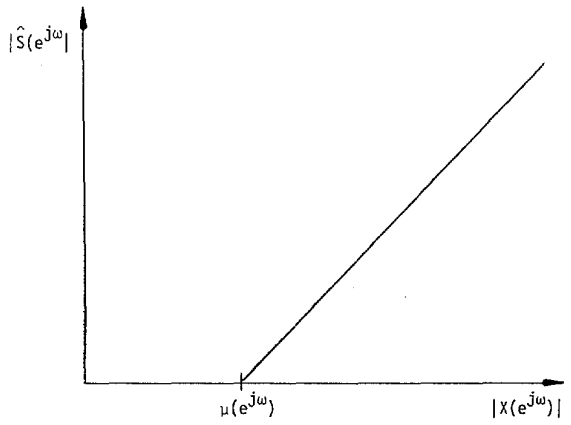

Fig. 1. Input-output relation between $X(e^{j\omega})$ and $\hat{S}(e^{j\omega})$.

change from frequency to frequency as well as from analysis time window to time window. The advantage of half rectification is that the noise floor is reduced by $\mu(e^{j\omega})$. Also, any low variance coherent noise tones are essentially eliminated. The disadvantage of half rectification can exhibit itself in the situation where the sum of the noise plus speech at a frequency ω is less than $\mu(e^{j\omega})$. Then the speech information at that frequency is incorrectly removed, implying a possible decrease in intelligibility. As discussed in the section on "Results," for the helicopter speech data base this processing did not reduce intelligibility as measured using the DRT.

G. Residual Noise Reduction

After half-wave rectification, speech plus noise lying above μ remain. In the absence of speech activity the difference $N_R = N - \mu e^{j\theta_n}$, which shall be called the noise residual, will for uncorrelated noise exhibit itself in the spectrum as randomly spaced narrow bands of magnitude spikes (see Fig. 7). This noise residual will have a magnitude between zero and a maximum value measured during nonspeech activity. Transformed back to the time domain, the noise residual will sound like the sum of tone generators with random fundamental frequencies which are turned on and off at a rate of about 20 ms. During speech activity the noise residual will also be perceived at those frequencies which are not masked by the speech.

The audible effects of the noise residual can be reduced by taking advantage of its frame-to-frame randomness. Specifically, at a given frequency bin, since the noise residual will randomly fluctuate in amplitude at each analysis frame, it can be suppressed by replacing its current value with its minimum value chosen from the adjacent analysis frames. Taking the minimum value is used only when the magnitude of $\hat{S}(e^{j\omega})$ is less than the maximum noise residual calculated during nonspeech activity. The motivation behind this replacement scheme is threefold: first, if the amplitude of $\hat{S}(e^{j\omega})$ lies below the maximum noise residual, and it varies radically from analysis frame to frame, then there is a high probability that the spectrum at that frequency is due to noise; therefore, suppress it by taking the minimum; second, if $\hat{S}(e^{j\omega})$ lies below the maximum but has a nearly constant value, there is a high probability that the spectrum at that frequency is due to low energy speech; therefore, taking the minimum will retain the information; and third, if $\hat{S}(e^{j\omega})$ is greater than the maximum, there is speech present at that frequency; therefore, removing the bias is sufficient. The amount of noise reduction using this replacement scheme was judged equivalent to that obtained by averaging over three frames. However, with this approach high energy frequency bins are not averaged together. The disadvantage to the scheme is that more storage is required to save the maximum noise residuals and the magnitude values for three adjacent frames.

The residual noise reduction scheme is implemented as

$$|\hat{S}_i(e^{j\omega})| = |\hat{S}_i(e^{j\omega})|, \quad \text{for } |\hat{S}_i(e^{j\omega})| \geq \max |N_R(e^{j\omega})|$$

$$|\hat{S}_i(e^{j\omega})| = \min \{|\hat{S}_j(e^{j\omega})| j = i-1, i, i+1\}, \quad \text{for } |\hat{S}_i(e^{j\omega})| < \max |N_R(e^{j\omega})|$$

where

$$\hat{S}_i(e^{j\omega}) = H_R(e^{j\omega}) X_i(e^{j\omega})$$

and

$$\max |N_R(e^{j\omega})| = \text{maximum value of noise residual measured during nonspeech activity.}$$

H. Additional Signal Attenuation During Nonspeech Activity

The energy content of $\hat{S}(e^{j\omega})$ relative to $\mu(e^{j\omega})$ provides an accurate indicator of the presence of speech activity within a given analysis frame. If speech activity is absent, then $\hat{S}(e^{j\omega})$ will consist of the noise residual which remains after half-wave rectification and minimum value selection. Empirically, it was determined that the average (before versus after) power ratio was down at least 12 dB. This implied a measure for detecting the absence of speech given by

$$T = 20 \log_{10} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\hat{S}(e^{j\omega})}{\mu(e^{j\omega})} \right| d\omega \right].$$

If T was less than -12 dB, the frame was classified as having no speech activity. During the absence of speech activity there are at least three options prior to resynthesis: do nothing, attenuate the output by a fixed factor, or set the output to zero. Having some signal present during nonspeech activity was judged to give the higher quality result. A possible reason for this is that noise present during speech activity is partially masked by the speech. Its perceived magnitude should be balanced by the presence of the same amount of noise during nonspeech activity. Setting the buffer to zero had the effect of amplifying the noise during speech activity. Likewise, doing nothing had the effect of amplifying the noise during nonspeech activity. A reasonable, though by no means optimum amount of attenuation was found to be -30 dB. Thus, the output spectral estimate including output attenuation during nonspeech activity is given by

$$\hat{S}(e^{j\omega}) = \begin{cases} \hat{S}(e^{j\omega}) & T \geq -12 \text{ dB} \\ cX(e^{j\omega}) & T \leq -12 \text{ dB} \end{cases}$$

where $20 \log_{10} c = -30$ dB.

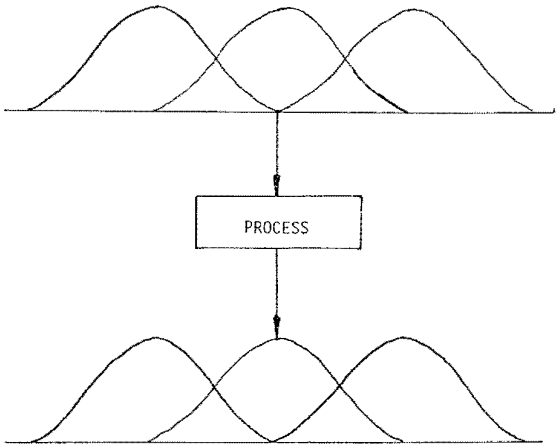

Fig. 2. Data segmentation and advance.

III. Algorithm Implementation

A. Introduction

Based on the development of the last section, a complete analysis–synthesis algorithm can be constructed. This section presents the specifications required to implement a spectral subtraction noise suppression system.

B. Input–Output Data Buffering and Windowing

Speech from the A-D converter is segmented and windowed such that in the absence of spectral modifications, if the synthesis speech segments are added together, the resulting overall system reduces to an identity. The data are segmented and windowed using the result [12] that if a sequence is separated into half-overlapped data buffers, and each buffer is multiplied by a Hanning window, then the sum of these windowed sequences adds back up to the original sequences. The window length is chosen to be approximately twice as large as the maximum expected pitch period for adequate frequency resolution [13]. For the sampling rate of 8.00 kHz a window length of 256 points shifted in steps of 128 points was used. Fig. 2 shows the data segmentation and advance.

C. Frequency Analysis

The DFT of each data window is taken and the magnitude is computed.

Since real data are being transformed, two data windows can be transformed using one FFT [14]. The FFT size is set equal to the window size of 256. Augmentation with zeros was not incorporated. As correctly noted by Allen [15], spectral modification followed by inverse transforming can distort the time waveform due to temporal aliasing caused by circular convolution with the time response of the modification. Augmenting the input time waveform with zeros before spectral modification will minimize this aliasing. Experiments with and without augmentation using the helicopter speech resulted in negligible differences, and therefore augmentation was not incorporated. Finally, since real data are analyzed, transform symmetries were taken advantage of to reduce storage requirements essentially in half [14].

D. Magnitude Averaging

As was described in the previous section, the variance of the noise spectral estimate is reduced by averaging over as many spectral magnitude sets as possible. However, the nonstationarity of the speech limits the total time interval available for local averaging. The number of averages is limited by the number of analysis windows which can be fit into the stationary speech time interval. The choice of window length and averaging interval must compromise between conflicting requirements. For acceptable spectral resolution a window length greater than twice the expected largest pitch period is required with a 256-point window being used. For minimum noise variance a large number of windows are required for averaging. Finally, for acceptable time resolution a narrow analysis interval is required. A reasonable compromise between variance reduction and time resolution appears to be three averages. This results in an effective analysis time interval of 38 ms.

E. Bias Estimation

The spectral subtraction method requires an estimate at each frequency bin of the expected value of noise magnitude spectrum μ_N:

$$\mu_N = E\{|N|\}.$$

This estimate is obtained by averaging the signal magnitude spectrum $|X|$ during nonspeech activity. Estimating μ_N in this manner places certain constraints when implementing the method. If the noise remains stationary during the subsequent speech activity, then an initial startup or calibration period of noise-only signal is required. During this period (on the order of a third of a second) an estimate of μ_N can be computed. If the noise environment is nonstationary, then a new estimate of μ_N must be calculated prior to bias removal each time the noise spectrum changes. Since the estimate is computed using the noise-only signal during nonspeech activity, a voice switch is required. When the voice switch is off, an average noise spectrum can be recomputed. If the noise magnitude spectrum is changing faster than an estimate of it can be computed, then time averaging to estimate μ_N cannot be used. Likewise, if the expected value of the noise spectrum changes after an estimate of it has been computed, then noise reduction through bias removal will be less effective or even harmful, i.e., removing speech where little noise is present.

F. Bias Removal and Half-Wave Rectification

The spectral subtraction spectral estimate $\hat{S}$ is obtained by subtracting the expected noise magnitude spectrum μ from the magnitude signal spectrum $|X|$. Thus

$$|\hat{S}(k)| = |X(k)| - \mu(k) \qquad k = 0, 1, \cdots, L-1$$

or

$$\hat{S}(k) = H(k) \cdot X(k), H(k) = 1 - \frac{\mu(k)}{|X(k)|} \qquad k = 0, 1, \cdots, L-1$$

where L = DFT buffer length.

After subtracting, the differenced values having negative magnitudes are set to zero (half-wave rectification). These

negative differences represent frequencies where the sum of speech plus local noise is less than the expected noise.

G. Residual Noise Reduction

As discussed in the previous section, the noise that remains after the mean is removed can be suppressed or even removed by selecting the minimum magnitude value from the three adjacent analysis frames in each frequency bin where the current amplitude is less than the maximum noise residual measured during nonspeech activity. This replacement procedure follows bias removal and half-wave rectification. Since the minimum is chosen from values on each side of the current time frame, the modification induces a one frame delay. The improvement in performance was judged superior to three frame averaging in that an equivalent amount of noise suppression resulted without the adverse effect of high-energy spectral smoothing. The following section presents examples of spectra with and without residual noise reduction.

H. Additional Noise Suppression During Nonspeech Activity

The final improvement in noise reduction is signal suppression during nonspeech activity. As was discussed, a balance must be maintained between the magnitude and characteristics of the noise that is perceived during speech activity and the noise that is perceived during speech absence.

An effective speech activity detector was defined using spectra generated by the spectral subtraction algorithm. This detector required the determination of a threshold signaling absence of speech activity. This threshold ($T = -12$ dB) was empirically determined to ensure that only signals definitely consisting of background noise would be attenuated.

I. Synthesis

After bias removal, rectification, residual noise removal, and nonspeech signal suppression a time waveform is reconstructed from the modified magnitude corresponding to the center window. Again, since only real data are generated, two time windows are computed simultaneously using one inverse FFT. The data windows are then overlap added to form the output speech sequence. The overall system block diagram is given in Fig. 3.

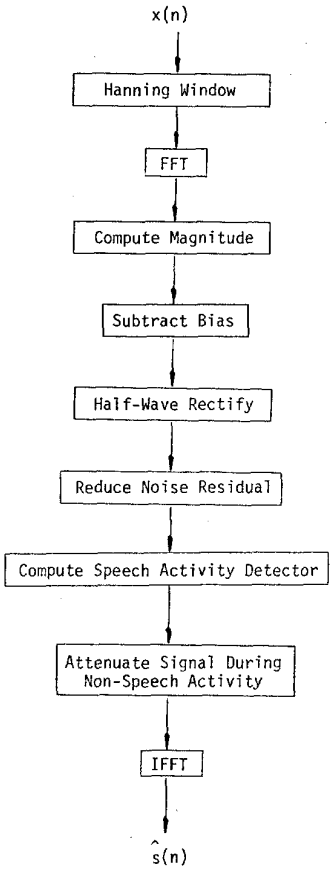

Fig. 3. System block diagram.

VI. Results

A. Introduction

Examples of the performance of spectral subtraction will be presented in two forms: isometric plots of time versus frequency magnitude spectra, with and without noise cancellation; and intelligibility and quality measurement obtained from the Diagnostic Rhyme Test (DRT) [11]. The DRT is a well-established method for evaluating speech processing devices. Testing and scoring of the DRT data base was provided by Dynastat Inc. [12]. A limited single speaker DRT test was used. The DRT data base consisted of 192 words using speaker RH recorded in a helicopter environment. A crew of 8 listeners was used.

The results are presented as follows: 1) short-time amplitude spectra of helicopter speech; 2) DRT intelligibility and quality scores on LPC vocoded speech using as input the data given in 2); and 3) short-time spectra showing additional improvements in noise rejection through residual noise suppression and nonspeech signal attenuation.

B. Short-Time Spectra of Helicopter Speech

Isometric plots of time versus frequency magnitude spectra were constructed from the data by computing and displaying magnitude spectra from 64 overlapped Hanning windows. Each line represents a 128-point frequency analysis. Time increases from bottom to top and frequency from left to right.

A 920 ms section of speech recorded with a noise-cancelling microphone in a helicopter environment is presented. The phrase "Save your" was filtered at 3.2 kHz and sampled at 6.67 kHz. Since the noise was acoustically added, no underlying clean speech signal is available. Fig. 4 shows the digitized time signal. Fig. 5 shows the average noise magnitude spectrum computed by averaging over the first 300 ms of nonspeech activity. The short-time spectrum of the noisy signal x is shown in Fig. 6. Note the high amplitude, narrow-band ridges corresponding to the fundamental (1550 Hz) and first harmonic (3100 Hz) of the helicopter engine, as well as the ramped noise floor above 1800 Hz. Fig. 7 shows the result from bias removal and rectification. Figs. 8 and 9 show the noisy spectrum and the spectral subtraction estimate using three frame averaging.

These figures indicate that considerable noise rejection has

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.