AUDIO ENHANCEMENT USING NONLINEAR TIME-FREQUENCY FILTERING

ROBERT C. MAHER

Digital Audio Signal Processing Laboratory, Department of Electrical and Computer Engineering
Montana State University, Bozeman, MT USA
rob.maher@montana.edu

Forensic audio recordings may contain undesired noise that can impair source identification, speech recognition, and other audio processing requirements. In this paper several custom analysis/synthesis algorithms are presented based on a time-varying spectral representation of the noisy signal. The enhancement process adapts to the instantaneous signal behavior and alters the noisy signal so that the enhanced output signal is higher in quality than the unprocessed noisy input signal. Nonlinear and time-varying filters operate on the spectral representation in order to retain features that are attributable to the desired signal, such as human speech, while removing the features that are more likely to be due to the noise contamination.

INTRODUCTION

When a recorded audio signal contains unwanted additive noise it is desirable somehow to enhance the perceived signal-to-noise ratio before playback [1, 2, 3, 4]. Assuming the noisy signal of interest is a digital data file, the enhancement can be performed off-line (not in real time) with a digital copy of the original without risk of damage to the evidence itself. Off-line processing also allows multiple passes through the data, the use of iterative algorithms, and the opportunity for subjective evaluation of the results.

A truly useful enhancement process needs to be *single-ended*, meaning that it must operate with no information available at the receiver other than the noise-degraded signal itself. Thus, it is necessary to devise an enhancement process that can adapt to the instantaneous signal behavior and alter the noisy signal in both the time domain and the frequency domain so that a listener or forensic examiner will rate the enhanced signal as both higher in quality and more useful than the unprocessed noisy signal.

The fundamental issue for single-ended noise reduction is that the problem is ill-posed. A desired signal, $s(t)$, is corrupted with unwanted additive noise, $n(t)$, resulting in the observed signal $x(t) = s(t) + n(t)$. With only the signal $x(t)$ available there is one equation with two unknowns, and therefore it is generally impossible to determine $s(t)$ from $x(t)$ unless some other information is available. Thus, in order to enhance the desired signal $s(t)$ while attenuating the unwanted $n(t)$, some means is needed to determine which features of the composite signal $x(t)$ are attributable to the noise and which are due to the desired signal. Once the unwanted features are somehow identified, some means is needed to reduce or remove those components from the composite signal. Finally, some adaptive control scheme is needed to adjust the detection and removal methods to compensate for the expected time varying behavior of the signal and noise.

An issue not considered in this paper is how to handle signal gaps, or "dropouts," which can occur if the signal is lost momentarily due to mechanical or electrical interruptions in the recording or playback systems. Dropouts can pose a serious problem for forensic audio enhancement. The signal enhancement process must detect the signal dropout and take some suitable action, either muting the playback momentarily or preferably synthesizing an estimate of the missing material [5].

1 BACKGROUND

Methods intended to reduce the audibility of unwanted additive noise have been proposed and studied for many years. The fundamental methods fall into two general categories: time-domain level detectors and frequency-domain filters.

1.1 Time-domain level detection

The key element of most time-domain methods is a user-specified signal level, or *threshold*, that indicates the likely presence of the desired signal. If the input signal level is currently below the threshold the input is assumed to contain only noise and a gain-controlled amplifier is used to reduce (*gate*) the output signal level. The gain reduction causes the output signal to be perceived as less noisy than the input signal. Conversely, if the input signal level is currently above

the threshold, the gain-controlled amplifier is set for unit gain and the input signal is passed through to the output. By continuously monitoring the input signal level with respect to the threshold, the output signal can be gated on and off as the input signal level varies. This sort of time-domain level detection system is variously referred to as a *squelch* control, a *dynamic range expander*, or a *noise gate*.

An example of a basic time-domain level-sensitive process is depicted in Fig. 1.

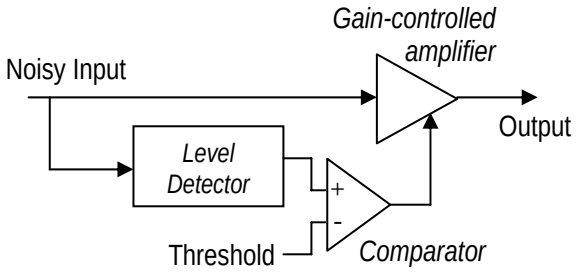

Figure 1: Time-domain level detection (noise gate).

Simple threshold gating does not remove noise when the desired signal is present: the gate is simply "open" when the threshold is exceeded. Thus, if the signal-to-noise ratio is poor even when the desired signal is present, the gate will not be of much benefit. Also, changing the gain between the "pass" mode and the "gate" mode must be done carefully to avoid audible noise modulation or gain pumping. Nevertheless, this time-domain method can be effective in many applications if the noisy input signal consists of a low-level background noise and a signal with a slowly varying amplitude envelope, such as speech.

Improvements to the time-domain noise gate can include careful control over the attack and release times of the gain-controlled amplifier, splitting the gating decision into two or more frequency bands, and devising a way to vary the gate threshold automatically as the background noise level changes with time. Noise gates are useful in many forensic processing situations.

1.2 Frequency-domain filtration

The popular frequency-domain method for signal enhancement is to use some form of *spectral subtraction*. The concept is to find an estimate of the noise spectrum (noise amplitude as a function of frequency), then subtract this noise estimate from the input signal spectrum, ideally leaving only the desired signal spectrum [6, 7, 8, 9].

To demonstrate spectral subtraction, consider the signal spectrum shown in Fig. 2. This example spectrum is *harmonic*, meaning that the energy is concentrated at a series of discrete frequencies that are integer multiples (*harmonics*) of a base frequency, or *fundamental*. In this example the fundamental is 100 Hz so the energy consists of harmonics at 100, 200, 300, … Hz. A signal with a harmonic spectrum is generally perceived to have a specific *pitch*, or musical note, to the human ear.

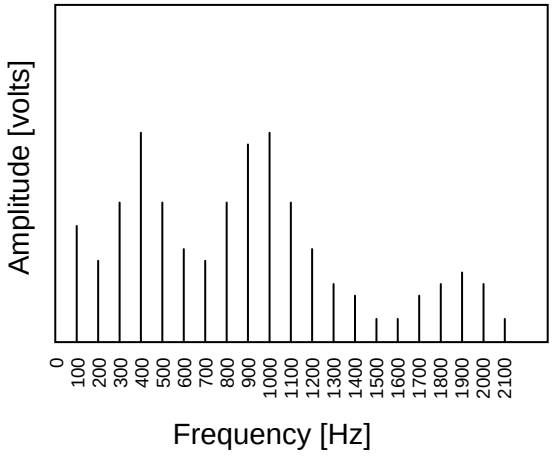

Figure 2: Example signal spectrum consisting of a 100Hz fundamental component and harmonics.

The example spectrum of Fig. 2 is intended to represent a clean, noise-free original signal that is subsequently contaminated with additive noise. An example of the noise spectrum is shown in Fig. 3. Note that unlike the discrete frequency components of the harmonic signal, the noise signal in Fig. 3 has energy distributed across the entire frequency range.

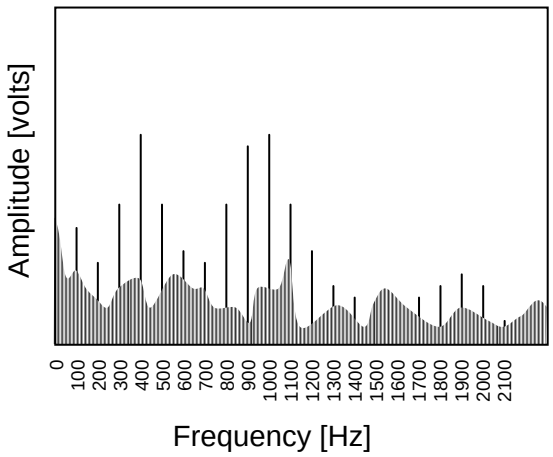

Figure 3: Example of noisy signal containing both the desired spectrum and the added noise spectrum.

A spectral subtraction system must *estimate* the noise level as a function of frequency. The noise level estimate is usually obtained during a "quiet" section of the signal, such as a pause between spoken words in a speech signal. The spectral subtraction process involves subtracting the noise level estimate, or spectral *threshold*, from the received signal so that any spectral

energy that is below the threshold is removed [7, 10]. The noise-reduced output signal is reconstructed from this subtracted spectrum. An example of the noise-reduced output spectrum for the noisy signal of Fig. 3 is shown in Fig. 4. Note that in this particular example some of the desired signal spectral components are below the noise threshold, so the spectral subtraction process has inadvertently removed them. Nevertheless, the spectral subtraction method can conceivably improve the overall signal-to-noise ratio if the noise level is not too high.

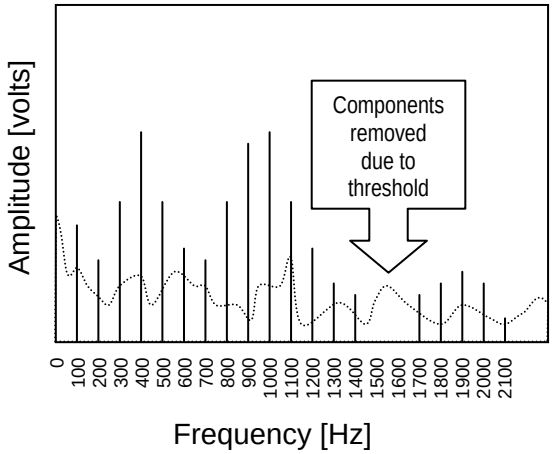

Figure 4: Example of the output spectrum after spectral subtraction.

The spectral subtraction process can cause various audible problems, especially when the actual noise level differs from the estimated noise spectrum. In this *mismatch* situation the noise is not perfectly canceled and the residual noise can take on a whistling, tinkling quality sometimes referred to as *birdie noise* or *musical noise*. Spectral subtraction also suffers from the fact that the noise spectrum will generally fluctuate rapidly from time to time and the desired signal spectrum itself will also generally change with time. If some of the desired partials are below the noise threshold at one instant in time but then peek above the noise threshold at a later instant in time, the abrupt change in those components can also result in audible birdie noise or other annoying burble or gargle sounds.

Practical spectral subtraction systems incorporate a variety of improvements, such as frequently updating the noise level estimate, switching off the subtraction in strong signal conditions, and attempting to detect and suppress the residual musical noise [11]. Thus, the quality and effectiveness of the spectral subtraction technique depends upon the particular forensic task to be accomplished and the corresponding processing requirements.

2 SPECTRAL NOISE FILTER CONCEPT

The proposed single-ended noise reduction method extends the time-domain level detection and the frequency-domain spectral subtraction concepts by providing the means to distinguish between the *coherent* behavior of the desired signal components and the *incoherent* (uncorrelated) behavior of the additive noise. The proposed technique is based on a short-time Fourier transform (STFT) analysis of the noisy input signal [12, 13, 14, 15]. The procedure identifies features that behave consistently over a short-time window and attenuates or removes features that exhibit random or inconsistent fluctuations. As a single-ended method the determination of noise vs. signal cannot be perfect, but for many important forensic signals (such as noisy speech) the process can be made sufficiently reliable to improve the output signal for subsequent analysis.

2.1 Signal Enhancement Features

The major features of the proposed signal enhancement system are as follows.

- The broadband noise reduction is implemented as a set of two-dimensional (2-D) filters in the frequency vs. time domain. Rather than treating the noisy signal in the conventional way as an amplitude variation as a function of time (one dimension), the proposed approach treats the noisy signal by observing how its spectral content evolves with time. In other words, the behavior of the signal is observed as a function of two dimensions, time *and* frequency, instead of just amplitude vs. time or amplitude vs. frequency. In this way the proposed technique shares some common features of a multi-band noise gate or a bandsplitting processor [10].
- Time-frequency 2-D filters with differing time and frequency resolutions can be used in parallel to match the processing resolution to the time-variant signal characteristics. This means that the expected variations of the desired signal, such as speech, can be retained and not unnecessarily distorted or smeared by the noise reduction processing.
- For speech signals, the method obtains an improvement in intelligibility by explicitly estimating and treating the voiced-to-silence, voiced-to-unvoiced, unvoiced-to-voiced, and silence-to-voiced transitions. Since spoken words contain a sequence of phonemes that include these characteristic transitions, correctly estimating the transitions helps avoid mistaking the desired fricative phonemes for undesired additive noise.

Thus, the proposed method includes a data-adaptive multi-dimensional (amplitude vs. frequency and time) filter structure that works to enhance spectral components that are narrow-in-frequency but relatively long-in-time (coherent), while reducing signal components that exhibit neither frequency nor temporal correlation (incoherent) and are therefore most likely to be the undesired additive noise. The effectiveness of this approach is due to its ability to pass the quasi-harmonic characteristics and the short-in-time but broad-in-frequency content of fricative sounds found in typical signals such as speech and music, as opposed to the uncorrelated time-frequency behavior of the broadband noise.

An example of the time-frequency behavior of a noisy speech signal is depicted in Fig. 5. Several notable (and typical) features are shown. The time segments with sets of parallel curves, or *tracks*, indicate the presence of voiced speech [2, 16]. The vertical spacing of the tracks varies with time as the fundamental frequency varies, but all the tracks are equally spaced, indicating that they are harmonics.

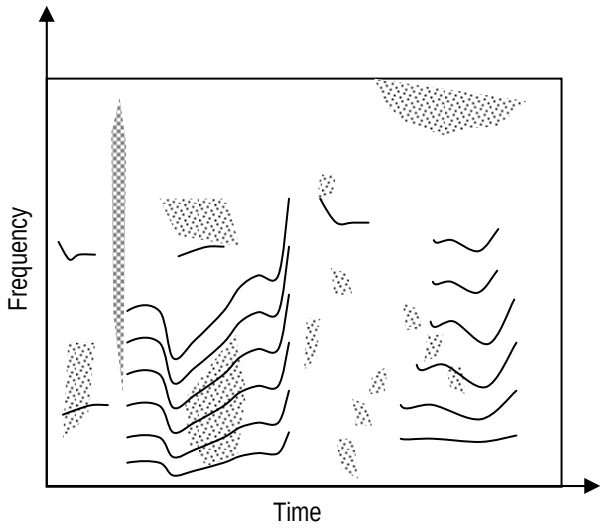

Figure 5: Noisy speech displayed as a frequency vs. time spectrogram.

The signal shown in Fig. 5 also contains many less distinct concentrations of energy that do not show the coherent behavior of the voiced speech. Some are short tracks that do not appear in harmonic groups while others are less concentrated incoherent smudges. These regions in the frequency vs. time representation of the signal are likely to be undesired noise because they appear uncorrelated in time and frequency with each other. However, there is a segment of noise that is narrow-in-time but broad-in-frequency that is also closely aligned with the start of a coherent segment. Because sequences of speech phonemes often include fricative-to-voiced transitions, it is likely that the alignment of the narrow-in-time and broad-in-frequency noise segment is actually a fricative sound from the desired speech. This identification is shown in Fig. 6.

2.2 Narrow and Broad Frequency Selectivity

This situation has led us to consider a time-frequency orientation in which two separate two-dimensional (time vs. frequency) filters are constructed. One filter is designed so that it preferentially selects spectral components that are narrow-in-frequency but relatively broad-in-time (corresponding to discrete spectral partials as found in voiced speech and other quasi-periodic signals), while the other 2-D filter is designed to pass spectral components that are broad-in-frequency but relatively narrow-in-time (corresponding to plosive and fricative consonants found in speech signals). This 2-D filter arrangement is depicted in Fig. 7.

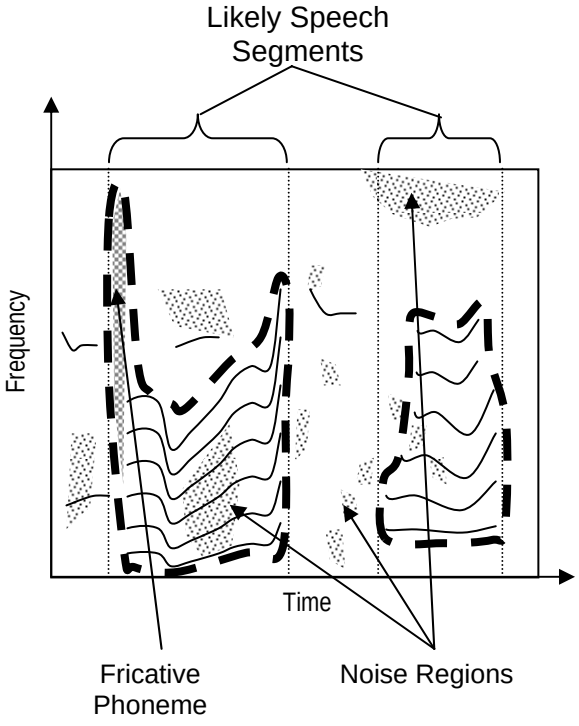

Figure 6: Noisy speech of Figure 5 with segments identified.

Each of the filters can be literally convolved over the entire time-frequency space or the processing strategy can involve a set of spectral tests and actions. In either case the filters emphasize the features in the frequency vs. time data that match the shape of the filter while attenuating the features of the signal that do not match. Note also that although the narrow and broad filters are shown as pure rectangles, the actual filters can be shaped with a smoothing window to taper and overlap the time-frequency response functions.

As mentioned above, the narrow and broad filter structures need not be implemented explicitly as 2-D digital filters: a frame-by-frame analysis and recursive testing procedure can also be used in order to minimize the computational complexity.

The relative energy distribution between the output of the narrow and broad filters can be used to determine the proportion of long-in-time and narrow-in-frequency components compared to the short-in-time and broad-in-frequency components.

2.3 Maintaining Voiced and Unvoiced Features

One problem with simply applying the 2-D filter concept is that common signals such as human speech contain noisy fricative and plosive components that are critical to speech intelligibility. To help improve the output signal intelligibility we have developed an iterative pattern detection process so that the fricative components are allowed primarily at boundaries between intervals with no voiced signal present and intervals with voiced components, since the presence and audibility of prefix and suffix consonant phonemes is a key feature for speech recognition. In other words, the behavior of the time-frequency filters includes some knowledge of the phonetic and expected fluctuations of natural speech and these elementary rules are used to aid noise reduction while enhancing the characteristics of the speech.

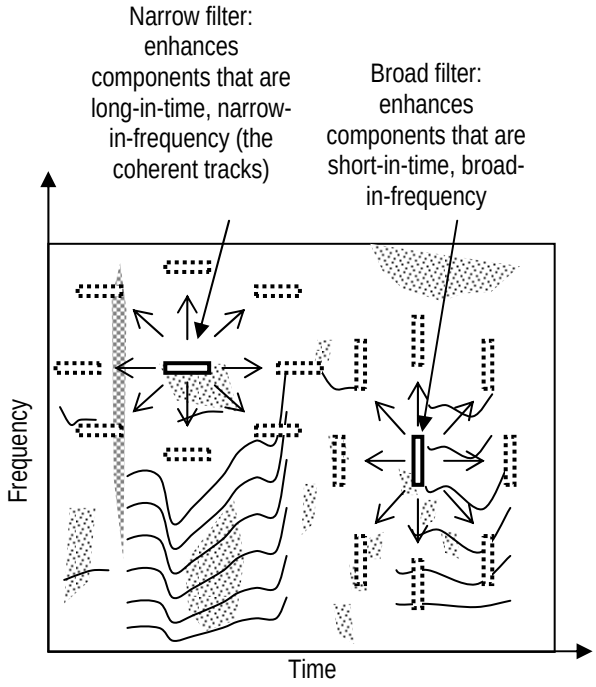

Figure 7: 2-D filtering concept.

Another useful refinement is to monitor the transient behavior of the noisy input signal since the transitions from voiced to unvoiced speech contribute to the intelligibility of the recovered speech signal. One useful technique is to switch to a shorter frame length during a transient segment, as is done in many perceptual audio coders [17].

The overall arrangement of the proposed noise reduction system is shown in Fig. 8. The high-resolution frequency filter and the high time resolution filter process the signal simultaneously. The reconstructed signal is created by mixing the filtered signals according to the detected structure of the speech signal.

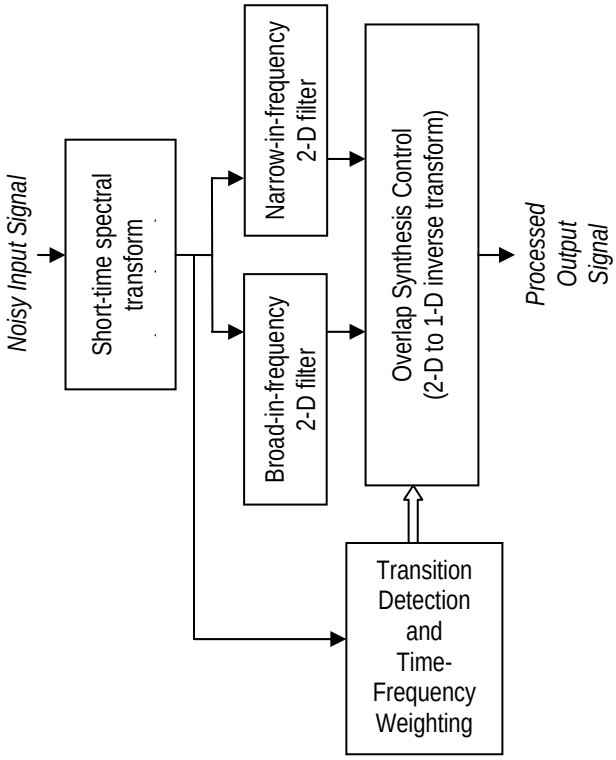

Figure 8: 2-D enhancement structure.

The *transition detection and time-frequency weighting* block shown in Fig. 8 contains the rules that determine the noise reduction strategy used to generate the output signal. For example, if the detection block identifies a strong signal from the narrow frequency filter, the current input signal segment is judged to contain a coherent signal such as voiced speech, and therefore this portion of the processed signal is emphasized in the output. If the prior signal segment contained a strong signal in the short-time filter, the detection block performs a transition to use the short-time filter output momentarily before going to the long-time filter in order to capture the prior fricative speech energy (if any) that preceded the voiced segment so that speech intelligibility is maintained.

3 SPECTRAL FILTER IMPLEMENTATION

The basic noise reduction processing is implemented as shown in Fig. 9. The noisy input signal is segmented into overlapping frames. The frame length may be fixed or variable, but without loss of generality a fixed frame length is assumed in this description. The overlap between frames is chosen so that the signal can be reconstructed by overlap-adding the blocks following the noise reduction process. A 50% or more overlap is appropriate [13]. The frame length is chosen to be sufficiently short that the signal within the frame length can be assumed to be stationary, while at the same time being sufficiently long to provide good resolution of the spectral structure of the signal. A frame length corresponding to roughly 20 milliseconds has been found to be appropriate for speech signals, since this length accommodates a 50 Hz fundamental frequency.

For each frame the noisy input data is multiplied by a suitable smoothly tapered window function (e.g., a hanning window) to avoid the Fourier truncation effects of an abrupt (rectangular) window. The windowed frame of data is zero-padded to be a power of 2 in length so that a standard radix-2 fast Fourier transform (FFT) can be used. The windowed data can be zero-padded to a longer block length if more spectral samples are desired for each windowed frame. The FFT computes the complex discrete Fourier transform of each windowed data block.

Next, the raw FFT data blocks are stored in a time-ordered first-in first-out queue. The queued sequence of FFT blocks—comprising the short-time Fourier transform (STFT)—is used in the two-dimensional narrow and broad filtering process. The magnitude of each FFT block is also computed, forming a sequence of spectral "snapshots" (a spectrogram).

The *evaluate* and *calculate* blocks in Fig. 9 process the raw FFT and FFT magnitude data from the short-time queues to determine the current composition of the input signal. In the case of speech input, the evaluation includes an estimate of whether the input signal contains voiced or unvoiced (fricative) components, whether the signal appears to be in a steady-state condition or undergoing a transition from voiced to unvoiced or from unvoiced to voiced, whether the signal shows a transition to or from a noise-only segment, and similar empirical rules that interpret the input signal conditions. For example, a steady-state voiced speech condition could be indicated by harmonics in the FFT magnitude data, while a transition from voiced to unvoiced would be flagged by a change from harmonic to non-harmonic spectral features during two or more FFT blocks.

The next step is to apply the narrow and broad 2-D filter processing. This can be accomplished either explicitly using a set of 2-D filters or implicitly by locating and retaining strong spectral features that persist for several FFT blocks while attenuating incoherent spectral features except when they occur just prior or just following a coherent section.

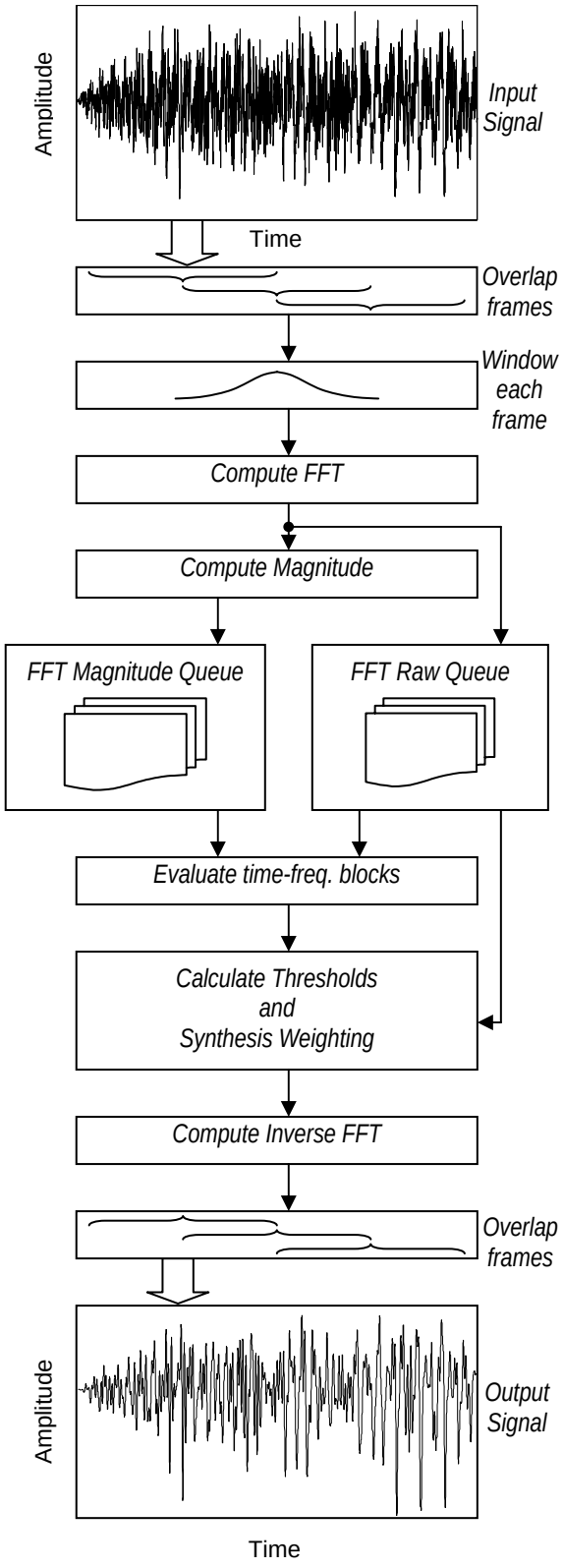

Figure 9: Overall STFT-based process.

The composite filtered output FFT data blocks are each sent through the inverse FFT and the resulting inverse FFT blocks are overlapped and added to create the noise-reduced output signal. The nature of the overlap-add resynthesis process can include a synthesis window for each block, although good results have been obtained by selecting a 50% or greater overlap in the input analysis frames.

4 PERFORMANCE EXAMPLES

The first example is a noisy AM broadcast signal containing the speech of a male talker and some background music. The noise is a broadband hiss and static that was present in the received signal.

A 10 second spectrogram of the noisy signal is shown in Fig. 10. Lighter shades correspond to higher signal levels. The harmonic partials of the voiced speech are visible as the light colored parallel tracks in the low frequency range, but the overall light gray background texture is due to the significant amount of broadband noise present in the signal. The subtle background music is visible as faint tracks in the time interval prior to 1 second and after 8.5 seconds. The speech in this example was mostly intelligible, but the broadband noise was sufficiently annoying and distracting to interfere with reliable transcription.

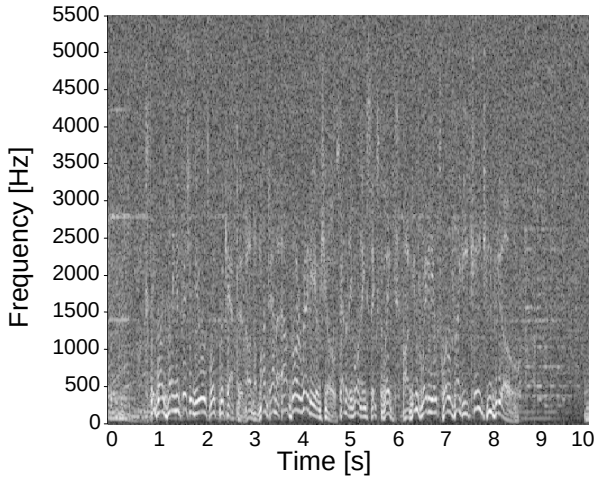

Figure 10: Example 1: Original noisy male speech and music signal.

The noisy signal of Fig. 10 was processed using the proposed spectral filtering method described in this paper. The resulting spectrogram of the output signal is shown in Fig. 11. The spectral structure of the speech and music is now enhanced, as seen by the reduced (darker) background noise level and the proportionally "higher contrast" appearance of the spectrogram. The audible quality and speech intelligibility of the output signal is noticeably improved compared to the original. The processed signal also seems more amenable to transcription or other forensic processing. However, the output signal does contain some undesirable characteristics including audible birdie noise in the high frequency range. The birdie effect can be seen as the gray speckle superimposed on the darker background in Fig. 11.

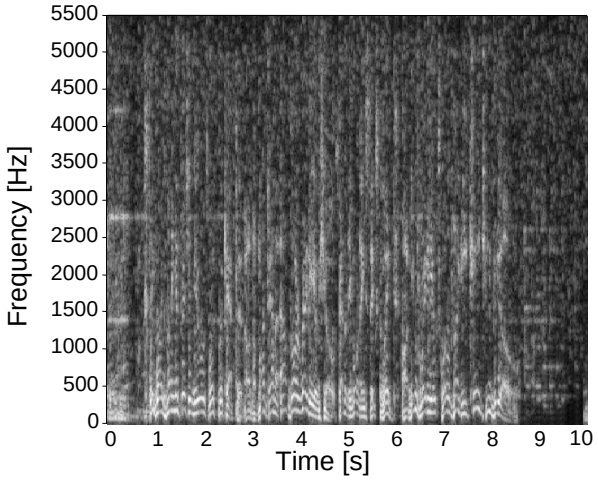

Figure 11: Output signal with reduced noise.

A second example of a noisy speech signal is shown in Fig. 12. In this example the processing strategy was adjusted to be more aggressive in removing suspected noise. The enhanced output shown in Fig. 13 demonstrates the relatively stronger presence of the speech components compared to the background noise level, and shows a greater degree of noise suppression between the words.

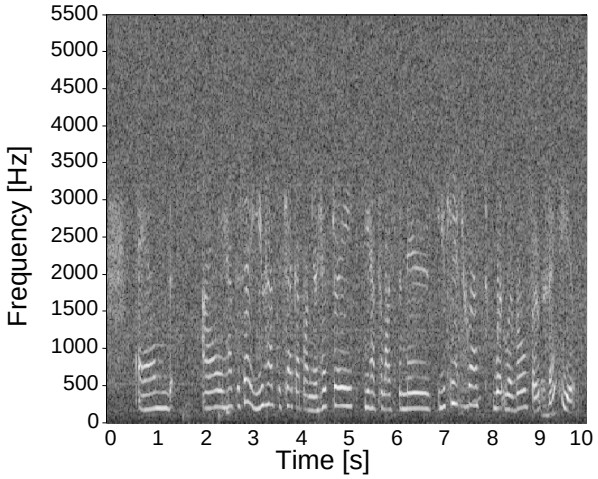

Figure 12: Example 2: Original noisy female speech.

The strong noise gate behavior can be seen as vertical black stripes in Fig. 13, corresponding to the absence of any output signal during those time intervals of the spectrogram. Note that despite the aggressive gating the leading and trailing fricatives of each word are clearly

present in the processed output, which we feel will help increase intelligibility.

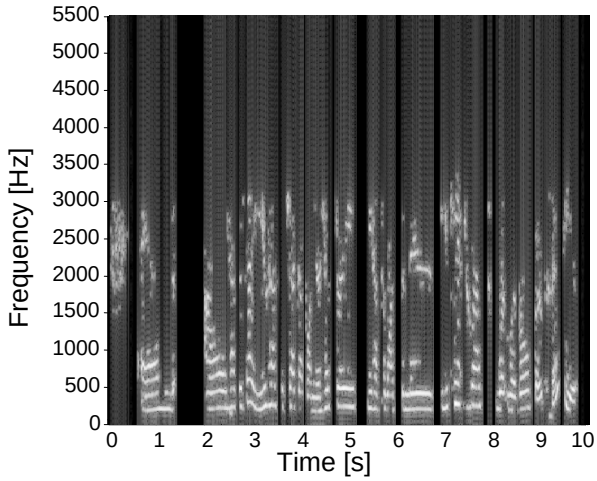

Figure 13: Output signal with reduced noise.

5 CONCLUSIONS

In many respects, the single-ended noise removal problem for audio signals has been studied *ad nauseam*, and it is reasonable to wonder if there are any remaining areas for improvement in performance. The general principles of sub-band separation and level-based gating have been proposed and used in commercial applications for many years [10, 18]. Nevertheless, we feel that the shortcomings of existing noise reduction techniques can be ameliorated to some extent by adopting the time-frequency orientation proposed in this paper. The use of two separate two-dimensional (time vs. frequency) filters, one designed so that it preferentially selects spectral components that are narrow-in-frequency but relatively broad-in-time (corresponding to voiced speech), while the other 2-D filter is designed to pass spectral components that are broad-in-frequency but relatively narrow-in-time (corresponding to plosive and fricative consonants), is a useful and productive strategy. The effectiveness of this approach is due to its ability to pass the quasi-harmonic characteristics of predictable signals (such as music and voiced speech), as opposed to the uncorrelated time-frequency behavior of the broadband noise.

The other key feature of the proposed system is an explicit procedure for treating unvoiced consonant sounds and fricatives through the use of a silence-to-unvoiced and voiced-to-unvoiced transition model. While most existing speech enhancement systems use thresholds and time constants that tend to smear the transitions at the onset or release of a spoken word, the initial or trailing consonant (fricative or plosive) is often degraded and this can lead to lower intelligibility ratings even if the quality of the noise reduced speech is judged to be good. The nonlinear and time-variant features of the proposed system are intended to expand the palette for broadband noise reduction in forensic applications.

REFERENCES

[1] M.R. Weiss, E. Aschkenasy, and T.W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Nicolet Scientific Corp., Final Rep. NSC-FR/4023 (1974).

[2] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall (1978).

[3] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604 (1979).

[4] M.R. Weiss and E. Aschkenasy, "Wideband Speech Enhancement (Addition)," Final Tech. Rep. RADC-TR-81-53, DTIC ADA100462 (1981).

[5] R.C. Maher, "A method for extrapolation of missing digital audio data," *J. Audio Eng. Soc.*, vol. 42, no. 5, pp. 350-357 (1994).

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-29, pp. 113-120 (1979).

[7] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter" *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-28, pp. 137-145 (1980).

[8] D.E. Tsoukalas, J.N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 479-514 (1997).

[9] S. Godsill, P. Rayner, and O. Cappé, "Digital Audio Restoration," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, eds., pp. 133-194, Kluwer Academic Publishers (1998).

[10] J. Moorer and M. Berger, "Linear-phase bandsplitting: theory and applications," *J. Audio Eng. Soc.*, vol. 34, no. 3, pp. 143-152 (1986).

[11] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345-349 (1994).

[12] J.L. Flanagan and R.M. Golden, "The Phase vocoder," *Bell System Tech. J.*, vol. 45, no. 8, pp. 1493-1509 (1966).

[13] J.B. Allen and L.R. Rabiner, "A unified approach to short time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564 (1977).

[14] M.R. Portnoff, "Time–frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 1, pp. 55–69 (1980).

[15] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonizing and other exotic audio modifications," *J. Audio Eng. Soc.*, vol. 47, no. 11, pp. 928–936 (1999).

[16] T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Prentice Hall (2002).

[17] K. Brandenburg, "Perceptual Coding of High Quality Digital Audio," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, eds., pp. 39-84, Kluwer Academic Publishers (1998).

[18] R.W. Carver, "Method and apparatus for reducing noise content in audio signals," United States Patent #3,989,897, November 2, 1976.