MINUET: MUSICAL INTERFERENCE UNMIXING ESTIMATION TECHNIQUE

Scott Rickard, Conor Fearon

University College Dublin, Dublin, Ireland
{scott.rickard,conor.fearon}@ee.ucd.ie

Radu Balan, Justinian Rosca

Siemens Corporate Research, Princeton, NJ, USA
{radu.balan,justinian.rosca}@scr.siemens.com

ABSTRACT

We propose a noise cancellation technique that performs robustly in the presence of poor channel estimates and channel synchronization errors. The technique is based on the assumption that the signals have a sparse representation in a chosen signal basis, in this case, the time-frequency domain. Moreover, we assume the components of the signal of interest that contain a majority of its power overlap with components of the interference signal containing negligible power. In case of speech mixed with music, this occurs because in the time-frequency domain both music and speech are sparse and the large magnitude coefficients rarely overlap. The robustness of the technique to channel estimation and synchronization errors is demonstrated experimentally on speech/music mixtures.

1. INTRODUCTION

The problem of cancelling an unwanted interference signal from a single mixture given an unfiltered version of the interference signal has been well studied. Many of these techniques, however, rely on precise channel estimates in order to cancel the interference. The classical noise cancellation techniques often fail to remove the interference, or even add more interference into the mixture, when phase errors in the channel estimates occur. When the channel changes suddenly, or when the reference interference signal and mixture interference signal lack synchronicity, the performance of the standard noise cancelling techniques suffers. Motivated by recent advancements in field of blind source separation, we propose here a noise cancellation technique which performs robustly in the presence of poor channel estimates and channel synchronization errors.

The DUET (Degenerate Unmixing Estimation Technique) algorithm presented in [1] and [2] and analyzed further in, for example, [3], [4], and [5], is a method for blind source separation in the degenerate case, that is when the number of sources is greater than the number of mixtures. In this situation, inversion of the mixing matrix is impossible and thus prevents demixing via mixing matrix inversion. The DUET method relies on the concept of approximate W-disjoint orthogonality which quantifies the non-overlapping nature of the time-frequency representations of the sources. This property is exploited to facilitate the separation of any number of sources from just two mixtures using the spatial signatures of each source.

Separation in the monaural case is considerably more difficult than in the binaural case given that the spatial cues arising from microphone separation, on which the latter relies, are notably absent with only one mixture. As seen in [6], [7] and [8], prior information on the nature of the sources is necessary to overcome this challenge.

Alternatively, in the monaural case, often side information is available to aid in separation. Such algorithms are known as adaptive noise cancelling techniques, and the algorithm analyzed in this paper falls into this class. Specifically, we consider the case where we have a single mixture, $x(t)$, consisting of a speech source of interest, $s(t)$, and an interfering musical signal, $n(t)$, where both $s(t)$ and $n(t)$ incorporate the impulse responses resulting from their respective transmission paths in the environment.

$$x(t) = s(t) + n(t) \tag{1}$$

Using a reference signal, $n'(t)$, which is the interference signal which has *not* been passed through an unknown time-varying filter, we want to recover the signal of interest from the mixture.

The established methods of noise cancelling in this situation involve adaptive filtering and are essentially variations on the scheme introduced in [9] and depicted in Figure 1.

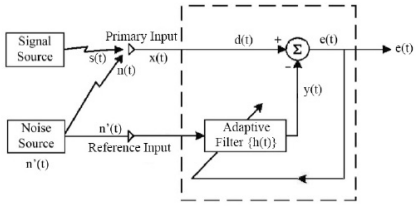

Fig. 1. Adaptive Noise Cancelling. The reference signal is filtered and subtracted from the mixture to produce an error signal that is used to control the filtering process

Adaptive noise cancelling requires very little a priori knowledge of the characteristics of either source since the adaptive filter uses an adaptive process to adjust its own coefficients. It does this such that the filtered reference (i.e. the output of the filter, $y(t)$ resembles, as closely as is possible, the interfering signal in the primary input. The reference signal is related to the interference in the primary input, $n(t)$, by a convolution with the impulse response of the environment, which we will denote $h_n(t)$,

$$n(t) = h_n(t) \star n'(t). \tag{2}$$

Thus, it is $h_n(t)$ which the adaptive filter must learn in order that the subtraction will remove as much of the interference as possible. It achieves this by using an adaptive algorithm which works to minimize in some way the error signal

$$e(t) = s(t) + n(t) - y(t) = s(t) + n(t) - h(t) \star n'(t) \tag{3}$$

by making adjustments to the adaptive filter $h(t)$. There exist a wide variety of recursive algorithms for adaptive filtering [10] and we compare the technique described later in this paper to two of the most established algorithms; normalised Least-Mean-Square (NLMS) variant and Recursive-Least-Squares (RLS).

This paper analyzes MINUET (Musical Interference Unmixing Estimation Technique), an adaptive noise cancelling algorithm introduced in [11], which utilizes similar principles to those of DUET and eliminates the interference using a binary time-frequency mask instead of the classic approach of Figure 1. The technique was originally developed to remove a musical interference signal from a voice/music mixture, but all that MINUET requires is that the signal of interest and interference are approximately W-disjoint orthogonal, a concept quantified in [2] and briefly discussed in the next section. The method consists of three steps. First, the mixture and the side information signal are roughly aligned so that sounds in each occur approximately at the same time. Second, an estimate of the relationship (spectral weights) between the instantaneous spectral powers of the side information signal and its presence in the mixture is calculated, for example, using a section of the mixture which contains little to no contribution from the desired signal but a relatively large contribution of the interfering signal. Third, a time-frequency mask is created comparing the weighted instantaneous spectral powers of the side information signal to the mixture instantaneous spectral powers. Time-frequency points which are likely dominated by the interfering signal are suppressed to remove the interfering source from the mixture.

The rest of the paper is organized as follows. In Section 2, we introduce MINUET and demonstrate how it can be used to solve the noise cancellation problem. Section 3 presents results of simple experiments to demonstrate MINUET's robustness with respect to phase errors. Finally, Section 4 contains conclusions and suggestions for further work.

2. MINUET

We can express the mixing in the time-frequency domain using the windowed Fourier transform. The windowed Fourier transform of x is defined,

$$F^W(x(\cdot))(\tau,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t-\tau)x(t)e^{-i\omega\tau}dt, \quad (4)$$

which we will refer to as $x(\tau,\omega)$. The mixture in the time-frequency domain is expressed,

$$x(\tau,\omega) = s(\tau,\omega) + n(\tau,\omega). \quad (5)$$

We will assume the filter process can be modelled as

$$n(\tau,\omega) = H_n(\omega)n'(\tau,\omega), \quad (6)$$

where $H_n(\omega)$ is the Fourier transform of $h_n(t)$. Mixing then becomes

$$x(\tau,\omega) = s(\tau,\omega) + H_n(\omega)n'(\tau,\omega) \quad (7)$$

Our goal is to create a time-frequency mask, $m(\tau,\omega)$, such that the mask preserves most of the desired source power,

$$\|m(\tau,\omega)s(\tau,\omega)\|^2 / \|s(\tau,\omega)\|^2 \approx 1, \quad (8)$$

and results in a high output signal to interference ratio,

$$\|m(\tau,\omega)s(\tau,\omega)\|^2 \gg \|m(\tau,\omega)n(\tau,\omega)\|^2. \quad (9)$$

Approximate W-disjoint orthogonality is embodied by Equations (8) and (9). That is, if a time-frequency mask exists such that it captures a large percentage of the power of the signal of interest without capturing a large percentage of power of the interference, then the signal of interest and interference are approximately W-disjoint orthogonal. For such a mask, converting $m(\tau,\omega)x(\tau,\omega)$ back into the time domain will be the interference free signal. Thus, our goal of estimating $s(t)$ can be achieved by determining an appropriate time-frequency mask $m(\tau,\omega)$.

MINUET uses a binary time-frequency mask of the form

$$M_\alpha(\tau,\omega) = \begin{cases} 1 & |x(\tau,\omega)| \geq \alpha|H(\omega)||n(\tau,\omega)| \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $H(\omega)$ is an estimate of the interference channel transfer function and α is parameter set to maximize intelligibility. Strict W-disjoint orthogonality, as defined in [1], allows no more than one source to have non-zero energy at every point in time-frequency space. Since assumptions based on this rigid definition are violated for speech, [2] introduces a measure of "approximate" W-disjoint orthogonality which, for the purposes of MINUET implies that the energy of one source dominates each time-frequency point. If we find one of these points in the representation of the mixture where the amplitude is α times greater than the amplitude of the corresponding point in the representation of the reference times the corresponding magnitude of the transfer function estimate, it is reasonable to assume that this energy must come from some source other than the interfering source, i.e. it must come from the speech signal. M_α, therefore, is turned on for all time-frequency points dominated by the speech signal. In this way, M_α is a binary mask which can be applied to the mixture in order to recover an estimate of the speech signal,

$$\hat{s}(\tau,\omega) = M_\alpha(\tau,\omega)x(\tau,\omega) \quad (11)$$

Converting $\hat{s}(\tau,\omega)$, back into the time domain produces the estimate of the signal of interest.

One way in order to estimate $H_n(\omega)$, is to locate regions of $x(\tau,\omega)$ which are dominated by $n(\tau,\omega)$. That is, we wish to find a set of (τ,ω), $\mathcal{S}$, such that $x(\tau,\omega) \approx n(\tau,\omega)$ for $(\tau,\omega) \in \mathcal{S}$. We then estimate $|H_n(\omega)|$ via,

$$H(\omega) = \frac{\int_{(\tau,\omega)\in\mathcal{S}} \left| x(\tau,\omega)\overline{n'(\tau,\omega)} \right| d\tau}{\int_{(\tau,\omega)\in\mathcal{S}} |n'(\tau,\omega)|^2 d\tau} \quad (12)$$

Clearly, $H(\omega)$ will be real-valued in this case, which is fine as we only require its magnitude in the mask generation Equation (10). One possible choice for $\mathcal{S}$ is the set of time-frequency points where $M_\alpha(\tau,\omega)$ is zero, as these are the points where the noise is likely to have dominated the speech. We can imagine an iterative batch technique where the mask estimation is performed for some initial guess for $H(\omega)$, and then the channel estimation and mask estimation are fed back into each other a number of times, each time updating the estimate of $\mathcal{S}$. Alternatively, an online version would update the current estimate for $H(\omega)$ based on $\mathcal{S}$ generated from $M_\alpha(\tau,\omega)$ up to the current time.

The MINUET technique differs from classical adaptive noise cancelling techniques which are sensitive to errors in the phase estimates of the filter and interfering signal and the synchronization of the side signal to the mixture. The proposed technique does not estimate the phase but is based on instantaneous time-frequency magnitude estimates resulting in the technique being more robust

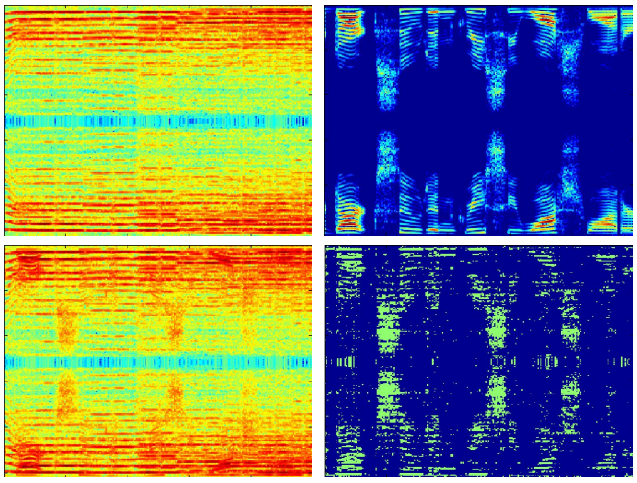

Fig. 2. Time-frequency representation of the reference signal (upper left), the original speech signal (upper right), the mixture of speech and reference (lower left), and the binary mask for $\alpha = 2$. The binary mask captures 81.1% of the energy of the speech, while improving the SNR of the mixture by 20.7 dB (from -7.8 dB to 12.9 dB).

to alignment errors. If we can ensure that the mixture and the reference are roughly aligned by some method so that sounds in each occur at approximately the same time, MINUET is robust enough to alleviate the need for perfect synchronization that is crucial for successful operation of the adaptive filtering methods.

This filtering scheme can be viewed as a thresholded form of the time-frequency formulation of the time-varying Wiener filter as discussed in [12], [13] and [14] which is an optimal filter designed to adapt to spectral change in nonstationary signals. This scheme can also be thought of as an adaptive hard thresholding scheme, where the threshold is $\alpha|H(\omega)n(\tau,\omega)|$.

Although the presentation here was done for continuous time signals, the application would be for sampled signals. In discrete time, the windowed Fourier transform would be a windowed DFT (discrete time Fourier transform) and the estimates of $H_n(\omega)$ would be finite sums over discrete time points for each frequency center.

For illustration purposes, let us assume that $H_n(\omega) = 1$, $\forall\omega$, and look to a sample speech/music mixture. Images of the time-frequency domain representations of a music signal, a speech signal, their mixture, and $M_\alpha(\tau,\omega)$ with $\alpha = 2$ are shown in Figure 2. The similarities between the music and the mixture are clearly evident as the speech signal present in the mixture is of a very low amplitude. The fact that $M_\alpha(\tau,\omega)$ matches so well with the speech shows the approximate W-disjoint orthogonality of the signals.

3. EXPERIMENTS

We present here some simple experiments which demonstrate the robustness of MINUET to synchronization errors. One issue unresolved at this point is the selection of a value for the parameter α. For the purposes of the experiments in the following section, a value of 2 has been used for this parameter throughout.

For the first experiment, we imagine a system where the reference signal has lost synchronicity with the mixture and test time-frequency masking versus subtraction-based noise cancelling methods. The music in the mixture is simply a perfect copy of the music delayed by a certain number of samples. We fix the filter used by both MINUET and in the conventional noise canceller to be the unit impulse response with no delay and do not allow either algorithm to adapt. Thus, this test evaluates the robustness of the removal step of the algorithms to phase errors in the channel estimates. All signals were sampled at 16kHz and normalised to unit energy. Each datapoint in the graphs depicting the results represents the average of 100 tests corresponding to mixtures created from speech signals taken from the TIMIT database mixed with classical or pop music. Figure 3 displays the results in SNR improvement for both MINUET and the subtraction based schemes as a function of synchronization error.

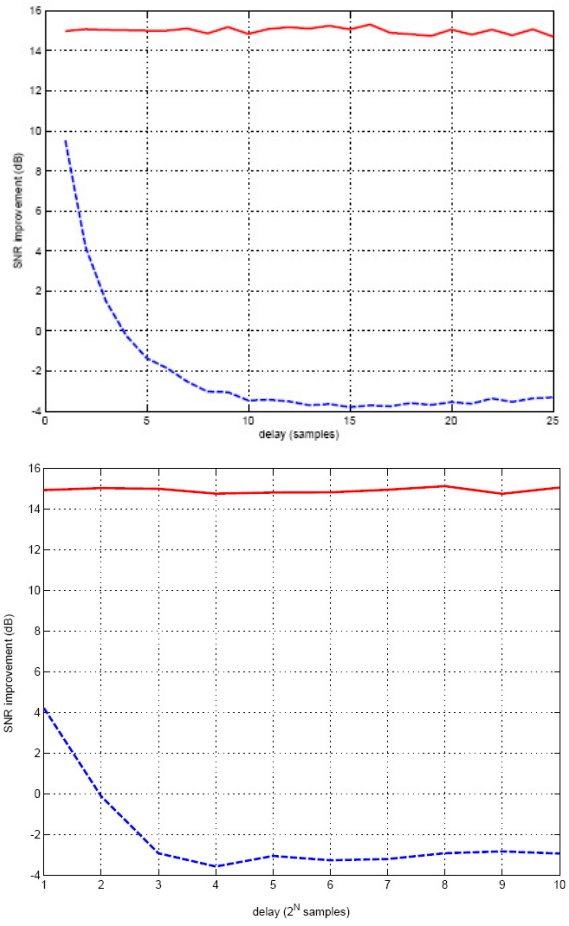

Fig. 3. Algorithm robustness alignment errors. SNR improvement (dB) for MINUET (solid) and subtraction based noise cancellers (dashed) as a function of synchronization error sample shift $\{1, 2, \ldots, 25\}$ (upper plot) and $\{2^1, 2^2, \ldots, 2^{10}\}$ (lower plot).

It can be seen from Figure 3 that even when the reference is shifted by just one sample, the SNR improvement for the subtraction method falls dramatically below that for time-frequency

masking. Moreover, after about 10 samples, the subtraction method hits a noise floor of approximately -3dB which confirms that at this level of misalignment, subtraction effectively doubles the noise power present in the mixture. Meanwhile, the graphs clearly demonstrate MINUET's robustness to synchronisation errors in this environment, with a constant SNR improvement of 15dB even with a relatively large shift in the reference.

While SNR improvement is a standard performance measure, it is not well correlated with speech quality. [2] presents an alternative measure, one of approximate W-disjoint orthogonality, which is correlated with the perceived quality of speech. It can be defined via two other important performance criteria: the preserved-signal ratio (PSR) and the signal-to-noise ratio (SNR). The PSR is, for our purposes, the portion of energy of the speech signal preserved after noise cancellation. Clearly, PSR = 1 for the subtraction method since none of the speech signal is removed, only the interference. For MINUET, on the other hand, we have:

$$\text{PSR} := \frac{\|M_\alpha(\tau,\omega)s(\tau,\omega)\|^2}{\|s(\tau,\omega)\|^2} \tag{13}$$

SNR is defined in the usual way for the subtraction method while MINUET's SNR measure is:

$$\text{SNR} := \frac{\|M_\alpha(\tau,\omega)s(\tau,\omega)\|^2}{\|M_\alpha(\tau,\omega)n(\tau,\omega)\|^2} \tag{14}$$

We can now combine Equations (13) and (14) into the measure of approximate W-disjoint orthogonality as follows:

$$\text{WDO} := \text{PSR} - \frac{\text{PSR}}{\text{SNR}} \tag{15}$$

The results for the same tests as in Figure 3 are displayed in Figure 4 as WDO versus synchronisation error. Again, as expected, we note the rapid fall in performance of the subtraction based schemes even for small synchronization errors while MINUET is not even affected by large synchronization errors.

Next, we test the performance of NLMS and RLS along with that of MINUET in an environment with synchronization jitter. We model reference signal jitter by shifting the reference every N samples by just one sample. In these tests we allow all algorithms to adapt their channel estimate and measure their performance in response to reference jitter as outlined above, setting the value of N equal to 100. NLMS is used in our experiments as it adapts to non-stationarity in far fewer iterations to a result comparable with that of the regular LMS algorithm employed in [9]. RLS offers even faster convergence than NLMS along with smaller misadjustment. For both NLMS and RLS, we use MATLAB implementations from the MATLAB Filter Design Toolbox with 13 taps in each filter. For both adaptive filtering algorithms, given the value N, empirical optimum values were obtained for the step size, μ, of NLMS and the forgetting factor, λ, in RLS. These values were 0.64 and 1 respectively. For a full discussion of the adaptive filtering algorithms see [10]. Three experimental setups were used: (a) the reference signal was unfiltered (unity channel) and every 100 samples the reference signal was shifted forward or backward one sample with equal probability (b) the interference was first passed through a random 13-tap FIR filter and every 100 samples the reference signal was shifted forward one sample (c) the interference was first passed through a random 13-tap FIR filter and and every 100 samples the reference signal was shifted forward or backward one sample with equal probability. The results of the experiments for each of the algorithms are tabulated in Table 1, given in both SNR improvement and WDO.

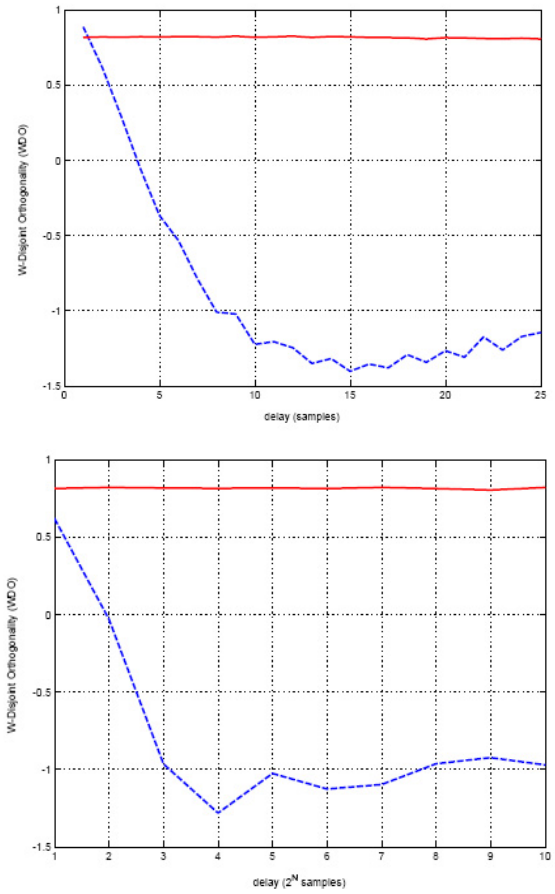

Fig. 4. Algorithm robustness alignment errors. WDO for MINUET (solid) and subtraction based noise cancellers (dashed) as a function of synchronization error sample shift $\{1, 2, \ldots, 25\}$ (upper plot) and $\{2^1, 2^2, \ldots, 2^{10}\}$ (lower plot).

Algorithm	SNR (dB)	WDO
NLMS	-0.76	-0.09
RLS	7.94	0.84
MINUET	14.34	0.73

(a) forward/backward jitter test for unity channel

Algorithm	SNR (dB)	WDO
NLMS	-0.84	-0.21
RLS	10.10	0.90
MINUET	19.27	0.55

(b) forward-only jitter test for random 13 tap channel

Algorithm	SNR (dB)	WDO
NLMS	-2.19	-0.66
RLS	-0.73	-0.18
MINUET	6.71	0.44

(c) forward/backward jitter test for random 13 taps channel

Table 1. Results of jitter tests averaged over 200 mixtures.

4. CONCLUSIONS

A method for eliminating an unwanted signal from a mixture via time-frequency masking was analyzed. Given a mixture of a signal of interest and unwanted interference, our goal was to eliminate the interfering signal to obtain an estimate of the desired signal. The signal of interest could be speech and the interference music, and the goal would be to eliminate the music from the mixture. The method requires side information, namely, it requires a signal with related instantaneous spectral powers to the unwanted signal. Such a signal is often available. For example in the scenario where the unwanted signal is a music signal which was played from a CD or tape, the original recording can serve as the side information signal. In the presence of synchronization errors between the side signal and the mixture, the performance of subtraction-based noise cancellation methods, such as NLMS and RLS, falls quickly as the misalignment grows. Such misalignment could be caused, for example, by varying playback speed of the reference or mixture recording. The performance of the time-frequency based masking technique presented here did not decrease when the side signal was misaligned to the mixture and such a technique may be well suited for applications where there is jitter in the alignment between the side signal and the mixture.

5. REFERENCES

[1] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *ICASSP*, volume 5, pages 2985–2988, 2000.

[2] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 2004. To appear.

[3] M. Baeck and U. Zolzer. Performance analysis of a source separation algorithm. In *Int. Conference on Digital Audio Effects*, September 2002.

[4] H. Viste and G. Evangelista. On the use of spatial cues to improve binaural source separation. Proceedings of the 6th Int. Conference on Digital Audio Effects, London, UK, 2003.

[5] N. Roman, D. Wang, and G. Brown. A classification-based cocktail-party processor. *Neural Information Processing Systems (NIPS*04)*, 2004.

[6] G. Cauwenberghs. Monaural separation of independent acoustical components. *Neural Information Processing Systems (NIPS*00)*, 2000.

[7] S. Roweis. One microphone source separation. *Neural Information Processing Systems (NIPS'00)*, pages 793–799, 2000.

[8] G.-J. Jang and T.-W. Lee. A probabilistic approach to single channel blind signal separation. *Neural Information Processing Systems (NIPS*02)*, 2002.

[9] B. Widrow, J. Glover, J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Ziedler, E. Dong, and R. Goodlin. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63:1692–1716, 1975.

[10] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, London, 1996.

[11] R. Balan, S. Rickard, and J. Rosca. Method for eliminating an unwanted signal from a mixture via time-frequency masking. Siemens Corporate Research Report, August 2002.

[12] F. Hlawatsch, G. Matz, H. Kirchauer, and W. Kozek. Time-frequency formulation, design and implementation of time-varying optimal filters for signal estimation. *IEEE Transactions on Signal Processing*, 48:1417–1432, 2000.

[13] T. Quatieri and R. Baxter. Noise reduction based on spectral change. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[14] T. Quatieri and R. Dunn. Speech enhancement based on auditory spectral change. In *ICASSP*, April 2002.