2017

APHONIC: Adaptive Thresholding for Noise Cancellation in Smart Mobile Environments

Ruairí de Fréin
Technological University Dublin, ruairi.defrein@tudublin.ie

APHONIC: Adaptive thresholding for noise cancellation in smart mobile environments

Ruairí de Fréin†

†Dublin Institute of Technology,
Institiúid Teicneolaíochta Bhaile Átha Cliath,
Ireland
web: https://robustandscalable.wordpress.com

in: 2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (2017). See also BibTeX entry below.

BibTeX:

```
@article{rdefrein16Aphonic,
author={Ruair\'{i} de Fr\'{e}in$^\dagger$},
journal={2017 IEEE 13th International Conference on Wireless and
Mobile Computing, Networking and Communications (WiMob) (2017)},
title={APHONIC: Adaptive thresholding for noise cancellation in smart mobile environments},
year={2017},
volume={},
number={},
pages={285--292},
month={Oct.},
keywords={Interference, Mobile communication, Noise cancellation,
Transforms, Time-frequency analysis, Mobile handsets, Urban areas},
ISBN={978-1-5386-3840-8},
doi={http://doi.ieeecomputersociety.org/10.1109/WiMOB.2017.8115847},
url={http://ieeexplore.ieee.org/document/8115847/}
}
```


document created on: February 27, 2018
created from file: rdefreinSEUNet.tex
cover page automatically created with CoverPage.sty
(available at your favourite CTAN mirror)

APHONIC: Adaptive Thresholding for Noise Cancellation in Smart Mobile Environments

Ruairí de Fréin
School of Electrical and Electronic Engineering
Dublin Institute of Technology
Ireland
Email: rdefrein@gmail.com

***Abstract*—We propose a signal-channel, adaptive threshold selection technique for binary mask construction, namely APHONIC, (AdaPtive tHreshOlding for NoIse Cancellation) for smart mobile environments. Using this mask, we introduce two noise cancellation techniques that perform robustly in the presence of real-world interfering signals that are typically encountered by mobile users: a violin busker, a subway and busy city square sounds. We demonstrate that when the power of the time-frequency components of the voice of a mobile user does not significantly overlap with the components of the interference signal, the threshold learning and noise cancellation techniques significantly improve the Signal-to-Interference Ratio (SIR) and the Signal-Distortion Ratio (SDR) of the recovered voice. When a mobile user's speech is mixed with music or with the sounds of a city square, or subway station, the speech energy is captured by a few large magnitude coefficients and APHONIC improves the SIR by greater than 20dB and the SDR by up to 5dB. The robustness of the threshold selection step and the noise cancellation algorithms is evaluated using environments typically experienced by mobile phone users. Listening tests indicate that the interference signal is no longer audible in the denoised signals. We outline how this approach could be used in many mobile voice-driven applications.**

***Index Terms*—Mobile Voice-driven Applications; Mobile Computing; Noise Cancellation; Human Computer Interaction; Blind Source Separation.**

I. Introduction

In 2016 20% of Google Android searches were made by voice. Voice has an increasingly important role to play in mobile search [1] and in related application domains; we refer to the success of Amazon Echo, Siri etc. Voice assistants are an increasingly viable interaction medium; however, voice, unlike more traditional interaction methods, triggers strong emotions because users have a lower tolerance for error using voice interfaces. Users do not like to repeat themselves and the blame for failing to process a command typically lies with the system. In the mobile world, the environment in which voice is used can have a strong effect on the success of the system. The effects of community noise on communications among other human activities are outlined by the World Health Organization in [2]. It is also true that in many cases relatively clean recordings of potential inteferer signals are available from auxiliary sensors.

We consider the problem of monaural noise cancellation, where prior information is available, in particular, a recording of ambient conditions. We observe a mixture of the desired speech signal and ambient noise, on a mobile device. Prior information could be made available on a peer-to-peer basis, similar in spirit to the manner in which resources are shared by intelligent messaging frameworks [3], or by leveraging measurements from acoustic noise pollution monitoring systems in urban environments [4]. For example, in a Mobile Cloud Computing (MCC) scenario, energy-efficient stochastic leader-selection algorithms may be used by mobile handsets to determine which sensor should supply the prior information [5]. In-network compression techniques can then be used to reduce the bandwidth usage associated with sharing this prior information (cf. [6], and related works [7]). Leveraging information from auxiliary sensors to improve human-machine interaction has been studied by the source localization and source separation communities; one focus has been on using time-frequency spatial signatures [8]. Recent works have considered collaborative localization and discrimination of acoustic sources in urban environment monitoring [9]. In this paper, we rely on the corrupted observation observed on a mobile device, along with a single unfiltered version of the interfering source to perform cancellation, as opposed to a microphone array. Traditionally recursive algorithms for adaptive filtering [10] have been applied to this task, for example, normalized Least-Mean-Square (NLMS) variant and Recursive-Least-Squares (RLS); however the authors of [11] illustrated that these approaches do not perform well when there are synchronization errors.

To ensure that APHONIC is computationally cheap, source separation methods that make the windowed Disjoint Orthogonality assumption, a property quantified for speech in [12], are examined and extended [11]. One challenge for many of these approaches is that a TF mask must be constructed to aid separation. In the stereo case [12], information about the relative placement of the microphones can be used to determine source support sets. This placement information can be built into the linear transform used: the synchronized Short-Time Fourier Transform queries the candidature of different TF transforms for generating sparse or WDO representations of multi-channel anechoic mixtures [13] by optimizing the transform for the sensor arrangement. However, in the monaural case [11] this placement information is not available. How to adaptively threshold a natural acoustic scene in order to separate out different signal components remains an open

problem. There is a need for an algorithm that can adaptively determine which source is present in which TF bin in the monaural case. Our primary contribution is that we present a solution to this challenge. We then demonstrate how this information can be used to inform two new monaural noise cancellation techniques, which target urban environments.

This paper is organized as follows. In Section II we present a model for supervised single channel demixing, which is suited to the urban noise cancellation challenge. In Section III we formulate an approach for estimating a demixing filter, given the support set of an interference signal. In Section IV we investigate the role of phase (or elevation) a criteria for estimating the support set of the interference. We construct a histogram of phase values, called the Elevatogram. We propose an algorithm for determining the support of the interference. In section V we provide a numerical evaluation of the approach and provide recommendations for our next steps.

II. Mixing Model and Problem Definition

The windowed Fourier transform of the continuous time domain signal, $x(t)$, is defined as

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

We use a Hamming window, $w(t)$, in this paper.

We observe a mixture signal on a mobile, which is defined as

$$x(\omega,\tau) = s(\omega,\tau) + n(\omega,\tau), \quad (2)$$

where $s(\omega,\tau)$ is the target signal that we want to extract and $n(\omega,\tau)$ is an interference signal that we want to remove. We do not have a clean version of the signal, $n(\omega,\tau)$, in order to aid us in this task; instead we have a reference version of this signal, $\hat{n}(\omega,\tau)$, which has been filtered by some process. For example, the filter $h(t)$ is the impulse response of the recording set-up. It could be that $h(t)$ scales or delays the signal. For example, an auxiliary sensor, closer to the source of the interference could record a clean version of the interference, which is scaled and delayed relative to the version on the mobile phone.

Given that we want to make use of the reference signal $\hat{n}(\omega,\tau)$ for denoising, in the time-frequency domain, a simple model for the interference signal is

$$n(\omega,\tau) = H(\omega)\hat{n}(\omega,\tau). \quad (3)$$

Putting this together, the observed mixture is approximated by

$$x(\omega,\tau) \approx s(\omega,\tau) + H(\omega)\hat{n}(\omega,\tau). \quad (4)$$

In this paper we will investigate the approach of time-frequency masking to remove the interference signal. A time-frequency mask in this setting has the form

$$M_\alpha(\omega,\tau) = \begin{cases} 1, & |x(\omega,\tau)| \gg f(|\hat{n}(\omega,\tau|) \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

It is defined as a function of the reference interference as opposed to the clean interference, which is not observed. Once a suitable mask has been selected we can then estimate the target source, $\tilde{s}(\omega,\tau)$, by multiplication,

$$\tilde{s}(\omega,\tau) = M_\alpha(\omega,\tau)x(\omega,\tau), \quad (6)$$

if the target and interference signals do not overlap in TF.

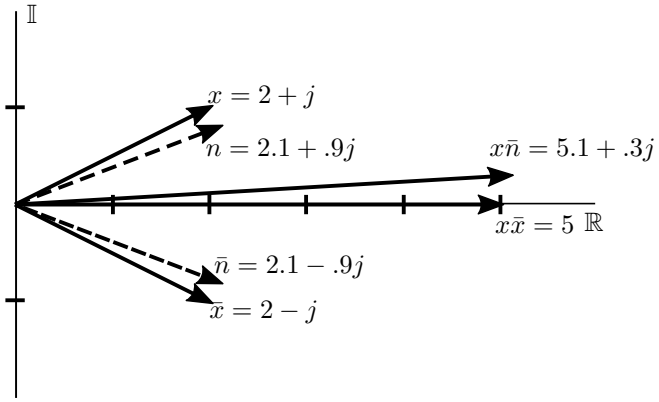

Fig. 1. Geometric Motivation: the mixture TF bin, the full-line straight arrow, is $x = 2 + j$. Its complex conjugate is also illustrated $x = 2 - j$. The calculation $x\bar{x} = 5$ yields a real value. The mixture is dominated by the interference signal, n, therefore it is approximately equal to x, e.g. $n = 2.1 + .9j$. It is illustrated using a dashed line. The product $x\bar{n}$ is elevated from the x-axis and it is slightly further from the origin than $x\bar{x}$. Introducing a scaling term h corrects the distance from the origin of the product $x\bar{n}$, e.g. $hx\bar{n}$.

Selecting a linear transform that aids separation is attractive. If a linear transform produces a compactly supported representation of speech for example, and the mixture components are independent, it is unlikely that they will activate the same frequencies at the same time. By using a windowed Fourier transform, we consider the frequency components of the signals in a *local-in-time* manner, which greatly increases the likelihood that signals do not overlap in TF. In short, we look to concentrate signal energy in some transform domain, if they are independent, it is likely that they will not overlap.

The advantages of the TF masking approach are that

- the signals are demixed in a computationally cheap way if the signals are not severely overlapped in TF (cf. Eqn. 6);
- we do not have to estimate the phase of the target signal if we mask in the TF domain. We can use the phase of the mixture.

Problem: The problem that we consider in this paper is how to select the mask for this task. We will focus our study on functions of the form

$$f(|\hat{n}(\omega,\tau)|) = \alpha|H(\omega)||\hat{n}(\omega,\tau)|. \quad (7)$$

Contribution: We propose an algorithm for detecting the support of the interference signal; for estimating the filter applied to it; and finally, for reconstructing the target signal.

III. Filter Estimation

We assume that mobile phone observations $x(t) \in L^2(\mathbb{R})$ are band-limited and sampled at a sufficiently high sampling rate. In this paper, a continuous time signal $x(t)$ is denoted

by $x_n = x(nT)$ in the discrete time domain, where T is the sampling period, the sampling rate is 16kHz, and the index n is drawn from the non-negative integers. The short time Fourier transform of x_n is denoted $x_{m,k}$, which denotes that the analysis window $w_n \in \mathbb{R}^N$ is positioned at sample mR, where R is the rational oversampling factor. The DFT size is denoted N. We use the parameters $R = \frac{N}{2}$, where $N = 1024$. We select a window which has the property that both the root mean square duration Δ_w and bandwidth Δ_W of the continuous time window $w(t)$ are finite. We use the absolute value of the filter in the discrete frequency domain in the remainder of this paper; we denote it $h_m = |H_m|$, where H_m is $H(\omega)$ in the discrete frequency domain.

We define the support sets of the source and interference signals to be Λ and Ω respectively. For example, the set Ω consists of the TF bins, $\{m, k\}$ where the interference signal is dominant. We desire a linear transform such that

$$\Lambda \cap \Omega = \emptyset. \tag{8}$$

The images of the sources under the TF linear transform provide approximate disjoint support –our results support this empirically. We first estimate h_m for all m.

Geometric Motivation: Consider one TF bin of the mixture, $x_{m,k}$. We call it x for notational simplicity. We denote its real and complex components $x = c+jd$ and illustrate it in the Argand diagram in Fig. 1. Similarly, the same TF bin of the interference is denoted $n = a+jb$. The square of the absolute value of x is an instantaneous estimate of the occupancy of the TF bin; it can be computed by $|x|^2 = x\bar{x} = c^2 + d^2 \in \mathbb{R}$.

If the interference is the only signal present in the TF bin and $h_m = 1$, then $x = n$, and therefore, $|x|^2 = x\bar{n} \in \mathbb{R}$, which is a real-valued quantity ($a = c$ and $b = d$). If the target source component is also present, then $x \neq n$. It follows that $x\bar{n} = ac+bd+j(ad-bc) \in \mathbb{C}$, which gives a complex-valued quantity, which has a non-zero elevation from $|x|^2$.

If the interference signal is dominant in the TF bin under consideration then $x\bar{n} \approx x\bar{x}$ and the angle of elevation of $x\bar{n}$ from the x-axis is small. We interpret the term $x\bar{x}$ as a squared-distance of x from the origin, along the real axis. The distance of the vector, $x\bar{n}$, resulting from the cross-product, from the origin can be computed by taking its absolute value, $|x\bar{n}|$.

If $h_m \neq 1$ we can introduce a scaling term for the interference TF point, e.g. hn, and estimate the coefficient h that ensures the two distances match $|xh\bar{n}| \equiv |x\bar{x}|$. We extend this idea to the set of TF bins where the interference is dominant. We consider the set of time bins, indexed by k, for the frequency bin m, where the interference signal is dominant and we denote this set, Ω_m. We introduce one coefficient for each frequency bin h_m and attempt to *fit* $|x_{m,k}h_m\bar{\hat{n}}_{m,k}|$ to $|x_{m,k}\bar{x}_{m,k}|$, for $\{m, k\} \in \Omega_m$.

Let us consider the m-th frequency bin. The term h_m denotes that this is the weight for m-th frequency bin.

$$g = \frac{1}{|\Omega_m|}\left(\sum_{k\in\Omega_m} |x_{m,k}|^2 + \epsilon\right) \tag{9}$$

Interpretation: The first term in function $\frac{1}{|\Omega_m|}\sum_k |x_{m,k}|^2$ is the average squared-distance of the mixture TF points from the origin. The second term in g is an error term

$$\epsilon = \sum_{k\in\Omega_m} |h_m\bar{\hat{n}}_{m,k}|^2 - 2\sum_{k\in\Omega_m} |h_m x_{m,k}\bar{\hat{n}}_{m,k}| \tag{10}$$

which computes the deviation around a squared distance. It is also an equally valid approach to express this function in terms of a deviation around $\frac{1}{|\Omega_m|}\sum_{k\in\Omega_m} |h_m\bar{\hat{n}}_{m,k}|^2$.

To minimize $g(h_m)$ with respect to h_m we compute

$$\min_{h_m} \epsilon, \tag{11}$$

we set the resulting term to zero, and solve for h_m, e.g.

$$\frac{\partial\epsilon}{\partial h_m} = 2h_m \sum_{k\in\Omega_m} |\bar{\hat{n}}_{m,k}|^2 - 2\sum_{k\in\Omega_m} |x_{m,k}\overline{\hat{n}_{m,k}}| = 0. \tag{12}$$

It follows that

$$h_m \sum_{k\in\Omega_m} |\bar{\hat{n}}_{m,k}|^2 = \sum_{k\in\Omega_m} |x_{m,k}\overline{\hat{n}_{m,k}}|, \tag{13}$$

which can be re-arranged to give the estimator:

$$h_m = \frac{\sum_{k\in\Omega_m} |x_{m,k}\overline{\hat{n}_{m,k}}|}{\sum_{k\in\Omega_m} |\bar{\hat{n}}_{m,k}|^2}. \tag{14}$$

Remark: It is important to point out that given that we are only interested in computing the distance between $x_{m,k}h_m\bar{\hat{n}}_{m,k}$ and the origin, the approach that we have proposed is simpler, in terms of notation and geometrical illustration, than expressing the problem using the traditional Squared Euclidean Distance approach. For example, consider $\hat{n}_{m,k} \in \mathbb{C}$ and $x_{m,k} \in \mathbb{C}$.

$$\sum_k |h_m\hat{n}_{m,k} - x_{m,k}|^2 \tag{15}$$

which may be expressed as

$$\sum_k (h_m\hat{n}_{m,k} - x_{m,k})\overline{(h_m\hat{n}_{m,k} - x_{m,k})} \tag{16}$$

which can be written as:

$$\sum_k h_m^2\hat{n}_{m,k}\bar{\hat{n}}_{m,k} - h_m\left(x_{m,k}\bar{\hat{n}}_{m,k} - n_{m,k}\bar{x}_{m,k}\right) + x_{m,k}\bar{x}_{m,k} \tag{17}$$

which has a slightly more involved geometrical justification; it involves the complex conjugate of both the mixture and the reference signal.

IV. Elevation Estimation via the Elevatogram

To estimate h_m we assumed that the members of the sets Ω_m were known a priori in Section III. This assumption was also a limitation of [11]. It is generally not valid; it motivates a second interesting challenge –determining in which TF bins the interference is dominant. What is the best criteria for deciding which TF bins should be in Ω_m?

A first assumption is that the reference interference is roughly aligned with the interference in the mixture. In keeping with our approach thus far, we appeal to Fig. 2 for

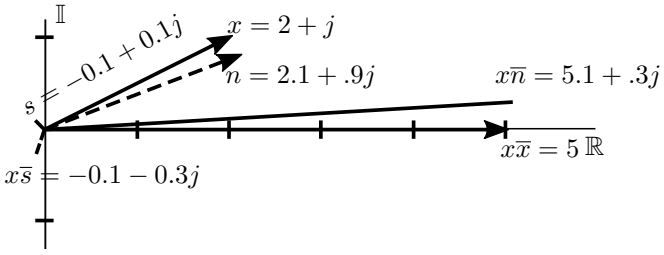

Fig. 2. Examining Elevation: the mixture TF bin, the full-line straight arrow, is $x = 2 + j$. The calculation $x\bar{x} = 5$ yields a real value. The mixture is dominated by the interference signal, n, therefore it is approximately equal to x, e.g. $n = 2.1 + .9j$. The interference is illustrated using a dashed-line arrow. The product $x\bar{n}$ is slightly elevated from the x-axis. It is slightly further from the origin than $x\bar{x}$. The target source, $s = -0.1 + 0.1j$, is illustrated with a full-line and no arrow head. The product, $x\overline{s}$ (illustrated by an arrow-head-less dashed line), has an elevation which is significantly larger than $x\overline{n}$. The support-set of n can be determined by considering the elevation of $x\overline{n}$.

inspiration. In the illustrated example, the difference between the mixture x and the reference signal n is due to the presence of the target signal, s, e.g. $s = x - n = 2 + j - 2.1 - .9j = -0.1 + 0.1j$.

Recall that (1) the product $x\overline{x}$ yields a real value (illustrated by the vector along the real axis in Fig. 2); (2) the product $x\overline{n}$ yields a vector with a small elevation from the x-axis if $n \approx x$ (vector without the arrow head in Fig. 2); (3) finally, a vector s which is not approximately equal to x causes the product $x\overline{s}$ to have a large angle of elevation (dashed vector without an arrowhead in Fig. 2). To determine the set Ω_m we search for the set of points $x\overline{n}$ which have a small angle of elevation. To this end we construct the *Elevatogram*.

We start by computing the angle of the product $x_{m,k}\overline{\hat{n}}_{m,k}$ with the positive real axis, $x_{m,k}\overline{x}_{m,k}$, which is written as

$$\varphi_{m,k} = \arg(x_{m,k}\overline{\hat{n}}_{m,k}) = \arctan\left(\frac{i}{r}\right) \tag{18}$$

where $r = \text{Re}\{x_{m,k}\overline{\hat{n}}_{m,k}\}$, $i = \text{Im}\{x_{m,k}\overline{\hat{n}}_{m,k}\}$ and we assume that $r > 0$. If this is not the case, we do not consider the TF bin.

Partition Construction: We must partition $\varphi_{m,k}$ into the source and interference support sets Λ_m and Ω_m for each m to compute the filter h_m.

The problem reduces to converting real-valued elevation data into binary-valued data, e.g. the signal components *interference* and *target source*. We assume that the angles of the TF points for each discrete frequency, m, contain two classes of points only –it follows a bi-modal distribution. We calculate the optimum threshold, t_m^*, between the two classes of points by separating the two classes in the sense that the intra-class spread is minimal. This approach can be interpreted as a 1-D discrete analog of Fisher's Discriminant Analysis [14] or threshold selection methods used in image processing [15].

We quantize the angles (Eqn. 18) which allows us to generate a histogram with L bins. More detail is given below. We call this object an Elevation histogram –Elevatogram. This object makes an exhaustive threshold search routine practical.

The problem is to partition the points in Argand diagram into the two sets of TF bins, Ω_m and Λ_m. We posit that the interference points will aligned with the x-axis. The target source set of points should have a non-zero elevation from the x-axis. How can we select an optimal threshold t_m for partitioning these sets of points?

To simplify this task we project all of the angles or elevations, $\varphi_{m,k}$ into the non-negative orthant by applying the non-linear operator, $[x]_\epsilon \leftarrow |x| + \epsilon$ with the value $\epsilon = 10^{-16}$,

$$\varphi_{m,k} \leftarrow [r_{m,k}]_\epsilon + j[i_{m,k}]_\epsilon. \tag{19}$$

In a second simplifying step we quantize each of the angle estimates $\varphi_{m,k}$ so that they are members of one of L quantization bins between 0 and $\frac{\pi}{2}$. This quantization step has the effect of making an exhaustive search routine computationally cheap. The quantization step-size is denoted Δ, it yields L angles, Δi, for $i = 0, 1, \ldots L-1$ and the value of the angle of each bin is φ_i.

The probability that a quantization bin is occupied is p_i. Given the threshold t_m, the probability of an angle being drawn from set Ω_m or Λ_m respectively, is

$$p_\Omega(t_m) = \sum_{i=0}^{t-1} p_i,\ p_\Lambda(t_m) = \sum_{i=t}^{L-1} p_i. \tag{20}$$

We do not indicate the threshold, t_m, or the frequency bin, m, in the remainder of this section, when the meaning is clear. The mean angle of each set is

$$\mu_\Omega(t_m) = \sum_{i=0}^{t_m-1} \frac{\varphi_i}{p_\Omega} p_i,\ \mu_\Lambda(t_m) = \sum_{i=t_m}^{L-1} \frac{\varphi_i}{p_\Lambda} p_i \tag{21}$$

where φ_i denotes the value of the i-th quantization level. The mean value of all of the quantized points is

$$\mu = p_\Omega \mu_\Omega + p_\Lambda \mu_\Lambda. \tag{22}$$

We will calculate the optimum threshold by separating points in Ω_m and Λ_m so that their *spread* is minimized, by considering the objective function

$$\begin{aligned} f &= p_\Omega(\mu_\Omega - \mu)^2 + p_\Lambda(\mu_\Lambda - \mu)^2 \\ &= p_\Omega p_\Lambda(\mu_\Omega - \mu_\Lambda)^2. \end{aligned} \tag{23}$$

The optimal threshold, t_m^* is computed by an exhaustive search over the values of i

$$f(t_m^*) = \max_{0 \le i \le L-1} f(i). \tag{24}$$

The support set of the interference, Ω_m, and source set, Λ_m, is computed by comparing the angles of each TF bin with the optimum threshold t_m^*, e.g.

$$\Omega_m = \{\{m,k\} | \varphi_{m,k} < t_m^*\}. \tag{25}$$

This support set gives rise to our first binary mask, the elevation mask, which is defined as

$$M_{m,k}^e = \begin{cases} 1 & \text{if } \{m,k\} \in \Omega_m, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Putting this threshold selection algorithm together with the filter estimation step in Section III, we propose two noise cancelling algorithms for removing an interference signal given an unfiltered version of the interference.

The first algorithm, Elevation Mask Denoising, combines steps 1 and 2 below. The second algorithm, Filter Mask Denoising, combines steps 1, 2 and 3.

1 Threshold Estimation - t_m
- Approximately align the reference interference with the mixture;
- Compute the discrete TF representation of the mixture, $x(t)$, and the prior information, $\hat{n}(t)$;
- Compute the angle subtended between the positive real axis and the product $x_{m,k}\bar{\hat{n}}_{m,k}$ in a counterclockwise direction, e.g. $\varphi_{m,k}$;
- Apply the partition construction step;
  - Project the angle back into the non-negative orthant using the non-linear operator (Eqn. 19);
  - Generate the histogram of the angles in the non-negative orthant;
  - Determine the optimal threshold by solving the optimization problem $\max_{0\leq i\leq L-1} f(i)$;
  - Compute the support sets Ω_m and Λ_m using t_m^*;

2 Elevation Mask Construction - M^e
- Construct the elevation mask using Eqn. 26.
- Remove the interference signal by applying M^e to (Eqn.6).

3 Estimate and Construct Filter Mask - M^f
- Estimate the filters h_m using the estimator in Eqn. 14 and the support set Ω_m
- Compute the binary mask (Eqn. 5) by computing the function in Eqn. 7 with $\alpha = 1$;
- Remove the interference signal by applying M^f to (Eqn.6).

V. Numerical Evaluation

In this section we demonstrate that the bi-modality assumption is justified and that it can be used to learn an appropriate separation threshold and its associated binary mask. We then examine the efficacy of combining the threshold with the filter learning step for a refined binary mask construction. Finally, we examine the interference removal step on a number of single channel mixtures constructed using data from the "Sixth Community-Based Signal Separation Evaluation Campaign" (SiSEC), which is available on-line here: https://sisec.inria.fr/home/bgn-2016/ data-sets. The SiSEC is a long-running denoising channel which looks to perform multi-channel denoising. In comparison with the work of this evaluation campaign, we restrict our experiments to the monaural case in this paper; we use prior information about the interfering source.

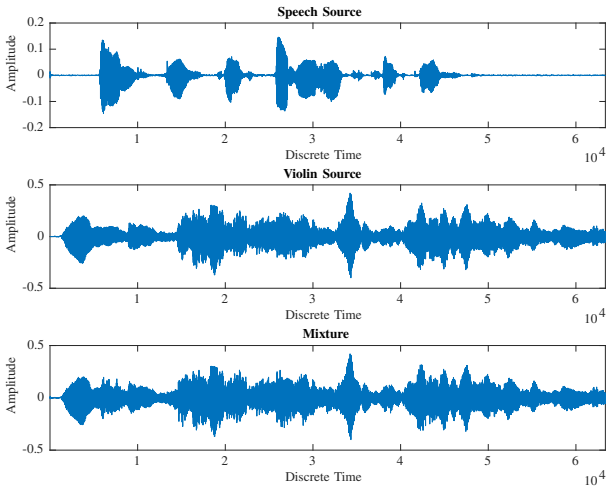

Fig. 3. Time Domain Signals: speaker (row 1); violin (row 2) and a mixture of the speaker and violin (row 3). We observe the mixture and a filtered version of the violin signal. The goal is to extract the speech.

We start by illustrating that the bi-modality assumption is realistic by considering a mixture of a speaker and a violin (in Fig. 3). The violin is the interference signal we want to remove. It is illustrated in row 2 of Fig. 3. The speaker is the target signal. It is illustrated in row 1. A mixture of both the violin and the speaker is illustrated in row 3 of Fig. 3. All mixtures examined in this paper are sampled at 16kHz. We use a clean, but delayed version of the violin signal as prior information.

Bi-modality Assumption and Threshold Evaluation: Fig. 4 illustrates that when the violin dominates in a range of frequencies the threshold selected is higher, because the approach has *correctly* decided that most TF bins in this range of frequencies correspond to the violin. Conversely, when the speaker is dominant, the selected threshold is lower. This is because two concentrations are present in the histogram: the higher concentration corresponds to the speaker which is not correlated with the interference and thus has a higher angle of elevation; the lower concentration corresponds to the violin signal, which is well-correlated with itself. There is also background noise.

Fig. 5 illustrates the threshold selected for all of the TF frequency bins in the speech-violin mixture. It summarizes the threshold selected when (1) the violin is dominant; (2) the speaker is dominant; and finally, (3) a mixture of both sources present, or no source is present. The threshold selected in each frequency bin is determined by the presence of speech. When speech is present, there is more energy present on the right-hand side of the angle histogram. When speech is not present, the energy is concentrated on the left-hand side.

Mask Evaluation: We use these thresholds, $t_m \forall m$, to construct the support sets –which are readily converted into binary masks– for the speech and the violin signal, e.g. Λ_m

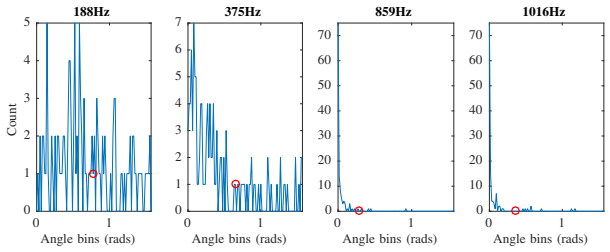

Fig. 4. Illustrating bi-modality: consider a mixture of a violin and speech in four frequencies in TF, e.g. 188, 375, 859 and 1016 Hz. Speech dominates the mixture at 188 and 375 Hz. The violin dominates the mixture at 859 and 1016 Hz. The violin histograms have their highest count at 0 radians. When the speech is dominant the histogram exhibits two concentrations. The higher (radian) concentration corresponds to the speech energy, the lower (radiant) concentration corresponds to the violin. The selected threshold (red dot) is lower when speech dominates.

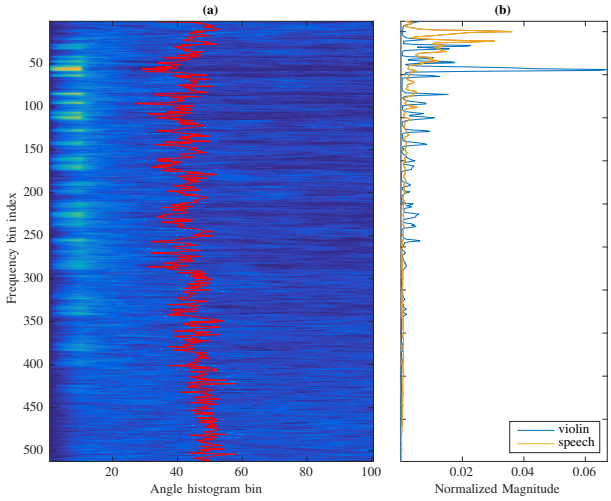

Fig. 5. The threshold estimation procedure is examined for a mixture of speech and a violin. The estimated threshold depends on dominance of the speech or violin in each frequency. (a): Histogram of angle of cross-product of mixture with clean interference signal the red line denotes the estimated threshold; (b): Comparison of the normalized sum of the magnitudes of the TF bins of the violin with the speech, a measure of the dominance of each signal in each frequency bin. The threshold estimated in each bin depends on the energy of each mixture component.

and Ω_m for all frequency bins. Then, we compare three types of binary mask: 1) an ideal binary mask, which is used as a benchmark and relies on full knowledge of both sources; 2) a mask constructed using the elevation threshold scheme. We call this mask M^e (for the speaker) and $M^{e,*}$ (for the violin) respectively, where the superscript denotes the elevation method; and finally 3) we refine our estimates of these masks by using the filter estimation procedure in Eqn. 5 and Eqn. 7. We denote these masks M^f and $M^{f,*}$ respectively, where the superscript denotes that these masks are constructed by using the elevation threshold scheme to form the support sets for the filter estimator.

The purpose of this evaluation is to compare the the efficacy of M^e and M^f with the best binary mask –the *ground-truth* or *ideal* binary mask. Given that our interference removal approach is based on binary masking, the best binary mask can be constructed if we have knowledge of both the clean speech and violin signal, by comparing the magnitude of each TF bin for the speech and the violin and adding a one to the mask if the speech is dominant, e.g.

$$I_{m,k} = \begin{cases} 1, & \text{if } |s_{m,k}| > |\hat{n}_{m,k}| \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

The mask for the violin, I^*, is constructed by inserting ones in the positions of zeros, and vice verse.

TABLE I
PERFORMANCE OF BINARY MASKS

	Speech	Violin
% Mask Preserved by M^e	54.2	90.6
% Energy Preserved using M^e	53	97
% Interfering Energy m^e	2.6	46.7
% Energy Preserved using I	95.5	99.25
% Mask Preserved by M^f	66	58
% Energy Preserved using M^f	67	70
% Interfering Energy M^f	30.99	26.89

Mask Performance: We consider a number of different performance measures. We then consider the masks qualitively. The first measure is the percentage Mask Preserved (MP), or the percentage of elements of the learned mask which are in common with the ideal binary mask. This is computed as follows for an arbitrary binary mask Y:

$$M_p(Y) = \frac{\sum_{m,k} I_{m,k} Y_{m,k}}{\sum_{m,k} I_{m,k}} \tag{28}$$

The second measure is the percentage Energy Preserved (EP) by using a given mask Y,

$$E_p(Y) = \frac{\sum_{m,k} Y_{m,k} |s_{m,k}|^2}{\sum_{m,kj} |s_{m,k}|^2}. \tag{29}$$

For example, for speech separated using the ideal mask we have $Y_{m,k} = I_{m,k}$. We tabulate these measures for the violin-voice mixture in Table I. Approximately 90% of the ideal binary mask for the violin is preserved by the elevation method for thresholding, and 54% of the speech ideal mask elements are preserved. Speech is compactly represented in the TF domain, and therefore, identifying 54% of the relevant TF bins is sufficient to accurately reconstruct the speaker. The elevation mask preserves 53% of the speech energy and 97% of the violin energy. For completeness, we also note the percentage energy of the violin that was incorrectly assigned to the speech and vice versa, in the *Interfering Energy* in Table I, which is computed for the mask Y as follows:

$$I_p(Y) = \frac{\sum_{m,k} Y_{m,k} |\hat{n}_{m,k}|^2}{\sum_{m,kj} |\hat{n}_{m,k}|^2}. \tag{30}$$

This measure indicates the effect of masking the violin signal using the speaker mask. To put these results in context we give the percentage EP for the ideal binary mask for the violin. The percentage EP using M^e for the violin is 97% which compares

favourably with the percentage EP by the ideal binary mask (99.25%). This result, coupled with the fact that the ideal mask preserves 95% of the energy of the speech signal, demonstrates that binary masking is a good approach for monaural interference removal. Good performance is achievable as $> 95\%$ of the TF bins for the two source do not overlap; however we temper this result with the caveat that 5% of the energy of the sources does overlap, and that speech and music signals have a compact support in TF –this 5% could be important.

Recall that in this task a reference version of the violin signal is used to help determine a good threshold for separating a mixture of speech and violin. In this case, TF bins which have an elevation which are close to the real axis are assigned to the violin. This is undesirable. When TF bins consist of a mixture of both the speech and the violin, our investigation supports the assertion that these TF bins are assigned to the violin support set, which in turn reduces the quality of the recovered speech signal.

Table I summarizes the performance of the mask constructed using the elevation method and the filter estimation step, M^f. This mask preserves more of the elements of the ideal mask. The percentage MP by M^f is 66% compared to the percentage MP of 54% obtained by the mask M^e. Similarly, the energy of the speech signal preserved by M^f is also increased. These increases come at the expense of separation performance. The violin signal is more audible in the denoised speech for the mask M^f compared to the mask M^e. This result is confirm by examining the percentage energy leaked by the masks.

We discuss the performance of the masks qualitatively by reference to Fig. 6. A clean segment of the TF representation of the speech is given in Fig. 6 along with a version produced by ideal masking. The ideal mask preserves the high-energy TF bins of the speech, however, TF bins with low energy are generally assigned to the violin. The elevation mask preserves high-energy TF bins of the speaker without introducing energy from the violin; however, fewer TF bins are preserved than by the Ideal mask. Finally, the M^f mask preserves the high-energy TF bins of the speaker, but preserves more lower energy high frequency bins than the elevation mask. This mask produces a better representation of the speech by introducing a significant amount of violin energy into the estimated signal. This is not desirable from a perceptual evaluation of denoising. This analysis underpins the quantitative evaluation in Table I.

Interference Removal in Urban Environments: We illustrate the performance of the estimator in Eqn. 14 combined with the threshold in an interference removal task in an urban-type environment. We consider the performance measures outlined in [16], which have been used in many subsequent multi-channel source separation challenges by the community (https://sisec.inria.fr/home/bgn-2016/). The measures we consider are the: Signal to Distortion Ratio (SDR); Source to Interference Ratio (SIR); and finally, the Sources to Artifacts Ratio (SAR). One benefit of these measures is that they do not depend on signal normalization or power which introduce error into Signal-to-Noise-Ratio estimates.

Table II evaluates the Blind Source Separation (BSS) scores

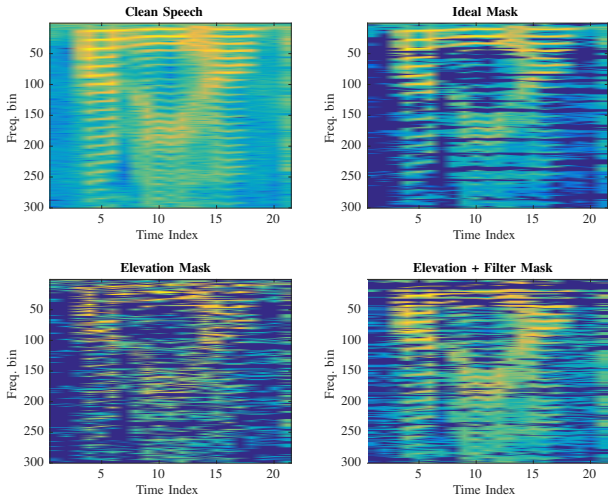

Fig. 6. Qualitative comparison using a short segment of the speech-violin mixture: Ideal Mask, Elevation Mask, Elevation and Filter Mask. We have illustrated the lower frequencies (first 300 frequency bins) and 25 consecutive time bins.

TABLE II
BLIND SOURCE EVALUATION TOOLBOX: TARGET SPEECH SOURCE IN THE PRESENCE OF NOISE.

	I	M^e	M^f	$\mathbf{1}$
Signal to Distortion Ratio	15.8832	5.0927	3.2876	-0.0008
Signal to Interference Ratio	22.2958	20.6979	6.8212	0.1321
Sources to Artifacts Ratio	17.0350	5.2506	6.6513	18.1516

obtained for the Speaker-Violin denoising task for the ideal binary mask, M^e, M^f and finally for a mask of all ones, which is denoted $\mathbf{1}$, which serves to illustrate the performance when no denoising is performed.

Regarding the Speech-Violin mixture, the SIR for the mask M^e is closer to that of the ideal mask I than the mask derived from the filter, M^f. This result agrees with the quantitative and qualitative analysis of these masks above. Moreover, the mask M^e produces a denoised signal that exhibits an improved SDR over M^f, but at the cost of introducing more artifacts in the denoised signal than the mask M^f. This result is supported by a comparison of the SIR and SDR of the masks. We now broaden our analysis to consider a wider range of signal ensembles. We examine mixtures of (1) a male speaker and a guitar; (2) a male speaker and a female speaker; (3) a female speaker and a moving subway car; and finally, (4) a female speaker in a city square environment. In each denoising task the first source is the target source and the second listed source is the interference. In these experiments, we assume that an unfiltered version of the interference signal is available as prior information as described by the our motivating scenario in Section I. Mixtures are created by normalizing each source signal and adding them (weighting them by one).

In each task the elevation mask M^e yields the best SIR. In all but one case, the best SAR and SDR is given by the mask M^f generated by the elevation method, coupled

TABLE III
MALE SPEAKER (TARGET) AND GUITAR (INTERFERENCE)

	I	M^e	M^f	$\mathbf{1}$
SDR	10.8883	2.6394	4.7032	-0.2941
SIR	20.7462	16.6687	7.5474	-0.1378
SAR	11.3983	2.9072	8.5898	17.3004

TABLE IV
MALE SPEAKER (TARGET) AND FEMALE SPEAKER (INTERFERENCE)

	I	M^e	M^f	$\mathbf{1}$
SDR	13.2361	3.4319	3.9832	-0.2368
SIR	23.0074	20.2969	6.0257	-0.0793
SAR	13.7416	3.5626	9.2089	17.2960

TABLE V
FEMALE SPEAKER (TARGET) AND SUBWAY CAR MOVING (INTERFERENCE)

	I	M^e	M^f	$\mathbf{1}$
SDR	15.3383	4.3769	2.9938	0.0024
SIR	21.6354	18.2969	4.2447	0.1531
SAR	16.5290	4.6205	10.3970	17.6096

TABLE VI
FEMALE SPEAKER (TARGET) AND CITY SQUARE (INTERFERENCE)

	I	M^e	M^f	$\mathbf{1}$
SDR	15.9713	5.8592	1.9853	-0.1768
SIR	24.1081	23.1886	4.0517	-0.0301
SAR	16.7122	5.9611	7.6439	17.6346

with the filtering method. In summary, the elevation mask, M^e, removes the interfering source, almost completely, but the resultant target signal is corrupted by this mask. Musical noise is introduced, but this artifact does not overly affect the comprehensibility of the recovered signals. Fewer artifacts are introduced by the mask M^f, but this mask achieves this by not removing some components of the interference signal. Listening tests confirm that the denoised target signals, resulting from the application of the elevation mask M^e, are cleaner. This result bears out the inherent trade-off in using binary masking as a denoising approach. To remove an interfering signal using hard masking we introduce artifacts into the recovered target signal (because the windowed disjoint orthogonality assumption is only approximately true). Some TF bins are assigned to one of the two sources when they should potentially be assigned to both. On the other hand the performance of binary masking is impressive and the resultant approaches are computationally cheap.

VI. CONCLUSION

Two methods for denoising a single-channel mixture via time-frequency masking, given a prior information about the interfering signal, were proposed and analyzed. Given a mixture of a target signal and an unwanted interference, our goal was to eliminate the interference without introducing artifacts. Prior information, in the form of a clean unfiltered recording of the interference was required. We proposed a method for automatically constructing signal support sets. Prior information about the signal supports sets is generally assumed for single-channel binary masking approaches. We then proposed two algorithms for interference removal. The first approach constructed a binary mask using the learned signal support set and gave the best interference removal. However, a bye-product of this approach was the introduction of musical noise into the recovered signal. The second approach, combined the support set information with a filter estimation step. The introduction of artifacts was reduced by this method, however, the recovered signal exhibited traces of the interfering signal. We posit that these denoising approaches may be well suited to applications where low computational complexity is a requirement, and prior information about the interference is available from other networked sensors, or by the user.

REFERENCES

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, ""Your word is my command": Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, A. Neustein, Ed. New York: Springer, 2010, ch. 4, pp. 61–90.

[2] B. Birgitta and L. Thomas, "Community noise," *Stockholm: Archives of the Centre for Sensory Research*, vol. 2, 1995.

[3] B. Shahriari and M. Moh, "Intelligent mobile messaging for urban networks: Adaptive intelligent mobile messaging based on reinforcement learning," in *2016 IEEE 12th Int. Conf. on Wireless and Mobile Computing, Networking and Communications*, Oct. 2016, pp. 1–8.

[4] D. Radu, C. Avram, A. Atilean, B. Parrein, and J. Yi, "Acoustic noise pollution monitoring in an urban environment using a vanet network," in *IEEE Int. Conf. on Automation, Quality and Testing, Robotics*, May 2012, pp. 244–248.

[5] R. Loomba, R. de Fréin, and B. Jennings, "Selecting energy efficient cluster-head trajectories for collaborative mobile sensing," in *2015 IEEE GLOBECOM*, Dec. 2015, pp. 1–7.

[6] R. de Fréin, "Learning convolutive features for storage and transmission between networked sensors," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Jul. 2015, pp. 1–8.

[7] ——, "Learning and storing the parts of objects: IMF," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Sep. 2014.

[8] R. de Fréin, S. T. Rickard, and B. A. Pearlmutter, *Constructing Time-Frequency Dictionaries for Source Separation via Time-Frequency Masking and Source Localisation*. Springer Berlin Heidelberg, 2009, pp. 573–80.

[9] M. Kushwaha, X. Koutsoukos, P. Volgyesi, and A. Ledeczi, "Acoustic source localization and discrimination in urban environments," in *12th Int. Conf. on Information Fusion*, Jul. 2009, pp. 1859–1866.

[10] S. Haykin, "Adaptive filter theory," 1996, London.

[11] S. Rickard, C. Fearon, R. Balan, and J. Rosca, "Minuet: Musical interference unmixing estimation technique," *Conf. on Inf. Sc. and Sys.*, pp. 1–6, 2004.

[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–47, Jul. 2004.

[13] R. de Fréin and S. Rickard, "The synchronized short-time-Fourier-transform: Properties and definitions for multichannel source separation," *IEEE Trans. Sig. Proc.*, vol. 59, no. 1, pp. 91–103, Jan. 2011.

[14] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pat. Ana. and Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Sys., Man., Cybernet.*, vol. 9, no. 1, pp. 62–6, Jan. 1979.

[16] "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–9, 2006.

Acknowledgment: This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/SIRG/3459