2016-11-11

Power-Weighted Divergences for Relative Attenuation and Delay Estimation

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

Scott T. Rickard Prof
Salesforce San Francisco, CA, USA

Recommended Citation

deFrein, R. & Rickar, S.T. (2016). Power-Weighted Divergences for Relative Attenuation and Delay Estimation. *IEEE Signal Processing Letters*, vol. 23, no. 11, pg. 1612-1616. doi: 10.1109/LSP.2016.2610481

Funder: IRC International Career Development Fellowship co-funded by Marie Curie Actions

Power-Weighted Divergences for Relative Attenuation and Delay Estimation

Ruairí de Fréin[†††] and S. T. Rickard[†††]

[†]KTH - Royal Institute of Technology, Stockholm, Sweden

[††]Dublin Institute of Technology, Ireland

[†††]Salesforce San Francisco, CA, USA

web: https://robustandscalable.wordpress.com

in: IEEE Signal Processing Letters. See also BibTeX entry below.

BibTeX:

```
@article{deFrein16Power,
author={Ruair\'{i} de Fr\'{e}in^\dagger$$^{\dagger\dagger}$ and S. T. Rickard^{\dagger\dagger\dagger}$},
journal={IEEE Signal Processing Letters},
title={Power-Weighted Divergences for Relative Attenuation and Delay Estimation},
year={2016},
volume={23},
number={11},
pages={1612-1616},
doi={10.1109/LSP.2016.2610481},
ISSN={1070-9908},
month={Nov},
url={http://ieeexplore.ieee.org/document/7570234/},}
```


document created on: March 14, 2018
created from file: double.tex
cover page automatically created with CoverPage.sty
(available at your favourite CTAN mirror)

Power Weighted Divergences for Relative Attenuation and Delay Estimation

*Ruairí de Fréin, *Member, IEEE,* Scott T. Rickard, *Senior Member, IEEE,*

***Abstract*—Power Weighted estimators have recently been proposed for relative attenuation and delay estimation in Blind Source Separation. Their provenance lies in the observation that speech is approximately Windowed-Disjoint Orthogonal (WDO) in the Time-Frequency (TF) domain; it has been reported that using WDO, derived from TF representations of speech, improves mixing parameter estimation. We show that Power Weighted relative attenuation and delay estimators can be derived from a particular case of a Weighted Bregman Divergence. We then propose a wider class of estimators, which we tune to give better parameter estimates for speech.**

***Index Terms*—Bregman divergence, Kullback Leibler, Itakura-Saito, relative attenuation estimation, relative delay estimation.**

I. Introduction

Consider a stereo an-echoic de-mixing problem which consists of M sources, $s_1(t), \dots s_M(t)$ and two mixtures $x_1(t) = \sum_{j=1}^{M} s_j(t)$ and $x_2(t) = \sum_{j=1}^{M} a_j s_j(t-\delta_j)$, where t is continuous time. The second mixture, $x_2(t)$, is the sum of attenuated (by α_j) and delayed (by δ_j) versions of the source signals, $s_j(t)$, observed by $x_1(t)$. The DUET algorithm has been successfully applied to the task of separating-out each source given the two observation signals $x_1(t)$ and $x_2(t)$ –it relies on the fact that speech sources are naturally partitioned in the Time-Frequency (TF) lattice [1]. It uses these partitions to separate sources. The partitions are constructed by using the spatial signature (α_j, δ_j) of each source. Using a linear transform that promotes separation in TF is crucial. DUET describes a special case of the more general echoic convolutive BSS mixing procedure [2], [3]. It attempts to emulate the sophistication of the human auditory system by solving the "cocktail party problem" [4].

The linear transform of choice is the discrete Short-Time Fourier Transform (STFT) [5], which is denoted $T : s_j[i] \in \mathbb{I} \mapsto \hat{s}_j[\omega, \tau] \in \mathbb{C}$ where i, ω, τ are the discrete time, discrete frequency and the window position, and the collection of square integrable functions is denoted $\mathbb{I}$. The number of frequency bins used is F and the analysis window is shifted by $\frac{F}{2}$ here. The STFT is invertable $T^{-1}(Ts) = s$; it is approximately true that the common support of two (or more) speech sources is the empty set (of TF bins (ω, τ)). This property is evaluated versus the number of sources and the TF analysis parameters in [1]. It is called WDO; it states that for all pairs of source signals

$$\hat{s}_j[\omega, \tau]\hat{s}_k[\omega, \tau] = 0, \quad \forall \tau, \omega, j, k, \text{ and } j \neq k. \tag{1}$$

R. de Fréin is with Institiúid Teicneolaíochta Bhaile Átha Cliath, Ireland (e: rdefrein@gmail.com). S.T. Rickard is VP for Data Science with Salesforce. Manuscript received June. EDICS: SAS-ICAB

An estimate of the j-th source signal can be obtained by constructing an indicator function, a *binary mask*, from a source's support set, Λ_j: (1) by setting $\mathbf{1}_j[\omega, \tau] = 1$ if $(\omega, \tau) \in \Lambda_j$ and $\mathbf{1}_j[\omega, \tau] = 0$ otherwise, and (2) by inverting the product of the j-th binary mask times one of the mixture signals. This yields the estimate $s_j[i] = T^{-1}(\mathbf{1}_j T x_1[i])$.

The challenge of determining the support set Λ_j is addressed by appealing to the properties of the STFT [5]. Instantaneous estimates of the relative attenuation and delay of the sources (cf. [1]) observed by $x_1[i]$ and $x_2[i]$, are obtained from

$$\alpha[\omega, \tau] = \left| \frac{\hat{x}_2[\omega, \tau]}{\hat{x}_1[\omega, \tau]} \right| \text{ and } \delta[\omega, \tau] = -\frac{F}{2\pi\omega} \angle \frac{\hat{x}_2[\omega, \tau]}{\hat{x}_1[\omega, \tau]}. \tag{2}$$

The support sets $\Lambda_j, 1 \leq j \leq M$ are determined by clustering the pairs of instantaneous estimates $(\alpha[\omega, \tau], \delta[\omega, \tau])$ and assigning each TF bin, (ω, τ), to the nearest of the M centroids, for example, using the Squared Euclidean Distance (SED). The cardinality of the support set is denoted $N = |\Lambda_j|$; the source index is clear from the context. The problem of learning the optimal spatial signature $\alpha_j^\star$ and $\delta_j^\star$ for the j-th source, is that of estimating the centre of the point-cloud of instantaneous parameter estimates corresponding to each source, which have been identified by a clustering procedure. To estimate the j-th source's spatial signature the authors of [1] simplified the stereo an-echoic mixing model in the TF bins Λ_j for the j-th source by treating it as a linear observations in noise model, where $W = e^{-j\frac{2\pi}{F}}$:

$$\begin{aligned} \hat{x}_2[\omega, \tau] =& \hat{s}_j[\omega, \tau]\alpha_j W^{\omega\delta_j} + \hat{n}_2[\omega, \tau], \\ \hat{x}_1[\omega, \tau] =& \hat{s}_j[\omega, \tau] + \hat{n}_1[\omega, \tau], \quad \forall (\omega, \tau) \in \Lambda_j, \end{aligned} \tag{3}$$

where $\hat{n}_1$ and $\hat{n}_2$ are iid white complex Gaussian noise signals with zero mean and variance σ^2, which represent the contribution of the other source signals in the support set of the target source. The Maximum Likelihood Estimators (MLE) derived in [1] minimized a scaled version of the log-likelihood

$$\sum_{(\omega,\tau)\in\Lambda_j} |\hat{n}_1[\omega, \tau]|^2 + |\hat{x}_2[\omega, \tau] - \alpha_j W^{\omega\delta_j} \hat{s}_j[\omega, \tau]|^2. \tag{4}$$

The resulting estimators performed poorly. This objective (Eqn. 4) neglects to account for the approximate WDO assumption which gave rise to DUET [1]. All TF bins, no matter how much they are corrupted by interfering signals, are given an equal weighting in the MLEs. In this paper we show that the family of weighted element-wise Bregman divergences produces better power weighted estimators.

II. Bregman Divergences

Divergences are distance-like functions that are non-negative and separable; however, they do not satisfy the

triangle equality and the symmetry axiom of a distance [6]. Some examples of divergences include the Kullback-Liebler Divergence (KLD), Itakura-Saito Divergence (ISD) [7] and SED, which we include as they are of particular historical relevance to speech. The Bregman divergence encompasses the SED, the KLD and the ISD [8]; we present our ideas using this divergence. We consider a generalized form of the KLD; its arguments are not restricted to points on the simplex.

Let $\mathcal{S}$ be a set. A divergence on $\mathcal{S}$ is a function $D : \mathcal{S}\times\mathcal{S} \mapsto \mathbb{R}_+$ satisfying $\forall(p,q) \in \mathcal{S}\times\mathcal{S}$

$$D(p||q) \geq 0, \quad D(p||q) = 0 \text{ iff } p = q. \tag{5}$$

The Bregman divergence is defined as follows: let $\mathcal{S}$ be a convex subset of a Hilbert space and $\Phi : \mathcal{S} \mapsto \mathbb{R}$ a continuously differentiable strictly convex function. The Bregman divergence $D_\Phi : \mathcal{S}\times\mathcal{S} \mapsto \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of non-negative real numbers, is defined as

$$D_\Phi(\mathbf{x}||\mathbf{y}) = \Phi(\mathbf{x}) - \Phi(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \nabla\Phi(\mathbf{y})\rangle \tag{6}$$

where $\nabla\Phi(\mathbf{y})$ is the gradient of Φ defined at $\mathbf{y}$ and $\langle\cdot,\cdot\rangle$ is the Hermitian dot product. A divergence on $\mathcal{S}$ is called an element-wise divergence if there exists a divergence d on $\mathcal{S}$ such that $\forall\mathbf{x} = [x_1,\ldots,x_N]^T$, and $\forall\mathbf{y} = [y_1,\ldots,y_N]^T$

$$D(\mathbf{x}||\mathbf{y}) = \sum_{n=1}^{N} d(x_n|y_n). \tag{7}$$

An example of such a divergence is the element-wise Bregman divergence. It is a subclass of Bregman divergences for which Φ is the sum of N scalar, continuously differentiable and strictly convex element-wise functions, e.g.

$$\forall\mathbf{x} = [x_1,\ldots,x_N]^T \in \mathcal{S}, \quad \Phi(\mathbf{x}) = \sum_{n=1}^{N}\phi(x_n). \tag{8}$$

Defining $D_\Phi(\mathbf{x}||\mathbf{y}) = \sum_{n=1}^{N} d_\phi(x_n|y_n)$ where $d_\phi = \phi(x) - \phi(y) - \phi'(y)(x-y)$ it is clear the divergence is element-wise.
Examples: The SED can be composed as a Bregman divergence by selecting $\phi(x) = x^2$ and then $d_\phi = \phi(x) - \phi(y) - \phi'(y)(x-y) = x^2+y^2-2xy$. The KLD can be composed by selecting $\phi(x) = x\log x$. It follows that $d_\phi = x\log\frac{x}{y} - x + y$. The ISD can be composed by selecting $\phi(x) = -\log x + x - 1$, and thus, $d_\phi = -\log\frac{x}{y} + \frac{x}{y} - 1$.

III. Weighted Element-wise Bregman Divergence

We focus on weighted element-wise divergences, which are a subclass of the Bregman family of divergences. This family is characterized by the fact that Φ is the sum of N scalar continuously differentiable and strictly convex element-wise functions and the arguments of $\Phi(x)$ are weighted by the elements of $\mathbf{w} = [w_1,\ldots,w_N]^T \in \mathbb{R}_+^N$, e.g.

$$\forall\mathbf{x} = [x_1,\ldots,x_N]^T \in \mathcal{S}, \quad \Phi_w(x) = \sum_{n=1}^{N}\phi(w_n x_n). \tag{9}$$

Then $D_{\Phi_w}(\mathbf{x}||\mathbf{y}) = \sum_{n=1}^{N} d_{\phi_w}(x_n|y_n)$ where $d_{\phi_w} = \phi(wx) - \phi(wy) - \phi'(wy)(x-y)$.
Examples: The weighted SED can be composed as an element-wise weighted Bregman divergence by selecting $\phi(wx) = (wx)^2$ and then $d_\phi = (wx)^2 + (wy)^2 - 2xyw^2$. Similarly, the weighted KLD is written as $\phi(wx) = (wx)\log(wx)$ and then $d_\phi = wx\log\frac{wx}{wy} - wx + wy$. And finally, the weighted ISD is written as $\phi(wx) = -\log(wx) + wx - 1$ and then $d_\phi = -\log\frac{x}{y} + \frac{x}{y} - 1$. Using weighted element-wise Bregman divergences opens up the possibility of (1) extending traditional clustering/estimation algorithms to data which is better modelled using an arbitrary member of the family of exponential distributions [8] and (2) encoding domain specific weights into the divergence.
Weighted estimators: Consider the case where we have a set of N instantaneous estimates of some parameter, for example, the instantaneous relative attenuation or delay estimates used by DUET. We arrange them in vector form $\mathbf{x} = \text{vec}_{1\leq n\leq N}(x_n) = [x_1,\ldots,x_N] \in \mathbb{R}^N$. We desire the centre of this point cloud of parameters, e.g. a scalar $c^\star \in \mathbb{R}$ with respect to an (a)symmetric divergence and the vector of weights $\mathbf{w} = [w_1,\ldots w_N] \in \mathbb{R}_+$, which encodes domain specific information about the reliability of the associated instantaneous parameter estimates $\mathbf{x}$. Using the notation $\odot$ to denote element-wise multiplication, we consider the solutions to this problem by minimizing a number of divergences,

$$c^\star = \min_c D_{\Phi_w}(\mathbf{w}\odot\mathbf{x}||c\mathbf{w}) = \min_c \sum_{n=1}^{N} d(w_n x_n|cw_n). \tag{10}$$

Thm 1: Given the weighted SED form of the element-wise weighted Bregman divergence, where $\phi(w_n x_n) = (w_n x_n)^2$,

$$D_{\Phi_w}(\mathbf{w}\odot\mathbf{x}||c\mathbf{w}) = \sum_{n=1}^{N}(w_n x_n)^2 + (w_n c)^2 - 2x_n c w_n^2; \tag{11}$$

the unique, optimum solution is $c^\star_{\text{SED}} = \frac{\sum_{n=1}^{N} w_n^2 x_n}{\sum_{n=1}^{N} w_n^2}$.
Thm 2: Given the weighted KLD form of the element-wise weighted Bregman divergence, where $\phi(w_n x_n) = w_n x_n \log w_n x_n$, and

$$D_{\Phi_w}(\mathbf{w}\odot\mathbf{x}||c\mathbf{w}) = \sum_{n=1}^{N} w_n x_n \log\frac{w_n x_n}{w_n c} - w_n x_n + w_n c; \tag{12}$$

the unique, optimum solution is $c^\star_{\text{KLD}} = \frac{\sum_{n=1}^{N} w_n x_n}{\sum_{n=1}^{N} w_n}$.
Thm 3: Given the weighted ISD form of the element-wise weighted Bregman divergence, e.g. where $\phi(w_n x_n) = -\log w_n x_n + w_n x_n - 1$

$$D_{\Phi_w}(\mathbf{w}\odot\mathbf{x}||c\mathbf{w}) = \sum_{n=1}^{N} -\log\frac{x_n}{c} + \frac{x_n}{c} - 1, \tag{13}$$

the unique, optimum solution is $c^\star_{\text{ISD}} = \frac{\sum_{n=1}^{N} x_n}{N}$. We can generalize the weighted SED, KLD and ISD cases by appealing to the weighted β-divergence, which gives rise to a family of power weighted estimators.
Thm 4: Given the β-divergence [9] form of the element-wise weighted Bregman divergence, $\phi(w_n x_n) = \frac{(w_n x_n)^\beta}{\beta(\beta-1)} - \frac{w_n x_n}{\beta-1} + \frac{1}{\beta}$, for $\beta \in \mathbb{R}\setminus\{0,1\}$, and

$$D_{\Phi_w}(\mathbf{w}\odot\mathbf{x}||c\mathbf{w}) = \sum_{n=1}^{N}\frac{(w_n x_n)^\beta + (\beta-1)(w_n c)^\beta - \beta w_n^\beta x_n c^{\beta-1}}{\beta(\beta-1)}; \tag{14}$$

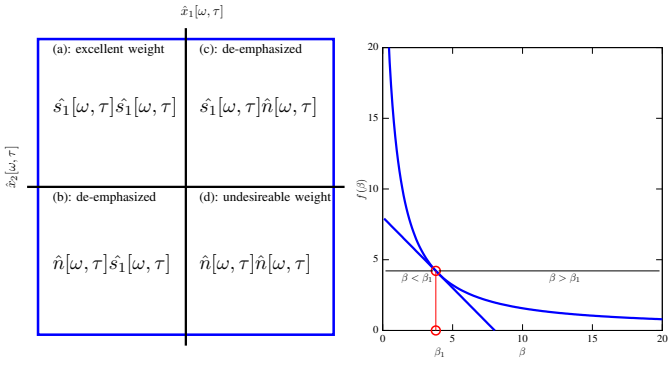

Fig. 1. **WDO Quadrant:** (a) when the source has high power the associated instantaneous estimate is heavily weighted; when the noise is present in one channel and is uncorrelated with the source ((b) and (c)), the corresponding instantaneous estimate has a small weight; (d) when the noise is dominant in both channels it generally has a much lower power than case (a).
Selecting β: The MSWE is the weighted sum of decreasing exponentials. The tangent line to the MSWE with slope -1 is illustrated. The MSWE and its tangent intersect at $\beta = \beta_1$.

the unique, optimum solution is $c^\star_\beta = \frac{\sum_{n=1}^N w_n^\beta x_n}{\sum_{n=1}^N w_n^\beta}$. We continue by demonstrating that the widely used DUET Blind Source Separation algorithm [1] is a special case of the weighted β-divergence family of estimators.

IV. Weighting function: Orthogonality in TF

The WDO property (in Eqn. 1), which was proposed in [1], suggests the following weighting scheme for measuring the orthogonality of the target source with the other interfering sources, in the observed mixtures $\hat{x}_1$ and $\hat{x}_2$, in the disjoint source support set Λ_j for the j-th source signal:

$$w_j[\omega,\tau] = |\hat{x}_1[\omega,\tau]\hat{x}_2[\omega,\tau]| \quad \forall(\omega,\tau) \in \Lambda_j. \tag{15}$$

When $\hat{x}_1[\omega,\tau]$ and $\hat{x}_2[\omega,\tau]$ are dominated by the target source, $\hat{s}_j[\omega,\tau]$ then $w_j[\omega,\tau] \approx |\hat{s}_j[\omega,\tau]\hat{s}_j[\omega,\tau]|$, which is large when the power of the source is large in that TF bin. When $\hat{x}_1[\omega,\tau]$ is dominated by the target source and $\hat{x}_2[\omega,\tau]$ is dominated by noise, if the source and the noise are uncorrelated, the associated weight is small $w_j[\omega,\tau] \approx |\hat{s}_j[\omega,\tau]\hat{n}_2[\omega,\tau]|$. When both channels are dominated by noise, then $w_j[\omega,\tau] \approx |\hat{n}_1[\omega,\tau]\hat{n}_2[\omega,\tau]|$. In this case, if the noise signals on both channels are uncorrelated, $w_j[\omega,\tau]$ is small. If there is significant correlation in the noise, we rely on the *disjointness* assumption; only the target source is active in Λ_j, and thus the power of the noise is generally much smaller than the power of the target source. Consequently, the corresponding weight is *relatively* small. Fig. 1 summarizes this argument.
Thm 5: DUET Power Weighted Estimators [1] minimize the weighted β-divergence, $D_{\Phi_w}(\mathbf{w} \odot \mathbf{x}||c\mathbf{w})$, between the instantaneous parameter estimates $\mathbf{x} = \text{vec}_{(\omega,\tau)\in\Lambda_j}(\alpha[\omega,\tau])$ and $\mathbf{x} = \text{vec}_{(\omega,\tau)\in\Lambda_j}(\delta[\omega,\tau])$, and the centre of the point-cloud of parameter estimates, $c = \alpha_j^\star$ and $c = \delta_j^\star$, where each estimate is weighted by its WDO, Eqn. 15.
Interpretation: When $\beta \to 1$ the DUET estimators minimize an objective that tends to the weighted generalized KLD. When $\beta = 2$ the DUET estimators minimize the weighted SED. When $\beta \to 0$ the DUET estimators minimize an objective that tends to the weighted ISD. *Remark:* Using element-wise weighted Bregman divergences has a number of advantages. (1) We have directly encoded WDO into the estimation problem. This produces estimators that directly weight the instantaneous estimates by a measure of orthogonality. The MLEs in [1] do not consider orthogonality but do consider the disjointness property (in the summation index). (2) The DUET power weighted estimators, e.g. $\beta = \frac{1}{2}, 1, 2$, are now derived in a straightforward way. (3) We can motivate the use of a range of new estimators for different statistical distributions and new application domains [10], by choosing an appropriate weighted divergence and weighting function [8], e.g. when $\beta = 2, 1, 0$ the distribution corresponds to a Normal, Poisson and Gamma distribution respectively. (4) We can empirically evaluate the correct estimators to use with speech by evaluating the performance of the estimators associated with a wide range of parametrized divergences.
Selecting the Weighting Function: Consider K typical relative delays (or attenuations) that a source can experience in a certain environment $\{c_k\}_{k=1}^K$. What β for the set of weighting functions, $\{\{w_{n,k}\}_{n=1}^N\}_{k=1}^K$ minimizes the Mean Squared Weighted Error (MSWE) between the N instantaneous parameter estimates, $x_{n,k}$ for the k-th relative delay, and the true parameter c_k, for all k? If the error is $\epsilon_{n,k} = c_k - x_{n,k}$ and the weighted error is $f_{n,k} = w_{n,k}^\beta \epsilon_{n,k}$, the MSWE is

$$f = \frac{1}{NK}\sum_{n,k} f_{n,k}^2, \quad \text{for } \beta > 0. \tag{16}$$

We normalize the weights so that they are bounded above by one, $0 < w_{n,k} < 1$, by setting $w_{n,k} \leftarrow \frac{w_{n,k}}{a_k}$, where $a_k = \max w_{n,k}$. Normalization has no effect on the estimators. The squared deviation $\epsilon_{n,k}^2$ is nonnegative. At least one error term is positive, $\sum_n \epsilon_{n,k}^2 > 0$.
Analysis: As β increases, $w_{n,k}^\beta, \forall n,k$ decreases, but never to zero. We cannot solve $\frac{\partial f}{\partial \beta} = 0$ for β. When β is large the MSWE emphasizes the largest weight to the extent that one weight may dominate the error, $f \approx \max f_{n,k}$. A trivial minimization strategy for Eqn. 16 is to set β to be a large positive value. This approach uses the instantaneous estimate corresponding to the largest weight as the estimate. This estimate may be incorrect. We trade-off (1) the number of significantly active weights $w_{n,k}^\beta$, the relative influence of the errors $\epsilon_{n,k}$ on our estimate, with (2) using β to prioritize large weights, e.g. estimates $x_{n,k}$ which we posit are more likely to be accurate because the source is relatively *more* dominant in that TF bin. We choose β by solving

$$f' = \left|\frac{\partial f}{\partial \beta}\right| = \left|\frac{2}{NK}\sum_{n,k}\epsilon_{n,k}^2\left(\frac{w_{n,k}}{a_k}\right)^{2\beta}\log\frac{w_{n,k}}{a_k}\right| = \theta. \tag{17}$$

The slope of the MSWE is $-\theta$ at β_θ. The magnitude of the derivative, f', decreases as β increases. When $\beta < \beta_1$ the rate of minimization of the MSWE is fast because $f' > 1$. When $\beta > \beta_1$ the rate of minimization of the MSWE is slow because $f' < 1$. The trade-off between these two regimes of f as a function of β, that achieves parity between the rate of improvement of the MSWE and the *participation* of the

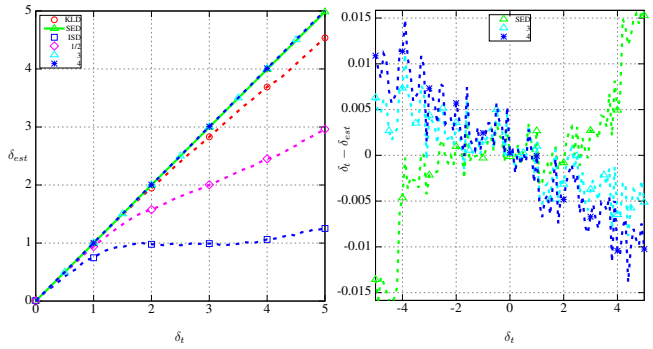

Fig. 2. The LHS illustrates the average relative delay estimates, δ_{est} obtained by using the KLD, SED, ISD and $\beta = \{.5, 3, 4\}$ for true relative delays, δ_t, in the range, $0 \leq \delta_t \leq 5$. The RHS illustrates, $\delta_t - \delta_{est}$, for the best performing estimators, the SED and $\beta = \{3, 4\}$ in the range $-5 \leq \delta_t \leq 5$.

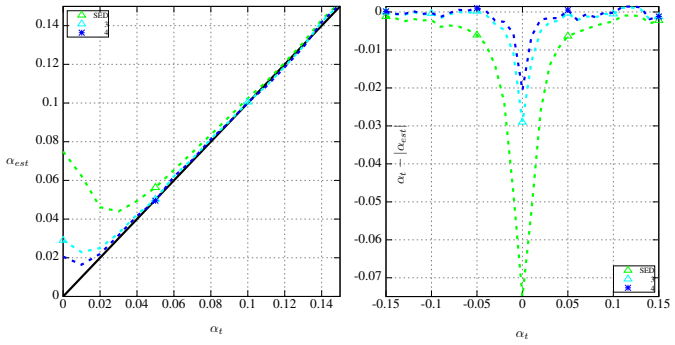

Fig. 3. The LHS illustrates the average relative attenuation estimates, a_{est} obtained by using the KLD, SED, ISD and $\beta = \{.5, 3, 4\}$ for true relative attenuations, α_t in the range $0 \leq \alpha_t \leq .15$. The RHS illustrates, $\alpha_t - |\alpha_{est}|$, for the best performing estimators, the SED and $\beta = \{3, 4\}$ in the range, $-.15 \leq \alpha_t \leq .15$.

weights is illustrated in Fig. 1 (RHS). We examine the case where the $\theta = \epsilon$, an arbitrarily small value. Choosing $\beta > \beta_\epsilon$ has limited pay-off in terms of MSWE reduction, but it limits the contribution of small weights.

V. Numerical Evaluation

We experimentally evaluate the element-wise weighted Bregman divergence estimators using speakers from the TIMIT database, which are sampled at 16kHz. We compute the STFT using a window of length 1024 samples. We examine the best divergence for each speaker using the approach in App. B of [1]. We generate mixtures using the source-interferer mixing model in Eqn. 3. This model is valid for the dominant TF points of one source according to [1]. We add iid Gaussian white noise to the dominant TF points of the target source on both channels. We adjust the noise energy to 9.87dB to model the effect of 4 interfering sources. In effect, we fix the number of sources as 5 and find the best β. We use the 0dB mask [1] to determine the set of dominant TF points Λ_j. This approach has the advantage that the results do not depend on any particular choice of interfering sources or mixing parameters, and we can compare the estimates with the ground-truth parameters. An evaluation of the MLEs was presented in [1].

The LHS of Fig. 2 illustrates the true relative delay, in the range $0 \leq \delta_t \leq 5$ samples, when $\alpha = 1$ (Fig. 3 illustrates the true relative attenuation, $0 \leq \alpha_t \leq .15$, when $\delta_t = 0$ samples)

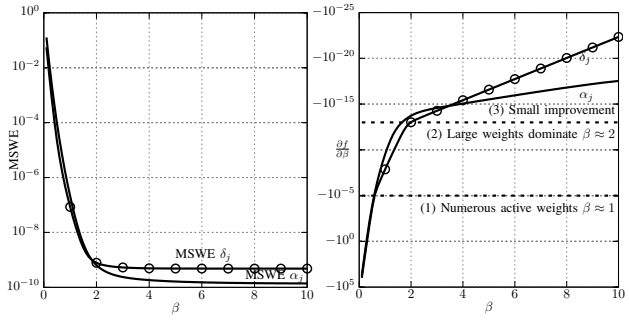

Fig. 4. **Selecting** β: The MSWE for δ_j and α_j (LHS) are illustrated. The derivatives of the MSWEs, illustrated on the RHS, increase rapidly in the range $\beta < 2$. They flatten rapidly when $\beta > 2$ which suggests that $\beta > 2$ is suitable for speech. Three regimes (1), (2) and (3) are indicated with text.

versus the average estimated relative delay δ_{est} for each of these true relative delays (and attenuations α_{est} respectively). Each figure plots the average estimate from 100 independently generated mixtures. The RHS panel in each figure zooms in on the three best estimators and plots the difference between the true and the estimated parameter. Fig. 4 illustrates the MSWE (Eqn. 16) for one TIMIT speaker for relative attenuation and delay estimation, by averaging the MSWE over $0 \leq \delta_t \leq 5$ and $0 \leq \alpha_t \leq .15$ respectively. We plot $\frac{\partial f}{\partial \beta}$ for each estimator beside this figure to illustrate the best trade-off for β. The $\beta = .5$, the KLD and ISD estimators perform poorly for both relative attenuation and delay estimation. These divergences correspond to a MSWE slope magnitude of $\theta \approx 1$. The best estimator for relative delay estimation (cf. Fig. 2) is parametrized by $2 < \beta < 3$, which corresponds to $\theta \approx 10^{-12}$ in Fig. 4. The most accurate relative attenuation estimates lie in the range $\beta > 4$, which corresponds to $\theta < 10^{-12}$. The slight difference in the range of best β trade-offs is due to the fact that the weights themselves are functions of the true parameters, which are slightly different for the relative delay and attenuation estimation. In summary, the role of β is highly influential in accuracy of these relative parameter estimators the range $\beta < 2$ for speech. The magnitude of the MSWE slope rapidly goes to zero above $\beta > 2$, irrespective of the normalization factor of the weights. The sensitivity of an estimator to values of $\beta > 2$ depends on the WDO of the source. We have demonstrated that $2 \leq \beta < 4$ gives good parameter estimates. This contradicts the suggested default choice by the authors of [1], e.g. $\beta = 1$.

VI. Conclusion

We have provided a justification for the form of the parameter estimators used by the DUET algorithm, by stating the optimization problem that gave rise to them. Our mathematical formalism is naturally extendable to other distributions. We demonstrated that selecting $2 \leq \beta \leq 4$ gave a good trade-off between estimation accuracy and the relative influence of the instantaneous estimates used.

References

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

[2] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP. J. Adv. Sig. Proc.*, vol. 19, no. 86484, 2007.
[3] A. Blin, S. Araki, and S. Makino, "A sparseness-mixing matrix estimation (SMME) solving the underdetermined BSS for convolutive mixtures," *IEEE ICASSP*, vol. 4, pp. 85–8, 2004.
[4] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acous. Soc. Am.*, vol. 25, no. 5, pp. 975–9, 1953.
[5] R. de Fréin and S. Rickard, "The Synchronized Short-Time-Fourier-Transform: Properties and Definitions for Multichannel Source Separation," *Sig. Proc., IEEE Trans.*, vol. 59, no. 1, pp. 91–103, 2011.
[6] A. Cichocki and S. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, pp. 1532 – 68, 2010.
[7] F. Itakura and F. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Int. Congr. Acoust.*, pp. 17 – 20, 1968.
[8] A.B.S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *J. of Mach. Learn. Res.*, vol. 6, pp. 1705 – 49, 2005.
[9] A. Basu, I.R. Harris, N. Hjort, and M. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, pp. 549 – 59, 1998.
[10] R. de Fréin, "Source separation approach to video quality prediction in computer networks," *IEEE Comms. Ltr.*, vol. 20, no. 7, pp. 1333–6, Jul. 2016.