ESTIMATING TIMING AND CHANNEL DISTORTION ACROSS RELATED SIGNALS

Colin Raffel, Daniel P. W. Ellis

LabROSA, Dept. of Electrical Engineering
Columbia University
{craffel, dpwe}@ee.columbia.edu

ABSTRACT

We consider the situation where there are multiple audio signals whose relationship is of interest. If these signals have been differently captured, the otherwise similar signals may be distorted by fixed filtering and/or unsynchronized timebases. Examples include recordings of signals before and after radio transmission and different versions of musical mixes obtained from CDs and vinyl LPs. We present techniques for estimating and correcting timing and channel differences across related signals. Our approach is evaluated in the context of artificially manipulated speech utterances and two source separation tasks.

Index Terms— Audio Recording, Optimization, Source Separation, Signal Reconstruction, Microphone Arrays

1. INTRODUCTION

There are a number of scenarios in which we may have several related audio signals that we would like to precisely align in order to fully characterize their relationship, or to isolate their differences. The signals may have a common source, but have been subjected to different processing (including independent additions), or a single acoustic event may have been recorded by multiple sensors yielding related but different signals. We consider the problem of estimating and correcting the relationship where both timing and channel have been modified.

In the simultaneous capture setting, an acoustic scene is recorded by separate devices. The sampling timebases of the resulting signals may not be synchronized and will frequently differ by several hundred parts per million, depending on the quality of the quartz oscillators used, which can amount to drifts of a second or more over longer recordings. Miyabe et. al. [2] considered this problem in the context of an ad-hoc microphone array composed of separate devices. After performing a coarse estimation of the sample rate offset, they apply optimal frame-level linear-phase filters to correct sub-sample timing drift. They found that compensating for the timing drift greatly improved performance in a source separation task.

Signals containing a common source also occur in the domain of blind source separation, where the separation algorithm may have no way to identify a fixed coloration of a separated source. To accommodate this, the BSS_EVAL toolkit [1] estimates an optimal linear projection of the output onto target (and optionally interference) components to obtain performance metrics invariant to fixed filtering, given the original clean signal as a reference.

This work was supported in part by NSF grant IIS-7015. The authors also thank Hacking Audio and Music Research for supporting preliminary work on this project.

Instrumental and a cappella "mixes" of pieces of popular music are often released alongside the complete original mix. These mixes contain only the non-vocal and vocal sources respectively (in contrast with instrumental and a cappella arrangements, which are non-vocal and vocal reinterpretations of the original composition). The comprehensive online music discography Discogs.com lists over 200,000 releases containing an instrumental mix but only about 40,000 which include an a cappella mix. The availability of these separated mixes are crucial in the creation and performance of some genres of music [3, 4, 5]. These instrumental and a cappella versions can also be used as ground-truth for vocal removal or isolation algorithms [6].

The disparity in the number of available instrumental and a cappella mixes suggests that it would be beneficial to have a general technique for removing or isolating the vocal track in a recording of a piece of music when only one or the other is available. A simple approach is proposed in [5] where an optimally shifted and scaled instrumental mix is subtracted from the complete mix in the time or frequency domain in attempt to obtain a (previously unavailable) a cappella mix. However, this approach does not cover the more general case where different mixes may be extracted from different media (e.g. vinyl records and compact discs), which results in a varying time offset as well as a media-specific channel distortion. In addition, the different mixes may have differing equalization and nonlinear affects applied [7], causing further channel distortion.

The process of isolating or removing vocals using an instrumental or a cappella mix can be viewed as a source separation problem with a great deal of prior information. While completely blind source separation has seen a great deal of recent research focus, systems incorporating prior information have also been developed. For example, vocalized imitations [8] and artificially synthesized approximations [9] of the source of interest have been used as priors to improve separation results. Similarly, exactly [10] and approximately [11] repeated patterns in a piece of music have been used to improve the extraction of the remaining varying components.

We can formalize each of these settings by letting $m[n], c[n] : n \in \mathbb{Z}$ be two discrete-time signals which are assumed to be sampled from bandlimited underlying continuous signals $c_a(t)$ and $m_a(t)$, and which have some content in common. We are interested in extracting information about their relationship, but there is undesirable timing and channel distortion applied to c relative to m. Assume that m was captured with a constant sampling period of T, while the sampling rate of c varies in time relative to m resulting in a time-varying offset. We denote $\phi[n]$ as the offset (in real-valued samples) of c relative to m at sample n. In the process of capturing these signals some channel distortion $\mathbf{D}$ relative to m was also applied to c, so that

$$c[n] = \mathbf{D}(c_a((n + \phi[n])T)$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

We are interested in estimating $\mathbf{D}^{-1}$ and ϕ^{-1} so we may remove the channel distortion and sample rate drift present in c.

2. PROPOSED SYSTEM

In this section, we describe a general system for estimating the functions $\mathbf{D}^{-1}$ and ϕ^{-1} described above. In particular, we model $\phi[n]$ as a piecewise linear function in a two-step process by first estimating any large-scale drift and then estimating local offsets between m and c. We then estimate $\mathbf{D}^{-1}$ in the frequency domain by minimizing a convex error function to obtain a complex filter which minimizes the residual between m and c. Finally, we optionally use Wiener filtering for post-processing when the distortion is substantially nonlinear.

2.1. Timing Offset

If the timing distortion caused by ϕ is particularly extreme (i.e. highly nonlinear), the problem of reversing its effect may be intractable. However, in the applications discussed in Section 1, the nonlinear characteristics of ϕ are relatively mild. For example, in the simultaneous capture setting, the primary contributor to ϕ is the recorder's clock drift, which will result in a predominantly linear function of n. As a result, we model ϕ as a piecewise linear function.

We first attempt to compensate for the global difference in effective sampling rate by resampling $c[n]$ as in [2]. A straightforward way to choose the optimal resampling rate f^* would be to maximize the cross-correlation

$$f^* = \arg\max_f \max_\ell \sum_n m[n]\mathcal{R}_f(c)[n-\ell]$$

where $\mathcal{R}_f(c) = c_a(fnT)$ denotes resampling c by a factor f. Unfortunately this problem is non-convex, so we perform a linear grid search over a problem-specific range of values of f close to 1 to obtain f^*.

Once we have obtained $c_\mathcal{R} = \mathcal{R}_{f^*}(c)$, we are left with the nonlinear effects of ϕ. We can estimate this function by computing the local offset (in samples) of $c_\mathcal{R}$ with respect to m at some regular interval. In this way, we can obtain a sequence $\mathcal{L}[k]$ denoting the offset of $c_\mathcal{R}$ with respect to m at sample k. We can choose $\mathcal{L}[k]$ by finding the lag ℓ which maximizes the cross-correlation between m and $c_\mathcal{R}$ in a small window around k. This process has also been used to estimate and model the effect of radio transmission on a clean signal, where the recording of the received signal exhibited some timing drift relative to the source [12]. Specifically, we set

$$\mathcal{L}[k] = \arg\max_\ell \sum_{n=k-W}^{k+W} m[n]c_\mathcal{R}[n-\ell]$$

where $\ell, W \in \mathbb{Z}$, and W controls the window over which we compute the unbiased cross-correlation.

This optimization is non-convex and must therefore also be solved using an exhaustive search. In practice, we constrain ℓ to be in a range $[-L, L]$ based on our experience of the largest offsets encountered. Computing the cross-correlation is relatively expensive, so, based on our assumption that ϕ is slowly-changing, we only compute $\mathcal{L}[k]$ every K samples so that $k = \{0, K, 2K, \ldots\}$. We then assume a linear interpolation for intervening values; although the computed values of $\mathcal{L}[k]$ will be integers, the interpolated values may be fractional. We can apply these offsets to construct $c_\mathcal{O}[n] = c_\mathcal{R}[n - \mathcal{L}[n]]$ where we use windowed sinc interpolation to calculate the non-integral sample values [13].

2.2. Channel Distortion

Our estimation of $\mathbf{D}$ is based on the assumption that it is a linear, time-invariant filter; fortunately, in our applications of interest this is a usable assumption. In the case of isolating or removing vocals using available a cappella or instrumental mixes, much of the nonlinearity of $\mathbf{D}$ will be caused by the relatively minor effects (dynamic range compression, excitation, etc.) applied during mastering. We therefore can approximately invert $\mathbf{D}$ by estimating $\mathbf{D}^{-1}$ as a complex filter H in the frequency domain.

To compute H, we can exploit the fact that our signals m and c (and therefore $c_\mathcal{O}$) will be dominated by the same signal components at least for a large number of time-frequency points (i.e., those in which the additional components have zero or low energy). Thus we are looking for an H which makes m very close to $c_\mathcal{O}$ over as much of the signal as possible. If we denote $M[k]$ and $C_\mathcal{O}[k]$ as the kth frame of the short-time Fourier transform of m and $c_\mathcal{O}$ respectively, an intuitive approach would be to solve

$$H^* = \arg\min_H \sum_k |M[k] - H \odot C_\mathcal{O}[k]| \tag{1}$$

where $\odot$ indicates element-wise product and $|\cdot|$ indicates both magnitude and $L1$ norm computation. This effectively requires that the difference between M and $C_\mathcal{O}$ filtered by H is sparse in the frequency domain. The use of an $L1$ norm also results in the objective being less sensitive to outliers (compared to e.g. an $L2$ norm), which is important when we expect there to be components of m not in c or vice-versa. This approach has also been used for speech dereverberation [14]. This objective function is a sum of independent convex functions of each term $H[i]$ of H and is therefore convex and can be solved efficiently. In practice, we use the L-BFGS-B algorithm [15] for minimization.

Once we have computed H^*, we can apply it to $C_\mathcal{O}$ in the frequency domain for each frame k to compute $C_\mathcal{F}[k] = H^* \odot C_\mathcal{O}[k]$ from which we can obtain $c_\mathcal{F}[n]$ by computing an inverse short-time Fourier transform. If we are interested in the components of m which are not in c (as in the source separation case), we can now obtain their approximation by computing $\hat{s}[n] = m[n] - c_\mathcal{F}[n]$.

2.3. Post-Processing

In cases where $\mathbf{D}$ is nonlinear and/or estimation of $\mathcal{L}$ is inaccurate due to the interfering components, the estimation procedures described above may not be able to exactly invert their effects leading to residual interference in $\hat{s}$. However, provided that m and $c_\mathcal{F}$ are closely aligned in time, we can suppress components of $c[n]$ which remain in $\hat{s}[n]$ using Wiener filtering. Specifically, if $\hat{S}$ is the short-time Fourier transform of $\hat{s}[n]$, let

$$R = \frac{1}{\tau}(20\log_{10}(|\hat{S}|) - 20\log_{10}(|C_\mathcal{O}|) - \lambda)$$

where τ is the Wiener transition and λ is the Wiener threshold, both in deciBels. R is negative for time-frequency cells where $C_\mathcal{O}$ is large relative to $\hat{S}$. Thus, we can compute

$$\Omega = \frac{1}{2} + \frac{R}{2\sqrt{1+R^2}}$$

and we can further suppress compoents that align to energy in $C_\mathcal{O}$ by computing the inverse short-time Fourier transform of $\hat{S} \odot \Omega$.

3. EXPERIMENTS

To test the effectiveness of this approach, we carried out three experiments covering the applications mentioned in Section 1. First, we reversed synthetic resampling and filtering applied to speech utterances to mimic the conditions encountered in simultaneous capture settings. We then tested our technique for vocal isolation (i.e. extracting the vocals from the full mix) and vocal removal on real-world music data in both mildly and substantially distorted situations.

3.1. Synthetic Speech Data

In the simplest case, neither the timing distortion nor the channel distortion will be nonlinear. This closely matches a scenario when independent recorders are used to capture a dominant acoustic source. Since this matches our assumptions, we expect to be able to undo such distortion almost perfectly. To test this assertion, we generated 100 recordings by concatenating independent sets of 10 sentences from the TIMIT corpus [16]. We then resampled each recording by a random factor in the range $[.98, 1.02]$ and convolved it with a randomly generated 10-point causal filter h of the form

$$h[n] = \begin{cases} 1, & n = 0 \\ e^{-n} r[n], & 0 < n < 10 \\ 0, & n > 10 \end{cases}$$

where each $r[n] \sim \text{Normal}(0, 1)$ is a Gaussian-distributed random variable with mean 0 and variance 1.

For each of our synthetically distorted recordings, we estimated $\mathbf{D}^{-1}$ and ϕ^{-1} using our proposed system. Because ϕ is strictly linear, we did not estimate $\mathcal{L}$ in this case (i.e., we set $\mathcal{L}[k] = 0 \; \forall k$). All utterances were sampled at 16 kHz and all short-time Fourier transforms were computed with 16 ms Hann-windowed frames taken every 4 ms. To evaluate our estimation of ϕ, we calculate the percentage error in our optimal resampling factor f^*. We can also determine the extent to which we were able to reverse the effects of ϕ and $\mathbf{D}$ by comparing the RMS of the residuals $m[n] - c[n]$ and $m[n] - c_{\mathcal{F}}[n]$.

Our system recovered the resampling factor exactly in 72 out of 100 cases; on average, the error between the estimated resampling factor and the true factor was 1.6%. The average RMS across all recordings of the residual $m[n] - c[n]$ was 0.174, while the average RMS of $m[n] - c_{\mathcal{F}}[n]$ was only 0.0162. The system had more difficulty estimating the filter in the 28 cases where the resampling factor was not estimated correctly; in these cases, the average RMS of $m[n] - c_{\mathcal{F}}[n]$ was 0.057. This suggests that even when ϕ is not recovered exactly, we are still able to produce a reasonable estimate of h. The frequency response H of a random filter and its estimate $\hat{H}$ using the procedure outlined in Section 2.2 are shown in Figure 1.

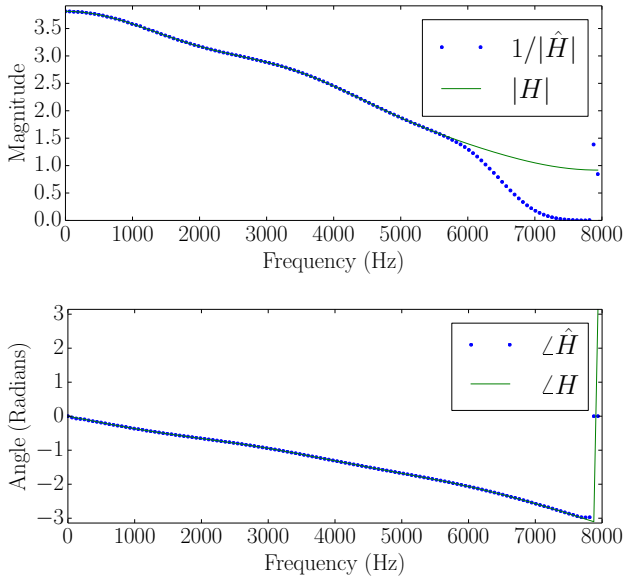

Fig. 1. Magnitude and phase response of an example of a randomly generated filter h, generated as described in Section 3.1, alongside the estimated filter $\hat{h}$. The deviations at high frequencies arise because the speech signals have virtually no energy in these regions.

3.2. Digital Music Separation

To test the importance of estimating ϕ and $\mathbf{D}$ in a real-world scenario, we focused on the isolation and removal of vocals from music signals using an instrumental or a cappella mix where all signals are sourced from the same compact disc. In this setting, we do not expect any timing distortion ϕ because the signals should be derived from the same sources without any steps likely to introduce timing drift. As a result, we may be able to achieve good vocal isolation or removal by simply subtracting the two signals at an appropriate single time offset. However, differences in the processing applied to the different mixes may make $\mathbf{D}$ substantial, making the estimation of $\mathbf{D}^{-1}$ useful.

We extracted 10 examples of instrumental, a cappella , and full mixes of popular music tracks from CDs to produce signals sampled at 44.1 kHz. In order to compensate for any minor clock drift caused during the recording of these signals, we estimated the optimal resampling ratio f^* over a range of $[.9999, 1.0001]$. We then estimated the local offsets every 1 second by computing the cross-correlation over 4 second windows with a maximum allowable offset of 100 ms. Finally, we computed the the optimal channel filter H^* using short-time Fourier transforms with Hann-windowed 92.9 ms frames (zero-padded to 186 ms) computed every 23.2 ms.

For each track, we estimated ϕ and $\mathbf{D}$ of the instrumental and a cappella mix with respect to the original mix $m[n]$ to obtain $c_{\mathcal{F}}[n]$ and computed $\hat{s}[n] = m[n] - c_{\mathcal{F}}[n]$ to isolate or remove the vocals respectively. Because we are assuming that there may be timing and channel distortions in both the a cappella and instrumental mixes, we also estimate the distortion in the to the "true" source $s[n]$ to obtain $s_{\mathcal{F}}[n]$. Wiener filter post processing was not needed or used in this setting. The frequency response of a typical estimated channel distortion filter H^* is shown in Figure 2.

To measure the performance of our separation, we used SDR (signal-to-distortion ratio) [1]. SDR computes the energy ratio (in deciBels) of the target source relative to artifacts and interference present in the estimated source. To examine the individual contributions of estimating ϕ and $\mathbf{D}$, we computed the SDR of both $m[n] - c_{\mathcal{O}}[n]$ and $m[n] - c_{\mathcal{F}}[n]$, and subtracted the SDR of $m[n] - c[n]$ to obtain an SDR improvement for each condition. All SDRs were computed relative to $s_{\mathcal{F}}[n]$. Figure 3 shows these results, where each line represents the SDR trajectory for a single example. Both timing and filter estimation gave separate improvements in most cases, indicating both are necessary for these data, but there is substantial variation among individual tracks.

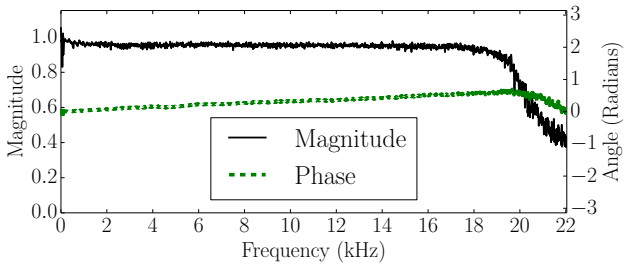

Fig. 2. Magnitude and phase response of a typical filter estimate H^* of the channel distortion between an a cappella mix and the full mix. The linear trend in the phase response indicates a sub-sample offset.

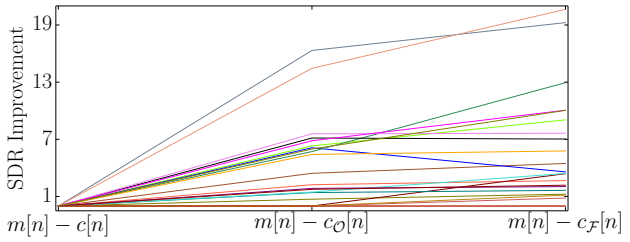

Fig. 3. Improvement of SDR (in dB) due to inverting timing and channel distortion. The SDR for each example at each stage was normalized by the SDR of $m[n] - c[n]$ to show the relative improvement caused by each step.

3.3. Vinyl Music Separation

A more challenging application for our technique arises when trying to isolate or remove vocals using an instrumental or a cappella mix which has been recorded on a vinyl record. The signal captured from a vinyl recording will vary according to the playback speed, needle, and preamplifier circuit which results in substantial timing and channel distortion. We carried out an experiment similar to Section 3.2 except that the instrumental and a cappella mixes used to extract and remove the vocals were sourced from vinyl recordings. Both the original mixes and the reference signals were extracted from compact discs to minimize distortion present in our ground truth. Note that there will be some timing and channel distortion of our reference signal relative to the original mix (as described in Section 3.2) but the distortion present in the compact disc format is insubstantial compared to that of the vinyl format.

To obtain digital representation of the vinyl recordings, we digitized the playback of the record at a sampling rate of 44.1 kHz. The original mix and ground-truth signals were extracted directly from CDs also at 44.1 kHz. As above, we first estimated the resampling ratio f^* which optimally aligned the vinyl signal to the original mix, except here we allowed for ratios in the range [0.98, 1.02]. We then estimated the local offsets using the same process and parameters as in Section 3.2. As expected, the resulting local offset sequences $\mathcal{L}[k]$ were often non-linear due to variance in the turntable motor speed. An example of the short-time cross-correlation is shown in Figure 4.

Once the signals were aligned in time, we estimated the optimal complex filter H^* using the same procedure as in Section 3.2. However, due to the substantial nonlinearities present in vinyl recordings, the resulting sequence $c_{\mathcal{F}}[n]$ did not sufficiently cancel or isolate

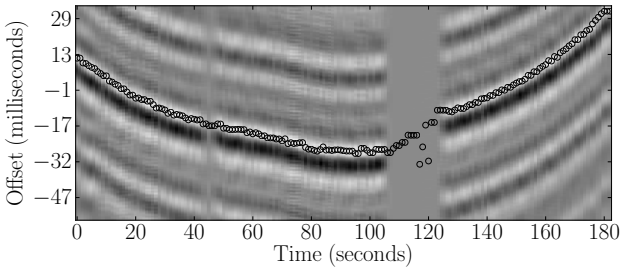

Fig. 4. Local cross-correlation of m against $c_{\mathcal{R}}$. Lighter colors indicate larger correlation, with black circles indicating the maximum correlation. The grey region between 105 and 125 seconds corresponds to a portion of c which has low energy.

the vocals when subtracted from $m[n]$. Thus, we further applied the Wiener filter post-processing of Section 2.3, based on short-time Fourier transforms with 46 ms Hann-windowed frames computed every 12 ms, and using a threshold $\lambda = 6$ dB over a $\tau = 3$ dB transition.

We carried out this procedure for 14 tracks, 7 each of vocal isolation and removal. The resulting SDRs are presented in Table 1. In general, our approach was extremely effective at removing vocals. For reference, typical SDRs achieved by state-of-the-art blind source separation algorithms (which are disadvantaged because they do not exploit any prior information) are around 3 dB [6, 11]. The SDRs for the vocal isolation examples were generally lower, which is likely due to the more varied frequency content of the instrumental component we are trying to remove. As a result, we also computed the SDR for the vocal extraction examples after high pass filtering the extraction with a 24 dB/octave filter with cutoff set at 216 Hz, as is done in [11]. This improved the SDR by about 1 dB in all cases.

Task	Mean $\pm$ SDR
Vocal Removal	11.46 $\pm$ 3.59 dB
Vocal Isolation	5.14 $\pm$ 1.69 dB
Vocal Isolation (Filtered)	6.37 $\pm$ 1.46 dB

Table 1. Mean and standard deviations of SDR values for vocal removal and isolation using instrumental and a cappella mixes sourced from vinyl records.

4. CONCLUSION

We have proposed a technique for estimating and reversing timing and channel distortion in signals with related content and proved its viability in settings of varying difficulty. In particular, we approximated the timing distortion with a piecewise linear function by computing local offsets and estimated the channel distortion with a complex frequency-domain filter found by solving a convex minimization problem. All of the code and data used in our experiments is available online so that the proposed techniques can be easily applied to any situation where precise alignment and channel distortion reversal of related signals is needed.[1]

[1] `http://github.com/craffel/remixavier`

5. REFERENCES

[1] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "BSS_EVAL toolbox user guide - revision 2.0," Tech. Rep. 1706, IRISA, April 2005.

[2] Shoji Makino Shigeki Miyabe, Nobutaka Ono, "Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[3] Peter Manuel and Wayne Marshall, "The riddim method: aesthetics, practice, and ownership in jamaican dancehall," *Popular Music*, vol. 25, no. 3, pp. 447, 2006.

[4] Philip A Gunderson, "Danger mouse's grey album, mash-ups, and the age of composition," *Postmodern culture*, vol. 15, no. 1, 2004.

[5] Hung-Ming Yu, Wei-Ho Tsai, and Hsin-Min Wang, "A query-by-singing system for retrieving karaoke music," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1626–1637, 2008.

[6] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbyněk Koldovskỳ, Guido Nolte, Andreas Ziehe, and Alexis Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): Audio source separation," in *Latent Variable Analysis and Signal Separation*, pp. 414–422. Springer, 2012.

[7] Bob Katz, *Mastering audio: the art and the science*, Taylor & Francis US, 2007.

[8] Paris Smaragdis and Gautham J Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 69–72.

[9] Joachim Ganseman, Paul Scheunders, and S Dixon, "Improving plca-based score-informed source separation with invertible constant-q transforms," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2634–2638.

[10] Sean Coffin, "Separation of repeating and varying components in audio mixtures," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[11] Zafar Rafii and Bryan Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1-2, pp. 73–84, 2013.

[12] Daniel P. W. Ellis, *RENOISER - Utility to decompose and recompose noisy speech files*, http://labrosa.ee.columbia.edu/projects/renoiser/, 2011.

[13] Julius Smith and Phil Gossett, "A flexible sampling-rate conversion method," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1984, vol. 9, pp. 112–115.

[14] Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel D Lee, "Blind channel identification for speech dereverberation using l1-norm sparse learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 921–928.

[15] Dong C Liu and Jorge Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[16] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.