ESTIMATING TIMING AND CHANNEL DISTORTION ACROSS RELATED SIGNALS

Colin Raffel and Daniel P. W. Ellis
craffel@ee.columbia.edu, dpwe@ee.columbia.edu

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University
in the City of New York

Abstract

Many problems involve multiple signals whose relationship is of interest but have been differently captured. As a result, the otherwise similar signals may be distorted by fixed filtering and/or unsynchronized timebases. We present techniques for estimating and correcting timing and channel differences across related signals.

Timing Distortion

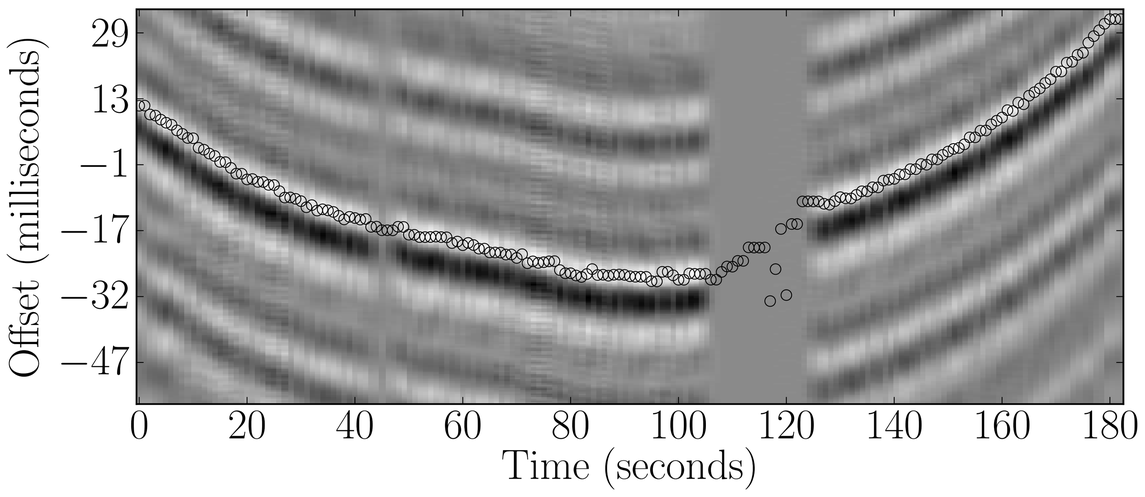
First, we find the resampling ratio f^* which maximizes the cross-correlation between two signals m and c whose timebase is relatively distorted:

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

where $\mathcal{R}_f(c)$ denotes resampling c by a factor of f. In practice, we perform a linear grid search over a problem-specific range of values of f close to 1 to obtain f^*. Second, we estimate local offsets between m and $c_\mathcal{R} = \mathcal{R}_{f^*}(c)$ by maximizing their local cross-correlation over a window of size W:

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

Values of $\mathcal{L}[k]$ are found by exhaustive search. We constrain ℓ to be in a range $[-L, L]$ based on our experience of the largest offsets encountered. Because the timing distortion tends to change slowly over time 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. Finally, we apply the local offsets to $c_\mathcal{R}$ to obtain $c_\mathcal{O}$.



Channel Distortion

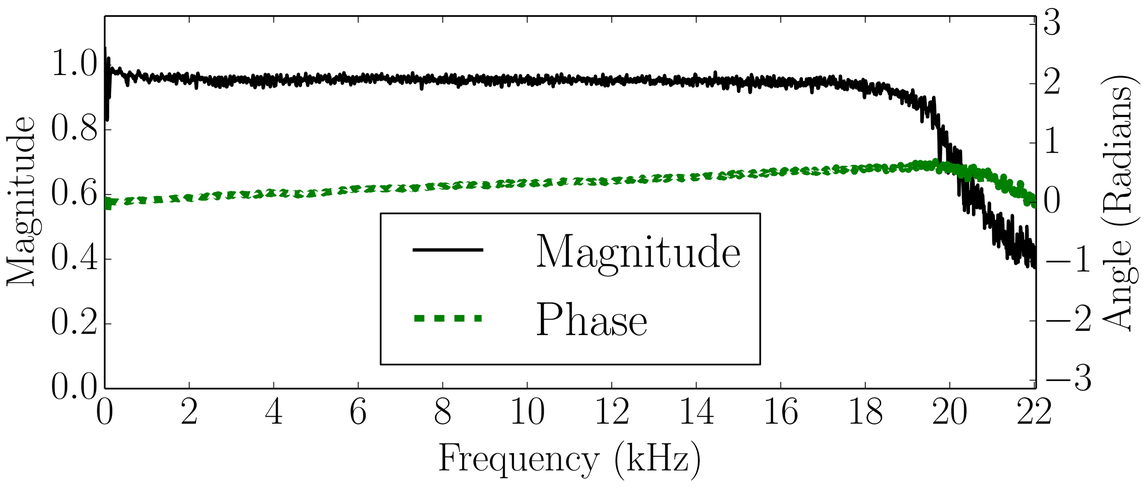
We approximate the channel distortion as a complex filter H which we estimate by optimizing

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

where $M[k]$ and $C_\mathcal{O}[k]$ are the kth short-time Fourier transform of m and $c_\mathcal{O}$ respectively. This objective can be solved by convex optimization. Once we have computed H^*, we can compute

$$C_\mathcal{F}[k] = H^* \odot C_\mathcal{O}[k]$$

from which we can obtain $c_\mathcal{F}[n]$.



Post-processing

When the channel distortion is nonlinear, we can apply Wiener filtering to further suppress its effects. If $\hat{S}$ is the short-time Fourier transform of $m[n] - c_\mathcal{F}[n]$, let

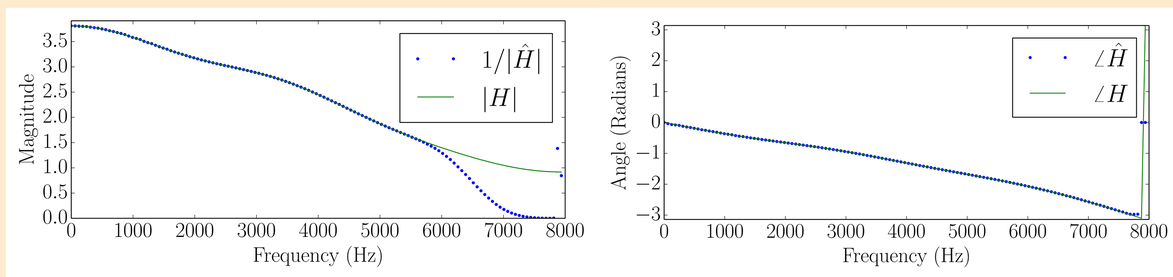
$$R = \frac{1}{\tau}\left(20\log_{10}(|\hat{S}|) - 20\log_{10}(|C_\mathcal{O}|) - \lambda\right)$$

which we can use to suppress the channel distortion by computing

$$\hat{S} \odot \left(\frac{1}{2} + \frac{R}{2\sqrt{1+R^2}}\right)$$

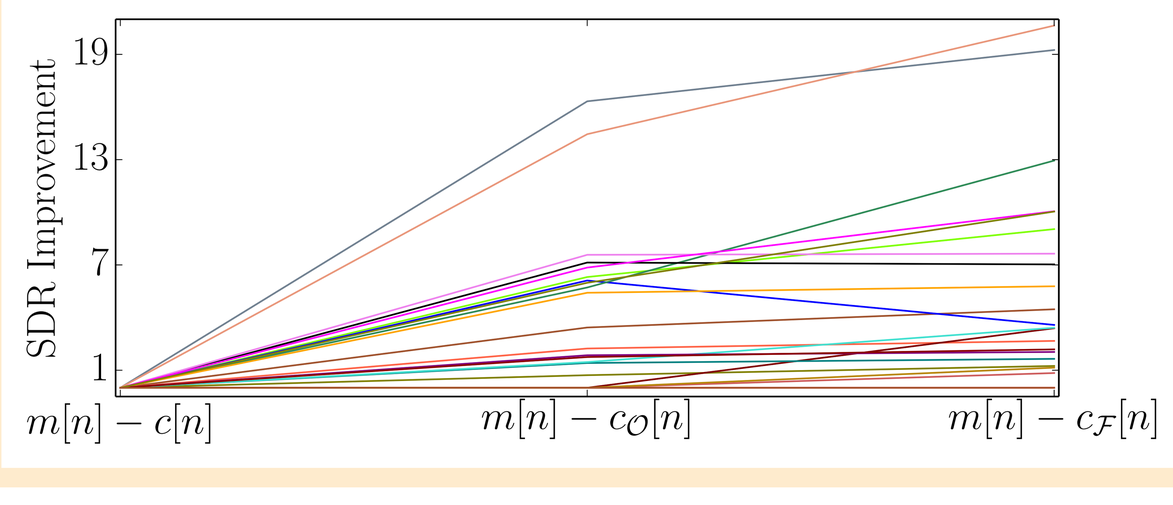
Experiment 1: Synthetic Speech Data

To test the simple case where neither the timing nor the channel distortion is nonlinear, we resampled 100 speech utterances by a random factor in the range $[.98, 1.02]$ and convolved them with a randomly-generated causal FIR filter. We then re-estimated the resampling factor and filter using the approach described above. 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.



Experiment 2: Digital Music Separation

We extracted 10 examples of instrumental, a cappella, and full mixes of popular music tracks from CDs. For each track, we estimated the timing and channel distortion 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. We also estimate the distortion in the to the "true" source $s[n]$ to obtain $s_\mathcal{F}[n]$. Finally, 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.



Experiment 3: Vinyl Music Separation

We also considered the case where the instrumental and a cappella mixes used to extract and remove the vocals were sourced from vinyl recordings. 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. Both the original mixes and the reference signals were extracted from compact discs to minimize distortion present in our ground truth. We carried out this procedure for 14 tracks, 7 each of vocal isolation and removal.

Task	Mean SDR
Vocal Removal	11.46 ± 3.59 dB
Vocal Isolation	5.14 ± 1.69 dB