Signal Component Estimation in Background Noise

R. R. Violanda, H. van de Vooren, R. A. J. van Elburg, and T. C. Andringa
Artificial Intelligence Department, University of Groningen
Groningen, The Netherlands, Email: r.violanda@ai.rug.nl

Introduction

Auditory scenes in a natural environment consist typically of multiple sound sources. From this the human auditory system can segregate and identify individual audible sources with ease even if the received signals are distorted due to noise and transmission effects [1]. Systems for computational auditory scene analysis (CASA) try to achieve human performance by decomposing a sound into coherent units and then grouping these to represent the individual sound sources in the scene [2-4]. Some CASA systems aim to estimate masks that enclose all regions dominated by a single sound source in the time-frequency (TF) plane [2]. This task becomes increasingly difficult when high levels of noise are present.

The aim underlying our approach to CASA is to estimate the physical properties of the source which give rise to its signal components. We define a signal component (SC) as a connected trajectory in the TF plane with a positive local signal-to-noise ratio (SNR). Ideally, if the sound stemming from a source is represented in the TF plane, the source's characteristics carried by the sound are reflected as a pattern of signal components [5]. Hence, by correctly tracking and grouping the signal components, it is possible to retrieve the physical properties that shaped the source signal.

In this paper, we present two methods of estimating signal components. The first method uses both the energy and phase information derived from the time-frequency representation of the signal. The second method uses solely the energy of the signal. We compare the effectiveness of the two methods in the extraction of target sound signals from noisy backgrounds. Our results show the importance of the often neglected phase information in TF analysis.

Methods

The most common pre-processing step in CASA is the transformation of sound signals represented by a time-series into their TF representation. Such a representation can be obtained using transformation techniques such as filterbanks [6] and short-time Fourier transform (STFT) [7]. As expected from the uncertainty relation between time and frequency these transformations cause a smearing of the energy in the TF plane. This introduces difficulties in tracking the evolution of frequency and energy and limits the possibilities to extract the physics of the sound production event. In the STFT transformation, optimizing the window parameter allows fine tuning of the resolution in either frequency or time, however this reduces the general applicability. Therefore we use phase information because it is less sensitive to the choice of the window parameters [8].

In addition it allows for the retrieval of instantaneous frequency (IF) and group-delay corrected time (GDCT) representation which can be used to remap the energy into narrower bands [9-10]. In the IF and GDCT representations the dominant TF bins will be close to the frequencies and temporal positions of the sound source and show a a reduced sensitivity to transmission effects. The local dominance of a signal component at a TF bin is determined by measuring the spread of the IF or GDCT over its neighboring TF bins [11], with a small spread indicating strong dominance. We exploit this phenomenon for the extraction of the signal components.

Signal component estimation using energy and phase information (SCEP)

To use phase properties in the extraction of signal components, it is necessary to calculate the phase as accurately as possible from the signal. To do this, we use the phase calculation method described by Auger et. al [12]. We apply two STFTs to the time series signal. The first STFT uses a Gaussian function $w(t)$ as a windowing function giving the time-frequency representation $X(t,f)$ where t and f refer to discrete time and frequency indices, respectively. The second STFT uses the time-derivative $dw(t)/dt$ of $w(t)$ as a windowing function giving the time-frequency representation $X'(t,f)$.

Based on the complex amplitudes $X(t,f)$ and $X'(t,f)$, the instantaneous frequency (f_{ins}) and group-delay corrected time (t_{gdc}) can be calculated as follows:

$$f_{ins}(t,f) = f + \frac{1}{2\pi}\Im\left\{\frac{X(t,f)X'(t,f)}{|X(t,f)|^2}\right\} \quad \text{[Hz]} \quad (1)$$

$$t_{gdc}(t,f) = t - \sigma^2\Re\left\{\frac{X(t,f)X'(t,f)}{|X(t,f)|^2}\right\}, \quad \text{[s]} \quad (2)$$

where σ is the standard deviation of the Gaussian window used in the STFT, and $\Re$ and $\Im$ denote the real and imaginary parts, respectively. The calculated f_{ins} corresponds to the frequency content of the sound source which dominates a few neighbouring frequency bins in the TF plane. The t_{gdc} corresponds to the actual temporal positions of pulses and fast chirping signals in the TF plane. The dominance of a source signal components can be determined by measuring the spread of the f_{ins} over a frequency range and the spread of t_{gdc} over a temporal duration.

We calculated spectral and temporal degrees of dominance (DoD) as

$$DoD_f(t,f) = -\log\left(\frac{\sum_{f'=f-\Delta f}^{f'=f+\Delta f}\left(f_{ins}(t,f')-f\right)^2\left|X(t,f')\right|^2}{\sum_{f'=f-\Delta f}^{f'=f+\Delta f}\left|X(t,f')\right|^2}\right) \quad (3)$$

$$DoD_t(t,f) = -\log\left(\frac{\sum_{t'=t-\Delta t}^{t'=t+\Delta t}\left(t_{gdc}(t',f)-t\right)^2\left|X(t',f\right|^2}{\sum_{t'=t-\Delta t}^{t'=t+\Delta t}\left|X(t',f\right|^2}\right). \quad (4)$$

Equation (3) defining the spectral degree of dominance was originally used by Nakatani et. al to estimate the pitch in a speech [11]. We extended their formulation to the temporal aspects of the signal which leads to our definition of the temporal degree of dominance in equation (4). Equations (3) and (4) yield high values if the spread of f_{ins} and t_{gdc} are small which indicates the presence of a dominant source. Signal components are extracted by applying a threshold of one standard deviation above the mean of all DoD points in the TF plane on the DoD matrices and subsequently identifying connected regions with high DoD values.

Signal component estimation using energy peaks (SCE)

To show the importance of phase in signal component estimation, we present an alternative method of extracting the signal components in the TF plane that only uses the energy information. This method is a simplified version of the commonly used sinusoidal modelling technique in speech and music analysis [7]. To extract the signal components, a spectrogram is obtained by applying an STFT to the time-domain signal. Energy peaks in the spectrogram are obtained by a peak tracking algorithm applied independently in frequency and temporal direction . The peaks are linked together to form signal components if the distance between peaks satisfy a certain criterion.

Sound Resynthesis

To determine how well the extracted signal components represent the original sound, we transform it back to a time series representation. This allows us to listen and compare the resynthesized sound to the original sound. The resynthesis uses the signal components location in the TF plane as a mask. To obtain the resynthesized sound, an inverse Fourier transform is applied to the masked complex amplitudes followed by and overlap-and-add operation [13].

Experiment

We determine the efficacy of both signal component extraction methods by testing it in a denoising experiment. We measure how much of the target signal can be retrieved by both signal component extraction processes . To test the applicability of the signal component extraction to a wide range of sounds, we used 100 different recordings of environmental sound sources as targets. These recordings were used in Gygi's studies on environmental sound [14]. Based on Gygi's categorization, the 100 sound samples were divided into 3 groups namely tonal sounds (harmonic category), impact sounds and continuous sounds [14]. All target sounds are mixed with 14 different levels of pink noise with signal-to-noise ratios (SNR) ranging from 25 dB to -10 dB.

Signal components were extracted using both SCEP and SCE. The results of the initial tests showed that signal components extracted from pure noise (generated pink and white noise) are often short in duration (for tonal behavior) or span only a few frequencies (for pulse-like behavior). We used this characteristic to remove noise by discarding signal components shorter than a threshold. The remaining signal components were labeled as target sound and used as mask for the resynthesis. The resynthesized sounds were then compared to the original target sounds to evaluate the amount of target sounds retrieved from the noise mixture.

The performance of extracting the target signal from the noisy mixture is computed using

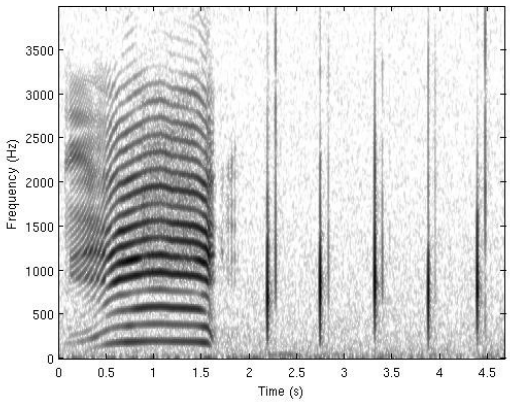
$$\text{Performance} = \frac{\langle x_t, x_r \rangle}{\left|\langle x_t, x_t \rangle\right|^{1/2}\left|\langle x_r, x_r \rangle\right|^{1/2}}, \quad (5)$$

where x_t is the time-series of target signal and x_r is the resynthesized signal based on the extracted signal components. The performance, has a value between 0 (uncorrelated signals) and 1 (perfectly matched signals).

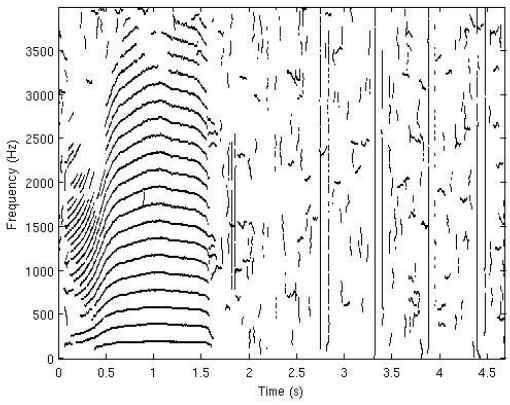
Results

To illustrate the signal extraction algorithm, we concatenated two sound files from sound sources that exhibit very different spectro-temporal behaviors. A spectrogram of this concatenated sound sample is shown in Figure 1a. The spectrogram shows the tonal structure of a mooing cow sound and the pulse-like structure of footsteps sound. Pink noise is also added to the concatenated sounds with 25 dB SNR. The pink noise can be seen as an energy gradient in the spectrogram, i.e., darker shades (high energy) in the low frequency part and lighter shades (low energy) in the high frequency part. The signal components extracted using energy and phase (SCEP) are shown in Figure 1b. The signal components of the cow sounds are the horizontal trajectories seen in the TF plane. It can also be observed that these trajectories make up a harmonic structure which is common property of many animal vocalizations. The vertical structures are the signal components of the footsteps sound. The signatures of footstep sounds can be identified as two subsequent pulses due to the dynamics of the impact of the heels and toes and a periodic repetition of these vertical structures. Figure 1c shows the signal components extracted with the energy peaks (SCE). Both SCEP and SCE capture the structures of the cow mooing sound and the footsteps sound. But after applying SCE there are still some signal components of the pink noise present. This indicates that SCEP removes noise signal components more efficiently than SCE, while nevertheless capturing the signal components of the relevant sound sources effectively.

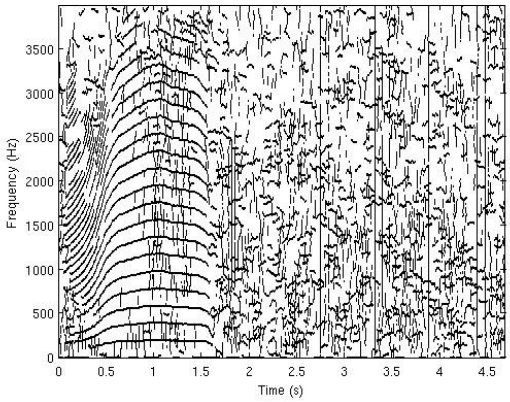
The performance of SCEP and SCE were evaluated separately on the 3 categories. Figure 2a depicts the performances of the SCEP (circles) and the SCE (squares) with respect to SNR using tonal sounds as targets. SCE has



(a)

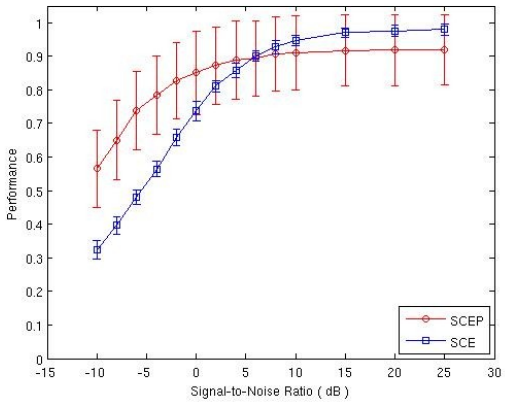


(b)

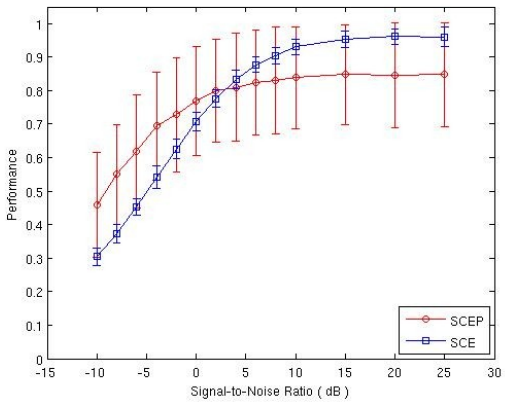


(c)

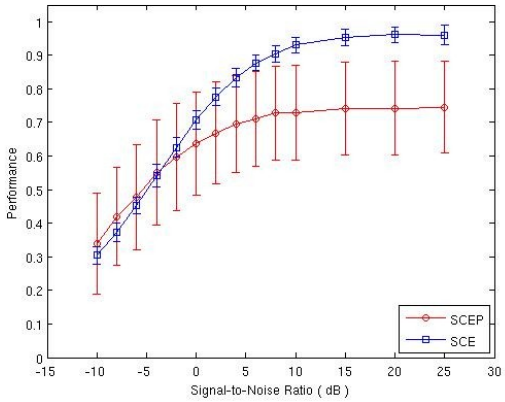
Figure 1. (a) A spectrogram of the concatenated cow mooing sound and footsteps sound. Dark shades correspond to high energies. (b). Signal components extracted using SCEP. (c). Signal components extracted using SCE.



(a)



(b)



(c)

Figure 2. Target sound extraction performances of the SCEP and SCE algorithm for different signal-to-noise ratios. (a) Tonal sounds as targets; (b) impact sounds as targets; (c) continuous sounds as targets. SCEP are represented by blue lines with squares and SCE are represented by red lines with circles.

higher mean performance at higher SNR, but decreases faster compared with the SCEP performance. Even at -10 dB SNR the SCEP performance is high (mean performance is well above 0.5). Figure 2b depicts the performances when impact sounds were used as targets. Still SCE outperforms SCEP at higher SNR, but again its performance decreases faster as SNR decreases. These sounds exhibit predominantly a spectro-temporal behavior similar to noise. The performances of both SCEP and SCE are comparable at lower SNR values, however, SCE has better performance as compared to SCEP at higher SNR. Out of the three classes of target sounds, SCEP achieves the best performance on tonal sound category and least performance on the continuous sound category. SCE, on the other hand, has almost the same performance rating for all the 3 classes. At high SNR values SCE always has a higher performance as compared to SCEP.

Discussion

We presented a method for extracting signal components from sound sources. This method is based on the energy and phase information calculated from a time-frequency analysis of the signal. For comparison an alternative signal component extraction solely based on the energy is also presented. Both methods were tested in a denoising experiment which measures the amount of target signal retrieved from a noisy mixture. Initial tests on clean artificially generated sound shows that both methods work equally well at higher SNR.

When the sound samples are mixed with noise a performance difference is observed. As depicted in Figure 1b and Figure 1c, SCE retains more of the noise as compared to SCEP. These unfiltered noise signal components cause a higher performance of SCE (a bias) over SCEP at high SNR since the majority of sound signal used in the experiment already contain a fair amount of noise prior to the addition of pink noise. SCEP effectively removes signal components of both the intrinsic noise and pink noise. Therefore SCEP based resynthesized sounds lacks the presence of intrinsic noise and thus leads to a lower performance score than SCE when both are compared to the original target sound. We also tested both methods on clean speech sound samples and found that both perform equally well at high SNR indicating that the bias is indeed caused by the presence of intrinsic noise.

With increasing noise level a decrease in performance is observed. In SCEP this decrease is mainly due to the interference between the noise and the target signal components. The interference degrades the coherence of the target's signal components which introduces discontinuities in the trajectories. These discontinuities fragment the signal component. The resulting signal component fragments will eventually be removed with the noise signal components. This problem is more severe for sounds in the continuous class which already exhibit a noise-like behavior and are therefore more sensitive to this interference effects. This explains the reduced performance of SCEP on continuous sounds.

In conclusion, a method of extracting signal components that uses both energy and phase information from the signal is more reliable compared to a method that uses energy alone, especially in noisy conditions. The methods work best on signals that have prominent tonal or a pulse-like structures. Tracking and analyzing the development of signal components constitutes a promising preprocessing step to reveal the physics of sound producing events.

Acknowledgement

HvdV is supported by the Dutch Technology Foundation STW. RvE and TA are (partly) supported by the Dutch Companion Project funded by the Dutch agency for innovation and sustainable development SenterNovem (IS053013).

References

[1] A.S. Bregman, Auditory Scene Analysis. MIT press, 1990.

[2] D. Wang and G. J Brown, Computational Auditory Scene Analysis. IEEE Press, 2006

[3] T.C. Andringa, Continuity Preserving Signal Processing, PhD Thesis. University of Groningen, The Netherlands, 2003.

[4] T.C. Andringa, and M.E. Niessen," Real-world sound recognition: A recipe," Proceedings of LSAS 2006 Greece. (2006), 106-118.

[5] K. van den Doel, and D. K. Pai, "The sounds of physical shapes," Presence-Teleoperators and Virtual Environments. **7** (1998), 382–395.

[6] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," Journal of the Acoustical Society of America. **101** (1997) ,412-419

[7] X. Serra and J. O. Smith, III, "A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," Computer Music Journal, **14** (1990),12-24

[8] P. Guillemain and R. Kronland-Martinet, "Characterization of acoustic signals through continuous linear time-frequency representations", Proceedings of IEEE. **84** (1996), 561-855

[9] T.J. Gardner and M.O. Magnasco, "Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations," Journal of the Acoustical Society of America. **117** (2005), 2896-2903

[10] S.A. Fulop and K. Fitz, "Algorithm for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," Journal of the Acoustical Society of America. **119** (2006), 360-371

[11] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," Journal of the Acoustical Society of America. **116** (2004),3690-3700

[12] F. Auger and P. Flandrin, "Improving readability of time-frequency and time-scale representations by the reassignment method," IEEE Transaction on Signal Processing. **43** (1995), 1068-1089

[13] P. R. Cook, Real Sound Synthesis for Interactive Applications, A K Peters, Ltd., 2004.

[14] B. Gygi, G. R. Kidd and C. S. Watson," Similarity and categorization of environmental sounds," Perception and Psychophysics. **69** (2007), 839-855