ACOUSTIC SCENE ANALYSIS BASED ON POWER DECOMPOSITION

Alain de Cheveigné, Michael Slama

CNRS, Université Paris 5, Ecole Normale Supérieure,
29 rue d'Ulm, F-75230, Paris, France,
Alain.de.Cheveigne@ens.fr

ABSTRACT

A method is proposed for the analysis of acoustic scenes. The contribution of each competing source is suppressed on the basis of harmonic structure or cross-sensor correlation, in such a way that other sources may be estimated. Successive suppression of sources allows the scene to be characterized. In the limit of purely periodic sources and no noise, the method provides accurate estimates of fundamental frequency and spectral content. In the presence of noise or imperfect periodicity, it offers likelihood functions from which the spectral content of sources may be inferred using Bayesian methods. The method is an alternative to more familiar spectral and spectro-temporal methods of acoustic scene analysis. An advantage is that precise analysis is possible using short temporal windows.

1. INTRODUCTION

Cherry [1] coined the expression "cocktail party effect" to describe our ability to understand the overlapping speech of multiple speakers, a task which human listeners still perform better than machines [2]. Many efforts have attempted to replicate our skills of Auditory Scene Analysis (ASA) [3], in what are known as Computational Auditory Scene Analysis (CASA) systems. Pioneering work was done by Parsons [4] and Weintraub[5]; more recent efforts are reviewed by Cooke and Ellis [6] and others [7, 8, 9]. In addition to perception-inspired approaches, recent progress has been made using statistical and machine-learning techniques (e.g. [10, 11]).

These techniques usually operate on a *spectro-temporal* representation produced either by a short-term Fourier transform (STFT) or an auditory filterbank. Performance is crucially dependent on the choice of analysis parameters such as window shape and size, or filter shape and bandwidth. There is a well-known tradeoff between temporal and spectral resolution [12]. The best choice of parameters is signal-dependent, and values appropriate for one task may be less effective for another. No representation is optimal for all scenes: any given choice is necessarily suboptimal for some scenes(this situation may be alleviated by the use of overcomplete bases, e.g. [13]).

This paper proposes instead to use representations based on running power and autocorrelation functions (ACF) as a basis for scene analysis. The goal is to "make sense" of the scene: extract signal parameters such as fundamental frequency (F_0) or spectral envelope, or allow a pattern-matching algorithm to access voices within the scene (for example to perform automatic speech recognition, or content-based retrieval of music or multimedia). In this paper the observable input is a single-channel signal, but processing principles can be extended to multichannel processing.

2. INGREDIENTS

This section introduces the required ingredients and notations. Given a signal $x(t)$, the *running autocorrelation function* is defined as:

$$\langle x \rangle_t(\tau) = (1/W) \sum_{i=t+1}^{i+W} x(i)x(i-\tau) \quad (1)$$

where τ is the lag, t indexes the time at which the calculation was made, and W is the size of the integration window. W determines the amount of temporal smoothing applied to the statistic. The definition differs from that of the more common *short-term autocorrelation function*:

$$(1/W) \sum_{i=t+1}^{i+W-\tau} x(i)x(i-\tau) \quad (2)$$

by the upper limit of the integration. This latter definition is more common than the first, because it is related to the *short-term power spectrum* and can be efficiently implemented based on the FFT. Its main drawback is that the amount of temporal smoothing varies with τ and vanishes as τ approaches W. In Eq. 1 the integration time is instead independent of τ, and the running ACF can thus be calculated for arbitrary τ. This paper uses the definition of Eq. 1. The running *power* of the signal indexed by time t is defined as:

$$||x||_t = \langle x \rangle_t(0) \quad (3)$$

Given a candidate period T, we define the *periodic-aperiodic power decomposition* of x using the following notation:

$$\begin{aligned} x^{+T}(t) &= [x(t) + x(t-T)]/2 \\ x^{-T}(t) &= [x(t) - x(t-T)]/2. \end{aligned} \quad (4)$$

These are simply the time-domain comb-filtered versions of x that sum up to x:

$$x^{+T} + x^{-T} = x.$$

They constitute a partition of its *power* in the sense that:

$$||x^{+T}||_t + ||x^{-T}||_t = (||x||_t + ||x||_{t-T})/2. \quad (5)$$

The right hand side is the average of two estimates of the power. This decomposition is useful for at least three reasons: (a) If x is periodic with period T, then $x^{+T} = x$ and $x^{-T} = 0$. In this sense, we can interpret x^{+T} and x^{-T} as the "periodic" and "aperiodic" parts of x respectively. (b) From Parseval's relation it follows that the same decomposition applies to short-term power spectra: $X^{+T}(\omega)_t + X^{-T}(\omega)_t = (X(\omega)_t + X(\omega)_{t-T})/2$, where $X(\omega)_t$ designates the short-term power spectrum calculated at t. In other words, the power spectrum of x is split into "periodic" and "aperiodic" parts.This can also be understood by noting that Eq. 4 defines filters with transfer functions $H^{+T}(\omega) = cos^2(\omega T)$ and $H^{-T}(\omega) = sin^2(\omega T)$ that sum to one for all ω. (c) The decomposition can be generalized to more than two terms, as discussed below.

The *squared difference function* (SDF) is defined as:

$$d_t(\tau) = (1/W) \sum_{i=t+1}^{i+W} [x(i) - x(i-\tau)]^2 = 4||x^{-\tau}||_t. \quad (6)$$

This is simply the Euclidean distance between windows of size W shifted by τ. It is related to the running ACF by:

$$d_t(\tau) = ||x||_t + ||x||_{t-\tau} - 2\langle x \rangle_t(\tau). \quad (7)$$

To the extent that the first two terms are constant (W large or period multiple), SDF and ACF offer the same information.

We will need quantities such as $\langle x \rangle_t(\tau)$ or $d_t(\tau)$ for arbitrary values of t, τ, and W. Computational efficiency is not our main focus, but it is worth saying a few words about the issue. The running ACF may be implemented by FFT as the cross-correlation between windows of size W and $W + 2\tau_{MAX}$, where τ_{MAX} is the maximum required lag. An FFT-based formula is efficient if the statistic is calculated sparsely in time, but for higher frame rates an incremental formula may be more efficient, as illustrated here for power:

$$||x||_t = ||x||_{t-1} - x(t)^2 + x(t+W)^2.$$

Similar formulae exist for the running ACF and SDF. If $\langle x \rangle_t(\tau)$ has been calculated for the required range of t and τ, then other useful statistics can be derived cheaply according to formulae such as Eq. 7. By carefully arranging these precalculated values, it is possible to derive values for arbitrary W at cost ($o(\log W)$). Quadratic interpolation formulae can be used to derive values for non-integer values of t, τ, and W.

To summarize, the ingredients required by the methods to be described include power, running ACF, SDF, and derived statistics that can be calculated at a reasonable computational cost.

3. F_0 ESTIMATION

3.1. Single F_0 estimation

Supposing the observed signal $x(t)$ is quasiperiodic, its period at time t can be found by searching over lags τ for a minimum of $||x^{-\tau}||$. This is the basis of the YIN method of fundamental frequency estimation that is described and evaluated in [14].

3.2. Multiple F_0 estimation

Supposing the observed signal $x(t)$ is the sum of two or more quasiperiodic voices, their multiple F_0 values can be estimated iteratively. An initial estimate T of one voice is derived from a single-voice algorithm and a time-domain comb-filter tuned to T is applied to suppress that voice so that the others can be more easily estimated. The operation is repeated for each voice in the mixture. Concretely, a method such as YIN is applied to the mixture to obtain T, and then applied again to the comb-filtered signal x^{-T} to obtain a second estimate T'. The comb filter is then tuned to T' to refine the estimate of T, or else the method may be applied to $x^{-T,-T'}$ to estimate a third period, etc. This is the basis of the MMM method of multiple F_0 estimation described in [15, 16].

4. HARMONIC POWER DECOMPOSITION

The stage is now set for the analysis of acoustic scenes containing periodic or quasiperiodic voices. Important information is usually lost in the mixing operation from which observable signal(s) are derived, and no general solution exists to extract all component voices in all cases. Instead, we should aim to extract the partial information that *is* available in each case, and then piece it together using model-based techniques, e.g. within a Bayesian framework. Some ways to do so are outlined here.

4.1. Single periodic voice

If the observed signal x is periodic and the period T is an integer multiple of the sampling period, the power spectrum can be derived *exactly* from the DFT of x over a T-sized window. Alternatively, its ACF can be derived from Eq. 1 if $W = T$

(or W large). The initial period estimate requires a signal segment of about $2T$ (see [14] for a more precise discussion). Spectral estimation then needs a segment of T, so accurate analysis of this simple scene requires a signal chunk of only about $2T$. If T is not a multiple of the sampling period, interpolation techniques may be applied.

4.2. Periodic voice plus noise

If the observed signal x is the sum of a periodic signal s of period T and a noise signal n, phase and amplitude uncertainties prevent perfect estimation. However relatively accurate estimates can be provided for the statistics of n and the *distribution* of s if T is known. Applying the "periodic-aperiodic" decomposition, we know that: $n^{-T} = x^{-T}$ because s^{-T} maps to zero. It follows that the short-term power spectrum $N^{-T}(\omega)_t$ can be accurately obtained from the observed signal.

On the other hand $N^{+T}(\omega)_t$ is inextricably confounded with $S(\omega)$, and thus neither $N(\omega)$ nor $S(\omega)$ can be measured. However if we suppose that the process that produces n has a "smooth" spectrum, we can use $N^{-T}(\omega)$ as an approximation of $N^{+T}(\omega)$, and thus characterize the "noisy" part of the signal. Furthermore, $N^{+T}(\omega)$ can be used to characterize the error incurred if $S(\omega)$ were approximated by $X(\omega)$. More precisely, if the noise power follows a distribution for which the mean is a sufficient statistic, we can use $N^{+T}(\omega)$ to parametrize the *likelihood* of $S(\omega)$ given the observation.

If T is estimated from x, the effect of estimation error must be taken into account. This involves two steps: (a) derive the expected distribution of T given a noisy observation and (b) translate error on T in terms of error on $N(\omega)^{-T}$. The likelihood function is adjusted accordingly. The model may also handle variations over time of a periodic signal, e.g. amplitude variations [15].

4.3. Two periodic voices

Suppose that the observed x is the sum of two periodic signals s and s' with periods T and T'. The "periodic-aperiodic decomposition" can be extended to split x into four terms:

$$\begin{aligned}
x^{+T,+T'}(t) &= [x(t)+x(t-T)+x(t-T')+x(t-T-T')]/4 \\
x^{+T,-T'}(t) &= [x(t)+x(t-T)-x(t-T')-x(t-T-T')]/4 \\
x^{-T,+T'}(t) &= [x(t)-x(t-T)+x(t-T')-x(t-T-T')]/4 \\
x^{-T,-T'}(t) &= [x(t)-x(t-T)-x(t-T')+x(t-T-T')]/4 \quad (8)
\end{aligned}$$

where $x^{+T,+T'(t)}$ denotes $x(t)$ filtered by a cascade of two comb filters. Indeed the terms sum up to x:

$$x = x^{+T,+T'} + x^{+T,-T'} + x^{-T,+T'} + x^{-T,-T'} \quad (9)$$

and their powers sum up to that of x:

$$\begin{aligned}
&||x^{+T,+T'}||_t + ||x^{+T,-T'}||_t + ||x^{-T,+T'}||_t + ||x^{-T,-T'}||_t \\
&\quad = (||x||_t + ||x||_{t-T} + ||x||_{t-T'} + ||x||_{t-T-T'})/4. \quad (10)
\end{aligned}$$

This decomposition is useful in because of how each term depends on components of the mixture. The second term $x^{+T,-T'}$ depends only on s, and the third $x^{-T,+T'}$ depends only on s'. The fourth depends on neither, and it is *zero* if the signals are perfectly periodic. The first term depends on both, and represents power that cannot be assigned to either source given the observation. Again, thanks to Parseval's relation, Eq. 10 applies also to power spectra.

It is interesting to note the duration of the interval that supports this analysis. Initial F_0 estimates require on the order of $3T$ (supposing $T > T'$). Spectral estimation then takes on the order of $2T$. Accurate analysis of the scene is thus possible with a signal interval of only about $3T$. We discuss below what can be gained if a longer interval is available.

4.4. Two periodic voices with noise

Suppose now that $x = s + s' + n$, where s and s' are periodic sources with known periods T and T', and n is noise. The previous decomposition can again be applied. The power of the noise is split among the four coefficients; in particular $||x^{-T,-T'}||_t$ is no longer zero. As in the single-voice-with-noise case, this term can be used to estimate the amplitude and spectrum of the noise. That allows us to parametrize the likelihood of the sources given the observation.

5. GENERALIZING THE APPROACH

Similar techniques can be applied to a larger number of sources (in a single channel), as well as to a larger number of channels (e.g. multiple microphones). A strength of the method is that it requires short windows, and can thus track time-varying signals. Typically a source (voice, instrument) may go through intervals of stability, during which it is amenable to accurate processing, and intervals of transition, during which estimates are degraded. A reasonable strategy is to assign stronger *weight* to information gathered during the former than during the latter. Measures of unaccounted power ($||x^{-T}||_t$, $||x^{-T,-T'}||_t$) can be used to locate islands of reliability.

Longer intervals may be used to enhance the analysis. For example in the one-voice-plus-noise case, a frame of 3-periods duration allows the decomposition of Eq. 8 to be applied with $T' = T$. Among the four terms obtained, the first, $x^{+T,+T}(t)$, implements a relatively narrow "harmonic sieve" tuned to T, while the other terms gather power that does not pass the sieve. This can be understood in the spectral domain: the transfer function associated with the first term $H^{+T,+T}(\omega) = cos^4(\omega T)$ has relatively narrow peaks. Longer periods of stability allow yet-sharper selectivity.

6. DISCUSSION

This paper presented a framework for acoustic scene analysis based on the harmonic structure of one or more sources. The aim is to extract as much useful information as possible from observations, subject to the fact that some information is lost in the mixing. We assumed a single observable signal (single microphone), but similar principles can be applied to the analysis of multichannel observations (multiple microphones).

Analysis can be performed on very short frames of data, which is useful for tracking rapidly varying signals. In an ideal situation (no noise, perfect periodicity over the analysis frame), analysis produces an accurate but possibly incomplete estimate of component spectra. In the presence of noise or imperfect periodicity, it produces instead an estimate of the *distribution* of target values conditional on the observation.

Power-based analysis offers an alternative to standard spectral methods that involve an initial STFT applied to windowed frames of data. The *size* and *shape* of the analysis window are "nuisance parameters" that are hard to choose, given the tradeoff between spectral and temporal resolution. Here, the main parameter is the duration W of the integration window, that determines the temporal stability of the analysis but not its spectral resolution.

"Periodic-aperiodic" decompositions (Eqs. 4, 8) are analogous to a power spectrum in the sense that they obey a form of Parseval's relation (Eqs. 5, 10). They have the useful property that they can isolate parts that do *not* depend on a source. It has been argued that this is an essential trait of a good representation for acoustic scene analysis [17].

A key premise to addressing the "cocktail party problem" is that mixing destroys information, and thus the problem cannot be solved with full generality. Rather, one must rely on regularity and redundancy of the sources or scene to form models that can be constrained by the piecemeal information that is derivable, sometimes with perfect accuracy, from the observed signals. Bayesian methods provide a good framework for this purpose, but they are usually applied to short-term spectral representations. We suggest that they should be applied to the representations that we have sketched out here.

Acknowledgements

This research was funded in part by the S2S^2 initiative of the European Community. This paper was prepared while on leave at Shihab Shamma's lab at the Institute of Systems Research, University of Maryland.

7. REFERENCES

[1] E.C. Cherry, "Some experiments on the recognition of speech with one, and with two ears," *J. Acoust. Soc. Am.*, vol. 25, pp. 975–979, 1953.

[2] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, vol. 22, pp. 1–16, 1997.

[3] A. S. Bregman, *Auditory scene analysis*, MIT Press, Cambridge, Mass., 1990.

[4] T.W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911–918, 1976.

[5] M. Weintraub, *A theory and computational model of auditory monaural sound separation*, Ph.D. thesis, Stanford, 1985.

[6] M. Cooke and D. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Comm.*, vol. 35, pp. 141–177, 2001.

[7] G.J. Brown and D.-L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech enhancement*, J.Benesty, S. Makino, and J.Chen, Eds., pp. 371–402. Springer, New York, 2005.

[8] D.-L. Wang and G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, IEEE Press / Wiley, 2006 (in press).

[9] P. Divenyi, *Perspectives on Speech Separation*, Kluwer, New York, 2004.

[10] S. Roweis, "One-microphone source separation," in *Advances in NIPS*, p. 609616. MIT Press, Cambridge MA, 2000.

[11] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Workshop on statistical and perceptual audio processing*, Jeju, Korea, 2004.

[12] D. Gábor, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, pp. 591–594, 1947.

[13] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, pp. 863–882, 2001.

[14] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.

[15] A. de Cheveigné and A. Baskind, "F0 estimation of one or several voices," in *Eurospeech*, 2003, pp. 833–836.

[16] A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, pp. 175–185, 1999.

[17] A. de Cheveigné, "Separable representations for cocktail party processing," in *Forum Acusticum (FA2005)*, Budapest, 2005, pp. 1527–1531.