A wavelet-based sinusoid model of sound for auditory signal separation

Daniel PW Eide & Barry L. Vercoe
Music & Cognition Group, MIT Media Lab
Cambridge MA, 02139
dweg@media.mit.edu, bvl@media.mit.edu

Separation of monaurally superimposed sounds presents a paradox: it is seemingly effortless for human listeners to distinguish the components, but incommensurately difficult to build machines that imitate this skill. Psychology researchers have discovered many rules and regularities in human audition that should be useful to a computerized signal separator. They generally describe the reassembly of sounds after a loosely specified initial decomposition into time-frequency "atoms". These atoms must be large enough to reduce spectral data to a manageable quantity, but limited enough to avoid coding features of more than one sound into a single element.

The tracks generated by the McAulay/Quatieri sine wave model are promising candidates for this role. Unfortunately, their narrow-band short-time Fourier transform inevitably averages together the spectral characteristics of several pitch cycles. In order to preserve the small-scale timing information known to be critical for signal fusion, and also to build a system that more closely resembles the human ear, we use a low-Q wavelet transform instead of the FFT as the basis for the sine wave model. The resulting invertible representation is a good perceptual analog, and has promising applications in even detection and signal separation, as well as for other abstract analysis-synthesis processes.

INTRODUCTION
In artificial intelligence research - the simulation of human behavior by computers - an early and recurrent surprise is that "easy things are hard" [Mins86], i.e., the tasks that we perform without conscious effort are the hardest to reproduce. Many examples of this are found in perception, since much of the work in mediating the outside world into our consciousness is done by special-purpose neural structures to which introspection has no access. Auditory signal separation (defined in more detail below) is a case in point: it is impossible for us to listen to mixtures of sound except by parsing them into the separate sources. Most computer systems processing the same signals are incapable of hearing it as anything but a single entity. If we try to reproduce it, we find ourselves with a very difficult job.

A solution to auditory signal separation would be very beneficial. In general, to make computers more useful we are trying to make them respond to stimuli in the same way as people. Despite the progress in speech recognition systems, most recognizers are defeated by interfering background noise of a kind that would present little challenge to people. In the music industry, a system that could separate out different instruments in a mix would find many applications in the examination and improvement of existing recordings. But a more important reason to build such a system is what we would learn about the nature of auditory signal processing. This new insight would have profound implications for music and other sound design.

ABOUT SOURCE SEPARATION
To start with we need to try and define this problem more rigorously. When we listen to a song, we often feel that we can choose to listen to different components of the music - just the singer or just the accompaniment. This gives us an indirect definition of signal separation, as the decomposition of a complex sound into separate parts as it would appear to a human listener. If we try to do away with this subjective, human aspect of our definition we run into trouble. Information theory tells us that we cannot always work backwards from the sum of two signals to the individual components. So the perceptual system that appears to accomplish this for us must be employing various assumptions and restrictions about the signals.
it is extracting, and, since it usually gets things right, these assumptions must hold for 'real-world' sounds. Our problem in building a signal separator is to discover these constraints and how they are applied.

We decided to work from just a single channel of sound, and to ignore cues not intrinsic to the sound itself (such as spatial location or visual hints). The experience of listening to monophonic recordings tells us that separation is still quite easy from such a signal. There has been quite extensive psychoacoustical research into how people form auditory images [Breg90]. This includes factors affecting fusion (the perception of simultaneous energy in different frequency bands as a single sound), and streaming (the perception of separate, sequential sound events as belonging 'together'). Modelling both of these effects is necessary for signal separation. The basis for image formation applicable to our situation are:

a) Simultaneous onset (and offset)
b) Common amplitude envelopes
c) Common frequency modulation (shown to be particularly powerful in [McAd84])
d) Harmonic relationships
e) For streaming, proximity in time and frequency

The above qualities are defined in reference to individual sinusoidal components, the typical building blocks of psychoacoustical experiments. This set of rules seems appropriate for machine simulation, but real sounds do not trivially decompose into the sinusoidal components of this type - particularly sounds of several sources mixed together. Part of our work can be interpreted as an attempt to break down real sounds into a domain where these rules can be applied.

We are seeking to build a general solution to signal separation that works over the full range of sounds that people can handle. This extends the problem from simply dividing the signal energy to include deciding how many parts to attempt to extract, and detecting when particular contributions start and stop. Thus a complete signal separator needs an auditory event detector, as observed in [Mue91]. Sadly, we are only at the beginning of building even this prerequisite.

Previous work in signal separation has sometimes taken such a psychoacoustical approach (e.g. [Jaba90]). Other researchers have tried to exploit particular attributes of the signals they are separating such as harmonic structures, particularly co-channel speech [Que90] or musical duets [Mue90].

AN AUDITORY REPRESENTATION OF SOUND

By considering the brain as an information processing system, we can think about the pattern of stimulation excited in the brain in response to some event as the 'internal representation' of that event in the brain. Although this representation is hypothetical, it is the goal of our model. We may hope that the representation simplifies various dimensions of the sound, and that it will be somehow optimized to ease fundamental perceptual operations like signal separation (see [Tarte81]). We can work towards this representation by two methods - firstly by using what direct evidence we have from physiology and psychoacoustics and, when that is exhausted, by trial and error development of more speculative ideas.

Ultimately we are interested in generating any sounds separated, so attention has been given to the invertibility of our model at each stage. While most representations cannot be perfectly inverted, a good test of the validity of a given model is that the lost information does not have much influence on the nature of the sound - the imperfect reconstruction still sounds like the original.

THE SIGNAL SEPARATION MODEL

The current analysis/synthesis system is shown in figure 1. This diagram shows how the system might operate to separate two signals mixed into a common channel. It does not represent the functionality mentioned above required to decide on the number of sources.

The first stage of analysis is the constant-Q transform, simulating the frequency transformation effect of the cochlea [Glas90]. This is a bank of bandpass filters, implemented by direct c-simulation (FIR filters). Q' is the ratio of filter center frequency to bandwidth, making it constant implies each filter will have a slightly different bandwidth (dependent on center frequency), unlike FFT-based transforms which are characterized by constant bandwidth. Although this is computationally expensive, we felt it was crucial to reflect the varying time resolution across frequency observed in real ears. In particular, since small period modulations can be critical for separation, wide bandwidth, fast filters are crucial to detect and localize this
kind of information.

The output from the filter bank is a complete, fully irreversible representation of the input sound; indeed, in order to raise subsequent stages the filters are overlapped and thus the representation is oversampled.

The next stage aims to reduce the volume of information by removing spectral details incidental to the perceived quality of the sound. This is done by taking each instantaneous spectrum from the filter banks and characterizing it by just its local maxima. The frequencies, as well as magnitudes and phases, of each spectral peak are carefully interpolated and retained; all information pertaining to the energy in between peaks is discarded.

The psychoacoustical justification for this stage is a little less rigorous. Many "critical-band" phenomena suggest a lateral inhibition mechanism early in cochlea processing [Sham89]. More concretely, analysis-synthesis systems employing peak picking have generally been successful [McAus, Scott89].

Rather than seeking to process each instantaneous peak separately, in the next stage they are grouped into tracks. Each track represents the trajectory of some spectral energy over a period of time, both in frequency and amplitude. The psychoacoustic phenomenon being modelled here is continuity [Breg90], but it is eminently reasonable that a sound continuous in this sense will most likely have arisen from a single source.

We now have the input sound represented as a relative small number of distinct time-frequency trajectories of signal energy. These are our candidates for the 'sinusoidal' decomposition upon which to apply the psychoacoustical principles of source formation listed earlier. This is purpose of the next block, track correlation and segregation (currently in an early stage of development). The set of all tracks from the sound is partitioned into groups that appear to have a common origin by virtue of common start times, amplitudes or frequency modulation.

Each subset of tracks from this process can then be passed to a separate resynthesis to reconstruct the acoustic signal. At present, this is done in a literal manner by using each track to control a sine wave oscillator. This gives good results; however, alternative methods that more closely reverse each stage of analysis are also being investigated.

RESULTS

Figure 2 shows the time-frequency tracks overlaid on the constant-Q spectrogram of a sound. The sound starts with a guitar chord, then a voice comes in at time 0.7. Looking at the tracks after the entry of the voice, we can see that the first three or four harmonics have been resolved by the narrow, low frequency filters. At higher frequencies, however, several of the voice's harmonics fall into a single filter of the bank. This gives tracks that exist at the location of the formant peaks - not necessarily coincident with a harmonic.

This is encouraging since it reflects the important qualities of voice for people; fundamental pitch and formant location, but not details of individual high harmonics.

Since the high harmonics have such rapid variations it is impossible in this figure to see how well they exhibit the common modulations we have discussed. However, it is quite clear that the tracks arising from the guitar sound at the start of the sound are 'hidden' by the much stronger voice harmonics coming in over the

![System block diagram](image)

ICMC 88
top. This reflects results of signal separation experiments to so far: quieter, background sounds can be removed effectively. However, as it stands there is no way to counteract masking of quiet sounds. This is something we hope to address, as described below.

FUTURE WORK

Now that the problems with this scheme for signal separation are apparent, we are considering solutions. The next development will be a system that makes estimates of each spectral frame based on the tracks to date. This estimated complex spectrum is subtracted from the measured spectrum, and the residue is analyzed. By fitting the expected deviations resulting from simple peak movements, the existing tracks can be updated. Any remaining information corresponds to signal not yet accounted for, with a much better immunity to strong but well-established foreground energy. This of course mimics the well known adaptation of sensory organs.

Another area for further investigation is in processing and modifying sounds in this domain. There is some evidence that limited neural bandwidth might allow smoothing of the resolved tracks without audible effect [Math47], with implications for sound data compression. It should be possible to detect and correct situations where tracks cross and have adverse phase interactions. Finally, this mapping of sonic features has exciting musical possibilities such as cross-synthesis which we hope to develop.

REFERENCES

[Duda90] RO Duda, RF Lyon, M Slaney (1990) "Correlagrams and the separation of sounds" Proc Asilomar conf on sig, sys & computers
[Math47] RC Mathes, RL Miller (1947) "Phase effects in monaural perception" JASA 19(5)
[Quat90] TF Quatieri, RG Daniewicz "An approach to co-channel talker interference suppression using a sinusoidal model for speech:" IEEE TV ASSP 38(1)