presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, May 1992

A Perceptual Representation of Sound for Auditory Signal Separation

Daniel PW Ellis & Barry L Vercoe
Music & Cognition Group, MIT Media Lab, Cambridge, MA 02139
dpwe@media.mit.edu, bv@media.mit.edu

Building machines that emulate the kinds of acoustic information processing that human beings take for granted has proved unexpectedly difficult; the human auditory system is extremely sophisticated in its adaptation to the sounds of the real world, and uses an impressive array of features as cues to organization and interpretation.

As more of these cues become known through psychoacoustical experiment, it becomes feasible to program computers functionally to mimic human perception of sound. Concentrating on the task of distinguishing and separating individual superimposed sonic sources, such as a singer and accompaniment, we have built a perceptually-sufficient invertible representation that analyses sound down to a relatively small number of 'atoms'. Psychoacoustic rules of stream formation can then be applied to these atoms to simulate human source separation.

We describe the nature of this representation and how we propose to use it as the basis for source separation by computer.

1. INTRODUCTION

In artificial intelligence research - the simulation of human cognitive abilities by computers - a 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 hardware to which introspection has no access. Distinguishing the sounds of simultaneous events (auditory signal separation) is a case in point: In most cases, it is impossible for us to listen to mixtures of sound *except* by parsing them into the separate sources; computer systems processing the same signal tend to treat the mixture as a single entity.

A solution to auditory signal separation would have several distinct benefits: Perhaps most importantly, the insight gained into the nature of perceptual information processing would have great impact on machine information processing, particularly that oriented towards human users: As we gain a better understanding of what gives each sound its particular qualities, we can achieve more precise ends with greater efficiently. A second area of benefit is the endeavor to produce machines that respond and interact in the same ways as people to create the 'ideal' user interface. Clearly such machines must be able to make sense of acoustic information, perhaps in adverse conditions, with success comparable to human listeners; despite progress in speech recognition, most recognizers are easily defeated by modest background noise. 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.

In the next section we discuss in more depth the problem of source separation and what is known about how people solve it. We then argue, in section 3, for a categorical analysis of complex sounds as the best approach to the problem. In section 4, we describe the representation we have developed as satisfying many of these requirements, and show typical results of this analysis in section 5. How source separation will be implemented using this representation is outlined in section 6. Finally, we will describe the immediate next stages of development.

2. ABOUT SOURCE SEPARATION

Although introduced in the preceding section, we need to define this problem more rigorously. When we listen to a recording of music, for instance a song, we often feel that we can choose to listen exclusively to individual parts - just the singer or just the accompaniment, for example. 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 work backwards from the sum of two signals to the individual components without additional information. So the perceptual system that appears to accomplish this for us must be employing various assumptions and restrictions about the signals it is extracting. Since our senses usually give a useful analysis of the sound-causing processes, the assumptions employed are proven successful for 'real-world' sounds. Our problem in building a computerized signal separator is to discover these constraints and how they are applied.

We should note at this stage the distinction between the exact sound emanating from a given source and the percept it arouses in the mind of the listener. Even in ideal listening circumstances, there are many details of the actual sound that do not influence the percept (frequency components above 20 kHz are an extreme example). When several sounds are mixed it is almost inevitable that certain features will be masked or corrupted; however, in many cases their presence may be 'reasonably' inferred, and this 'filling-in' will be performed by the auditory periphery in a manner completely transparent to the listener. Thus, even when we believe we have heard a particular detail, we should be aware that such a percept may not be based on any direct evidence.

There is a large body of psychological research into the circumstances under which human listeners combine acoustic energy into a single perceived event (*fusion*) or distribute it between different perceived sources (*segregation*). [Breg90] comprises a very thorough survey of these results. The important cues leading to acoustic energy being assembled into a single source are:-

a) Common spatial origin as established by binaural processing or pinna-related coloration;
b) Common onset: Sonic events tend to be treated according to the 'old plus new' heuristic, that any change in the sound is due to the addition of a new source;
c) Other kinds of 'common fate', such as simultaneous offset or correlated frequency and amplitude modulation;
d) Harmonic frequency relationships i.e. common period;
e) Proximity in time and frequency;
f) Matching previously-learned patterns.

We decided to work from just a single channel of sound and ignore spatial information; the experience of listening to monophonic recordings tells us that different instruments may still easily be distinguished in such a signal. We also decided to put aside the problem of learning or experience required to pick out familiar sounds, since unfamiliar sounds may be easy separated in many circumstances. This eliminates the first and last items from our list, leaving us with features intrinsic to the sound itself upon which to base our processing.

Our objective is a general solution to signal separation that works over the full range of sounds experienced by people. This broader problem includes deciding how many parts to attempt to extract, and detecting when particular contributions start and stop. Thus a complete signal separator requires an auditory event detector.

Previous work in signal separation has sometimes taken such a psychoacoustic approach (e.g. [Duda90]). Both [Mell91] and [Cook91] have built large systems that attempt to apply Bregman-style rules for source formation. Other researchers have tried to exploit particular attributes of the signals they are separating such as harmonic structures, particularly for co-channel speech [Quat90] or musical duets [Mahe89].

3. REQUIREMENTS OF AN AUDITORY REPRESENTATION OF SOUND

Ultimately we are interested in regenerating any sounds separated, so attention has been given to the invertibility of our model at each stage. While many 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 heuristics of source formation described above are typically defined in reference to individual sinusoid components, the building blocks of many psychoacoustical experiments. This set of rules seems appropriate for machine simulation, but real sounds do not trivially decompose into the sinusoid components of this type - particularly sounds of several sources mixed together. The work described in this paper can be interpreted as an attempt to break down real sounds into a domain where these rules can be applied.

Such an analysis should represent a sound as a collection of *perceptually atomic elements* - components of the sound that can be safely assumed to belong to just one source, and to require no further division during perceptual processing. These elements will have the following properties:-

- They will correspond to contiguous energy concentrations in time-frequency. This basic assumption, that sounds have some duration and stability in time, is central to any attempt at source separation. The emphasis on the frequency axis reflects the known structure of the ear.
- In order to preclude the possibility that an element will turn out to be derived from two different sources (and thus fail atomicity), each element should have smooth, steady properties over its duration.
- Ideally, we would like elements of our representation to have a one-to-one correspondence with perceptual features of the sound, where such features are only vaguely defined. This is an argument that the elements should be as large as possible without violating the previous conditions, rather than just breaking down a sound into the smallest units of processing, and that, taken alone, each element should be recognizable (or at least plausible) as a component of the whole sound.

To analyze a sound into such elements implies the imposition of considerable structure on the sound signal, and the idea of having many distinct elements, each with individual significance, may be considered a conversion from a continuous distribution of acoustic energy into a *symbolic* representation, where the distribution is summarized as comprising a particular set of discrete sonic features. There are two principle advantages of such a categorical representation, which is common in cognitive models: when a nontrivial portion of the input signal has been identified as belonging to a single element, it is possible to evaluate interesting parameters for that element; very small elements will have very few meaningful parameters. The second advantage relates to efficiency: If our strategy is to perform some task such as separation on the complete sound, one decision must be made for each element of that sound. If we can contrive to make these elements larger and less numerous, the number of decisions, and presumably the computational load, is reduced.

4. THE SIGNAL SEPARATION MODEL

The current analysis/synthesis system is shown in the block diagram below (figure 1). The analysis and synthesis are based upon the McAulay-Quatieri Sinusoid Transform [McAu86], with the major modification of the fixed-Q analysis filterbank; hence we shall refer to it as the Constant-Q Sine-Wave Model or CQSWM. 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.

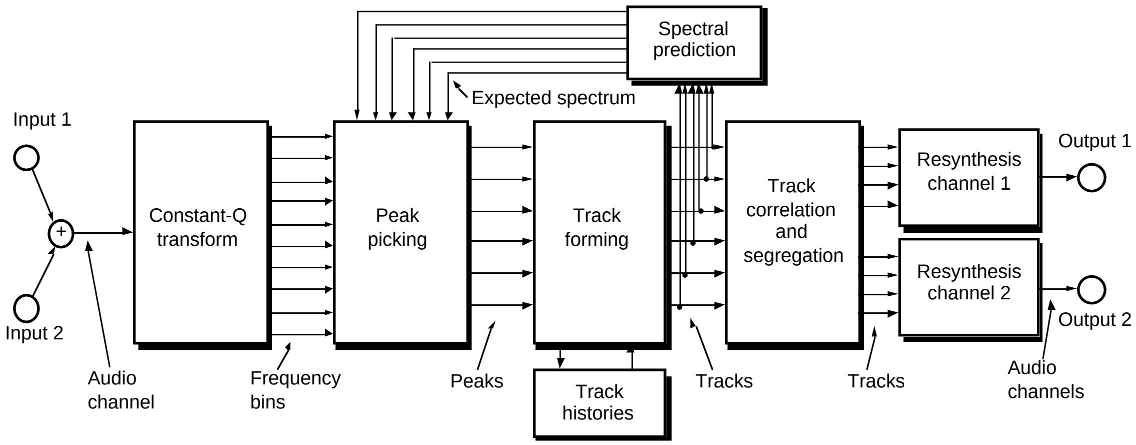

figure 1: block diagram of the Constant-Q Sine Wave Model (CQSWM)

The first stage of analysis is the constant-Q transform, simulating the frequency transformation effected by the cochlea [Glas90]. This is a bank of bandpass filters, implemented by direct FIR evaluation. '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 extract this kind of information. The varying bandwidth means each filter bin requires a different sampling rate to preserve all its information; this propagates through all the subsequent stages, thus the model is a multirate system.

The output from the filterbank is a complete, fully invertible representation of the input sound; indeed, in order to ease subsequent stages the filters are overlapped and thus the time-frequency plane is oversampled. The next stage reduces the volume of information by characterizing the spectrum at each instant by its maxima alone. This aims to preserve the perceptually important parameters while discarding information of little or no significance. The frequencies, as well as magnitudes and phases, of each spectral peak are carefully interpolated and retained, but nothing more.

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]. Regardless of theoretical foundation, we can observe that analysis-synthesis systems employing peak picking have been very successful [McAu86], [Serr89].

The feedback loop that enters the peak-picking block is a prediction of what will be coming out of the filterbank based on the analysis of the sound so far. This is an as yet unimplemented method of improving the sensitivity of peak-picking beyond that normally achieved. When we look at the local maxima in a spectrum, a large, broad peak may obliterate smaller peaks nearby that would otherwise have been visible. The later stages of the model give us a detailed description of what seems to be happening in the sound at a given time. From this description, we should be able to make a good estimate of the spectrum at the next time slot. By analyzing the *difference* between our 'most likely' predicted spectrum and the actual measured spectrum, we can detect the energy that was previously masked, as well as updating our model where the prediction made small errors.

Peaks at a given frequency in adjacent time slots are organized into tracks, describing the evolution of the acoustic energy concentration at that frequency. This reflects the basic assumption of continuity mentioned above.

We now have the input sound represented as a finite number of distinct time-frequency trajectories, each tracing a local maximum in the signal energy. These are our candidates for the 'sinusoidal' decomposition upon which to apply the psychoacoustical principles of source formation listed earlier. The next block, track correlation and segregation, serves to represent such processing. Although this has yet to be implemented, our intentions will be discussed in detail below.

Each subset of tracks from this process can be separately resynthesized into an acoustic signal. This is efficiently achieved in a literal manner by using each track to control a sine wave oscillator. This gives very satisfactory results, and is much simpler than alternative methods we considered that more closely reverse each stage of analysis.

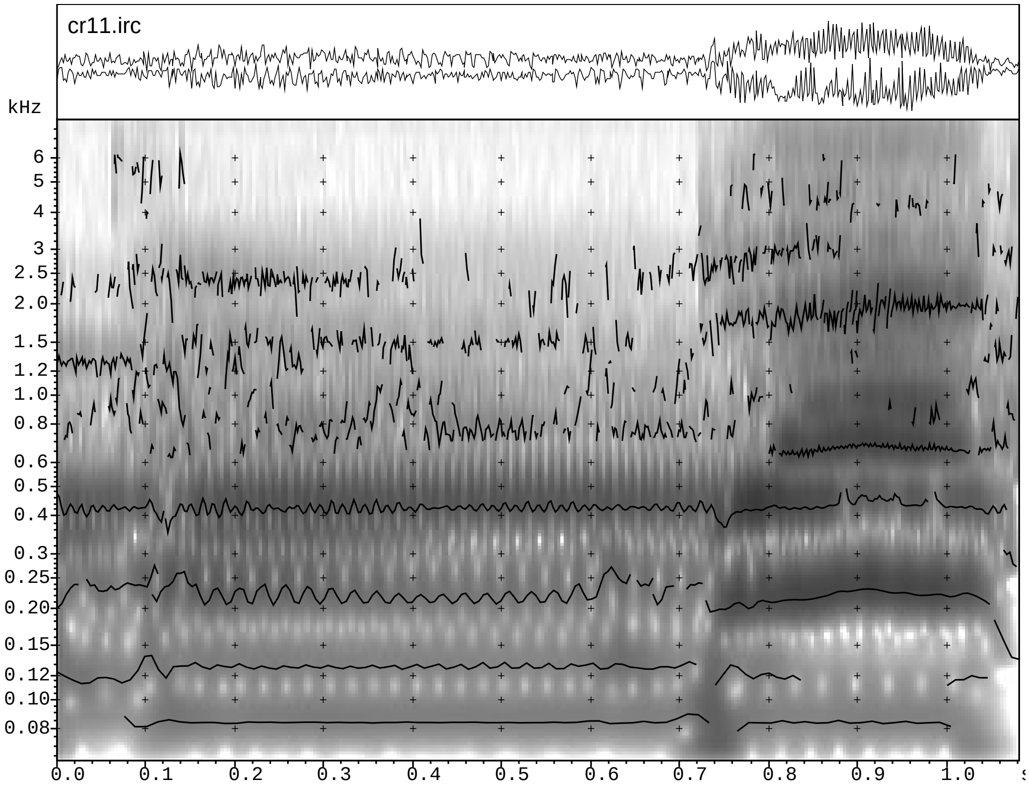

figure 2: Example of CQSWM analysis. Time axis is left to right; top pane shows waveform, lower pane shows filterbank output as grayscale with the tracks from the model overlaid.

5. THE NATURE OF THE REPRESENTATION

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. On this diagram, time goes from left to write, labeled in seconds along the bottom. The upper pane shows the magnitude waveform of the sound - a conventional view. The lower pane represents the output from the filterbank. The vertical axis is frequency labeled in Hertz. The scale is logarithmic, meaning that each uniformly-overlapped constant-Q filter occupies the same height. The gray-scale in the pane shows the magnitude of the output of each filter across time. Excepting the phase, which is not shown, this energy distribution is a complete representation. The black lines overlaid on the gray scale are the tracks of the model. We can see that they trace the ridges in energy concentrations in the filterbank output. The only information stored for each track is the magnitude and phase of the filterbank output directly beneath it, so we can see that this is a very incomplete description, despite its near-perfect reconstruction quality.

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, i.e. fundamental pitch and formant location, but *not* details of individual high harmonics.

The wide bandwidth of the high-frequency filters means that tracks derived from these filters have a correspondingly high bandwidth of modulation, which is difficult to show on the fixed timescale of the diagram. This wideband modulation means that a single track can represent quite a rich sound - such as a formant transition. Consequently, only a small number of simultaneous sinusoids (three to six per octave) is necessary, although the information describing each one may be considerable.

We can see 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 top. This is an example of the masking referred to above, which we hope to escape by use of expected-spectral feedback.

6. USING THE REPRESENTATION FOR SOURCE SEPARATION

Now that we have explained the nature of the representation, we can describe how it is to be used for signal separation. The basic strategy is to form links between tracks that seem likely to have a common origin based on the particular heuristics of source separation, described in section 2. These linked tracks are then clustered into sets presumed to correspond to as many sources as are necessary to describe the sound. This clustering will be achieved by an iterative scheme that seeks to maximize within-cluster links and minimize between-cluster links.

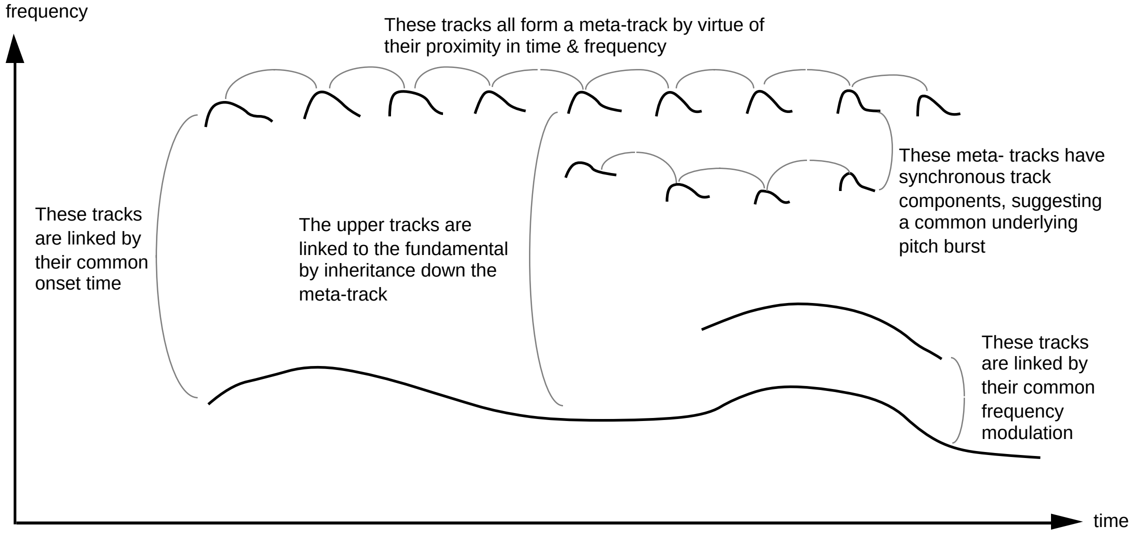
frequency
These tracks all form a meta-track by virtue of their proximity in time & frequency
These tracks are linked by their common onset time
The upper tracks are linked to the fundamental by inheritance down the meta-track
These meta- tracks have synchronous track components, suggesting a common underlying pitch burst
These tracks are linked by their common frequency modulation
time

figure 3: diagram of grouping principles applied to tracks

Figure 3 gives some examples of these links being made on a sketch of some tracks using the same axes as the lower panel of figure 2. In this diagram, we see some tracks that might arise from a formant-filtered pseudoperiodic sound such a voiced speech. The first rule to apply is common onset; the two leftmost tracks start at nearly the same time, and are thus candidates for originating in the same source. However, the high-frequency track shows the very rapid modulation we expect in this region of the spectrum, and has consequently been broken up into a series of short-duration tracks. The proximity and similarity of these tracks must be identified, and in fact they should probably be assembled into a higher-level structure that we will call a meta-track, which will be more directly comparable to the low frequency fundamental track, and will link to it on the basis of both common onset and offset.

The second train of short duration tracks does not start or end with the other tracks, but it does show alignment of the component short tracks with the longer train (presumably because they are sitting on the same pitch pulses). This synchrony, amplified over repeated instances between the same two meta-tracks, serves to link these two.

The remaining track shows no synchrony of any kind. However, it does parallel the frequency modulation contour of the fundamental track, reflecting a fixed harmonic relationship to it perhaps, and this is another common-fate cue that can link it in to the whole set.

Although this diagram is highly simplified, and does not have to deal with any of the problems of interference between different signals or noise corruption, it serves to illustrate how the empirical rules of source formation could be implemented computationally.

The CQSWM track model has other applications less taxing than full source separation. Without actually identifying the correct destination of every last track, it may be easy to segregate the principle energy components of distinct events, and report useful parameters for these events even from such incomplete descriptions. This kind of analysis can be used for *performance extraction*, where the control stream or score of a musical performance can be reconstructed from a recording alone. The symbolic nature of the representation also provides for a new kind of very precise signal-dependent perceptually-salient filtering (such as extracting particular formant transitions or emphasizing particular signals above noise). More exotic modifications of interest to computer musicians are also made possible in this new domain.

7. CONCLUSIONS AND FUTURE WORK

We have presented the problem of automatic source separation as one of analyzing sound into a representation amenable to the application of heuristics discovered through psychological experiment. After discussing various ideal qualities for such a representation, we have described our specific implementation, and how we will use it for separation. At this stage, there are a number of conspicuous obstacles to deal with. Issues we expect to address in the immediate future are:-

- Improve the sensitivity of the track model by techniques such as the expected-spectral-feedback described above. At present, there is an implicit assumption of superposability for track representations that simply doesn’t hold. If the feedback works as envisaged, this assumption would become more reasonable.
- Reduce still further the information in the representation. We have observed the very wide bandwidth of modulation for the high-frequency tracks. Yet such tracks often have only modest apparent perceptual content. One interesting possibility is the smoothing or subsampling of these tracks, on the assumption that the neural pathways connected to the corresponding very broadband haircell resonators are probably incapable of transmitting all the detail they are supplied. If correct, this could lead to a considerable datarate gain. (Some evidence relating to amplitude-modulation envelope discrimination appears to support a smoothing argument - see [Math47]). More mundane reduction could be achieved by pruning away tracks which are likely to be masked or otherwise considered insignificant. Again, the discrete nature of the representation makes tone-on-tone masking rules etc. relatively easy to apply.
- When tracks can be sorted and resynthesized into sources by hand satisfactorily, we will concentrate on the simulation of the higher level cognitive processes that apply the heuristics of source formation, as described in section 6.

We hope to progress in these areas in the coming year and are optimistic about the quality of perceptual emulation we will be able to demonstrate.

ACKNOWLEDGEMENT

This work has been supported from several sources. In this regard, we would particularly like to thank for their generosity : ESL, Nichidai, The Commonwealth Fund of New York and The Television of Tomorrow Consortium.

REFERENCES

[Breg90] AS Bregman (1990) *Auditory Scene Analysis*, Bradford Books MIT Press

[Duda90] RO Duda, RF Lyon, M Slaney (1990) “Correlograms and the separation of sounds” Proc Asilomar conf on sigs, sys & computers

[Glas90] BR Glasberg, BCJ Moore (1990) “Derivation of auditory filter shapes from notched-noise data” Hearing Research 47

[Mahe89] RC Maher (1989) “An approach for the separation of voices in composite music signals” PhD thesis, U Illinois Urbana-Champaign

[Math47] RC Mathes, RL Miller (1947) “Phase effects in monaural perception” JASA 19(5)

[McAu86] RJ McAulay, TF Quatieri (1986) “Speech analysis/resynthesis based on a sinusoidal representation” IEEE Tr ASSP 34

[Mell91] DK Mellinger (1991) “Event formation and separation in musical sound” PhD thesis, CCRMA, Stanford U

[Mins86] Marvin Minsky (1986) *The Society of Mind*, Simon and Schuster

[Quat90] TF Quatieri, RG Danisewicz “An approach to co-channel talker interference suppression using a sinusoidal model for speech” IEEE Tr ASSP 38(1)

[Serr89] X Serra (1989) “A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition” PhD thesis, Stanford U

[Sham89] S Shamma (1989) “Spatial and temporal processing in central auditory networks” in *Methods in neuronal modelling,* MIT Press