REAL-TIME AUDITORY MODELS

Daniel Pressnitzer, Dan Gnansia
Equipe Audition, LPE CNRS UMR 8581
Ecole Normale Supérieure
29 rue d'Ulm, F-75230 Paris Cedex 05
{Daniel.Pressnitzer, Dan.Gnansia}@ens.fr

ABSTRACT

The peripheral auditory system is a complex ensemble of mechanical and neural structures that have a profound influence on how we perceive sounds. As more became known about the physiology of these structures, computational models emerged to simulate their functions. Auditory models are now widely used in psychoacoustics to predict phenomena such as masking, loudness, pitch, roughness, etc. A variety of implementations exist, each tuned to a specific need. In this paper, we argue that a real-time implementation of simplified auditory models can open up a new range of usages, both for music (estimation of perceptual attributes, visualisation, analysis-resynthesis) and research (simulation of hearing impairments). We have developed an implementation of such models in Pure Data, starting with an auditory filterbank.

1. A BRIEF OVERVIEW OF THE PERIPHERAL AUDITORY SYSTEM

1.1. Anatomy and physiology

A sound is a pressure variation that reaches our ears. It has to be converted into neural activity in the brain before we can perceive it. The outer, middle and inner ear perform successive stages of this initial transduction, which shapes the features available to perception.

The outer ear and middle ear convert the aerial vibrations into mechanical ones, which are then communicated to the fluids of the cochlea. Running along the cochlea is the basilar membrane, an elastic structure where travelling waves are provoked by the mechanical vibrations. The travelling waves have an envelope that depends on the frequency content of sounds: the displacement is maximal near the base of the membrane for high-frequency tones, and near the apex for low-frequency tones.

Lying on top of the basilar membrane is the organ of Corti. This organ contains inner and outer hair-cell. When the basilar membrane is set into motion, the hairs of hair-cells are deflected and there is an increased probability of neural discharge in the corresponding auditory nerve fibres. Inner hair cells mostly transmit information to the brain, whereas outer hair cells mostly receive information from the brain.

1.2. Tonotopic and temporal coding

The cochlea thus performs a dual coding of the features of incoming sounds. The basilar membrane encodes mechanically the frequency content of sounds with its profile of displacement. This first code has been termed tonotopy, as it codes tone with place.

The second code is a temporal one. The probability of discharge in the auditory nerve follows the phase of displacements on the basilar membrane. The precise temporal structure of the displacement at a particular place is thus encoded in the discharge patterns of corresponding nerve fibres. This code is called phase locking.

1.3. Non-linear processes

There are many sources of non-linearities in the cochlea. Outer hair-cells, for instance, are thought to participate in an active feedback loop modifying the local properties of the encoding. Such non-linearities play a crucial role in the exquisite sensitivity and selectivity of the normal-hearing auditory system.

2. AUDITORY FILTERBANKS

2.1. Gammatone filters

There is a whole field of research devoted to cochlear modelling. The aim can be to better understand the structure of the cochlea described above, by accurate mechanical modelling, or to reproduce its signal processing characteristics on a functional level. Within this functional approach, an important tool has emerged in the form of the auditory filterbank.

Tonotopic and temporal coding in the cochlea provide the basis for some sort of time-frequency analysis by the auditory system. An auditory filterbank, as the name indicates, is a set of filters that try to reproduce the particularities of this time-frequency analysis. The precise shape and parameters of the filters can vary from model to model, but they are all fitted to perceptual or physiological measures. We will here consider the "gammatone" auditory filterbank. The shape of gammatone filters were chosen to fit perceptual masking experiments [1]. The hypothesis behind the model was that auditory filters can be viewed as somewhat independent processing channels, each with a given centre frequency and selectivity. In the case of

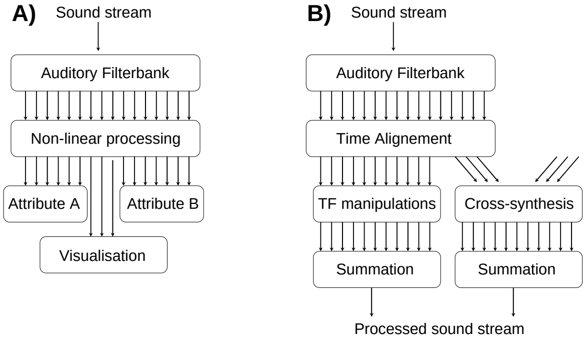

Figure 1. Summary of the proposed applications for a real-time auditory filterbank. A): the filterbank is used as a common front end to estimate auditory attributes - loudness, pitch, roughness, brightness, etc., or to visualise the signal B): a resynthesis is done after some processing in the auditory time-frequency domain (see text for details).

masking, a signal is detected when a fixed signal-to-noise ratio is exceeded in at least one auditory filter.

The gammatone auditory filterbank does not capture the important non-linear processes of cochlear mechanics. Some models include a non-linear component that affects the shape of the “filters” according to the input signal level. In the present approach, we will rather approximate the overall effects of non-linearities by introducing processing modules after the auditory filterbank.

2.2. Real-time implementation

In the context of real-time processing, the choice of implementation matters. We are currently investigating the gammatone implementation proposed by Hohmann [2]. It has two main advantages for our purposes: it is computationnaly efficient, and it allows for a resynthesis with minimal distortion in almost real-time. One problem with resynthesis at the output of a gammatone filterbank is that differences in bandwidth of the filters introduce various temporal delays. A time realignment is thus necessary if a resynthesis is to be made by summing up the frequency channels. Hohmann has shown that, with his implementation, near perfect resynthesis can be obtained with delays of 4 to 8 ms. This is acceptable for real-time processing.

We have implemented this filterbank in the Pure Data programming environment. The idea is to provide a free and readily available auditory modelling library, in the form of a set of patches and external objects. The user has access to the auditory filterbank, but also to a set of post-processing objects that all use the filterbank’s output. Another auditory front-end has been implemented recently in the context of query-by-humming systems [3]. Our approach is similar, with an additional effort to have low-latency and phase alignment of the filters to allow resynthesis as well as analysis (Figure 1).

3. ESTIMATION OF PERCEPTUAL ATTRIBUTES

Auditory models have been used to reproduce the results of a range of psychoacoustical experiments. For each experiment, the models have been tuned to provide a satisfactory fit to the dataset. As a result, many different auditory models exist that differ in their implementation and parameters. Most models, however, can be mapped with more or less distortion onto a common architecture : an auditory filterbank modelling the time-frequency analysis performed in the cochlea, a non-linear stage, and some post-processing.

The main features of several auditory attributes can be captured by using this common architecture. Naturally, the predictions of our post-processing modules won’t provide a perfect fit to any of the experimental datasets. On the other hand, as they come after a plausible front-end, they should capture the important trends of the data with the added benefit of a continuous, real-time estimation.

3.1. Loudness

Loudness, or subjective sound intensity, is a notoriously difficult attribute to measure and to model. The existing models compute spectral excitation patterns, submit them to non-linear compression to estimate partial loudnesses, and then integrate across the frequency range.

An approximate estimation of loudness can be built from the output of a gammatone filterbank. The instantaneous power present in each filter, computed by squaring the filter output, provides a time-varying estimate of spectral excitation patterns. Each channel can then be submitted to a compressive power law function, and the sum of all channels provides an approximation of perceived loudness. The type of temporal integration that should be applied to the real-time loudness estimates is an interesting question that goes beyond the scope of this paper.

3.2. Pitch

Many different types of sound can produce a pitch, and none of the simple representations (waveform, long-term spectrum) can predict perceived pitch in all possible cases. Auditory models of pitch have thus flourished to provide accounts of pitch perception. The time-based models predict pitch from periodicities present in the different auditory channels.

A pitch model can be implemented by adding a running autocorrelation module after the real-time filterbank output. A running cancellation (subtract a delayed version of the signal) can also replace the autocorrelation to pick up periodicities [4]. The cancellation approach has already been implemented in real-time, but it operates up to now on the raw waveform of the sounds to be analysed. Adding the filterbank will make the estimation more robust to noise: the channels in which a clear periodicity is detected can be weighted more in the pitch estimate, and the channels contaminated by noise can be ignored.

3.3. Roughness

Roughness is an auditory attribute of sound that is thought to contribute to musical consonance and dissonance. As auditory roughness is present for chords or intervals that are considered dissonant by Western tonal harmonic theory, it has been suggested that it provides a sensory basis to the theory. Many cognitive aspects of music perception related to cultural knowledge are combined with roughness to define musical consonance, however. In music that does not follow tonal syntax, roughness has been shown to be correlated with tension and release movements [5].

Auditory roughness can be estimated, to a first approximation, by the amount of envelope fluctuation within auditory channels (Figure 2). A roughness model can be implemented by following the auditory filterbank output with half-wave rectification, bandpass filtering, and rms computation. The bandpass filtering after half-wave rectification is a form of envelope extraction, as the signals in auditory channels are band-limited. This filtering should be adjusted to extract preferentially the amplitude modulation frequencies that were found experimentally to produce most roughness.

3.4. Timbre

Musical timbre has many perceptual dimensions, some related to the temporal evolution of sounds, some related to their spectral content. The auditory filterbank provides a unified front-end to compute these dimensions. For instance, it has been repeatedly found that a salient dimension is that of “brightness” or, simply put, richness in higher harmonics. Brightness can be computed in real-time as the centre of gravity of the power at the output of auditory filters. This processing algorithm proposes a solution to the conundrum of selecting a frequency and amplitude scale to compute the spectral centre of gravity, the usual correlate of brightness.

Brightness is but one of many ways to summarise the spectral shape of complex sounds. Many more can be invented to efficiently describe a given sound corpus [6]. Any sound descriptor based on spectral content should benefit from using an auditory filterbank as its front-end, as the filterbank provides a spectral resolution derived from perceptual experiments.

3.5. Musical applications

It is not in our domain of expertise to discuss in depth the musical applications that stem from a real-time implementation of auditory attributes estimation [1] . However, it is easy to imagine that the continuous estimations could be used to feed-back into a musical performance. The loudness, roughness or brightness estimates of a sound stream (instrumental or electronic) could be used as real-time interaction parameters with a synthesis or processing engine.

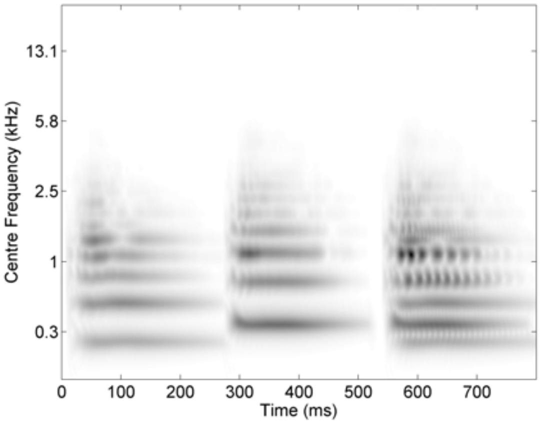

Figure 2. Auditory filterbank output for a trumpet playing a C4, a F#4, and the two notes added together to form a tritone chord. In the chord, the first three spectral components are resolved. The following components interact within auditory channels, thereby producing beats and roughness. This example illustrates the visual representations produced by an auditory filterbank. As the frequency resolution parallels perceptual data, it is adapted to the estimation of auditory attributes such as roughness.

4. AUDITORY REPRESENTATIONS OF SOUND SIGNALS: WYSIWYHEAR

Visual representations of acoustic signals have proven instrumental in our understanding of musical sound. Most commonly used are time-frequency representations, such as the short-term Fourier transform, and many musicians and acousticians have developed intuitions to read them efficiently. Time-frequency representations somewhat mirror the dual encoding performed by the peripheral auditory system. Auditory representations aim to refine this link between representation and auditory encoding.

An all-important parameter for time-frequency representations is the trade-off between temporal and spectral precision for the analysis. In the spectrogram, the trade-off is arbitrarily chosen and it is the same on the whole frequency range. In the auditory filterbank approach, this trade-off is specified by the width of the filters, themselves fitted to perceptual data.

This fit has important consequences. For instance, the presence or absence of beats between adjacent spectral components depends on the selectivity of the filters recruited by the components. Thus, as the width of auditory filters change with centre frequency, a same absolute frequency difference will produce beats in a given frequency regions but no audible beats in another. A visual representation based on an auditory filterbank will automatically reproduce this perceptual feature (Figure 2). It is hoped that such auditory representations will get closer to the “What You See Is What You Hear” goal, and that a real-time implementation will help experimenting with them.

[1] An investigation of the musical relevance of auditory models (real-time or not) can be found in the work of the Leman group, e.g. [7]

5. SOUND PROCESSING AND SYNTHESIS

The chosen implementation of the auditory filterbank allows for a signal reconstruction that is perceptually indistinguishable from the original [2]. It is therefore possible to imagine manipulations on the time-frequency representation discussed in the previous section, and a resynthesis of the resulting sounds. The advantage of using an auditory-inspired representation is that all the parameters accessible for manipulation are expected to have a real perceptual effect. It is again beyond the scope of this paper to enumerate all possible time-frequency manipulations, we just chose two possible examples.

5.1. Cross-synthesis

Different features of the filterbank outputs have different perceptual correlates. The distribution of energy across channels define the spectral profile of the sound, while the fine temporal structure in the channels define pitch and transients. A cross-synthesis algorithm can be imagined that imposes the spectral envelope of a sound on another, while leaving the precise pitch variations and characteristic transients unaffected. To this effect, an estimation of the energy in each channel can be made on one sound and imposed on the other by weighting the channels in the resynthesis. The energy estimation can be smoothed over a relatively long time to improve the quality of the spectral envelope estimation.

Another dichotomy has been shown between the temporal envelope in auditory channels and the fine structure. The envelope carries most of the speech intelligibility, while the fine structure is crucial to the perception of pitch and melodies [8]. Swapping envelope and time-structure within auditory channels provides another type of cross-synthesis. The algorithm is similar to the one above with shorter time-constants for energy estimation.

5.2. Time-frequency granular synthesis

The output of the auditory filterbank are continuous time-signals. If sliding temporal windows are applied in each channel, time-frequency grains are obtained. The grains can be manipulated to provide a sort of granular synthesis that operates within auditory channels. The temporal manipulations may introduce spectral artifacts, that can be removed by applying the auditory filterbank again before resynthesis (at the cost of temporal precision).

6. SIMULATIONS OF HEARING IMPAIRMENT

Another important application of real-time auditory models is the possibility to simulate hearing impairments. A typical cochlear hearing loss implies an elevation of thresholds that is not uniform over the audio range. With the auditory filterbank, it is possible to adjust precisely the thresholds (as a gain) in each channel. But hearing losses also affect the non-linear aspects of cochlear processing. To simulate the loss of non-linear compression, a method is to apply an envelope expansion after an auditory filterbank [9]. A real-time implementation of such a simulation would be a powerful tool to explain to the general public what it actually means to have a hearing loss.

Finally, it is also possible to simulate to some extent the nature of the information that is transmitted to cochlear-implant patients. The technology of cochlear implants consists in stimulating electrically the auditory nerve of profoundly deaf patients in order to restore some aspects of auditory function. Simulations of the resulting auditory percept can be obtained by replacing the fine structure in auditory channels by band-limited noise. Again, such a simulation in real-time could be a pedagogical tool or a way for normal-hearing scientists to develop intuitions about the specific problems encountered by implant patients.

7. REFERENCES

[1] Patterson, R.D. “Auditory filter shape derived with noise stimuli” *J Acoust Soc Am*, 59:640-654, 1976.

[2] Hohmann, V. ”Frequency analysis and synthesis using an auditory filterbank”, *Acta Acustica united with Acustica*, 88:433-443, 2002.

[3] De Mulder, T. et al. “Recent improvements of an auditory model based front-end for the transcription of vocal queries” *ICASSP 2004*, IV:257-260, 2004.

[4] de Cheveigné, A. and Kawahara, H. “YIN, a fundamental frequency estimator for speech and music”, *J Acoust Soc Am*, 111:1917-1930, 2002.

[5] Pressnitzer, D. et al. “Perception of musical tension for non-tonal orchestral timbres and its relation to psychoacoustic roughness” *Perception and Psychophysics*, 62:66-80, 2000

[6] Pachet, F. and and Zils, A. “Automatic Extraction of Music Descriptors from Acoustic Signals”, *Proceedings of ISMIR 2004*, 2004.

[7] Leman, M. “Auditory Models in Music Research”, *Journal of New Music Research special issue*, 23, 1994.

[8] Smith, Z.M., Delgutte, B., and Oxenham, A. “Chimaeric sounds reveal dichotomies in auditory perception” *Nature*, 416:87-90, 2002.

[9] Moore, B.C.J. *Perceptual consequences of cochlear damage*. Oxford University Press, 1995.