Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

IMMERSION AND CONTENT — A FRAMEWORK FOR AUDIO RESEARCH

Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Finland
matti.karjalainen@hut.fi

ABSTRACT

Audio technology is rapidly expanding to various directions. In addition to recording, transmission, storage, and reproduction of sounds it supports practically unlimited modification and generation of sounds and their properties. It is possible to create ever more immersive virtual soundscapes. Audio-related information is also considered not only as signals or data anymore: sound can be analyzed more and more deeply, approaching its content. The focus of this article is to discuss a framework in which modern audio signal and information processing could be placed to see more clearly the explored and unexplored realms of research. This is the author's personal framework that has helped in shaping research in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology, and hopefully some of the ideas and visions could be useful also to others working in the field.

1. THE FRAMEWORK

The rapidly developing technologies in multimedia, mobile communications, and Internet access, utilizing digital audio, open possibilities to entirely new kinds of applications. There are tendencies of both integration of the existing techniques and diversification along with new directions. In this changing world of audio it is necessary to search for views and perspectives — kind of orientation roadmaps — to be able to navigate to the future.

Figure 1 shows a diagram that the author has used as a framework for orientation in audio and virtual acoustics research. The framework has both a vertical and a horizontal organization which provide means to model a variety of applications in sound acquisition, storage, transmission, reproduction, analysis, and synthesis.

The horizontal (left to right) dimension corresponds to traditional audio techniques of sound signal acquisition by microphones, direct or coded transmission and storage, and playback through loudspeakers or headphones. The vertical dimension is related to levels of abstraction of sound between surface presentation and content-based representation. Relations between vertical representation levels are based on analysis and synthesis processes.

These two dimensions complement each other. In traditional audio the horizontal dimension has been utilized routinely while content has been considered by humans only. Along with more advanced information technology the processing of content of sound is becoming a part of systems and applications.

2. DEVELOPMENT OF AUDIO APPLICATIONS

The development of sound-related technology started with the ability to control physical sound environments (such as rooms, auditoria, concert halls) and sound generation devices (such as musical instruments). The emergence of audio (and speech) technology was based on electroacoustic and electronic inventions to transduce, transmit, and store, and play back sound signals. In this *reproduction* problem the objective is to reproduce sounds in such a way that the listener can have an experience similar enough to the original sound event and environment. For information transfer purposes many features of original sound can be omitted or degraded as far as the important properties, such as intelligibility of speech, are preserved. In music, however, the general goal is to reproduce all aspects that improve sound quality and contribute to an enjoyable or influential experience. In the framework of Fig. 1 this means transmission of channel $A_{\text{in}} \rightarrow A_{\text{out}}$ with good fidelity. Transducers (microphones, loudspeakers), amplifiers, and recording devices play a major role here.

Electronic engineering then provided means for generating signals that never existed in acoustic reality. First in simple synthetic signals, *electronic music* gradually enabled creating more and more rich sound compositions. This synthesis path $P_{\text{out}} \rightarrow S_{\text{synth}}$ was first controlled by manual control devices. In parallel to this, sound signals were decomposed into their constituents by devices such as the channel vocoder, whereby a signal may be transmitted or stored in parametric form, i.e., analysis by $S_{\text{in}} \rightarrow P_{\text{in}}$ cascaded with synthesis by $P_{\text{out}} \rightarrow S_{\text{synth}}$. The important finding here was that by selecting a proper set of signal parameters the formal quantity of information may be reduced considerably without degrading the perceived quality too much. As a complement to synthesis, the development of signal analysis tools for $S_{\text{in}} \rightarrow P_{\text{in}}$, such as spectrum and spectrogram analysis, greatly contributed to the understanding of acoustics, speech, and audio.

The next big step started along with *digital signal processing* (DSP) when applied to audio and speech. Improved accuracy and perfect storage were first utilized in Compact Disc and studio/laboratory systems. DSP provided means to control the details of audio signals and together with general purpose computers DSP improved the generation of synthetic sounds in *computer music* [1]. Now it was possible to realize sound synthesis from representations of musical structure and content, i.e., path $C \rightarrow P \rightarrow S$. The same happened with speech where synthetic speech based on modeling of the human speech production was successful.

In the recent years *spatial audio* or spatial sound reproduction, often called 3-D audio, auralization, etc., has been an act-

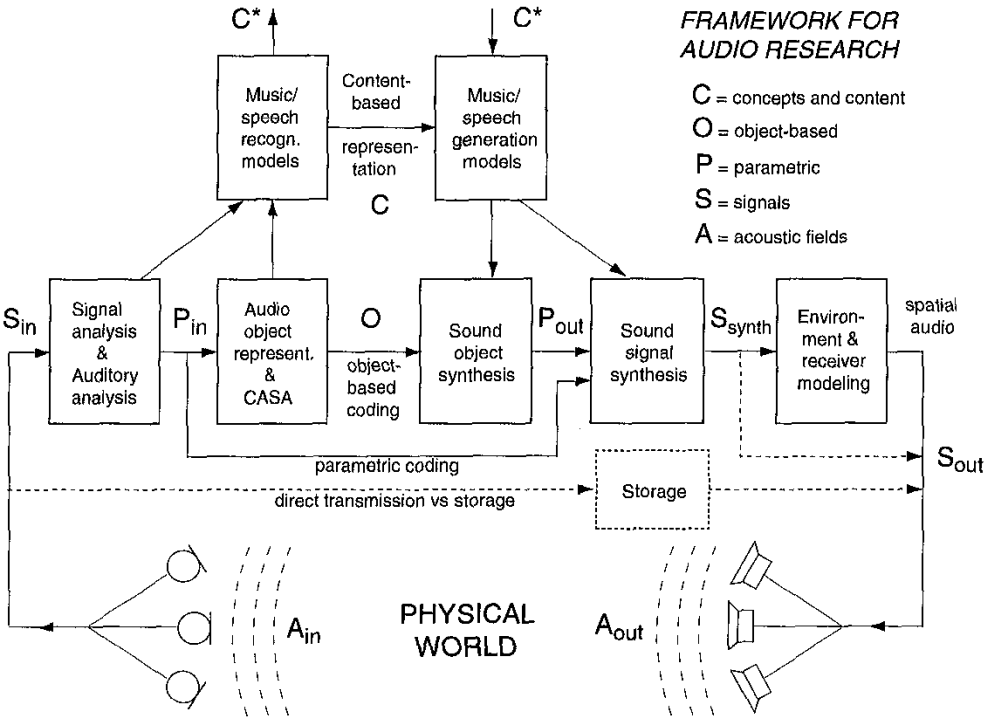

Figure 1: *A framework for audio research and development.*

ive and successful area of R&D. It is related to the signal path $S_{\text{synth}} \rightarrow S_{\text{out}} \rightarrow A_{\text{out}}$ in Fig. 1. Already early in audio technology two-channel stereo and multichannel experiments were carried out to improve the spatial quality of reproduced sound. The objective of current spatial audio is to render sound fields that provide *immersion* [2] into a virtual sound environment [3], being as realistic or expressive as possible.

An area of audio research and development which is still very young or just emerging is structural and content-based analysis. Model-based parametric representation of sound is already on its way, especially for synthesis and *structured audio* [4]. Here we can speak also about object-based representations of audio signals where the analysis path is $S_{\text{in}} \rightarrow P_{\text{in}} \rightarrow O$ and synthesis path $O \rightarrow P_{\text{out}} \rightarrow S_{\text{synth}}$, see Fig. 1. Advanced low-level audio analysis has been based for example on *auditory modeling* in the form of *psychoacoustic models* which are used extensively in perceptual audio coding [5] and also in objective sound quality measurements [6, 7]. Real content analysis $S_{\text{in}} \rightarrow (P_{\text{in}}) \rightarrow C$, however, is just beginning. The challenge is taken recently in CASA, computational auditory scene analysis [8, 9, 10], where the higher-level human auditory functions are the target of simulation.

2.1. Future directions

It is obvious from the discussion above that the main unexplored territory in the diagram of Fig. 1 is advanced audio analysis. This needs both powerful and efficient signal analysis techniques for parameter and feature analysis and simulation of the human auditory functions. Auditory modeling and CASA may have a special role since the relevance of a signal representation is typically based on how we perceive this information.

Interesting and extremely challenging tasks of advanced audio analysis are, e.g., *content analysis*, *automatic music transcription*, and *content-based coding*. All such tasks vary widely in task difficulty. For example automatic transcription of a musical instrument played in isolation, non-polyphonic, without background noise, may be an easy case, while real polyphonic orchestral sounds can be impossible to decompose into an analytic representation, even by the best human analytic listener.

Audio content analysis is found important in future multimedia since it may help, e.g., to navigate in the huge resources of audio data, using the Internet (MPEG-7). Automatic music transcription is a task that can be compared with speech recognition: we may make computers "understand" musical structures. When this becomes possible, there is a short way to audio "coding" by content and conceptual structures. A music recording could be automatically represented for example by common notation whereby notes as objects may include parameters such as the instrument and its playing features, acoustical environment, rule-based relations between such objects, and other related information. Synthesis and decoding techniques for playback largely exist already. Such object-based representation of audio signals makes them very flexible for almost unlimited modifications and reconstructions.

Combined with advanced modeling and rendering of virtual acoustic environments, content-based processing will lead to kind-of *deep immersion* where the perceptual scene is very realistic and we can explore it by concept and content.

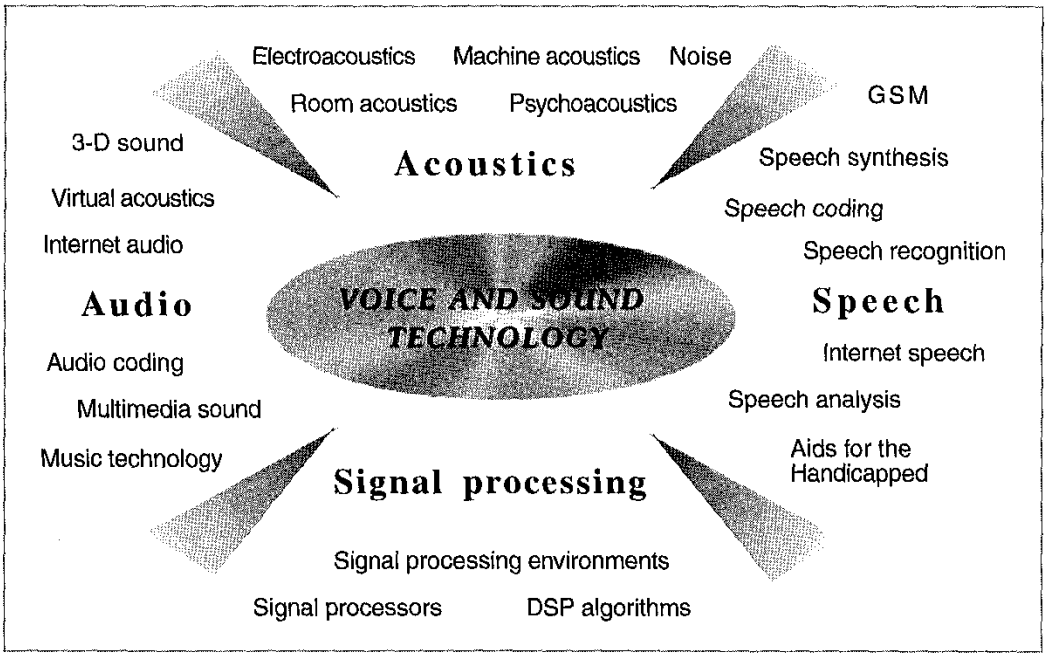

Figure 2: *Research activity map of the HUT Laboratory of Acoustics and Audio Signal Processing.*

3. CONCEPTUAL DISCUSSION

In order to clarify some basic audio concepts, as used by the author in this article, the following list is compiled:

Immersion: to render soundscape presentations in such a way that the listener has an impression of being entirely within a realistic sound environment. A typical meaning of this concept could be called *surface immersion* where only sensory perception aspects are created by the technical system. The term *deep immersion* could be used to refer to such a virtual environment where the system is able to do advaced interactive content-based processing of virtual soundscapes.

Structure, (*structural representation*): audio signals and environments are represented by decomposing them into object-like entities and their parameters. The structure may represent the physical sound source or the perceptual structure of the source.

Content, (*content-based processing*): refers to concepts that are natural to the thinking and language of subjects. Notice that the difference between content and structure is vague (as it is also between structure and simple parametric representation of a signal).

Virtual acoustics is used here as a general term to cover acoustic aspects of virtual reality. Virtual acoustics may consist of virtual sound sources, virtual acoustic environments, and virtual sound receivers (e.g., listeners). In some cases the goal is not to create virtual acoustics entirely independent of the current physical environment but just to modify existing acoustics. This may be the case for example when improving a concert hall by electroacoustic means. We may call this *active acoustics* or *artificial acoustics*.

4. WHERE DO WE FOCUS ON AT HUT

The Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology is a research and education unit that is active in acoustics, audio, speech, and related signal processing. Combining these four fields in an integrated way in a medium-sized academic unit is found challenging but it is also a special strength.

The framework of Fig. 1 has been used as one of our planning tools for research projects and general orientation. Figure 2 illustrates the activity map of current research topics.

Acoustics: Fundamentals of *physical acoustics* is considered important without which understanding audio and speech is difficult. A topic of increasing importance is *engineering psychoacoustics* [11] since practically in all sound and voice related applications the perception of sound or the effect of sound on a listener is essential.

Audio: Topics related to modern audio technology are the most active present areas of research, including *3-D audio* and *spatial sound reproduction*, both loudspeaker-based and headphone-based [12, 13] techniques. *Model-based sound synthesis* (physical modeling) is another topic of success [14, 15] in the 1990s. When the modeling of musical instruments, room acoustics, and the sound as received by a listener are integrated together, in such a way as is done in the DIVA project [16], these components make complete virtual acoustic environments (also supported with visual rendering). This can be done with moderate realism in real time and with more fidelity when extra computer time is allocated.

One important area of research is auditory modeling, both monaural [17] and binaural [18, 19], and its applications. A new

topic in audio research is structural and content-based analysis and representation of audio signals [20].

Speech: Some speech processing topics, such as speech analysis and synthesis, have a resemblance to their counterparts in audio. For example, long tradition in speech synthesis studies [21, 22] has helped in model-based sound synthesis of musical instruments. The focus of speech studies has recently been on developing speech databases and database systems. A similar challenge may be met in audio when advanced structural and content-based audio processing will be developed.

Signal processing: Most modern audio and speech applications require good knowledge and skills in digital signal processing. Signal processing is more like a tool for implementation or research than an independent topic of research in the HUT Acoustics Laboratory. However, some topics such as fractional delay filters [23] and frequency warped signal processing [24] have started a life of their own.

5. SUMMARY AND CONCLUSIONS

A framework for audio research has been presented in this paper that is used by the author as an orientation tool for project planning. The framework in Fig. 1 depicts different functionalities that horizontally are related to acquisition, transmission and storage, and rendering of acoustic fields. A current trend here is to develop *immersive* systems where perceptual reality or expression is emphasized. The vertical dimension in the diagram is related to *content*: analysis and synthesis between surface presentation of signals and deep representation of inherent structure or content of audio.

Based on the audio research framework and concept analysis, an activity map of acoustics, audio, speech, and DSP in the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology is presented.

6. REFERENCES

[1] C. Roads, *The Computer Music Tutorial*. MIT Press, Cambridge, Ma, 1996.

[2] D.. McLeod, U. Neumann, C. L. Nikias, and A. A. Sawchuk, "Media Immersion — Integrated Media Systems," *IEEE Signal Processing Magazine*, vol. 16, no. 1, Jan. 1999.

[3] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Academic Press, Boston, 1994.

[4] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured Audio: Creation, Transmission and Rendering of Parametric Sound Representations," *Proc. IEEE*, vol. 86, no. 5, May 1998.

[5] K. Brandenburg and M. Bosi," Overview of MPEG Audio: Current and Future Standards for Low-Bit-Rate Audio Coding," *J. Audio Eng. Soc.*, Vol. 45, No 1/2, 1997.

[6] J. G. Beerends, "Audio Quality Determination Based on Perceptual Measurement Techniques," in *Applications of Digital Signal Processing to Audio and Acoustics* (M. Kahrs and K. Brandenburg, eds.), Kluwer Academic Publishers, 1998.

[7] Special Issue on Sound Quality, *Acta Acustica*, vol. 83, no. 5, Sept/Oct., 1997.

[8] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, Cambridge, Ma, 1990.

[9] D. Ellis, *Prediction-driven Computational Auditory Scene Analysis*. PhD Thesis, MIT, June 1996.

[10] Special Issue on Auditory Scene Analysis, *Speech Communication* 27, April 1999.

[11] M. Karjalainen, "An Overview of Technically Oriented Psychoacoustics in Finland," in *Proc. of Symposium Psychoacoustics in Industry and Universities*, Eindhoven, January, 1997.

[12] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *J. Audio Eng. Soc.*, vol. 47, no. 4, 1999.

[13] V. Pulkki, "Virtual Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, Vol. 45, No 6, 1997.

[14] V. Välimäki, J. Huopaniemi, M. Karjalainen, and Z. Janosy, "Physical Modeling of Plucked String Instruments with Application to Real-Time Sound Synthesis," *J. Audio Eng. Soc.*, vol. 44, no. 5, pp. 331–353, May 1996.

[15] M. Karjalainen, V. Välimäki, and T. Tolonen, "Plucked-String Models, from the Karplus-Strong Algorithm to Digital Waveguides and Beyond," *Computer Music Journal*, vol. 22, no. 3, 1998.

[16] T. Takala, R. Hänninen, V. Välimäki, L. Savioja, J. Huopaniemi, T. Huotilainen, and M. Karjalainen, "An Integrated System for Virtual Audio Reality," *100th AES Convention*, Copenhagen, May 1998. Reprint 4229.

[17] M. Karjalainen, "A New Auditory Model for the Evaluation of Sound Quality of Audio Systems," *Proc. IEEE ICASSP-85*, Tampa, 1985.

[18] J. Backman and M. Karjalainen, "Modelling of Human Directional and Spatial Hearing Using Neural Networks," *Proc. IEEE ICASSP-93*, Minneapolis, 1993.

[19] V. Pulkki, M. Karjalainen, and J. Huopaniemi, "Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model," *J. Audio Eng. Soc.*, vol. 47, no. 4, 1999.

[20] M. Karjalainen and T. Tolonen, "Multi-Pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis," *Proc. IEEE ICASSP'99*, Phoenix, Arizona, 1999.

[21] M. Karjalainen, U. Laine, and R. Toivonen, "Aids for the Handicapped based on SYNTE 2 Speech Synthesizer," *Proc. IEEE ICASSP-80*, Denver, Co, 1980.

[22] M. Karjalainen, T. Altosaar, and M. Vainio, "Speech Synthesis Using Warped Linear Prediction and Neural Networks," *Proc. IEEE ICASSP-98*, Seattle, Washington, pp. 877–880, 1998 May 12-15.

[23] T. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unit Delay — Tools for Fractional Delay Filter Design," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, January 1996.

[24] M. Karjalainen, A. Härmä, U. K. Laine, and J. Huopaniemi, "Warped Filters and Their Audio Applications," *Proc. WASPAA-97*, Mohonk, New Paltz, 1997.