Psychoacoustic contributions to new techniques for the representation of and interaction with multi-media content

Armin Kohlrausch[1],[2], Jeroen Breebaart[1], Martin McKinney[1],
Steven van de Par[1], Janto Skowronek[1]

[1] *Philips Research Laboratories, Prof. Holstlaan 4, NL-5656 AA Eindhoven Email: armin.kohlrausch@philips.com*
[2] *Technische Universiteit Eindhoven, Human Technology Interaction, P.O. Box 513, NL-5600 MB Eindhoven*

Introduction

This contribution to the 'Vorkolloquium: Von der Psychoakustik zur Mensch-Maschine Kommunikation' of the DAGA meeting in 2005 tries to bridge the two poles covered in the title of the colloquium, by showing how psychoacoustic knowledge supports developments in user-system interaction (Mensch-Maschine Kommunikation). It gives some examples from the domain of multi-media technologies, based on recent work performed by the authors at the Philips Research Laboratories in Eindhoven, The Netherlands. By emphasizing our own contributions, the choice of topics is of course highly selective and subjective, and due to the space limitations, even these few chosen topics cannot be covered in depth. The interested reader can read more about present-day research in this area in the contributions to the session 'Music Processing' elsewhere in these proceedings.

The development of psychoacoustics in the past 100 years has always been tightly coupled with technological developments which required detailed knowledge about the functioning of the human hearing system. To give an early example: Around 1914, the Bell Laboratories started a comprehensive research program on hearing and speech, with the goal to improve the design of telephone systems. Also nowadays, psychoacoustic knowledge significantly influences (multi-)media technologies. A well-known example is perceptual audio coding which allows to represent digital audio content at strongly reduced bit rates (compared, e.g., to the CD) without audible degradation. Although it appears that the approaches which have lead to the MPEG1 standard in 1992 (of which the layer 3 subpart became later known as MP3 format) are reaching a limit, techniques based on parametric descriptions of (part of) the music signals enable additional strong bit rate reductions compared to, e.g., the MP3 format. The first two examples in this contribution deal with such parametric coding applications which are based on our earlier modelling work of monaural and binaural psychoacoustic phenomena. A comparatively new area is the automatic analysis of audio content, e.g. for purposes of classification, or the transcription of audio signals into musical scores. Such techniques allow to derive and add 'semantic' information to an audio file and will certainly play a great role in enabling an easier and more intuitive interaction which digital media contents. An example of our work in this area will end this contribution.

Perceptual noise substitution

One important concept in perceptual audio coding is to remove signal redundancies during encoding. A difficult signal type for this redundancy removal are noise-like signals. On the one hand, these signals have a high signal entropy and thus require a high bit rate in the encoder. On the other hand, these signal types have a much lower 'perceptual entropy', meaning that humans have great difficulties to distinguish between different statistical realizations of noise with the same spectro-temporal envelope. Based on these observations, we decided to develop and implement an algorithm that automatically identifies and substitutes noisy signal parts in different spectro-temporal regions. In contrast to other approaches, e.g. [1], we applied a perceptual model in order to compute a decision variable, that determines whether such a noise substitution will be audible or not. Based on that decision a substitution algorithm was developed and different audio signals were processed. The algorithm makes use of the fact that the modulation spectrum of noise is characteristically different from that of harmonic signal components. The audio signal is first decomposed into critical bands, as in the human cochlea. Each bandpass filtered signal is then rectified and low pass filtered to model the hair-cell characteristics. From the resulting signal, the power spectrum is computed and the spectral values are combined according to the bandwidths used in the modulation filterbank model proposed by Dau et al. [2]. This internal representation of the original audio signal is compared to that of another signal, in which a specific time-frequency section of the audio signal is replaced by noise. The distance in internal representation is used as a measure of perceptual similarity between original and noise-substituted signal, and if the similarity reaches a specific criterion, noise substitution is allowed. In [3] it was shown how well this approach works for stationary signals. In the meanwhile, we have extended it to nonstationary signals. In order to reach high perceptual quality for such signals, it is necessary to identify signal *transients*, because no noise substitution should be performed for such signal components.

Parametric stereo coding

Another example of a hybrid audio coding scheme is the extension of a traditional mono coder with a parametrization of spatial parameters. The first description of such algorithms became known as Binaural Cue Cod-

ing ([4, 5]). This scheme initially focussed on representing spatial *localization* parameters, but did not include parameters related to the spatial ambience. The BCC schemes are able to capture the majority of the sound localization cues, but suffer from narrowing of the stereo image and spatial instabilities, suggesting that this technique is mostly advantageous at low bit rates.

In the parametric stereo coding scheme developed jointly by Philips and Coding Technologies, extensive use was made of our experience in binaural modelling [6]. This knowledge influenced the *choice* of spatial parameters, the way the parameters are *quantized* and the spectro-temporal resolution used in the spatial analysis at the encoder side and also in the resynthesis stage. Our perceptual evaluation has revealed that for headphone playback, a spatial parameter bit stream of 5 to 8 kbit/s is sufficient to reach a quality level that is comparable to popular coding techniques currently on the market (i.e., MPEG1 layer 3). Furthermore, in the course of the standardization process within MPEG4 it has been shown that a state-of-the-art coder such as aacPlus benefits from a significant reduction in bit rate without subjective quality loss if enhanced with parametric stereo. A detailed description of the parametric stereo coding approach can be found in [7]. Presently, the experience gained in the development of the BCC and the parametric stereo coding schemes is combined to develop a parametric multi-channel coder in the context of MPEG4. At the time of this writing (April 2005), the merger of the proposals by Fraunhofer/Agere and by Philips/Coding Technologies has been selected as reference model 0.

Audio Content Evaluation

The previous two examples described algorithms which allow an efficient representation of the audio/music waveform itself. The next example addresses the problem of how the audio content can be characterized semantically, by generating so-called metadata directly from the audio waveform. Such metadata could, e.g., be used for the distinction between music and speech signals, for the classification of music into specific genres, or they could just indicate the average beat rate and rhythm-related descriptors. Other applications of such analyses are the automatic transcription of the waveform into a musical score, which requires the analysis of note onsets, a multipitch analysis allowing to determine the harmonic relations between the simultaneously played notes, and a timbre analysis enabling instrument recognition. In a more general way, one might expect that such algorithms will provide automatic methods to filter, process and store music data, which will support human users in interacting with large collections of (digital) audio content. Most audio classification systems combine two processing stages: feature extraction followed by classification. From a literature study we concluded that progress in classification performance could be made by developing more powerful features, rather than by building new classification schemes [8]. Thus, we focussed our work

on *features* for classifying audio and music. In a recent study, we have compared the two feature sets most commonly used, based on low-level signal properties and on Mel-Frequency Cepstral Coefficients (MFCC), with two new feature sets based on our psychoacoustic knowledge, and evaluated their performance in the classification of a set of general audio classes and a set of popular music genres.

The analysis, described in detail in [8], revealed that perceptually-informed features, and in particular information derived from the temporal variation of these features, lead to an overall improved classification performance, compared to the performance reached with so-called low-level signal features.

Conclusion

In this contribution, a number of examples were given of how quantitative knowledge about human auditory perception contributes to applications in the area of multimedia technology. In particular in the area of audio content evaluation, we foresee that the combination of signal processing expertise, detailed understanding of the human perceptual and cognitive abilities, and musicology knowledge is the key requisite for further technological improvements. This is thus one of the application domains in which psychoacoustics will have a lasting impact on 'user-system interaction'.

References

[1] D. Schulz. Improving audio codecs by noise substitution. J. Audio Eng. Soc. **44** (1996), 593–598.

[2] T. Dau, B. Kollmeier and A. Kohlrausch. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am **102** (1997), 2892–2905.

[3] J. Skowronek and S. van de Par. Automatic noise substitution in natural audio signals. CFA/DAGA'04, Strasbourg, March 2004, 1101–1102.

[4] F. Baumgarte and C. Faller. Binaural cue coding - part I: Psychoacoustic fundamentals and design principles. IEEE Trans. SAP **11** (2003), 509–519.

[5] F. Baumgarte and C. Faller. Binaural cue coding - part I: Schemes and applications. IEEE Trans. SAP **11** (2003), 520–531.

[6] J. Breebaart, S. van de Par and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model setup. J. Acoust. Soc. Am. **110** (2001), 1074–1088.

[7] J. Breebaart, S. van de Par, A. Kohlrausch and E. Schuijers. Parametric coding of stereo audio. Eurasip J. Appl. Signal Proc. (2005), in press.

[8] M. F. McKinney and J. Breebaart. Features for audio and music classification, 4th International Symposium on Music Information and Retrieval, Oct. 2003, Baltimore, Maryland.