Power Optimised Digital Filterbank as Part of a Psychoacoustic Human Hearing Model

Frank Poppen, Milan Schulte and Wolfgang Nebel

OFFIS Research Institute
Oldenburg, Germany

www.offis.de

ABSTRACT

The psychoacoustically motivated filter-bank design introduced in this paper was systematically examined for power reduction potential by applying methodologies and tools supporting power analysis and optimisation at different levels of abstraction. The application of selected tools from the tool suite that was developed in the European project POET, led to a 62 % reduction of power compared to the initial non-optimised design. We show that optimising integrated circuits for power early in the design-phase at high abstraction levels uncovers potential for dramatic savings of power consumption.

Table of Contents

Table of Figures

1 Introduction

Our society is strongly and increasingly communication-oriented. Focusing on sound and speech, many people experience severe limitations in their activities, caused either by hearing loss or by poor environmental conditions. Sophisticated speech and sound processing algorithms to support a better understanding are based on models of human hearing. One example is the Oldenburger Perception Model [1, 2]. Integrating these models into low power, wearable hearing devices is the key to a barrier-free communication society.

The Gammatone Filter Bank (GFB) as part of the Oldenburger Perception Model has been chosen as a benchmark for a framework developed in the European project POET (Power Optimisation for Embedded Systems). POET was funded by the European Union's program IST (Information Society Technology) for a period of three and a half years. The main objective was to develop a new design methodology and tool suite for power estimation and optimisation in heterogeneous embedded System on Chip (SoC) designs. The key innovation of the approach is to enable design-space exploration for low power system architectures, algorithm optimisations and system partitioning. POET tools manage and optimise all major contributors to power dissipation in large SoC designs such as ASICs, cores and processors, memories, communication and I/O interfaces.

The GFB was an ideal test case for this tool flow since the algorithm was specified in high level ANSI-C and already implemented in RTL-VHDL. Only traditional methodologies without power awareness were used. Starting from algorithmic level a complete redesign was performed.

In the following Chapter 2, we introduce part of the POET design flow that made the design optimisation possible. The original design itself is introduced in Chapter 3. The core of this work is Chapter 4. It contains the optimisation strategies and results. We close in Chapter 5 with our conclusions.

2 Tool Chain

Part of the tool chain developed within the POET project is coarsely depicted on the left side of Figure 1. The system is being defined in either ANSI-C or SystemC. ChipVision's power estimation and optimisation tool ORINOCO® [8] instruments the code with functions to record the dataflow activity during execution of the specification. This information serves as input to the tool together with a library of functional units (FU). Included are power models of e.g. adders, subtractors, multipliers, registers, etc., specially generated for the chosen target technology. In this case, Virtual Silicon's standard cell library for a 18 µm process from UMC has been selected. After scheduling, allocation of FU

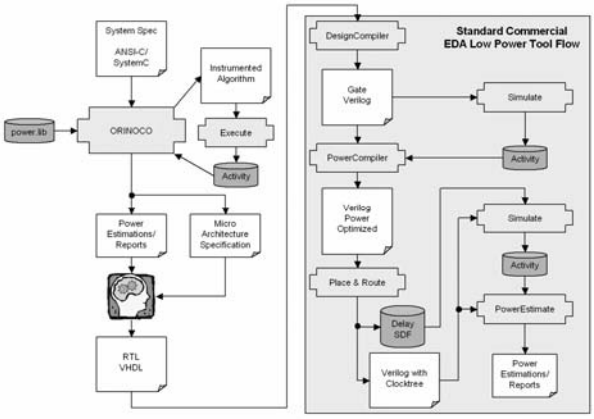

Figure 1 Overview Design Flow.

resources and power-aware binding of operations [3-6] to these, the tool generates detailed reports on power dissipation and power efficient architectures. This supports the architecture designer, in writing a power-efficient RTL description or even making changes at algorithmic level to optimise power.

For synthesis, Synopsys' DesignCompiler is being used together with their optimisation tool

PowerCompiler. The right side of Figure 1 shows the used constellations to connect the tool chain. Other compositions are possible, like e.g. executing HDL-simulations at register transfer level to save simulation time accepting lower accuracy.
First a standard synthesis is executed. The resulting gate-level netlist is being simulated to receive signal activity information stored in a file of the format SAIF. HDL-simulations have been performed with Mentor's ModelSim. Within PowerCompiler the SAIF information is being annotated to the gate level-netlist to perform a power aware resynthesis of the design. The result is a power optimised netlist. Floor-planning has been accomplished by Cadence's FirstEncounter. After place & route wire load delays are extracted with Fire&Ice and exported to a SDF format file. Another gate-level simulation – now including the clock-tree and detailed delay information – traces the signal activity of the design with high accuracy. The tools of the standard commercial EDA flow are in principle replaceable by alternatives from other EDA companies.

3 Design

The DUO (Design Under Optimisation) is the GFB as part of a human hearing model named Perception Model depicted in Figure 2. The GFB reproduces the behaviour of the human basilar membrane which acts like an array of band pass filters. The model in Figure 2 is not complete since the audio signal reaches the membrane by passing the outer and inner ear first. These can be modelled very easily by a simple digital filter that is not shown.

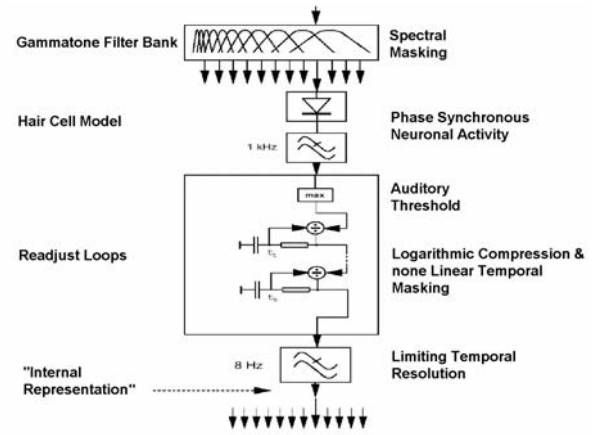

Figure 2 Perception Model of Human Hearing.

Different parameters like e.g. sampling rate, number of channels and centre frequencies for the GFB are possible. The DUO includes two times (left and right ear) 30 IIR (Infinite Impulse Response) filters with centre frequencies from 73 Hz to 7.7 kHz. The audio stream is sampled at a frequency of 16 kHz.
All 60 filters are implemented by the same dataflow depicted in Figure 3. The only difference between each filter is the 30 sets of three coefficients per filter pair (left and right ear).

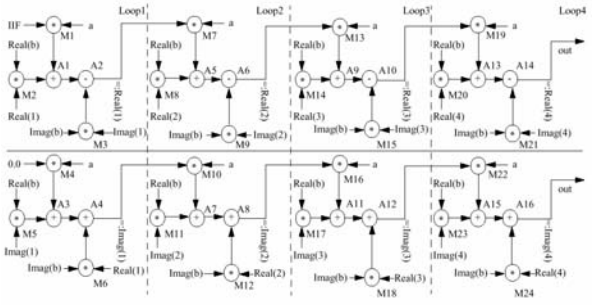

Figure 3 Dataflow of one GFB channel.

The original design approach was mainly area driven. Power aspects where completely ignored. Due to the low sampling frequency of the audio signal, it is not necessary to implement all filters in parallel. It is even possible to execute all operations on the architecture as depicted in Figure 4 with only one multiplier and one adder-subtractor instance.
The temporary, complex filter values are being stored in a single memory with 480 words at 24 bits. Prior to processing a filter, the temporary results are being copied into a register file. If required, this architecture makes it easy to unroll further instances of the data path in parallel with one global RAM. This way the latency can be reduced as required by future demands. The DUO's cellcount is 5,895. The area of the floor plan shown in Figure 5 is 743,476 µm². It takes

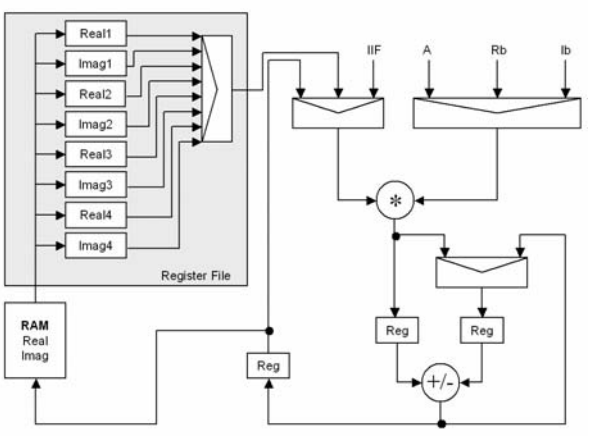

Figure 4 Architecture prior to optimisation.

32 clock cycles at a frequency of 50 MHz to compute one filter. Power dissipation is estimated to be 22.49 mW by Synopsys' PowerCompiler tool. This gate-level estimate includes detailed wire load and delay information back annotated from the place & route back end tool.

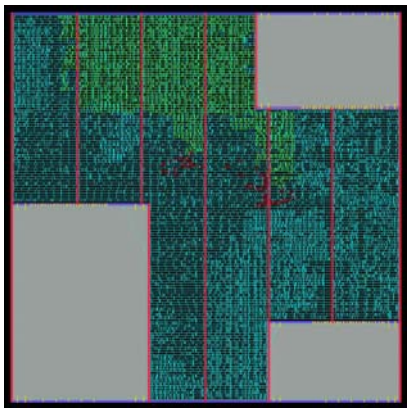

Figure 5 Floor plan of original design. Light green cells belong to multiplier.

The floor plan of Figure 5 also includes an iGFB (inverse GFB) that is part of the benchmark and transforms the GFB output from the frequency domain back into the time domain after applying variable gain values. The iGFB contains two memories placed on the upper and lower right of Figure 5. During analysis, the POET tools showed a relatively small optimisation potential for the iGFB. So we decided against optimisation and manual reimplementation of its architecture.

4 Results

The results of Table 1 that are depicted in Figure 6, show the initial power values of the non-optimised design. Further results of design optimisation experiments are denoted in Table 3. For time and effort reasons, the previous flow in introduced in Chapter 2 hasn't been traversed for all design experiments all the way down to P&R. For these designs, gate-level estimations are available only. Such experiments without P&R information available have to be compared to the values in the row "Gate-Level" and the column "DC" (synthesised with DesignCompiler) of Table 1. Experiments with a complete flow need to be compared to values in the row "Place & Route" of the same column. This implies that the values 17.69 mW and 22.49 mW respectively account for 0 % power-savings in the following discussion. The first statement to be made here is a visible increase in power due to the power dissipation of cells integrated during clock tree synthesis. This increase is about 27 %.

The results of the first experiments are shown in the same Table 1 in the column "PC". Denoted is the standard approach to optimise a design with Synopsys PowerCompiler. Even though the tool includes clock gating and operand isolation optimisation techniques [7] during synthesis, these features have not been included to reduce time and effort for the experiments. The already reasonable improvement of 20 % (compare with Figure 8) can be expected to be raised to around 45 % using clock gating. This guess is based on experiences gathered in the POET project but verifying this conjecture has to be part of future work.

Again, inserting the clock tree increases power after P&R, but compared to the originally P&R design power remains optimised by around 20 %. It can be concluded that P&R is important to evaluate absolute values for power dissipation. To make relative comparisons of designs P&R can be ignored to reduce time and effort for the experiments.

Design	Ref.	DC	PC
Gate-Level	Cell	15.48 mW	12.69 mW
	Net	2.21 mW	1.44 mW
	Clock	n.n.	n.n.
	Leakage	3.49 µW	3.26 µW
	Total	**17.69 mW**	**14.13 mW**
Place & Route	Cell	15.10 mW	12.21 mW
	Net	2.02 mW	1.32 mW
	Clock	5.37 mW	4.14 mW
	Leakage	3.57 µW	2.57 µW
	Total	**22.49 mW**	**17.67 mW**

Table 1 Estimated power dissipation of original design synthesised with DesignCompiler (DC) and PowerCompiler (PC).

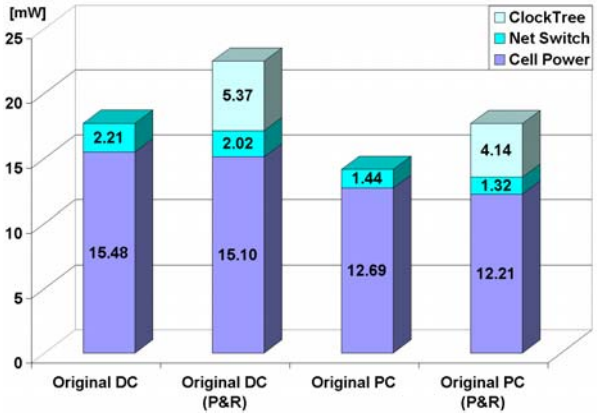

Figure 6 Estimated power dissipation of original design synthesised with DesignCompiler (DC) and PowerCompiler (PC).

In the following, we continue to improve the design starting from these initial savings. One experiment handles clock frequency scaling. At this point we have to make a statement, that is especially important for this low power design methodology:

Average power dissipation, as denoted in Table 1 is a common metric to compare power efficiencies of designs. As a matter of fact this can be misleading since e.g. simply doubling execution time will cut in half average power. Halving the clock frequency usually results in a doubled execution time, but no energy is being saved. In the special case of the GFB benchmark, reducing or increasing execution time can not be achieved by modifications to the clock frequency since the time is being externally framed by the audio signal's sampling frequency of 16 kHz. Reducing or raising the clock frequency will not change the execution time, but it will influence the number of clock cycles that are available to process one sample. Thus Table 1 already contains a valid metric to compare power/energy savings. Nevertheless, for completeness the values have been converted to the according energy values of Table 2 considering the execution time for 500 samples being processed. This results in a period of 0.03 seconds plus reset of the design.

Design	Ref.	DC	PC
Gate-Level	Cell	499 µJ	409 µJ
	Net	71 µJ	46 µJ
	Clock	n.n.	n.n.
	Leakage	112 nJ	105 nJ
	Total	**570 µJ**	**455 µJ**
Place & Route	Cell	487 µJ	394 µJ
	Net	65 µJ	43 µJ
	Clock	173 µJ	133 µJ
	Leakage	115 nJ	83 nJ
	Total	**725 µJ**	**570 µJ**

Table 2 Equivalent energy consumption of original design synthesised with DesignCompiler (DC) and PowerCompiler (PC).

Table 3 includes all results (average power dissipation, as well as, energy values) of each optimisation experiment listed in its own row starting with the design named Orinoco. The experiment is named after ChipVision's high level power estimation and optimisation tool ORINOCO®. Starting from system level the tool supports the design engineer to improve the design to 9.95 mW, reducing power by almost 44 % right away. The savings for each experiment are visualised in Figure 8.

The focus for the Orinoco experiment was to reduce power dissipation in the DUO's FUs. For the chosen technology, dynamic power dissipation overshadows leakage power significantly (compare with Tables 1, 2 and 3). So the main objective was to reduce switching activity.

Neighbours in correlated bit vectors like audio streams differ little in the Most Significant Bits (MSB) since signal changes are slow. Mostly the Least Significant Bits (LSB) add to the dynamic activity. Unfortunately, for the original design the multipliers input values became completely uncorrelated since all correlation is being destroyed by mapping every operation to a single

multiplier instance. This effect is visualized in Figure 7. The four additions executed on one adder a) belong to two audio streams – marked black and white.

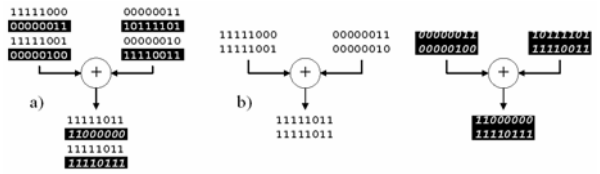

Figure 7 Keeping correlated bit vectors together to reduce activity and power.

Each stream for itself has only few bit-changes from one vector to the second. Mixing the two streams evidently causes high activity in the FU. When the correlation is maintained by mapping each stream to its own resource as shown in Figure 7 b), activity is low and dynamic power is reduced.

The newly designed architecture preserves correlation as much as possible. E.g. one set of filter constants A, Rb and Ib is valid for two successive filter channels: left and right. After finishing processing one filter the constants remain valid and no signal changes at the multipliers input are required. This reduces activity, and therefore power, dramatically.

	Cell	Net	Clock	Leak	Total
Orinoco	8.98 mW 289 µJ	0.97 mW 31 µJ	n.n.	3.04 µW 98 nJ	**9.95 mW 320 µJ**
Orinoco Pow	8.03 mW 259 µJ	0.74 mW 24 µJ	n.n.	2.76 µW 89 nJ	**8.77 mW 283 µJ**
Orinoco Pow, Freq	7.39 mW 238 µJ	0.74 mW 24 µJ	n.n.	2.76 µW 98 nJ	**8.13 mW 262 µJ**
Orinoco Pow, Freq, Volt	6.99 mW 225 µJ	0.60 mW 19 µJ	n.n.	41.6 µW *note	**7.63 mW 244 µJ**
Orinoco Pow, Freq, Volt, P&R	7.12 mW 229 µJ	0.68 mW 22 µJ	0.61 mW 20 µWs	42.3 µW *note	**8.45 mW 271 µJ**

Table 3 Estimated power/energy improvements of experiments with PowerCompiler (Pow), reduced frequency (Freq) and reduced Voltage (Volt).

*Note: Voltage scaling effects have been estimated using "worst case" library of standard cells: Voltage reduced by 10%, Temperature 125°C. The latter is responsible for high leakage and should be ignored.

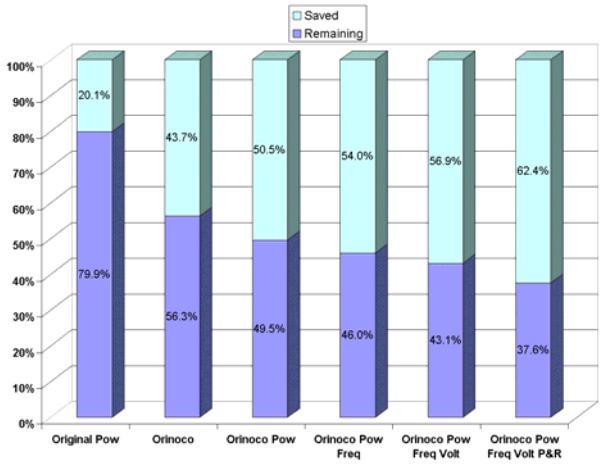

Figure 8 Power-savings of experiments with PowerCompiler (Pow), reduced frequency (Freq) and reduced Voltage (Volt).

At the same time, the concept of a register file to buffer filter values has been discarded in favour to splitting up the memory into two instances of two-port RAM. This way the register count could be reduced dramatically at the cost of no longer being able to parallelize the data path.

This changed the architecture to the structure depicted in Figure 9. The new design contains six multipliers, three adders and one subtractor. The cellcount is 11,250 gates. The area of the floor plan shown in Figure 10 is 1,180,134 µm². This is 59 % larger then the original design. It takes 20 clock cycles at a frequency of 50 MHz to compute one filter instance.

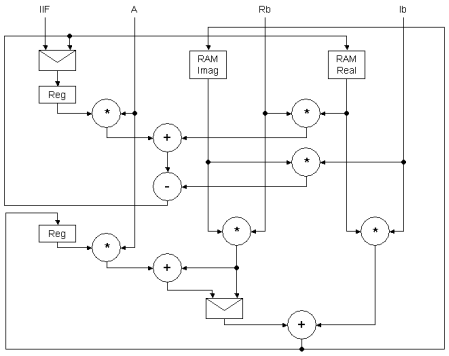

Figure 9 Redesigned architecture of GFB.

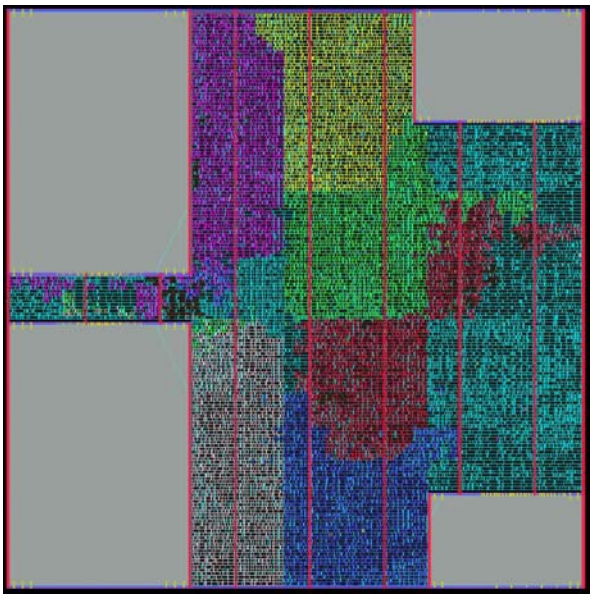

Figure 10 Floor plan of optimised design. Different colours show six multipliers.

Like in the first experiment lower-level power optimisation techniques can be applied to this high level optimised design. The experiment that produced the data for the second row of Table 3 (Orinoco Pow) includes the same gate-level optimisations of PowerCompiler as applied to the original design. This resulted in extra savings of almost 7 %. One might have expected to receive the previous result of an extra 20 % of power-savings but both methodologies – those of ORINOCO® and PowerCompiler – are based on reducing signal activity within designs. The higher level optimisation methodology offers more options for this approach and already harvests part of the potential for the lower level tool.

Due to parallel execution and chaining of operations the new architecture became 1.6 times faster than the original design. Once again we want to emphasize that calculations do not finish early. The external sampling frequency does not allow this. Faster means that processing one audio sample takes less clock cycles (20 instead of 32) raising the idle time the GFB has to wait until the next audio sample is available. With this speedup the design is now idle for more then 50 % of the time. This enables us to reduce the clock frequency of the optimised architecture down to 25 MHz. The estimated power values are denoted in the third row of Table 3 named Orinoco Pow, Freq. Another 4 % of power is saved, which is less than one might have expected. To complete the picture, it has to be noted that the experiment was executed without P&R and that the clock tree that profits the most of a reduced clock frequency is not included yet. We want to refer to the last experiment that has been separated from the other results by a triple line and that includes P&R and therefore clock tree information. These values imply an effective saving of about 9 % in power for reducing the clock frequency.

This last and the fourth experiment share the same architecture. For this design experiment the voltage has been reduced by 10 %. This was possible because the critical path within the designs logic had a positive slack of close to 40 % of the new clock period. By reducing the voltage power can be saved on the cost of extra delay. About 3 % of power was saved. For this experiment the available worst-case models of the standard cell library have been applied instead of the typical case models. Next to a reduced voltage from 1.8 V to 1.62 V and a longer delay this also resulted in an increased junction temperature from 25 °C to 125 °C. Even though this increase in temperature is unwanted for the experiment it could not be avoided due to missing appropriate standard cell models. Temperature does not have an effect on dynamic power but on leakage. This explains the dramatic increase in leakage power which is not the effect of voltage scaling and therefore should be ignored.

Even though P&R of this last design increases the power (compare results of row 4 and 5 of Table 3), the saving depicted in Figure 8 goes up by 5 %. This is not an error. One should keep in mind that the results of P&R have to be related to the P&R values of the original design and can not be compared to none P&R values. So compared to the value denoted in Table 1 the final saving is 62.4 %. As already mentioned above, this last 5 % is the effect of frequency scaling and the included clock tree after P&R.

5 Conclusions

The overview on our experiments shown in Figure 8 shows that we are able to reduce the power dissipation of a psychoacoustically motivated filter-bank design by 62.4% by applying high and low level low power methodologies and tools. This value still leaves ample room for improvement as has been mentioned before. Not all methodologies of the standard commercial tools have been exploited since they are widely used and their virtue is known. No deeper insight in this technology could be expected and we decided to save on time and effort instead. Nevertheless, it is to be expected that clock gating and operand isolation will reduce power even more. On top of that it would also be possible to reduce the supply voltage below 1.68 V. Missing standard cell models did not allow an estimate with the tools but we want to make a rough extrapolation:

According to [10] the increase of delay Δ which sets the clock frequency f depends on the supply voltage V_{dd} and the threshold voltage V_t:

$$\frac{1}{f} \sim \Delta \sim \frac{V_{dd}}{(V_{dd} - V_t)^{\delta}}$$

The dimensionless constant δ is the technology dependent saturation velocity index that can be derived from our experiments of Chapter 4. From this we can assume for the used standard cell technology that the voltage can be decreased to 1.52 V. The following widely accepted formula implies that the dynamic power consumption $P_{dynamic}$ of a CMOS circuit is proportional to the switching activity α, capacitive load C, clock frequency f, and the square of the supply voltage V_{dd}.

$$P_{dynamic} = \alpha \, f \, C V_{dd}^2$$

Using this dependencies and the results of our experiments, we can extrapolate a final power-saving potential of 65 % (7.9 mW) without clock gating and operand isolation.

The high level estimation and optimisation tools of the POET project enable fast analysis of alternative algorithms. This way, alternatives like different filter designs, utilizing tolerances in the model of human hearing, modifying filter constants, using different bit widths, etc. can be evaluated at system level without manual, time consuming RTL-HDL recoding and synthesis. This potential is available on top of the results reported in this document, stating that further optimisations are likely to be possible. This shows the dramatic power-saving potential uncovered by high level power estimation and optimization tools included early in the design-phase of integrated circuit design.

6 References

[1] Dau, T.; 1996 “Modeling Auditory Processing of Amplitude Modulation” Dissertation, University of Oldenburg, Medical Physics, Oldenburg, Germany

[2] Hohmann, V., 2002 “Frequency Analysis and Synthesis using a Gammatone Filterbank” Acustica / acta acustica 88(3): 433-442

[3] L. Kruse, E. Schmidt, G. Jochens, A. Stammermann, A. Schulz, E. Macii, and W. Nebel, Feb. 2001 “Estimation of Lower and Upper Bounds on the Power Consumption from Scheduled Data Flow Graphs” IEEE Trans. On Very Large Scale Integration (VLSI) Systems, Vol. 9, No. 1

[4] L. Kruse, Okt. 2001 “Estimating and Optimizing Power Consumption of Integrated Macro Blocks at the Behavioral Level” Dissertation, University of Oldenburg, Computer Science, Oldenburg, Germany

[5] A. Stammermann, D. Helms, M. Schulte, A. Schulz, and W. Nebel, Nov. 2003 “Binding, Allocation and Floorplanning in Low Power High-Level Synthesis” Proc. ACM/IEEE Int. Conference on Computer Aided Design, San Jose

[6] E. Macii, M. Pedram, F. Somenzi, 1998 “High-Level Power Modeling, Estimation, and Optimization” IEEE Transactions on Computer-Aided Design, Vol. 17, No. 11

[7] Synopsys Inc., “Synopsys Online Documentation V-2004.06”

[8] ChipVision, “ORINOCO-DALE 2005.1.1 User Guide”, www.chipvision.com

[9] P. Babighian, L. Benini, E. Macii, January 2005 “A Scalable Algorithm for RTL Insertion of Gated Clocks based on Observability Don't Cares Computation,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 24, No. 1, pp. 29-42

[10] Flavius Gruian, 2002 “Energy Centric Scheduling for Real-Time Systems” Dissertation, department of computer science, Lund University