LOW-POWER IMPLEMENTATION OF AN HMM-BASED SOUND ENVIRONMENT CLASSIFICATION ALGORITHM FOR HEARING AID APPLICATION

Rong Dong, David Hermann, Etienne Cornu, and Edward Chau

AMI Semiconductor Canada Company
611 Kumpf Drive, Unit 200, Waterloo, Ontario, Canada N2V 1K8
phone: +1 (519) 884-9696, fax: +1 (519) 884-0228, email: tracy_dong@amis.com
web: www.amis.com

ABSTRACT

Automatic program switching is a future trend for digital hearing aids. To realize this function, a solution for sound environment classification is required. This paper presents an HMM-based sound environment classifier that is implemented on a low-power DSP system designed for hearing aid applications. Our experimental results show that it is capable of distinguishing four sound sources (i.e. speech, music, car noise, and babble) with more than 95% accuracy rate and consumes only 0.225 mW of power.

1. INTRODUCTION

Digital hearing aids perform more complex signal processing and offer greater flexibility than analog hearing aids. Digital hearing aids allow clinicians to adjust a variety of parameter settings to meet the individual needs. Since hearing aid wearers are exposed to various listening environments, the signal processing and the parameter settings of hearing aids should adapt to listening environments. In multi-program hearing aids, the parameter settings in each program are optimized for a specific listening environment. The users can manually select the program depending on the current listening environment. As a more user-friendly and inconspicuous alternative, hearing aids with more intelligent capabilities should sense the current listening environment and automatically switch programs. This requires an algorithm to classify the sound environment by analyzing the audio signal. In this paper, we will present a DSP solution for the sound environment classification.

A classification algorithm generally consists of a feature extraction scheme and a classification scheme. The feature extraction scheme computes the most discriminative information from the input signal. Amplitude modulation is a typical feature used in sound environment classification [1, 2, 3, 4]. Due to the strong modulation in speech signals, amplitude modulation is particularly effective for clean speech detection. However, it is not useful in distinguishing noise and music signals [2]. Other features such as amplitude onset [3, 4] and harmonicity [4] are also used to measure the physical attributes of the sound sources. More recently, spectrum-based features, such as mel-frequency cepstral coefficients (MFCC), have been used for sound environment classification [5, 6]. This approach is motivated by their successful usage in speech recognition. With just a few coefficients, MFCC-based features are able to represent the perceptually relevant part of the spectrum and its temporal variation. Based on these features, the second step in a classification algorithm, the classification scheme, performs the actual classification on the feature space. In sound environment classification, heuristic classification approaches were used in the early works [1]. Recently, statistical classification approaches were introduced in this field, including Hidden Markov Model (HMM) [4], discrete HMM [5], and Gaussian Mixture Model (GMM) [6]. Statistical approaches are much more powerful than the heuristic approaches. However, the computational demand of statistical approaches has always been considered prohibitive for hearing aid applications.

In this work, we present a sound environment classifier aimed at tracking the changes in the sound environment. It is implemented on an ultra-low power and miniature DSP system. We use MFCC-based features and HMM classification to take full advantage of the signal processing functions provided by the DSP system. The efficiency and capability of this solution will demonstrate the feasibility of accommodating an HMM-based sound environment classifier in a real-time digital hearing aid application.

The rest of the paper is organized as follows. Section 2 gives an overview of the Ezairo 5900 DSP system upon which the algorithm is implemented. Section 3 describes the HMM-based sound environment classification algorithm and the implementation details. Section 4 discusses the system evaluation results. Finally, Section 5 summarizes the conclusions of the current work and proposes future works.

2. EZAIRO 5900 DSP SYSTEM

The environment classification algorithm is implemented on our new DSP based digital hearing aid system-on-chip. The DSP system uses an asymmetric, dual-core architecture. The chip is fabricated in 130 nm semiconductor technology for ultra-low power consumption and small physical die size.

Figure 1 gives a top-level overview of the DSP system. The entire system is centered around two processing cores: a general-purpose fixed-point digital signal processor called CFX and a configurable signal processing accelerator called HEAR. The CFX DSP is a dual-MAC, 24-bit core. It contains four 56-bit accumulators, four 24-bit data registers, twelve address registers, and other program control registers. Supported by a flexible parallel instruction set, the DSP can

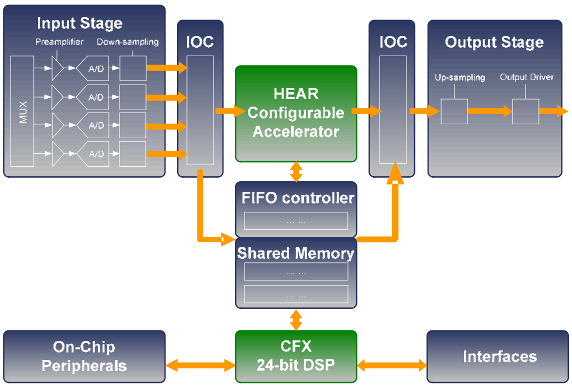

Figure 1 – Ezairo 5900 DSP system block diagram

execute up to four computation operations and two data transfers in one cycle.

The second processing core, the HEAR accelerator, provides a variety of signal processing functions including FIR/IIR filtering, filterbank operations, FFT, statistical operations and vector operations. These functions are highly optimized yet configurable. The developer can specify the configuration parameters and address parameters to fit each application's specific needs and minimize the function execution overhead. With these fundamental building blocks, it greatly reduces the development time of a new algorithm. A sequence of functions can be organized into a function chain or multiple function chains. The developer schedules the launch time of each function chain. While the accelerator is executing a function chain, the DSP can perform other tasks.

The two cores communicate by sending interrupt signals and shared memory. The accelerator is specialized in common arithmetic or signal processing functions, whereas the DSP is designed for controlling the program and the peripheral interfaces. The DSP may also be used for other customized processing that cannot be handled on the accelerator. Optimum system efficiency can be achieved by carefully balancing the computation load of the two cores.

Other than the processing cores, peripheral components and interfaces, the system also contains an input/output controller (IOC), a FIFO controller, analog audio inputs and analog audio outputs. The IOC manages the input/output data flow in the background without loading the cores. The FIFO controller provides hardware-based, configurable FIFO buffers to store the input, output and intermediate processing samples. They can also be used as software controlled circular buffers. The audio input and output circuits allow four input channels and one output channel, and are optimized to achieve high-fidelity audio performance.

3. THE HMM-BASED SOUND ENVIRONMENT CLASSIFICATION FRAMEWORK

To take advantage of the dual-core system, the sound environment classification algorithm is implemented in a parallel processing framework.

Figure 2 illustrates the processing flow of this framework. MFCC feature extraction is comprised of a series of signal processing blocks (white boxes in the figure) executed on the two processing cores. The regular signal processing

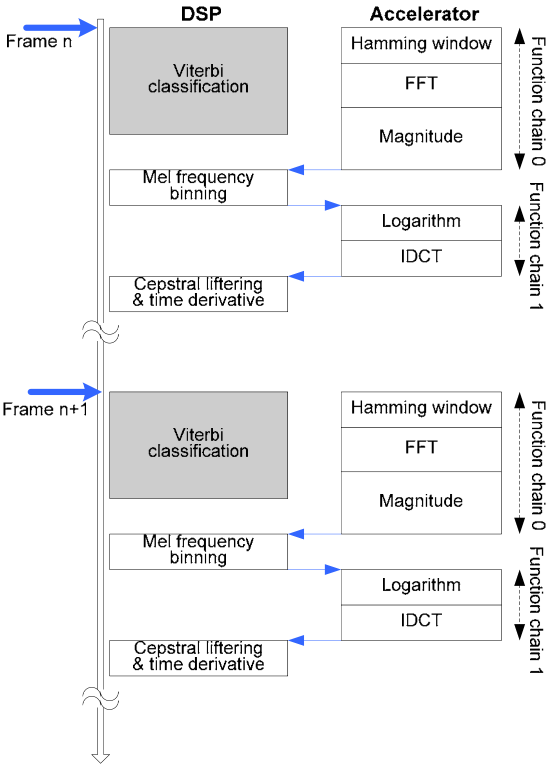

Figure 2 – Algorithm processing flow (not to scale)

and vector operations are executed on the accelerator. The customized function blocks are implemented on the DSP. HMM classification is performed using the Viterbi algorithm [7] which is executed on the DSP (grey box in the figure). The audio acquisition, feature extraction and the classification are pipelined to maximize the parallelism [8]. In other words, Viterbi classification is performed on the features of the previous frame when the accelerator is computing the features of the current frame. In the meantime, the IOC is collecting the audio samples for the next frame. The pipeline processing introduces one frame (8 ms) delay in the classification response. Assuming the sound environment does not change rapidly, this amount of delay is acceptable for this particular application.

3.1 MFCC feature extraction

The input signal is sampled at a 16 kHz sampling rate. Following the analog-to-digital (A/D) conversion, the input signal is divided into successive frames. This operation is managed by the IOC and the FIFO controller in the background. Each frame is 8 ms long with 4 ms overlapping. When a frame is ready, the DSP core launches the first function chain for MFCC feature extraction.

The first function chain begins with 128-point Hamming windowing followed by 128-point Fast Fourier Transform (FFT). Since the real input has a conjugate symmetric FFT, only the first 64-band FFT outputs are generated. Next, the magnitude of each FFT point is computed using a complex vector magnitude function.

Once the magnitude is obtained, the first function chain is completed. An interrupt is sent to the DSP to start mel frequency binning. The purpose is to map the 64-band linear

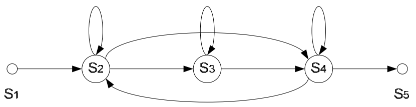

Figure 3 – Structure of the HMM model

FFT spectrum to an 18-band nonuniformly spaced, mel-scale spectrum. After mel frequency binning, the DSP launches the second function chain on the accelerator.

The second function chain begins with a vector logarithm function to calculate the logarithm of the mel-spectrum bin energy. This function gives a table-lookup based approximation of the base-2 logarithm. Next is an inverse discrete cosine transform (IDCT). Depending on the desired number of MFCC coefficients, only the first few orders of IDCT outputs are calculated (the zero-order coefficient is excluded). Therefore it is more efficient to implement IDCT as a matrix multiplication operation.

After the second function chain, the DSP performs cepstral liftering followed by time derivative. Cepstral liftering rescales the MFCC coefficients to approximately the same magnitude in order to improve the precision of the fixed-point implementation. The liftered MFCC coefficients are stored in a circular buffer, from which the first-order delta coefficients are derived.

3.2 HMM classification

The classifier uses HMMs to model the environment sound sources. As illustrated in Figure 3, each model consists of five states including three emitting states, one entry state and one exit state. Each emitting state is associated with a single mixture, multivariate Gaussian probability distribution, which is specified by a mean vector and a diagonal covariance matrix. A transition matrix defines the transition probability between the states. A negative insertion penalty is applied at the transition across the models. The Gaussian probability distributions and the transition matrix are estimated during an offline training process carried out using the HTK toolbox [9]. The value of the insertion penalty is determined by empirical analysis of simulation results.

The Viterbi classification is performed in real-time on the DSP. The objective is to compute the log-likelihood of the states frame-by-frame and determine the most likely sequence of models and states that produce the observed feature vectors [7]. Note that the log-likelihood is a monotonically decreasing value over time. As such, in a fixed-point implementation, it will eventually underflow. To prevent this in practice, we simply normalize the log-likelihood in each frame by subtracting its maximal value across all the models. So the maximal log-likelihood is always 0 at any time.

In Gaussian probability computation, the variables have very different dynamic ranges, which is another challenge for the fixed-point implementation. To achieve the best numerical precision, we used a C simulation to estimate the dynamic range of each variable and identify an optimal fixed-point representation for each variable.

4. SYSTEM EVALUATION

Our experiments use a sound database for training and test.

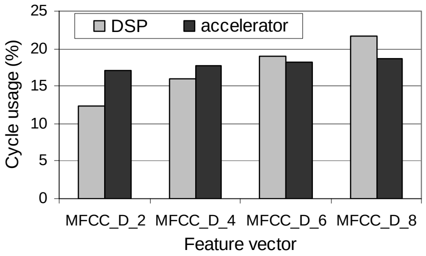

Figure 4 – Cycle usage for different feature vector size

Table 1 – Breakdown of cycles spent on the individual functions (configured with MFCC_D_4 feature vector)

Functions executed on the DSP	Cycle count
Viterbi classification	1177
Mel frequency binning	409
Cepstral liftering & time derivative	46
Total	1632
Functions executed on the accelerator	**Cycle count**
Hamming window	133
FFT	521
Magnitude	888
Logarithm	151
IDCT	104
Total	1797

The database includes 26.5 minutes of training data and 9.2 minutes of test data. All signals are sampled at 16 kHz. The sounds were chosen to represent the four sound sources: speech, music, car noise and babble. The material has certain variety to yield a general model for each class.

As described earlier, MFCC and delta-MFCC coefficients are used as the features. To determine the optimal feature vector size, we explore the tradeoffs concerning computation load, memory usage, and classification accuracy.

To compare the computation load with different sizes of feature vectors, we measured the cycle usage at 2.56 MHz clock frequency and plotted the data in Figure 4. In addition, Table 1 shows a break-down of cycle usage on a function-by-function basis to give some insight into how the cycles are used in detail. As we can see from Figure 4, the algorithm uses less than 20% of the computational resources at this clock frequency. The remaining cycles can be used to execute other function blocks typically required in a hearing aid application. The computation is split between the DSP and the accelerator. As shown in Table 1, Viterbi classification accounts for most of the computation on the DSP. Since the size of the feature vector determines the dimension of the Gaussian distribution, increasing the size of the feature vector leads to significant increase of the cycle usage on the DSP. However, on the accelerator, the two major parts of the computation, FFT and magnitude computation, are independent of the size of the feature vector, thus the increase of the cycle usage on the accelerator is relatively small. From Figure 4, we can see the computation load on the two processors is more balanced with MFCC_D_4 feature vector (4 MFCC and 4 delta-MFCC coefficients) or MFCC_D_6 feature vector (6 MFCC and 6 delta-MFCC coefficients). Therefore the system tends to be more power efficient under these two configurations, when considering only the environment

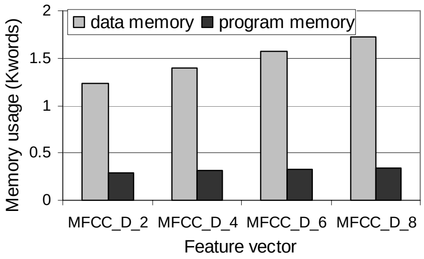

Figure 5 – Memory usage for different feature vector size

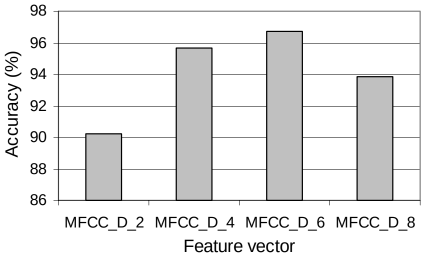

Figure 6 – Classification accuracy for different feature vectors

classification algorithm.

The DSP system has 12.75 Kwords data memory and 14 Kwords program memory available on the chip. The classification process only requires a small amount of memory. Figure 5 illustrates the memory usage of the classification process for different feature vector size. Because the HMM models are stored in the data memory, we see a notable increase in data memory usage as the size of the feature vector increases, whereas the impact on the program memory usage is negligible.

Finally, Figure 6 shows the classification accuracy obtained with different sizes of feature vectors. With the MFCC_D_4 feature vector, the accuracy is higher than 95%. Upgrading from the MFCC_D_4 to the MFCC_D_6 feature vector only provides marginal improvement on accuracy. However, with MFCC_D_8, the accuracy is even degraded. This is usually because the high-order MFCC coefficients contain the information of the fine spectrum structure, which is not useful for characterizing the noise signals.

Based on our analysis, MFCC_D_4 is a possible configuration that minimizes power consumption while keeping a high accuracy. To perform the sound environment classification with this particular configuration, the entire DSP system consumes only 0.225 mW of power.

5. CONCLUSIONS AND FUTURE WORKS

This paper presents an HMM-based sound environment classification framework for hearing aid applications. The algorithm has been successfully implemented on our latest Ezairo 5900 DSP system optimized for hearing aid applications.

The asymmetric dual-core DSP system is centered around a configurable signal processing accelerator and a dual-MAC general-purpose DSP. The configurable accelerator provides highly optimized signal processing functions that typically involve regular computations and data parallelism. The DSP allows customized, non-regular signal processing and program controls, and also supports instruction-level parallelism. In addition to the two cores, other units, such as the IOC and the FIFO controller, manage the input/output data flow in the background without interrupting the cores. With this architecture, we are able to pipeline the audio acquisition, the feature extraction and the classification to maximize the parallelism and the power efficiency.

An experimental analysis was performed to determine the optimal combination of feature coefficients in terms of trading off accuracy versus computation and memory load. The evaluation results show that higher than 95% accuracy rate can be achieved with a feature vector of 4 MFCC and 4 delta-MFCC coefficients. At 2.56 MHz clock frequency, the classifier takes less than 20% of the computational cycles available. Therefore it leaves enough computational resources for the other function blocks in a typical hearing aid application. With the 130 nm semiconductor technology, the DSP chip consumes 0.225 mW of power to perform the classification.

Note that the classifier in the current implementation is intended to identify non-mixed sources. In practice, the listening environment tends to be a mixture of multiple sound sources. Future work will involve the detection of a mixture of signals in a noisy background.

REFERENCES

[1] J. M. Kates, "Classification of background noises for hearing-aid applications," *Journal of Acous. Soc. Am.*, 97 (1), pp. 461–470, January 1995.

[2] V. Hamacher, J. Chalupper, J. eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, 18, pp. 2915–2929, 2005.

[3] D. J. Schum, "Noise Reduction in Hearing Aids: What Works and Why," *Audiological Research Document*, Oticon, Inc., April, 2003.

[4] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound Classification in Hearing Aids Inspired by Auditory Scene Analysis," *EURASIP Journal on Applied Signal Processing*, 18, pp. 2991–3002, 2005.

[5] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," *Journal of Acous. Soc. Am.*, 115 (6), pp. 3033–3041, June 2004.

[6] S. Ravindran and D. V. Anderson, "Audio classification And Scene Recognition and for Hearing Aids," in *ISCAS 2005*, Kobe, Japan, May 23-26. 2005, pp. 860–863.

[7] L. R. Rabiner, "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.

[8] E. Cornu, N. Destrez, A. Dufaux, H. Sheikhzadeh, and R. Brennan, "An ultra low power, ultra miniature voice command system based on hidden Markov models," in *Proc. ICASSP 2002*, Orlando, FL, May 2002, vol. 4, pp. IV-3800 – IV-3803.

[9] Speech Vision and Robotics Group, *HTK Tool Kit*. Cambridge University Engineering Department, Cambridge, UK, http://htk.eng.cam.ac.uk.