Sound Classification Using Summary Statistics and N-Path Filtering

Daniel Villamizar*, Daniele Battaglino†, Dante G. Muratore*, Reza Hoshyar‡ and Boris Murmann*
*Department of Electrical Engineering, Stanford University, California, USA
†NXP Semiconductors, Mougins, FR & EURECOM, Biot, FR
‡Texas Instruments, Santa Clara, California, USA

***Abstract*—Always-on sound classification is a desirable but power-intensive function for a variety of emerging Internet of Everything applications. This work explores the accuracy-complexity tradeoff by using summary statistics for classifying semi-stationary sounds. Compared to contemporary solutions including deep learning, this approach requires one to three orders of magnitude fewer parameters and can therefore be trained over ten times faster. We propose a mixed-signal design using N-path filters for feature extraction to further improve energy efficiency without incurring a large accuracy penalty for a binary classification task (less than 2.5% area reduction under receiver operating characteristic curve).**

***Index Terms*—acoustic environment recognition, acoustic textures, baby cry detection, speaker identification, Internet of Everything, N-path filter, passive mixer**

I. Introduction

The task of sound classification is a topic of active research due to its broad application space. Recent advances in deep learning have shown super-human performance for these tasks but their accuracy comes at the cost of high complexity, high computational loads, and large training datasets. For Internet of Everything (IoE) systems, it is important to perform the classification in real time, using low power, and in an always-on fashion [1], [2], [3]. To stay within their power budgets, systems that leverage large deep learning models must therefore resort to cloud processing at the cost of high latency and privacy issues. These issues motivate the development of systems that can produce the inference results locally.

To implement local inference at low power consumption, it is necessary to restrict model parameter count as much as possible. Further, for emerging applications where only limited training data is available, it is imperative to have a model that can quickly converge on trained parameters. Finally, it is desirable to benefit from application-specific simplifications when there is *a priori* knowledge about the signals of interest. To meet these conditions, we study a classifier that employs summary statistics as efficient and flexible features and present a comparison to other classifiers designed for the same tasks. We show that in some cases the classifier can train 10x faster while being 1000x smaller than a deep learning network and only exhibiting moderate accuracy loss. We then propose a CMOS chip implementation designed to perform the required signal processing at ultra-low power while maintaining flexibility for a variety of applications.

The rest of this paper is organized as follows. Section II introduces the summary statistics feature set. Section III presents a learning rate comparison to a convolutional neural network designed for the same task. Section IV compares the size and computational complexity of the proposed classifier with published work on two additional sound classification datasets. Section V introduces a mixed-signal approach to further increase the computational efficiency of the classifier and presents initial simulated results.

II. Classification with Summary Statistics

Distinguishing sounds that are not strictly stationary but exhibit semi-stationary properties constitutes a useful task for many IoE systems. These types of signals lend themselves to processing techniques that exploit their structure. Examples of sounds that fall in this category include motors, rain, a crowded room, insects, baby cry, sirens, applause, and even voice timbre. Extracting useful information from these types of signals in a computationally-efficient manner requires a feature set designed with this purpose in mind. For this work, we chose to evaluate a bio-inspired analysis based on the work of McDermott and Simoncelli [4], where the authors propose a set of statistics that represent human auditory cortex functionality. Subsequent work tested these statistical sound properties on a classification task in [5] and showed experimental results, but did not compare performance and computation cost to other techniques on common datasets. We further show that an important benefit of this technique is that it maps into an efficient circuit implementation in CMOS technology (see Section V).

The summary statistics (SS) feature set is composed of four main categories: (1) audio sub-band central moments, (2) sub-band correlations, (3) modulation-band power, and (4) modulation-band correlations. The mathematical definition of these features is described in [4]. Here, we only illustrate the resulting signal chain as summarized in Fig. 1.

We evaluated the feature set with multiple models and concluded that the performance of the classification is dominated by the mapping expressiveness of the features and not the classifier model used. We therefore chose a Support Vector Machine (SVM) model employing a linear kernel to ensure

low inference computation cost.[1] We therefore refer to the classifier as SS-SVM. For this model choice, the number of learned parameters are simply given by $M \times D$ where M is the number of classes and D is the number of features used. Performance tradeoff comparisons between this classifier and recent published work are summarized in the following sections.

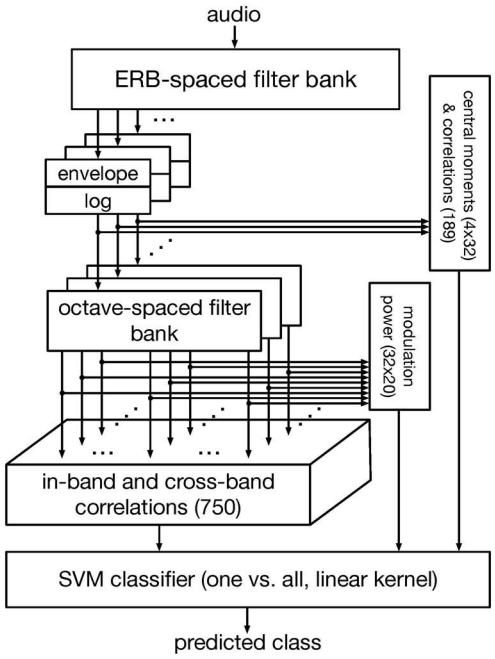

Fig. 1. SS-SVM classifier diagram. Features are defined in [4] and are flattened and concatenated as inputs to an SVM classifier.

III. Learning Rate Comparison

Deep learning techniques are increasingly being used for the tasks of environmental sound detection. For example, out of the top-10 submitted entries to the 2016 and 2017 DCASE challenges [6], [7], fifteen use ConvNets either exclusively or in combination with other feature extracting techniques.

A similar task of practical application (but lower complexity) is that of detecting baby cry events. In order to compare the tradeoffs between deep learning and the summary statistics classifier we designed a ConvNet for baby cry detection. The input to the network is a 5 second mel-spaced spectrogram of 40 frequency bins and 250 time frames of 20 ms duration computed from the raw audio signal. The network is summarized in Table I. A ReLU activation function was used after max pooling for all convolutional layers. Similar networks have been reported in [8], [9], [10], [11].

[1]The SVM model was implemented using the `liblinear` package using one-vs.-all learners for multi-class tasks. Regularization hyper-parameter was optimized for each training run. Experiments using RBF or quadratic kernels did not yield appreciable gains. Other classifiers tested include Gaussian Discriminant Analysis and Multinomial Logistic Regression which also did not perform significantly different than the SVM model used.

TABLE I
ConvNet employed for baby cry detection

layer name	output size	layer parameters
input	250×40	
conv1	84×14	5×5, 8, stride 1, batch norm 3×3 max pool, stride 3 dropout $p = 0.1$
conv2	29×6	5×5, 16, stride 1, batch norm 3×3 max pool, stride 3 dropout $p = 0.2$
output	1×1	1024-d, fc, softmax dropout $p = 0.5$
FLOPs		12.5×10^6
Parameters		2.89×10^6

For the comparison we used the dataset described in [11]. The observed ROC curve is shown in Fig. 2. The learning curves for both systems were reported in Fig. 3 to understand why the SS-SVM classifier slightly outperforms the ConvNet. For this analysis, the classifiers were trained on progressively larger subsets of the training dataset and the error rate was calculated for each. The SS-SVM classifier shows faster learning which is consistent with the fact that it possesses much fewer parameters than the ConvNet (2.9M vs. 2k) and is less likely to suffer from overfit. For the task evaluated, the ConvNet requires 10x more training data than the SS-SVM classifier to achieve 10% error rate (see Fig. 3). This is an important tradeoff in applications that have limited training data or require fast online learning (e.g., reinforcement learning) and should be considered when choosing inference systems.

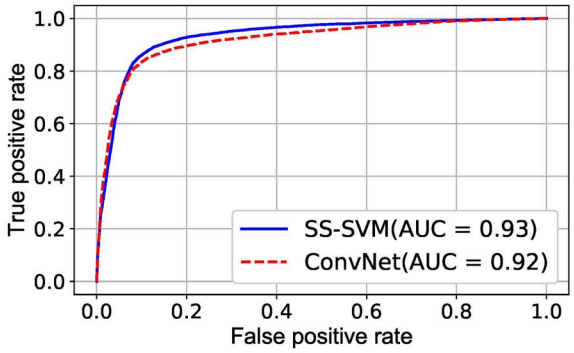

Fig. 2. ROC curve for both classifiers evaluated on the Baby Cry dataset. Accuracy performance is essentially equivalent.

IV. Computation Cost and Complexity Comparison

To understand the tradeoffs in computational cost and complexity, we evaluate the SS-SVM classifier on two additional datasets. For all evaluated tasks, the number of features are of the order of the number of training samples, thus also being likely to suffer from overfit.

A. Environmental Sound Classification

For this task we evaluated on the datasets published in [12], [13], [14], [8], [15], [10], namely DCASE 2016 and

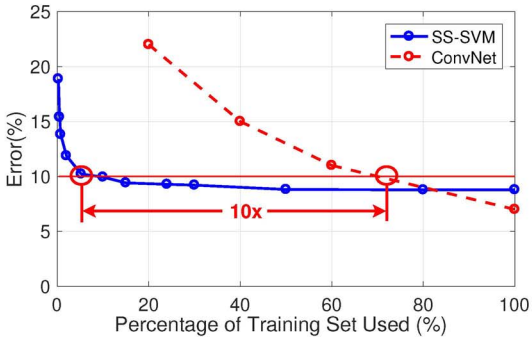

Fig. 3. Error vs. training samples used for both classifiers evaluated on the Baby Cry dataset. The learning rate of the SS-SVM classifier is much higher while the ConvNet is likely to outperform the SS-SVM with more training data.

ESC. These are classification problems with 10-50 classes and therefore require systems with more information capacity than the binary classification task of baby cry. The performance vs. cost tradeoff of the SS-SVM and several recently published deep learning classifiers is summarized in Fig. 4. The plot also illustrates the relative sizes of the parameter count for the different classifiers. While the deep learning methods for complex tasks are clearly superior in accuracy, this analysis is useful for determining the cost of using such approaches over smaller and lighter classifiers.

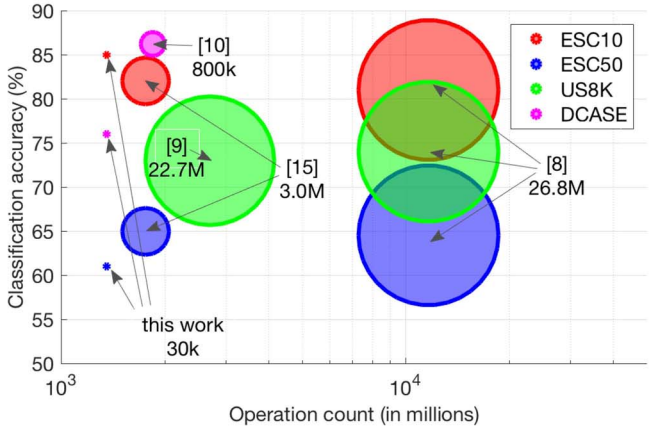

Fig. 4. Comparison of classifiers on different environmental sound datasets. Size of the circles represents the model parameter count. Numbers in brackets are the citation references and the parameter count is labeled below. Operation count is defined as a multiply-add operation and was calculated for a single inference performed on the DCASE 2016 dataset for 30 s clips sampled at 16 kHz. The data points for the US8K dataset [13] were added as an indirect comparison through the work in [8]. Comparison is drawn between systems without resorting to data augmentation.

B. Speaker Identification

For the task of speaker identification we evaluated the SS-SVM on the TIMIT dataset [16]. The SS-SVM slightly outperforms the ConvNet approach in [17] at a fraction of the model parameters and lies within 3% of the highest-performing systems that employ GMM models with other hand-crafted feature sets [18], [19]. Table II summarizes this comparison.

TABLE II
ACCURACY (IN %) OF CLASSIFIERS ON THE TIMIT DATASET FOR SPEAKER IDENTIFICATION.

Reference	Accuracy	Params ($\times 10^6$)	Method
Reynolds [18]	99.5	0.95	MFCC+GMM
Stadelman [19]	100		
Lukic [17]	97	275.97	ConvNet
This work	97.2	1.08	SS+SVM

V. HARDWARE IMPLEMENTATION

To further increase the energy efficiency of audio classifier systems, some of the feature extraction processing functionality can be implemented in the analog domain, close to the source of the audio signal. This general idea has already been investigated in prior research. The work in [2] presents a 6 nW front-end but limits the signals of interest to fixed tones under 500 Hz and to limited dynamic range. Other work has shown that analog filtering is efficient in performing frequency analysis for the purpose of audio classification. The work in [3] presents a 710 nW front-end and [20] demonstrates a 380 nW front-end, but in both cases the filter banks are based on g_m-C topologies, which suffer from limited configurability because the filter bandwidths and center frequencies depend on both the absolute and relative accuracy of capacitors and bias currents. Additionally, in [3] the features must be learned on a chip-to-chip basis by a training step, which is undesirable for mass-produced devices.

We propose an approach that overcomes the above-described limitations while being particularly suitable for the extraction of summary statistics. The envisioned system is described in Fig. 5. It follows a mixed-signal approach, where the different sub-blocks are partitioned between analog and digital implementations. Because the feature-set extraction can be approximated using passive switched capacitor circuits, this approach promises to be energy efficient while also offering a simple means for frequency tuning via the system clock and on-chip clock dividers.

The subband filter bank is composed of analog N-path filters where their respective center frequencies can be set by their switched capacitor clock rate and their bandwidths can be set by an adjustable baseband resistor value as described in [21]. Following each filter, we employ a direct-conversion mixer with a low-pass filter at baseband to extract the sub-band envelope signals depicted in the block labeled “envelope” in Fig 1. The resulting analog signal has a bandwidth of 200 Hz and can therefore be sampled by an analog-to-digital converter sampling at only 400 S/s. The remaining features can be extracted digitally at significantly lower energies than if they were extracted from the raw audio signal, which would be traditionally digitized at 16 kS/s (medium quality audio sample rate).

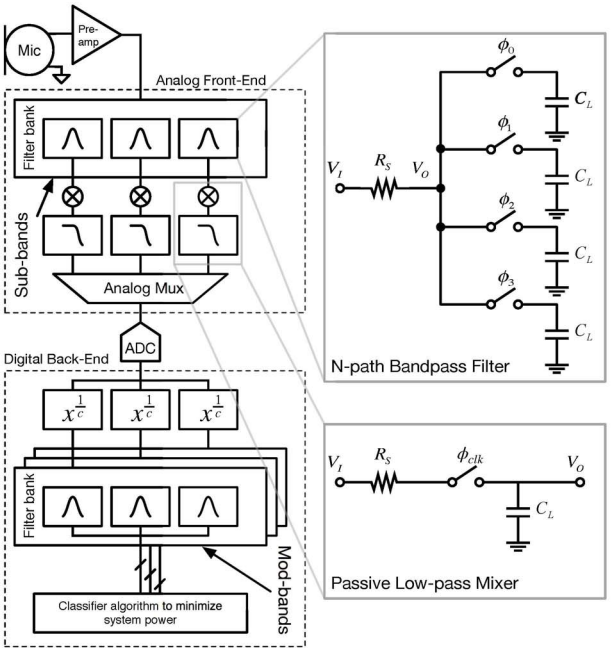

Fig. 5. Classifier system showing analog and digital signal chain partition for energy-efficient signal processing. Simplified single-ended circuit implementations of the analog processing is shown as well.

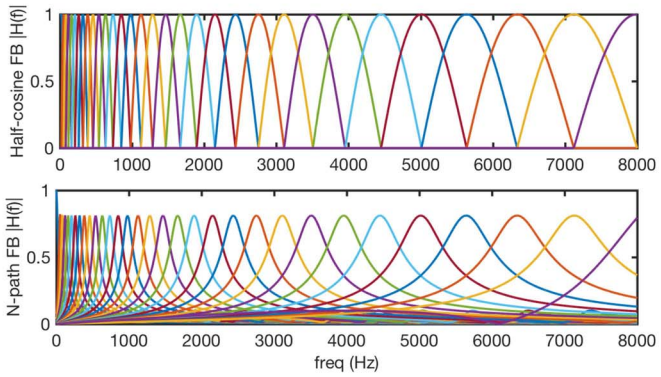

Fig. 6. Comparison between perfect reconstruction half-cosine filter bank and N-path implementation.

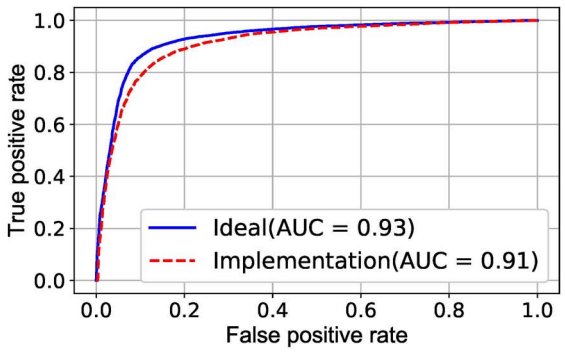

Fig. 7. ROC curves for baby cry classification using ideal vs. implemented feature extraction.

These implementation choices change the original signal processing reported in McDermott et al. [4]. The main differences are twofold. First, the bandpass filter banks, which were originally proposed as half-cosine orthogonal filter banks, are implemented as N-path filters with equivalent 3-dB bandwidths. The differences in filter transfer function magnitudes are illustrated in Fig. 6. Second, the envelope extraction step, which was originally computed as the magnitude of the analytic signal (i.e., Hilbert transform), is implemented as a passive direct demodulation combined with a low-pass filter at baseband. For simplicity, single-ended versions of these circuit implementation choices are also illustrated in Fig 5. In actual implementations, their differential counterparts would be used.

Using harmonic transfer matrix models [22] to approximate the N-path filter transfer function and a simplified time-domain model for the demodulation step, we evaluate the effectiveness of the implementation to achieve comparable classification performance. The resulting features are then passed to the same SVM classifier used in the original system described in Section II. From this simulation used on the task of baby cry detection, there was a small degradation observed in the ROC characteristics as shown in Fig 7.

The analog front-end shown in Fig 5 was designed in a CMOS 130 nm process in order to simulate the models used for the previous classification results at the transistor level. The efficiency of our proposed architecture is illustrated by the simulated power consumption summarized in Table III. The analog power is spent on level shifters and on driving their respective switches, while the digital power is separated to show how much is used for clocking versus supporting logic. We expect the actual power consumption of a full chip to increase between 20-30% to account for long-distance clock routing and other nonidealities. Our future work will evaluate the performance and power consumption through measurement of the fabricated chip.

VI. Conclusion

We have presented a quantitative analysis on a compact audio classification model with test error that converges about ten times faster than a deep learning model, while requiring one to three orders of magnitude fewer parameters for training. In some cases, the classifier accuracy is competitive with deep learning techniques and other engineered feature extraction, while consistently maintaining lower computation count. We have further demonstrated potential benefits of a CMOS mixed-signal circuit implementation to extract the same features and observed that the imposed simplifications do not significantly degrade the classification accuracy.

TABLE III
Simulated Power Consumption of Analog Front-End

Digital (nW)		Analog (nW)	**Total (nW)**
Clocking	Logic		
276	52	361	689

REFERENCES

[1] N. D. Lane and P. Georgiev, "Can Deep Learning Revolutionize Mobile Sensing?" in *International Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2015.

[2] S. Jeong, Y. Chen, T. Jang, J. M.-L. Tsai, D. Blaauw, H.-S. Kim, and D. Sylvester, "Always-On 12-nW Acoustic Sensing and Object Recognition Microsystem for Unattended Ground Sensor Nodes," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 261–274, Jan 2018.

[3] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 uW power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan 2016.

[4] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis." *Neuron*, vol. 71, no. 5, pp. 926–40, Sep 2011.

[5] D. P. W. Ellis, X. Zeng, and J. H. McDermott, "Classifying soundtracks with audio texture features," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[6] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2016.

[7] "Acoustic scene classification - dcase2017," http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results, [Online; accessed 24-October-2017].

[8] K. J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks," *IEEE International Workshop on Machine Learning for Signal Processing*, 2015.

[9] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar 2017.

[10] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017.

[11] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in *International Conference on Engineering Applications of Neural Networks*, 2017.

[12] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *ACM International Conference on Multimedia (MM)*, 2015.

[13] J. Salamon and J. P. Bello, "Unsupervised Feature Learning for Urban Sound Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[14] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[15] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," in *Conference on Neural Information Processing Systems (NIPS)*, 2016.

[16] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[17] Y. Lukic, C. Vogt, O. Durr, and T. Stadelmann, "Speaker Identification and Clustering Using Convolutional Neural Networks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[18] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, Aug 1995.

[19] T. Stadelmann and B. Freisleben, "Unfolding Speaker Clustering Potential," in *ACM International Conference on Multimedia (MM)*, 2009.

[20] M. Yang, C.-H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1μW voice activity detector using analog feature extraction and digital deep neural network," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2018.

[21] M. Darvishi, R. van der Zee, and B. Nauta, "Design of Active N-Path Filters," *IEEE J. Solid-State Circuits*, vol. 48, no. 12, pp. 2962–2976, Dec 2013.

[22] S. Hameed, M. Rachid, B. Daneshrad, and S. Pamarti, "Frequency-Domain Analysis of N-Path Filters Using Conversion Matrices," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 63, no. 1, pp. 74–78, Jan 2016.