REAL-TIME COLLABORATIVE MONITORING IN WIRELESS SENSOR NETWORKS

Visar Berisha, Homin Kwon, and Andreas Spanias

Arizona State University

ABSTRACT

In recent years, wireless sensor networks (WSN) have shown success in distributed real-time signal processing systems. In collaborative signal processing environments, each sensor is responsible for extracting pertinent information from the surrounding environment and transmitting it to other sensors and/or to the main processing station. Often times, the sensors operate under a number of constraints, such as limited processing power and low bandwidth. In this paper we propose a collaborative signal processing framework that is implemented in an acoustic monitoring scenario. A low-complexity voice activity detector and a gender classifier are implemented on the Crossbow sensor motes. A series of experiments are presented that characterize the performance of the algorithms under varying SNR conditions and in different environments.

1. INTRODUCTION

The development of independent, self-contained sensor devices, or motes, for use in wireless networks has made feasible distributed sensing systems. In these networks, each mote provides information about its surroundings to other motes and to the base station. In this scenario, the challenge is to develop distributed and collaborative methods that are optimized for the particular application and hardware platform [1]. Due to the power limitations on the hardware, all algorithms must be of acceptable complexity and provide adequate performance. The advantage of the collaborative environment lies in the fact that simple algorithms at the mote level can be combined to improve estimates and system performance. This paradigm differs from traditional frameworks in which a central sensor acquires and interprets the incoming data. Acoustic sensors are of particular interest in such scenarios because they provide a great deal of information about the environment. Human speech is perhaps one of the most revealing acoustic cues since it can convey important speaker information such as gender, age, or emotional state.

In this paper, we study low-complexity algorithms for acoustic scene recognition problems. More specifically, a voice activity detector (VAD) and a gender classification algorithm are presented for the purpose of integrating them in a low-power WSN. The algorithm runs at each sensor mote where a local decision is made. The individual decisions are then combined at the base station using a data fusion algorithm.

Efficient use of bandwidth in wireless sensor networks is of interest due to the constraints imposed by their size and power. In applications where an audio signal is to be transmitted from sensor to base station through a wireless medium, it is beneficial to discontinue transmission when speech is not present. As a pre-processor to such a system, a voice activity detector is required to make a decision as to whether or not the current frame should be transmitted. A VAD can also improve the performance of other speech classification algorithms. In addition to the VAD, we also present a simple gender classification algorithm that makes use of some of the same features that the VAD uses.

The challenge here is that the signal processing capabilities at the level of the mote are characterized by low precision and low clock rates. Hence signal processing computations are subject to a tight budget and are susceptible to numerical and transmission errors. Hence the VAD and the gender classifier are based on simple frame-by-frame energy and pitch measurements. The energy is estimated using a series of approximations in each frame; for pitch detection, a modified fixed-point version of the average magnitude difference function (AMDF) is developed and implemented at the level of the mote. A local decision is made based on these measurements and transmitted to the base station. At the base station, measurements are refined.

A commercial off the shelf (COTS) hardware platform was chosen as a test bed for our real-time experiments. The individual sensor nodes used were the MICAz^{TM} wireless motes and the MTS310CA^{TM} sensor boards from Crossbow^{TM}. The base station used was the MIB600^{TM} also from Crossbow [2]. The motes run TinyOS and are programmed using the nesC language [3]. The limitations and challenges of this system include limited processing power, drift in the clock rates, leaky buffers, and as mentioned before, finite word length effects. Keeping these limitations in mind, any algorithm involves clever numerical approximations and must be modular yet simple. The limited processing power necessitates the use of low-complexity functions that may make local decisions at the mote unreliable.

The major contributions of this study are the development of a generalized framework for a distributed sensing and classification system, the characterization of a low-complexity VAD, and the efficient implementation of a gender classification algorithm on a network of sensor motes. We note that implementation aspects in this network of sensor motes are by no means trivial and are hindered by lack of user-friendly development tools and documentation. We also note that the bit-by-bit processing and transmission at the sensor level requires knowledge of the nesC language and TinyOS system that differ considerably from typical DSP development suites.

This paper is organized as follows: Sections II provides an outline for a distributed classification system in which the motes collaborate with each other to form a final decision. In section III, a low-complexity VAD and gender classifier are described and results are provided that characterize their performance. In section IV we describe the hardware implementation of the algorithm and its complexity and section VII contains concluding remarks.

1-4244-0469-X/06/$20.00 ©2006 IEEE

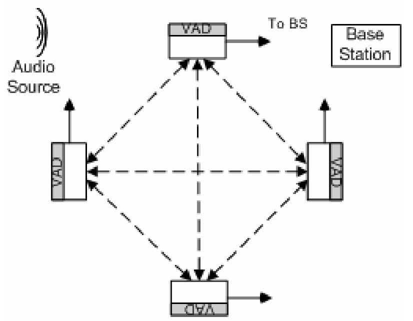

Fig. 1. The communication paths between individual sensor motes

2. COLLABORATIVE SIGNAL PROCESSING

In this paper we present a system for performing collaborative analysis, classification, and synthesis of real-time audio in a wireless sensor network. Consider the block diagram shown in Fig. 1. Each mote, represented by a block, is tasked with a particular sensing function and also sharing the information that it gathers with all other motes and the base station. In this scenario, at any point in time, any mote has all available information. In applications in which processing power is limited, it is of use to filter irrelevant data at the earliest stage in the process, thus preserving system resources. The local VAD at the sensor motes accomplishes this by only passing information pertinent to the task at hand. For example, in gender classification the pitch estimate during periods of voiced speech is often used as a feature. It makes sense in such scenarios to remove the unwanted input data (non-speech frames in the gender classification example) as early as possible in the classification process.

In typical classification problems based on audio inputs, a VAD is used to determine periods of speech activity. Within these periods, features are extracted and they act as inputs to pre-trained classifiers that make final decisions. A slightly adapted platform for use in wireless sensor networks is shown in Fig. 2. This figure shows the processing within each of the motes shown in Fig. 1. A local decision is made at mote *i* based on computed and received features and then transmitted to the base station. At the base station, a data fusion algorithm combines the data from each mote in order to form a final decision. Below we give a generalized mathematical framework for performing distributed classification.

Let $\boldsymbol{x}_i$ denote a frame of the acquired signal at mote *i* with SNR γ_i and VAD decision VAD_i. The mote makes use of the acquired signal, the SNR, and the VAD decision for that frame to extract a set of features, denoted by $\boldsymbol{f}_i$ using an extraction function $\Lambda^x{}_i$. This is shown in (1).

$$\mathbf{f}_i = \Lambda_i^x(\mathbf{x}_i, \gamma_i, VAD_i) \tag{1}$$

The resulting feature set is transmitted to the other motes and at the same time all other feature sets from other motes, depicted by $\boldsymbol{f}_j$, are received. Using (2) the features from the other motes and the SNR estimate from each mote are combined to form a final feature set, $\boldsymbol{f}^f{}_i$.

$$\mathbf{f}_i^f = \Lambda_i^f(\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n, \gamma_1, \gamma_2, ..., \gamma_n) \tag{2}$$

where $\Lambda^f{}_i$ is the input-output mapping between the local feature vectors and the final feature vector using the individual SNRs to appropriately weight the local features. This final feature vector is then used to make the local decision as shown in (3) using a classifier C_i.

$$d_i = C_i(\mathbf{f}_i^f) \tag{3}$$

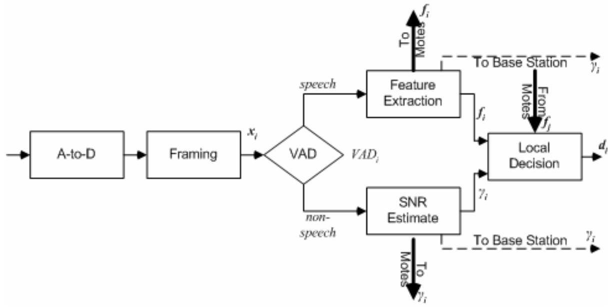

Fig. 2. Block diagram of the local classification algorithm located at each mote

The local decisions are then sent to the base station. At the base station, the decisions are combined using a fusion rule denoted by Λ_d as shown in (4). Fusion rules typically make use of the individual local decisions and the SNR estimates.

$$d^f = \Lambda_d(d_1, d_2, ..., d_n, \gamma_1, \gamma_2, ..., \gamma_n) \tag{4}$$

In the classification scenario, each mote is tasked with the extraction of a particular feature. It then transmits the computed feature to the other nodes so that each has the ability to make the most reliable local decision possible. In such a scenario, it is possible that due to channel errors the features may never reach the destination. Flexibility must be built into the network such that local decisions can be made with incomplete feature sets. This requires the design of new classifiers that can make acceptable decisions despite missing feature inputs. When a decision has been made for a particular frame of audio, each mote then transmits their decision to the base station where a global one is formed. Due to varying SNR levels around the motes and due to incomplete feature sets, the reliability of any local decision can vary at any point in time. Because of this, control information can also be sent along with the decision so that the base station can appropriately weight the incoming decisions when forming the global one.

This paradigm can also be used in a speech analysis/synthesis scenario. Speech compression is typically achieved by parametrizing a frame by an excitation signal and a system [9]. Although a number of low-complexity algorithms exist for performing this analysis (LPC-10), these are still of considerable complexity and are unable to execute in real-time on the selected hardware platform. One possible solution is to divide the complexity among a number of motes. For example, a group of motes can be tasked with performing the spectral envelope estimation while another group is responsible for obtaining an estimate of the excitation. By transmitting all parameters of a particular frame to all motes, any one mote at any point in time can have the ability of reconstructing the audio signal.

3. VOICE ACTIVITY DETECTION AND GENDER CLASSIFICATION

Within the collaborative model discussed in the previous section, we present a voice activity detector and a gender classifier for use in a distributed wireless sensor network. The VAD attempts to remove non-essential information from the incoming audio signal

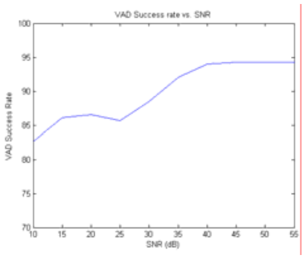

Fig. 3. VAD success rate at varying SNR conditions

so that the feature extraction for the gender classifier operates on pertinent parts of the incoming signal.

3.1. Voice Activity Detection

In this paper, we propose a low-complexity algorithm for voice activity detection for the purpose of implementing it in a WSN. Two important features that distinguish periods of voice activity and inactivity are the signal energy and the pitch. During voiced segments, the energy of the waveform is typically higher than during non-speech segments. In addition, the pitch estimate of segments of speech activity tends to stay constant assuming that the speaker does not rapidly vary their pitch. During non-speech segments, the pitch estimate typically varies greatly because of the lack of periodicity in the waveform. Because sensor motes have limited processing capability and precision, it is essential that the algorithm is simple yet robust with regard to numerical precision. Using low-complexity estimates of these two features, we can obtain a preliminary VAD decision at the sensor mote.

At the mote, the analog signal is first digitized by the A-to-D converter on the sensor. The resulting signal is then divided in frames of length *M* and an energy value for frame *m*, denoted by E_m is computed using (5).

$$E_m = \frac{1}{M}\sum_{k=0}^{M-1}|x_m(k)| \tag{5}$$

In addition to the energy, the pitch of the particular segment is also calculated using the average magnitude difference function (AMDF) shown in (6). Although the AMDF generally provides a modest estimate of pitch, it was selected over other methods because of its low computational complexity and its robust performance in the low precision environment of a sensor mote. We note again that a mote is an 8 bit processor operating at a low clock speed. In addition, the AMDF is sufficient for the purposes of the VAD algorithm because of the distinct patterns in the pitch contours between periods of voice activity and inactivity.

$$D_m(k) = \sum_{p=0}^{M-1}|x_m(p) - x_m(p+k)| \tag{6}$$

$$k = 0, 1, 2, 3...$$

Rather than considering the individual pitch estimates of the frames, a cumulative sum of the difference in pitch between successive frames is considered. This value increases during periods of non-speech activity and stays fairly constant during speech segments. The actual feature used as an input to the classifier is the slope of the cumulative sum for the last eight frames.

A multi-layer perceptron classifier was trained using the TIDIGITS database [6] with actual VAD values obtained using the VAD in [4]. In order to test the validity of the presented algorithm, unseen test data (data outside the training set) was processed by the algorithm with different AWGN levels. Fig. 3 shows the results for SNR levels ranging from 10 dB to 35 dB. The success rate is greater than 80% for all SNR levels with a maximum of 94% for clean speech.

3.2. Gender Classification

Within the framework described in the previous section, we present a simple gender classifier strictly based on the calculated pitch during voiced frames, motivated by the classifier used in [5]. This simple technique has shown success as a pre-processor for an automatic speech recognition system that contains both a male and a female model. The mote utilizes the pitch estimate obtained during the VAD decision in order to perform the classification.

A pre-trained classification tree was chosen to predict the gender of the speaker based on the features. Classification trees determine a set of logical if-then-else statements for prediction. Their advantage lies in their simplicity. The interpretation of the obtained results is simple since the output of the tree provides conditions on the independent variables for a particular classification output. Typical classifiers may yield complicated equations that must be programmed and computed in order to determine the value of the output based on the feature set. A classification tree, however, provides logical statements that can be programmed with a few if-then-else statements in nesC. Male and female speakers from the TIDIGITS [6] database were used to train the tree using the CART algorithm [7]. Informal experiments were conducted on data from the TIDIGITS database at different AWGN levels and the results ranged from 96% - 98%. These results are obtained from single speaker data consisting of numerical digits and higher error rates are to be expected in a more general implementation.

4. HARDWARE IMPLEMENTATION

4.1. Implementation challenges and preliminary results

There are a number of problems associated with implementing algorithms, such as those described above, in ad hoc sensor networks. The limited processing power, communications bandwidth, and storage capacity become problematic when processing speech and audio in real time. In this section, we present some of the challenges with implementing the above mentioned algorithms in hardware in addition to preliminary results. The selected hardware platform was the Crossbow MicaZ sensor mote and the MIB6000 base station. The implementation was done in fixed point in the nesC language. As is seen from the algorithmic descriptions in the previous sections, the main component of both the VAD and the gender classifier is the pitch detection algorithm, namely the AMDF. Because of its importance to both algorithms, the AMDF was individually tested on simpler, synthetic signals as well as on real time speech to validate its functionality.

Simple sinusoidal tones of known frequency were used as test signals and the sensing was performed by a single mote. The mote then analyzed the signal using the AMDF and the frequency was determined. This was done for a number of signals ranging from 200 Hz to 500 Hz. The experiment was conducted for varying distances from the audio source to the mote in a laboratory environment. The estimated lags were always within 3 of the ideal lags in this scenario. In a more interesting experiment, actual data

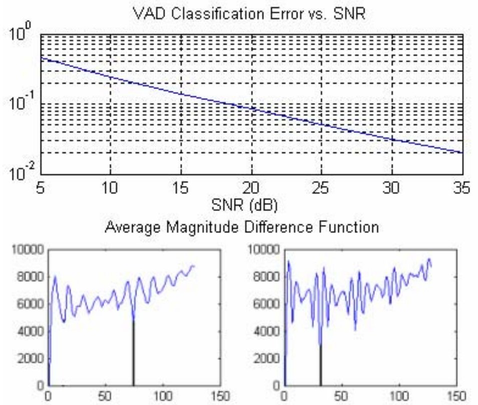

Fig. 4. (Top) VAD error at varying SNR conditions (Bottom) Two examples of the AMDF function plotted against the lag

was acquired at the motes and then transmitted back to the base station. The AMDF algorithm residing on the mote was simulated at the base station and the minimum lag of the AMDF was obtained. The simulated AMDF at the base station gives identical values to the one residing on the motes, therefore the computed AMDF is what the mote itself sees. In the top part of Fig. 4, we show an error vs. SNR curve for varying SNR levels. As the plot shows, the errors range from approximately 5% (for high SNR) to 30% for very low SNR. Below this plot, we show two sample AMDFs with the resulting minimum lag indicated by a dark line.

Even for the simple sinusoidal case, there is a slight variation between the computed AMDF lags and the ideal ones. There are a number of factors that contributed to these errors. The most prominent was the low quality signal produced by the microphone on the mote. High frequency noise distorted the waveforms, and, for the lower frequencies, the signals were attenuated. This accounts for the minor variations in the minimum lag as well as the fact that the algorithm could not detect the signal at lower frequencies (see the 200 Hz wave). In addition, the background noise in the testing environment, the fixed point arithmetic, and the simplicity of the algorithm necessitated by the mote limitations also contributed to some of the errors.

4.2. Complexity Analysis

Most algorithms developed for wireless sensor networks must take into account efficiency because of the limited resources available on the mote. In fact, a number of algorithms are available that attempt to reduce energy consumption in WSNs [8]. In an effort to maintain acceptable performance and speed, we have analyzed the complexity of the energy and AMDF calculations on the mote. Fig. 5 shows a chart of the different operations performed in the program and the corresponding clock cycles amount relative to the total. It is apparent that the operation requiring the most clock cycles is the AMDF. This is because of the nested loops (*$O(n^2)$* complexity) required to implement it. For future implementations of the algorithm, we are seeking less costly ways of computing pitch periods of acceptable quality. Taking into account all computations, for a single frame of 64 samples, a total of 11312 clock cycles are required, which translates to approximately 2.8 *msec* in actual time. In this analysis, additions, multiplications, and lower level subroutines that are called by TinyOS are not taken into account. Although the additions and multiplications may not change the total complexity of the algorithm significantly, the lower level subroutines can potentially have a large impact.

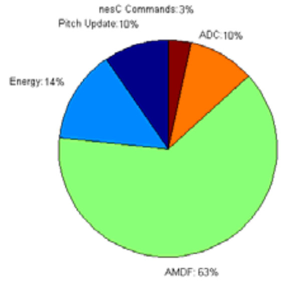

Fig. 5. Complexity analysis of the real-time implementation of the VAD algorithm on the Crossbow platform

5. CONCLUSION

A distributed acoustic sensing system for use in a wireless sensor network was presented in this paper. Within this framework, a voice activity detector and gender classifier were proposed and implemented on a limited power hardware platform. The proposed algorithm makes use of pitch and energy estimates at each frame in order to make a local decision, which is then shared with all other motes. Although the algorithm in this preliminary form shows some promise at varying SNR levels, further gain can be achieved. Due to the limitations imposed by the hardware, decreasing the complexity of the proposed system would be of benefit. Future work will focus on the reduction of complexity of the algorithm and the inclusion of other classification methods. Emotional state, age, and simple speech recognizers are all examples of potentially useful algorithms in monitoring scenarios. In addition, we are also currently studying an iterative data fusion algorithm to be implemented at the base station.

6. REFERENCES

[1]. M. Duarte and Y. H.Hu, "Vehicle Classification in Distributed Sensor Networks," *Journal of Parallel and Distributed Computing*, Vol. 64 No. 7, pp. 826-838, 2004.

[2]. J. L. Hill, D. E. Culler, "Mica: A Wireless Platform for Deeply Embedded Networks," *IEEE Micro*, Vol. 22, pp.12-24, Iss. 6, Nov/Dec 2002.

[3]. D. Gay, P. Levis, R. V. Behren, M. Welsh, E. Brewer, and D. Culler "The nesC Language: A Holistic Approach to Networked Embedded Systems", *In Proc. of PLDI 2003*, June 2003.

[4]. 3GPP TS 06.51 "AMR Wideband speech codec; Voice Activity Detector (VAD)," 2004.

[5]. W.H. Abdulla and N. K. Kasabov. "Improving Speech Recognition Performance Through Gender Separation", *Artificial Neural Networks and Expert Systems International Conference*, pp 218-222, Dunedin, New Zealand.2001.

[6]. R. G. Leonard, "A Database for Speaker-Independent Digit Recognition", *In Proc. of ICASSP 84*, Vol. 3, p. 42.11, 1984.

[7]. L.Breiman, J. H. Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Trees," Wadsworth, Pacific Grove, CA (1984)

[8]. R. Shah and J. Rabaey, "Energy Aware Routing for Low Energy Ad Hoc Sensor Networks", In Proc. of IEEE WCNC, Mar 2002.

[9]. A. Spanias, "Speech Coding: A tutorial review," 1994