High-resolution speech signal reconstruction in Wireless Sensor Networks

Andria Pazarloglou, Radu Stoleru, Ricardo Gutierrez-Osuna
Department of Computer Science, Texas A&M University
{andria, stoleru, rgutier}@cs.tamu.edu

***Abstract*—Data streaming is an emerging class of applications for sensor networks that has very high bandwidth and processing power requirements. In this paper, a new approach for speech data streaming is proposed, which is based on a distributed scheme. This scheme focuses on balancing the energy consumption among nodes in a sensor network by allowing low-resolution streams from multiple nodes to be fused at a central processing node in order to produce an enhanced resolution speech signal. Simulations and experimental results with real microphone signals are presented.**

I. Introduction

With the recent growth of Wireless Sensor Networks (WSNs), many advanced application areas have received significant attention. Apart from the traditional low data rate applications implemented in sensor networks, a need of supporting higher data rate applications, such as audio and video streaming, has appeared lately. Specifically, there is a lot of interest in real-time streaming for military surveillance purposes and for emergency situations [1]. The lack of bandwidth and limited sampling rate capabilities of sensors render the implementation of these applications challenging.

Among the requirements of streaming applications is the high volume of data that needs to be sent to the sink. The basic drawback of obtaining data from one particular node every time is that it could lead to an over-utilization of the radio of specific nodes. The radio is one of the most energy-inefficient component on a sensor and power supply is limited and battery recharging is almost inexistent. This eventually leads to depletion of the power supply of those nodes sooner than the other ones and reduces significantly the network lifetime.

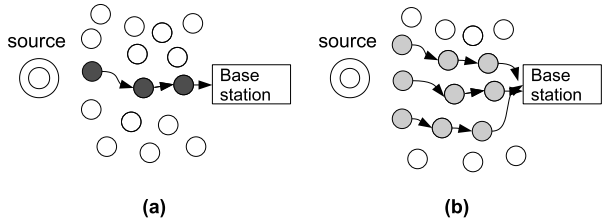

Fig. 1. Streaming schemes with (a) one and (b) multiple nodes listening and transmitting data

Another factor to take into consideration is that, in WSNs, the data from neighboring nodes is usually highly correlated. Consequently, the transmission of the gathered data of those nodes leads to redundant information. Our aim is to design a distributed scheme for data acquisition that balances energy consumption while exploiting spatial correlations between neighboring nodes. In the process, we also reduce the bandwidth and processing requirements for sensor nodes.

The contributions of the paper are the presentation of a novel approach for high resolution speech signal reconstruction and the demonstration of the proposed approach, through performance evaluation in simulations and a proof of concept hardware implementation.

The paper is organized as follows. Section 2 presents the related work. Section 3 details the design of the proposed scheme for high-resolution signal reconstruction. Section 4 presents and discusses performance evaluation results. Conclusions and future work plans are given in section 5.

II. Related Work

Current approaches for streaming voice data in WSNs, assign the responsibility of transmitting the data to a node that is closer to the source of the event, as in Figure 1(a). Mangharam et al. [1] implements a networking platform in order to support real-time voice streaming. Another attempt for streaming service support in order to support military surveillance applications is described in [2]. Common to both studies is the need to obtain understandable speech from WSNs with energy consumption constraints. Thus, microphone sample rates are lower than the normal 8kHz and also, Adaptive Differential Pulse Code Modulation (ADPCM) is used to encode the data and thus reduce the transmission data rate.

A high data-rate capturing system for sensor networks has been presented by Greenstein et al [3]. In this approach user-accessible devices (each capable of performing sophisticated signal processing) can tune the signal processing parameters of the sensors in order to reduce the amount of data to be transmitted.

III. Speech Signal Reconstruction from Asynchronous Data Sources

In our proposed scheme, the nodes that sense an event sample with low data rates and send the data through different routing paths to the sink, as shown in Figure 1(b). The sink receives the undersampled data from different nodes and reconstructs a high-resolution version of the original signal (Figure 2).

978-1-4244-2309-5/09/$25.00 ©2009 IEEE

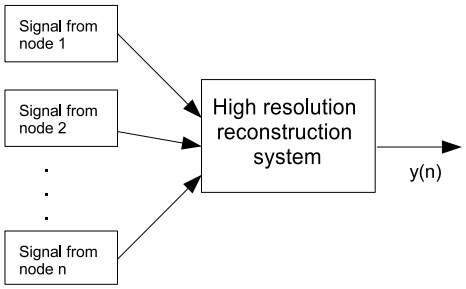

Fig. 2. Reconstruction scheme

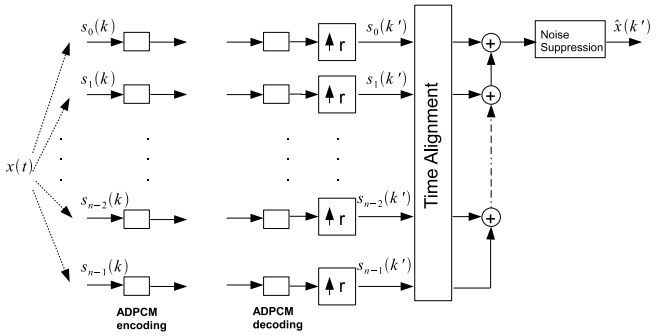

Fig. 3. Distributed Data Streaming and Signal Reconstruction

The major advantages of this scheme are that it distributes energy consumption in more than one node, thereby increasing the efficiency and the survivability of the WSN, and also shifts processing requirements from the nodes to the sink. At the same time, this scheme does not impose any demands for tight time synchronization, and instead exploits randomness in sampling times for each node to reconstruct a higher-resolution signal. On the contrary, if the received signals from multiple nodes were to be synchronized, reconstruction of a better quality signal would be unfeasible. Fortunately, this scenario is highly unlikely, since timing synchronization is really difficult to achieve.

A. Design of High-resolution Signal Reconstruction

In our approach, an optimal reconstruction [4] is not required, but rather an acceptable reconstruction. An acceptable sound signal is a signal that has some predefined qualitative characteristics, such as delivering a sound signal that is accurate enough to be understood by a listener. Given that the frequency range of a speech system is from 300 Hz to 3500 Hz, where the majority of signal energy being between 500 Hz to 2500 Hz, sampling rates used in existing applications are 4 kHz [1] and 2 kHz [2]. However, the sampling rate in our proposed system can be much lower depending on the number of the nodes that listen to the event at the same time.

Our reconstruction approach is based on the fact that streams from multiple sources with lower quality representations of the original signal provide different pieces of information about the original signal, provided that their sampling times are randomly shifted.

Figure 3 illustrates the sampling and reconstruction. The original signal is denoted by $x(t)$ and the discrete-time signals $s_i(k), i = 1, \ldots, n$ are noisy versions of $x(t)$ produced in each node after sampling (where k denotes the time indices at sensor node sampling rate f_s and k' denotes the time indices for the higher resolution signals at sampling rate $f_{s'} = f_s \cdot r$). Since the more efficient compression techniques (with compression ratio 12:1 or better) such as Mixed-Excitation Linear Predictive (MELP) and Linear Predictive Coding (LPC) are too complex for sensor nodes, we choose to compress $s_i(k)$ with 4-bit ADPCM which provides a compression ratio of 4:1.

On the reconstruction part, the received streams are first ADPCM-decoded and then aligned in time before being fused. Since time delays between signals from different nodes might not be integer multiples of their sampling rate, the signals are upsampled to a common sampling rate through interpolation.

Time delay estimation is performed from the resulting signals $s_n(k')$ in order to align them properly on the time domain. The output signal is then formed by superimposing the aligned signals and a denoising filter is applied.

B. Time-Delay Estimation

Here, the term time-delay estimation refers to the relative Time Difference Of Arrival (TDOA) between signals received at spatially separated sensors. Time-delay estimation is a common research topic in signal processing, where it is used as a first stage in applications such as speech enhancement and recognition, and source localization and tracking.

The most common method for the time delay estimation, between two waveforms x_1 and x_2, is the Generalized Cross-Correlation (GCC) function [5]:

$$\hat{d}_{GCC}(m) = \sum_{k=0}^{N-1} W[k] G_{x_1 x_2}[k] e^{j2\pi mk/N}$$

where

$$G_{x_1 x_2}[k] = E[X_1[k] X_2^*[k]]$$

is the cross-spectrum of x_1 and x_2, $X_i[k]$ is the discrete Fourier transform (DFT) of $x_i(n)$, N is the length of the DFT and $W[k]$ denotes the weighting function or prefilter.

In the case of a constant weighting function $W[k]$, the GCC function becomes a frequency domain implementation of the CC function (non-weighted cross correlation). Different solutions have been proposed to choose a suitable weighting function [5].

For the case of TDE in multichannel cases, one approach is to choose one channel as a reference and the TDOA for the rest of the channels can be estimated using a TDE method. More robust estimations can be obtained by exploiting the fact that more channels can be used as reference in a way to overcome errors caused by reverberation and noise.

C. Reconstruction from Low Resolution Signals

We considered two approaches for the superposition of the aligned signals. In the first approach, we interpolated the low

resolution signals and the average between them was taken

$$x(k') = \frac{1}{N}\sum_{n} s_n(k' - d_n)$$

where d_n is the delay for the signal s_n.

In the second approach, we increased the signal sampling rate with upsampling (e.g., fill the spaces between samples with zeros) as

$$s_n(k') = \begin{cases} s_n(k), \text{ if } k' = k \cdot r, k = 0, \pm 1, \pm 2, .. \\ 0, \text{ otherwise} \end{cases}$$

and the signals $s_n(k')$ are summed up as:

$$x(k') = \sum_{n} s_n(k' - d_n)$$

and zero samples were finally replaced through linear interpolation.

D. Noise Suppression

The noise suppression stage is the last stage in the signal reconstruction process, and is used to smooth the signal and remove possible background noise. The use of multiple microphones in observations of the same event, can give a better idea of the source signal and it has been an active area of research for the past decades. Noise suppression methods used in microphone arrays ([6], [7]) for speech recognition and enhancement are based on Wiener post-filtering techniques.

These methods are applied in the final reconstruction stage. However, in high noise environments, signals from multiple sensors can be enhanced so that subsequent steps (i.e., sampling rate increase, TDE) will perform better. A wide range of noise suppression techniques have been proposed. Among them the most popular are spectral methods, such as spectral substraction and Kalman filtering.

IV. Performance Evaluation

In this section we present performance evaluation results for the reconstruction of an original speech signal from multiple low resolution signals. This evaluation includes both simulation and actual recordings. We perform simulation of the reconstruction scheme using speech signals already acquired and present in simulator libraries, as well as using speech signals obtained in real laboratory experiments.

To evaluate the quality of a reconstructed signal, we measure the Normalized Root Mean Square Error (NRMSE) [8] and the PSNR (Peak Signal-to-Noise Ratio). The NRMSE is given by:

$$NRMSE = \sqrt{\frac{\sum_n (x_n - \hat{x}_n)^2}{E[(x_n - \mu(x_n))^2]}}$$

where x_n is the original signal, $\hat{x}_n$ is the reconstructed signal and μ denotes the mean value of the signal. The PSNR is given by:

$$PSNR = 10\log_{10}\frac{N x_{peak}^2}{\|x_n - \hat{x}_n\|^2}$$

, where x_n is the original signal, $\hat{x}_n$ is the reconstructed signal, x_{peak} denotes the maximum value of x and $\|x_n - \hat{x}_n\|^2$ is the energy of the difference between the original and reconstructed signal.

A. Simulations

In order to establish proof-of-concept, we first performed simulations. A speech signal sampled at 22 kHz sample rate and 16-bits per sample was used for the experiment[1]. We produced a number of undersampled versions of the original signal (as would have been acquired by a set of sensor nodes), with random delays. Finally, we added Additive White Gaussian Noise (AWGN) at different SNR levels.

1) Accuracy of Time Delay Estimation: Time-delay estimation was implemented using the non-weighted cross correlation. Low Signal to Noise Ratio (SNR) and the sampling rate increase method can change the performance of TDE. Figure 4 illustrates how the level of noise and different sample rates affect the result of the delay estimation. A signal having 1000 samples delay from the original 22 kHz signal was created. Additive White Gaussian Noise (AWGN) was added and both signals were downsampled with a rate of 5 and 10. In the next stage their sampling rate was increased to reach 22 kHz using interpolation. Time delay estimation for interpolated signals was accurate for SNR over 15 dB.

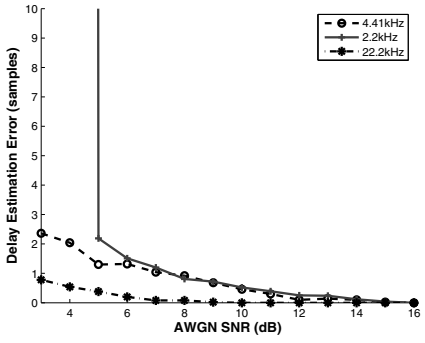

Fig. 4. Delay estimation error under different sampling rates and noise levels after interpolation (100 iterations).

Our experiments demonstrate that the non-weighted cross correlation using interpolated signals gives accurate time-delay estimates for high noise levels and a precision of 1 sample on average for SNR over 8 dB.

2) Quality of the Signal Reconstruction: Figure 5 shows a schematic of the process that was followed in the simulations. The time delay between each signal version that is created and the reference, S_k, is expressed with z^{-D_i}, where $i = 2, \ldots, n$. Noise suppression was performed on the undersampled signal representations using a bidirectional nonstationary Kalman filter.

Figures 6 show the reconstruction quality in terms of the NRMSE and the PSNR, under different noise levels. The two

[1]speech dft.wav with utterances from a male adult, as provided by the MATLAB library

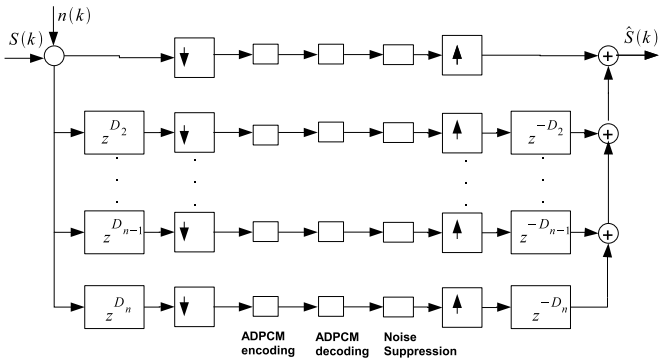

Fig. 5. Schematic for the simulations

approaches are applied for the reconstruction of a 7.35 k speech signal from five noisy undersampled versions and co pared against the speech signal produced after interpolati of one undersampled version. For this simulation, ADPC encoding was not applied.

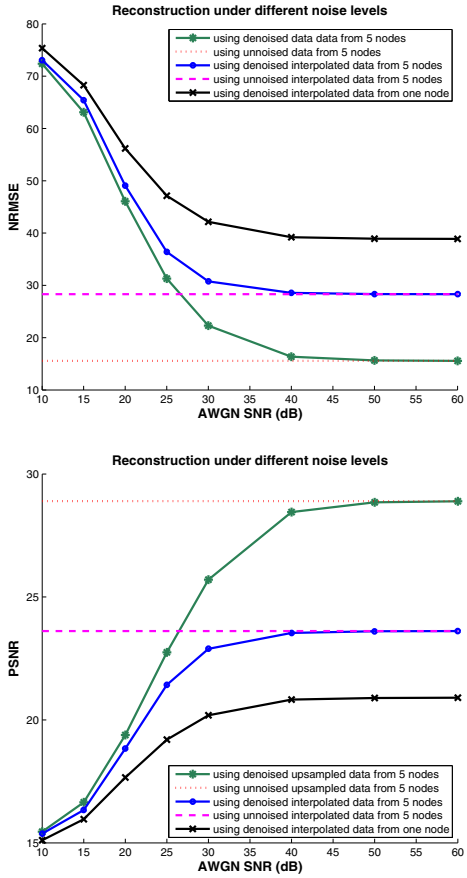

Fig. 6. Reconstruction quality (NRMSE, PSNR) using data from 1 and 5 nodes under different noise levels. One reconstruction approach is to take the average of the interpolated signals, while the other is to take the sum of the upsampled signals and fill the zero samples using interpolation.

Intuitively, the lower quality reconstruction using interpolated signals, can be justified from the fact that the actual samples contribute equally with the interpolated ones.

Figure 7 plots these signals (the first 1/6 portion of speech_dft.wav is plotted) for an SNR of 35 dB, as well as

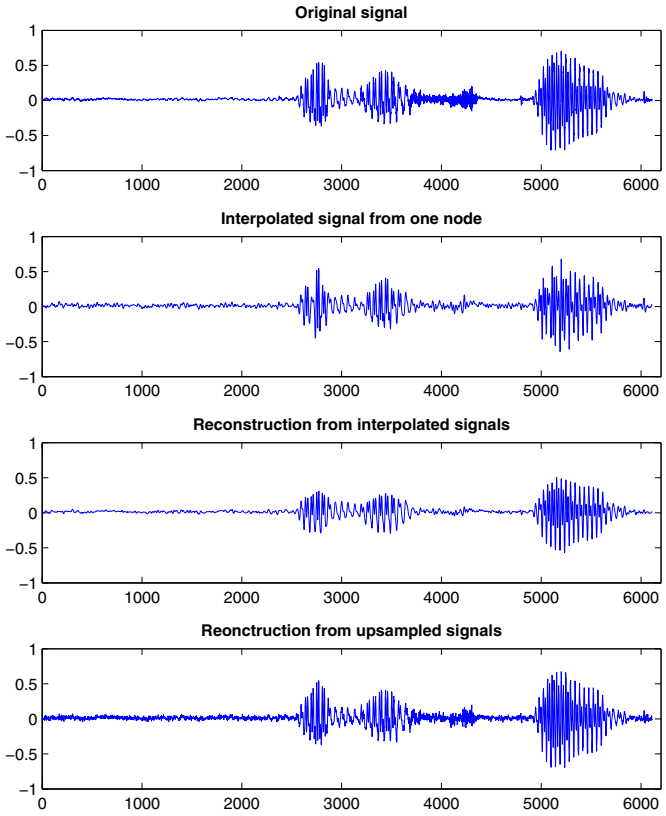

Fig. 7. Signal plots (35 dB SNR).

Figure 8 shows the sampling rate in Hz for each node, in accordance with the reconstructed signal quality. AWGN was added on the undersampled data to give an SNR of 35 dB.

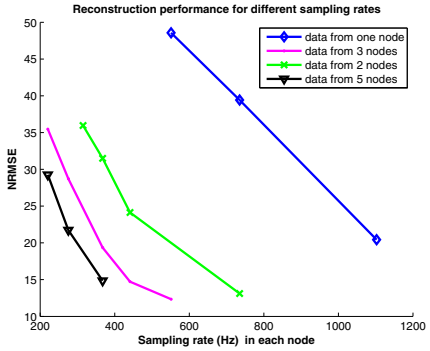

Fig. 8. Reconstruction quality using noisy data (SNR 35 dB) from 5, 3, 2, or 1 node in order to reconstruct a speech signal with 2205 Hz sample rate.

B. Experimental Results

Experiments with real recordings were performed in a computer lab. Speech signal acquisition was made from six

different positions placed in a straight line and having 1 cm distance between them. One computer speaker was placed at a distance of 50 cm of the line center. The recorded signal had 22050 Hz sample rate with 16-bit samples and one channel.

Timing delays were estimated using the six original signals, which were given as input to a non-weighted cross-correlation function; one of the signals was selected as a reference.

The six aligned microphone signals were downsampled by a factor of 20 to create 1102.5 Hz signals. The low rate signals were upsampled by a factor of 10, superimposed and interpolated to fill the zero samples. As a result, the output signal had a sampling rate of 11025 Hz.

NRMSE and PSNR were measured from an original microphone signal downsampled by a factor of 2, and were:

- 83.23 and 33.10 respectively for the reconstructed signal
- 135.25 and 28.92 for a 1.1025 Hz signal that had its rate increased using interpolation by a factor of 10
- 90.60 and 32.37 for a 5.5 kHz signal that had its rate increased using interpolation by a factor of 2

The above results show that it is feasible to obtain a clear signal after combining multiple lower-rate signals with non understandable content (such as 1.1 kHz speech signals).

V. Conclusion

We have outlined a new approach for speech data streaming based on a distributed scheme, that targets a balanced energy consumption among nodes in a sensor network. Among the advantages of this scheme is that it does not impose further processing or synchronization requirements for the sensor nodes. Moreover, the method takes advantage of the randomness of the sampling of each node. Another interesting aspect would be to implement a feedback algorithm that could change the recording phase of one or more nodes so as to have a more uniform distribution of the samples within a sampling period.

This approach opens a whole new field for research. A methodology must be developed to define which nodes should start recording based on their locations, taking into account the topology of the WSN. A more dense network could lead to more distributed sampling, leading to overall better network survivability.

In the experiments, the reconstruction method does not take into account the reliability of the signal itself. Samples that are raw data should be considered as more important in the reconstruction than samples that result of an interpolation function. This could lead to better quality in the reconstructed signal.

In addition, further research is required to find which TDE method is more appropriate to use for low rate speech signals and for specific environment conditions (i.e. reverberation, noise), since the non-weighted cross-correlation method is not robust when it accepts reverberant signals as input.

We discussed and evaluated two reconstruction approaches, one from upsampled signals and one from interpolated signals. In a real implementation, a hybrid solution that combines the best of both of them should be considered and evaluated, such as a solution that uses interpolated signals with weights.

Finally, the proposed approach has interesting implications for security-related applications. A multipath scheme where very low rate signals are circulated towards a sink could make eavesdropping considerably difficult, provided that low-resolution speech signals from each microphone are not understandable by a listener.

Our immediate goals are to evaluate the efficacy of our approach by developing a real sensor network implementation. In addition, we will investigate multipath routing in mobile environments and additional filtering schemes to enhance signal reconstructions.

References

[1] R. Mangharam, A. Rowe, R. Rajkumar, and R. Suzuki, "Voice over sensor networks," in *RTSS '06: Proceedings of the 27th IEEE International Real-Time Systems Symposium*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 291–302.

[2] J. Zhang, G. Zhou, S. H. Son, and J. A. Stankovic, "Ears on the ground: An acoustic streaming service in wireless sensor networks," in *IPSN '06*, 2006.

[3] B. Greenstein, C. Mar, A. Pesterev, S. Farshchi, E. Kohler, J. Judy, and D. Estrin, "Capturing high-frequency phenomena using a bandwidth-limited sensor network," in *SenSys '06: Proceedings of the 4th international conference on Embedded networked sensor systems*. New York, NY, USA: ACM, 2006, pp. 279–292.

[4] O. Jahromi and P. Aarabi, "Theory and design of multirate sensor arrays," *Signal Processing, IEEE Transactions on*, vol. 53, no. 5, pp. 1739–1753, May 2005.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[6] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, vol. 5, pp. 2578–2581, Nov. 1988.

[7] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[8] E.-B. Fgee, W. Phillips, and W. Robertson, "Comparing audio compression using wavelets with other audio compression schemes," *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, vol. 2, pp. 698–701 vol.2, 1999.