Voice Quality Improvement with Error Concealment in Audio Sensor Networks

Okan Turkes and Sebnem Baydere

Department of Computer Engineering
Yeditepe University, Istanbul 34755, TR
{oturkes,sbaydere}@cse.yeditepe.edu.tr
http://cse.yeditepe.edu.tr

Abstract. Multi-dimensional properties of audio data and resource-poor nodes make voice processing and transmission a challenging task for Wireless Sensor Networks (WSN). This study analyzes voice quality distortions caused by packet losses occurring over a multi-hop WSN testbed: A comprehensive analysis of transmitted voice quality is given in a real setup. In the experiments, recorded signals are partitioned into data segments and delivered efficiently at the source. Throughout the network, two reconstruction scenarios are considered for the lost segments: In the first one, a raw projection is applied on voice with no error concealment (V-NC) whereas the latter encodes a simple error concealment (V-EC). It is shown that with an affordable reconstruction, a comprehensible voice can be gathered even when packet error rate is as high as 30%.

Keywords: audio sensor networks, voice quality assessment, wireless multimedia sensor networks, voice coding, error concealment.

1 Introduction

Over the recent years, audio utilization in resource-constrained wireless networks has been a progressive subject. However, nodes composing these networks have inherently confined capabilities to handle streams generated. Hence, an affordable consistency has to be provided between stream delivery and limited resources. Another significant challenge is to receive an intelligible content at the network end-point. Accordingly, several applications related with audio sensor networks are developed [1,2,3,5] and several performance criteria are analyzed [13,12]. However, audio is a multi-dimensional function that no measure itself can accurately evaluate all of its aspects. So is the voice, which is a specific audio data type targeted in this study. Since related applications need a voice quality assessment (VQA), a great deal of data properties need to be analyzed as promptly as practicable. Despite the diversity of VQA methods [4,8,9], there is no standard measure which can evaluate several properties in company. Besides, many studies are not validated by real testbed experiments in spite of notable theoretical solutions. We elaborate on an objective VQA metric adapted from transmission rating factor defined in E-Model of ITU-T [6]. The proposed metric is modified according to the properties of the real environment targeted.

Y. Koucheryavy et al. (Eds.): WWIC 2012, LNCS 7277, pp. 307–314, 2012.

This study mainly focuses on data quality distortions caused by packet losses during voice transmission over a real wireless sensor network (WSN). Over 9,000 generated streams are transmitted over a multi-hop line. VQA of each delivery is analyzed with an experimental setup. During the transmission of segmented network packets, nodes accommodated with a simple buffering mechanism try to increase data integrity. As a follow-up work of [11], two scenarios are considered for lost segment reconstruction: First one set silence onto the lost samples in projected voices which maintains a raw form with no error concealment (V-NC). In the second one, lost segments are encoded with an averaging method (V-EC) based on a reconstruction between its neighbor packets. The results are evaluated for both scenarios in terms of several data and network parameters.

The rest of the paper is organized as follows. Voice coding and transmission model is presented in Section 2. VQA is issued in Section 3. Section 4 renders the transmission and evaluation environment. Section 5 discusses the system performance. Conclusion is given in Section 6 with comments on forward plans.

2 Voice Coding and Transmission

Transmission model consists of two types of nodes; Type 1 A_i, i=1, 2 ..., n_2, source node equipped with acoustic sensor on it and Type 2 S_{ij}, i=1, 2 ... n_1, j=1, 2 ..., n_2, node that is simple routing sensor. In this model, different network properties are analyzed with several data segmentation and transmission characteristics with regard to the presented error concealment (EC) schemes.

Partitioning of streams at the source node is necessary since the overall data cannot be fit within the limited network packet size. Besides, segmentation has to include low-cost steps in order to decrease processing delay. Size of a segment (s_w) should also be maximized as much as possible to decrease total number of generated packets, $n(p)$. In a particular transmission, $n(p)$ is determined with:

$$n(p) = \frac{f_s \times t}{s_w} \tag{1}$$

In this study, nodes are built up to hold $s_w = \{20, 40, 80\}$ amplitude values in a data segment. For a specific s_w, $n(p)$ varies according to the sampling frequency (f_s) and the duration (t) of a voice selected in the data set. We aim to examine the effect of s_w on voice quality when data and network characteristics differ.

Heavy number of data packets passed over to each inter-hop of the network struggles with transmission delay and bandwidth. To minimize pre-transmission processing delay, A_i utilize a simplistic mechanism which buffers the segments with corresponding packet indices into the data memory. Same buffering structure is also accommodated on S_{ij} in order to minimize the relay time.

In each successful packet transfer, its corresponding index is also gathered. Thus, unattained packets are determined at the end of a voice transfer. Hence, loss pattern for a transfer is generated and projection is applied for the lost segments by considering two construction schemes: The first scheme inserts s_w zero amplitude values into the location of each lost segment, thus maintaining

the raw form with no concealment (V-NC) over the data set. For the second scheme (V-EC), a lost segment is corrected by siting the arithmetic mean of the sample units gathered from the previous and the next successfully gathered packets onto the lost sample units. The details of the algorithm is given below:

Algorithm 2.1. Algorithm for V-EC

```
1: Read the received voice data, D^r.
2: Determine the indices of lost segments in ascending order.
3: for all lost segment do
4:    Find the starting location of the sample units going to be reconstructed in D^r.
5:    Find the starting locations for the preceding and the next segment of the lost segment
      being dealt. If the next or previous segment is also missing, refer to next or previous
      neighbor until a healthy one is found. If the lost segment being dealt is the initial
      segment of the data, assign zero amplitude values into the segment. If the lost segment
      being dealt is the last one of the data, assign zero amplitude values into the segment.
6:    Create a temporary array having a size of s_w.
7:    for s_w times do
8:       Sum up each value sample unit value of the neighbor segments.
9:       Store their arithmetic mean in the array.
10:   end for
11:   Locate the array to the location of the lost segment.
12: end for
13: Generate the overall constructed data at the sink.
```

3 Voice Quality Assessment

In this study, VQA is valuated by a simplified version of transmission rating factor (R-factor) of ITU-T, which is an objective metric that can be easily accommodated on sensor nodes. The parameters of the equation is given below:

$$R = R_0 - I_s - I_d - I_{e,eff} + A \tag{2}$$

This study treats the packet loss probability (P_{pl}) defined in $I_{e,eff}$ as the main impairment factor. P_{pl} is inversely associated to transmission success rate (TSR), thus a relationship between voice quality and TSR is wanted to be revealed. Besides, f_s of a data is associated to simultaneous impairment factor I_s and investigated with quantizing distortion unit (qdu) defined in I_s. The permitted interval for qdu starts from value 1, meaning that a complete data quantization is supplied. When the quantization is at the lowest, qdu ends at value 14. To specify a scale between f_s and qdu, we assume the maximum f_s utilized in the tests—16KHz has the complete quantization. Conversely, the minimum audible f_s that a human ear can sense—3KHz is set for the maximum distortion. For all f_s used in our data set, qdu grades are determined and corresponding I_s values are generated, as shown in Table 1. By setting other parameters to their default values, R-factor is simplified to the following function of f_s and P_{pl}:

$$R(f_s, P_{pl}) = 58.9843 - 95\frac{P_{pl}}{P_{pl}+1} + 2.0714 \times f_s \tag{3}$$

A value obtained with R-factor can be mapped to a Mean Opinion Score (MOS) which is a widely used subjective VQA method. It is simply determined

Table 1. qdu and I_s values according to several f_s

f_s (Hz)	qdu	I_s
16000	1	1.4136
11025	6.025	10.0949
8000	9	17.1918
6000	11	22.0516
4000	13	26.4005

by the perceptual grades of an experimental group of audience. Ranging from bad to excellent, MOS is identified among a numerical quality scale from 1 to 5, respectively. In this way, R-factor gives an advantage of a VQA in both objective and subjective manners. For example, 90, 70 and 50 as R-factor values are mapped to MOS values of 4.3 (excellent), 3.6 (fair) and 2.6 (bad), respectively.

The correlation among each inter-hop link quality is traced with the following signal-to-noise ratio (SNR) metric

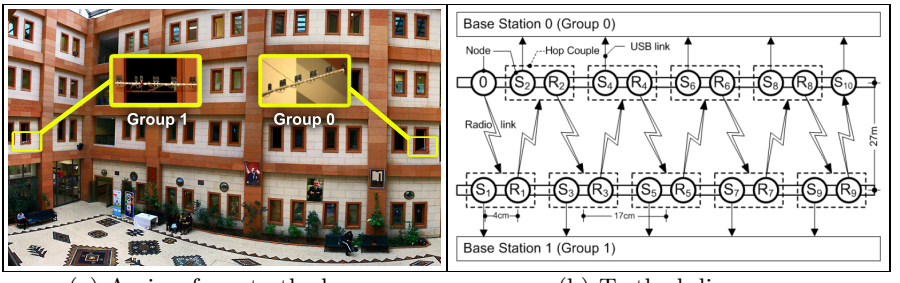
$$SNR(dB) = 10 \log_{10} \frac{|A_{signal}|}{|A_{noise}|} \tag{4}$$

where $|A_{signal}|$ is the total absolute values of the original amplitudes and $|A_{noise}|$ is of the difference between the original and reconstructed voices. SNR is utilized to examine the performance between the reconstruction techniques. Quality of both network and voice is assessed by the comparison of SNR and R-factor.

4 Experimental Setup

Actual transmissions are conducted inside a building with a large *atrium* as shown in Figure 1(a). A homogeneous testbed environment comprising a 10-hops network is constructed by 20 TMote Sky [10] sensor nodes. The nodes are associated with two groups, Group 0 and Group 1, which are lined up with a 28m distance in a parallel manner. The groups consist of five "hop couples", as depicted in Figure 1(b). TinyOS v2.1 with nesC v1.3 [7] is utilized to realize the data transfer. We have generated a voice data set comprising of simple invocatory commands each having a same fixed t. Each voice is generated with varying f_s listed in Table 2 and bit depths bd={8, 16}. Data $D^s{}_i$, i=1, 2 . . . , 8 are prerecorded with different f_s and bd via an acoustic sensor. Then, the samples with 8KHz/8bit are segmented and transmitted to the source serially, and then over the wireless transmission route with a fixed s_w in each unique test.

In our voice transfer scheme, each hop consists of two nodes called "hop couple". One of the nodes in each couple, called relay node (R_i, i=0, 1 . . . 9), is used to send the incoming data to the consecutive hop couple via radio link. Meanwhile, the other node, called snooping node (S_i, i=1, 2 . . . 10), is used to send the incoming data to the base station computer via USB link. To make hop based VQA with a wide variety of TSR, intermediate results in each hop



(a) A view from testbed area (b) Testbed diagram

Fig. 1. Testbed Environment

are recorded by snooping nodes. Nodes having IDs 0 and 10 are the source and the sink, respectively. In each test, the received data $D^r{}_i$, i=1, 2 . . . , 8 are saved at every hop with their corresponding packet indices. When a transmission of a voice data is over, unperceived data segments are determined. Corresponding mask files are generated for every s_w and f_s. Since f_s=8KHz is utilized in real tests, masks for lower and higher f_s are derived with down-sampling and up-sampling, respectively. With a conducive simulation, V-NC and V-EC are applied on the lost segments during projection over different voices in the data set.

Table 2. $n(p)$ according to s_w and f_s

		f_s (Hz)				
		4000	6000	8000	11025	16000
s_w	20	800	1200	1600	2205	3200
	40	400	600	800	1103	1600
	80	200	300	400	552	800

5 Performance Analysis

We have conducted 864 real voice transfer tests spreading to 10 days and gathered 22,800 voice loss patterns at 10 hops. The reflection of applying V-EC on the data gathered in comparison to V-NC can be clearly seen in Figure 2, which consists of nearly 18,000 SNR values calculated for all s_w. The distinction between each hop is noticed easily with color gradients for both V-NC and V-EC. For all s_w, values indicate that V-EC notably increases the quality. For s_w=40 and s_w=80, V-EC gets fair values in comparison to V-NC, but not so much higher as for s_w=20.

The graphs which show R-factor, SNR and TSR relationships in Figure 3 and Figure 4 include all the results projected to 8 different voice data with all f_s and *bd* in the simulation environment. It is seen that SNR values for the same voices

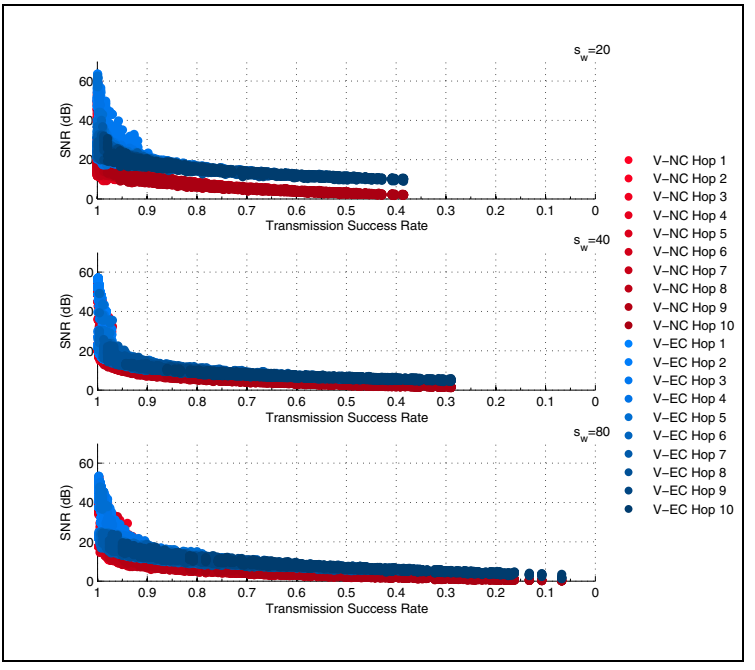


Fig. 2. SNR values of V-NC and V-EC algorithms wrt segment size

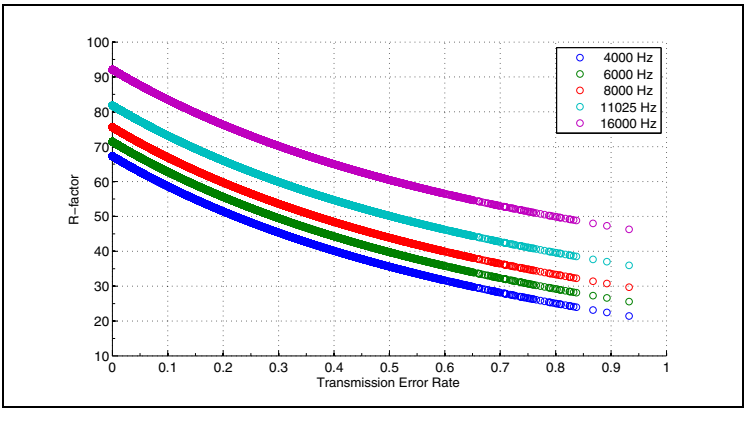


Fig. 3. Scatter plot of R-factor vs transmission error rate

with different *bd* versions are very similar to each other. Regardless of s_w or *bd* of a data transmitted, R-factor only depends on the overall TSR and f_s.

Figure 3 depicts the relation between R-factor and overall link quality. The graph shows that a comprehensible voice can be gathered when packet error rate is insured to be less than 30%. For a voice sampled at f_s=16KHz, R-factor value

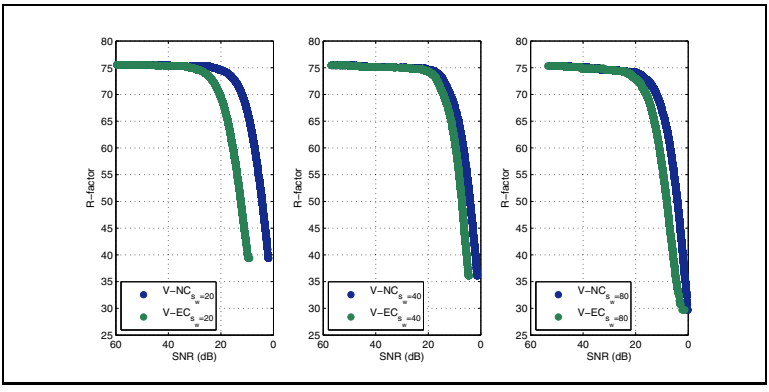


Fig. 4. Relationship between R-factor vs SNR

is nearly 70 even when TSR is 60%, which means that voice has a fair quality in terms of MOS. However, the decrease in f_s also decreases the metric values.

In Figure 4, the correlation between R-factor and SNR is depicted when f_s=8KHz. For both of the data resolutions—8 bit and 16 bit, SNR values for concealment algorithms on the overall data show resemblance. The effect of s_w can smoothly be seen on the data set. When s_w=20, increase in V-EC values are at maximum. Quality metrics for network and data intelligibility—R-factor and SNR visibly relate with each other.

6 Conclusion

In this study, a real wireless voice transmission testbed is established in order to disclose quality gradients of the continuous data being dispatched in a lossy multi-hop sensor network environment. The characteristics of the network against environmental factors are kept track of with several number of time-varying tests. The basic characteristics of voice data are essayed with different network properties. The results obtained after 9,000 real testbed transmissions reveal strong correlation between values obtained with the VQA metrics and TSR. The empirical results also showed that an affordable correction algorithm over the lost segments can provide a reasonable achievement in voice quality. We aim to concentrate on several EC algorithms and investigate their performances. Aside from error correction strategies, we aim to examine the effects of several factors defined in R-factor. Thus, data can be evaluated more concisely.

It is quite apparent that the network bandwidth must be efficiently used during voice transmission. Therefore, data characteristics for capturing and transmission must be kept as light as possible. However, the quality evaluation results clearly demonstrate that a transmitted stream becomes incomprehensible when the conventional characteristics of a voice—f_s and bd are lowered. In order to satisfy the affordance between network properties and data qualifications, the significance level of the information in voice data can be utilized. With these

considerations, a priority-based transmission scheme can be an exact solution for data integrity and validity. Several revisions and enhancements in both implementation and evaluation will pave the way for generating a complete voice transmission framework in Wireless Multimedia Sensor Networks.

References

1. Alesii, R., Graziosi, F., Pomante, L., Rinaldi, C.: Exploiting wsn for audio surveillance applications: The vowsn approach. In: 11th EUROMICRO Conference on Digital System Design Architectures, Methods and Tools, DSD 2008, pp. 520–524 (September 2008)
2. Azimi-Sadjadi, M.R., Kiss, G., Feher, B., Srinivasan, S., Ledeczi, A.: Acoustic source localization with high performance sensor nodes. In: Proceedings of SPIE, vol. 6562, 65620Y–65620Y–10 (2007)
3. Berisha, V., Spanias, A.: Real-Time Implementation of a Distributed Voice Activity Detector. In: Fourth IEEE Workshop on Sensor Array and Multichannel Processing, pp. 659–662 (2006)
4. Carvalho, L., Mota, E., Aguiar, R., Lima, A.F., de Souza, J., Barreto, A.: An E-Model Implementation for Speech Quality Evaluation in VoIP Systems. In: 10th IEEE Symposium on Computers and Communications, ISCC 2005, pp. 933–938 (2005)
5. Facchinetti, T., Ghibaudi, M., Anna, S.S.S., Pi, S.G.T., Goldoni, E., Savioli, A.: Real-Time Voice Streaming over IEEE 802.15.4, pp. 985–990. Packaging, Boston (2010)
6. International Telecommunications Union: Itu-t recommendation g.107 (2011), `http://www.itu.int/itudocr/itu-t/aap/sg12aap/history/g107/g107ww9.doc`
7. Levis, P.: Tinyos: An operating system for sensor networks (2006), `http://www.tinyos.net/tinyos-2.x/doc/pdf/tinyos-programming.pdf`
8. Li, L., Xin, G., Sun, L., Liu, Y.: QVS: Quality-Aware Voice Streaming for Wireless Sensor Networks. In: 2009 29th IEEE International Conference on Distributed Computing Systems, pp. 450–457 (June 2009)
9. Palafox, L.E., Garcia-Macias, J.A.: Wireless Sensor Networks for Voice Capture in Ubiquitous Home Environments. In: 2009 4th International Symposium on Wireless Pervasive Computing, pp. 1–5 (February 2009)
10. Telosb Crossbow: Telosb data sheet (2010), `http://www.willow.co.uk/TelosB_Datasheet.pdf`
11. Turkes, O., Baydere, S.: Voice Quality Analysis in Wireless Multimedia Sensor Networks: An Experimental Study. In: The International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP, pp. 317–322. IEEE (December 2011)
12. Wang, C., Sohraby, K., Jana, R., Ji, L., Daneshmand, M.: Voice communications over zigbee networks. IEEE Communications Magazine 46(1), 121–127 (2008)
13. Xu, J., Li, K., Shen, Y., Min, G., Qu, W.: Adaptive Energy-Efficient Packet Transmission for Voice Delivering in Wireless Sensor Networks. In: 2009 Sixth IFIP International Conference on Network and Parallel Computing, pp. 86–92 (October 2009)