AUTOMATIC RECOGNITION OF BLUETOOTH SPEECH IN 802.11 INTERFERENCE AND THE EFFECTIVENESS OF INSERTION-BASED COMPENSATION TECHNIQUES

Amr H. Nour-Eldin, Hesham Tolba and Douglas O'Shaughnessy

INRS-ÉMT, Université du Québec, Montréal, Québec, Canada.
{nour,tolba,dougo}@inrs-emt.uquebec.ca

ABSTRACT

In this paper, we investigate the ASR performance of speech transmitted over a noisy *Bluetooth* RF channel. Bluetooth shares its transmission channel with IEEE 802.11-based devices. Despite Bluetooth's frequency hopping scheme, our investigation shows that Bluetooth packet loss rates may reach up to 38% in unfavorable 802.11 interference conditions, and as Bluetooth uses a CVSD codec with *syllabic companding*, these packet losses not only manifest themselves as segments of missing speech upon CVSD decoding, but also as incorrect scaling of subsequent successfully received voice packets as CVSD step-size information is also lost. We investigate the effects of these degradations on the ASR performance of Bluetooth speech, and accordingly propose alternative CVSD decoder schemes employing *insertion-based* techniques for compensating for these effects. Results show that our proposed techniques improve ASR performance considerably while requiring only minor modifications to the current Bluetooth receiver.

1. INTRODUCTION

Bluetooth [1] is a recent (1999) low-power short-range wireless networking standard. Among its characteristics is that it handles both data and voice, low cost, universal interoperability, small size and negligible power consumption, facilitating its integration into portable low-power devices (e.g., cellular phones, laptop computers, PDAs). Bluetooth shares the 2.4 GHz ISM (Instrumentation, Scientific, and Medical) band with devices using the increasingly popular IEEE 802.11 standard for wireless networking [2], and thus suffers interference from 802.11 devices. Our investigation shows that Bluetooth packet loss rates may reach up to 38% in unfavorable 802.11 interference conditions. A Bluetooth packet is determined to be corrupt beyond repair, and accordingly discarded, upon failure of any of the two packet header error checking mechanisms [1, pp. 66–74]. Contrary to lost data packets which are retransmitted using an Automatic Repeat reQuest (ARQ) scheme, voice packets are not. The specifications rather point out that "measures have to be taken to fill in the lost speech segments" [1, p. 141].

Achieving the coexistence of IEEE 802.11 and Bluetooth has recently become the subject of some research [3]–[12], although packet error rates and collision modeling have received most of the attention [3]–[7]. In terms of mitigating the interference problem, only a few [8]–[10] have proposed practical solutions, but with the objective of improving 802.11 error rates rather than those of Bluetooth, mostly because of the perception that Bluetooth performance degrades only gracefully in interference, due to its frequency hopping scheme [11]. Moreover, these attempts have generally dealt with the interference problem from a network management and traffic control approach, requiring modifications to either the Bluetooth or 802.11 standard, or both. In terms of studying the interference effects on Bluetooth speech, mean opinion scores (MOS) were obtained as a function of the distance between the Bluetooth receiver and a WLAN transmitter in [12]; however, no measures were taken to replace the lost packets, effectively *splicing* the corrupt packets such that no gaps are left. This approach was evaluated in [13] to perform poorly (in terms of speech quality improvement), particularly for losses above 3%. No one, according to our knowledge, has actually investigated the ASR performance degradation of Bluetooth speech due to 802.11 interference, let alone improving it through compensation for the introduced distortions.

In contrast, much research has been directed towards the general problem of packet-loss recovery and *error concealment*. Traditionally, this research has concentrated on VoIP (Voice over IP) and wireless telephony applications, and has generally focused on the improvement of speech quality rather than ASR performance. Nonetheless, this well-researched area provides us with the ideology needed to compensate for the interference-caused distortions. A survey of such error concealment techniques can be found in [14, 15]. These techniques can be generally classified into sender-based and receiver-based schemes. Sender-based schemes involve higher computational demands and require considerable modifications to Bluetooth's packet format, while receiver-based schemes, although less performing, are efficient and require only minor modifications. Among receiver-based schemes, the insertion-based *silence substitution* and *packet repetition* are the most efficient and the simplest to implement in a real-time low-power Bluetooth environment, and hence, our choice of examining their effectiveness in conjunction with CVSD decoding.

Thus, we begin by presenting a thorough analysis of the speech degradation effects caused by 802.11 interference, based on which, we propose new CVSD decoding schemes employing insertion-based compensation techniques. Our results show that ASR improvements vary considerably depending on the scheme used, leading us to an optimal—from an ASR perspective—Bluetooth CVSD decoding scheme.

2. THE BLUETOOTH SPEECH TRANSMISSION AND ASR SYSTEMS

A Bluetooth speech transmission model based on that of [16] was constructed using Simulink®. The model simulates voice transmission between two Bluetooth-enabled devices, e.g., a mobile phone and a Bluetooth-enabled PC. In particular, we use the HV3 packet type [1, p. 59] (carrying 3.75 ms of speech), which performs

no forward error correction (FEC) on the payload. This specific packet type was chosen since we are interested in examining the effects of Bluetooth transmission on speech under least favorable conditions. A direct sequence spread spectrum (DSSS) 802.11b transmitter was added to the channel to model 802.11 interference. RF AWGN was also added to the channel together with a negative gain block to model path loss. The Bluetooth transmitter power level and the channel SNR and path loss were adjusted such that a maximum bit-error-rate (BER) of 0.1% is achieved at an input level of -70 dBm, the reference sensitivity level as specified by [1, p. 25], while the 802.11b transmitter power level and a similar channel negative gain were adjusted such that the power level at the Bluetooth receiver input is -20 dBm, the maximum operable level as specified by [1, p. 27]. Instead of splicing, which simply involves removing the corrupt packets from the packet stream and concatenating the remaining ones, thus disrupting the timing of the stream, silence substitution is performed by outputing zero-valued waveform samples while disabling the CVSD decoder for the duration of the lost packet(s), such that the most recent CVSD step-size is held. Upon reception of a new packet with uncorrupt header information, CVSD decoding resumes using the pre-loss step-size value.

A speaker-independent tri-phone HMM-based recognizer using MFCC parameters was also constructed for the continuous speech TIMIT task of 6146 words (some with multiple pronunciations), giving a word recognition correctness of 98.60% (i.e., a WER of 1.40%) for the 192 core test sentences defined on the TIMIT cdrom distribution.

3. PERFORMANCE IN 802.11 INTERFERENCE

As described above, silence substitution is performed upon failure of any of the packet header error checks, thus losing part of the transmitted speech; otherwise, the speech payload is passed to the CVSD decoder without any processing, on the assumption that it was received without any significant bit errors. However, even in the case of success of both header error checks, there is also the possibility that random bit errors may occur as a result of the 802.11 interference and/or the channel AWGN and path loss. In this latter case, the errors are passed undetected to the CVSD decoder. Fortunately, the CVSD decoder employing syllabic companding is quite robust against such random bit errors, which is why this speech codec was originally chosen for Bluetooth [1, p. 141]. Moreover, employing FEC for the payload bits (as in the Bluetooth packet types HV1 and HV2) would remove most of these bit errors if not all.

The effects of these two forms of errors, packet losses and random bit errors, can be investigated separately by examining the ASR performance and error statistics of speech transmitted under three interference conditions: no 802.11 interference, average interference, and worst-case interference. In the first case, the interferer is completely turned off; this represents the case where no 802.11 equipment is operating simultaneously in the same local area of the Bluetooth transceiver. The two latter interference scenarios were simulated with the interference set to ON at all times for the worst case, and alternating between ON and OFF at an equal rate for the average case interference. The 802.11 parameters (e.g., packet length, packet rate) needed for simulation of the average case interference were calculated according to [2, sec. 7.1.2 and 15.2.2]. Table 1 shows the average *raw* BER, *residual* BER, frame error (i.e., packet loss) rate (FER), and the corresponding ASR rates for the 192 test utterances. The raw BER is calculated over all received bits (including header and speech payload bits) regardless of the result of the header error checks, while the residual BER is calculated for the speech payload bits only in the case of success of the header error checks.

Interference	Raw BER	Res. BER	FER	ASR
OFF	0.10	0.11	0	97.58
average	9.16	1.73	21.33	86.11
worst	18.09	0.58	38.64	76.42

Table 1. Average error and ASR rates (%) for the three interference conditions.

Table 1 shows that in the absence of interference, not a single packet was discarded upon header error check, a direct consequence of the Bluetooth specification's BER constraints. Moreover, only 0.11% of the encoded speech bits were incorrectly received due to random bit errors caused by the channel AWGN and path loss. Due to the robustness of the CVSD decoder, these bit errors have a minimal effect during speech reconstruction as shown by the recognition performance (97.58% down from 98.60%).

As the level of interference increases, more bit errors occur as the Bluetooth hopping transmission falls in the 802.11 bandwidth while the interferer is ON. Consequently, more frame losses occur as shown in rows 2 and 3 of Table 1. The increase in the ratio of residual BER to the raw BER for the average interference case (18.89%) compared to that of the worst case (3.21%), is explained by the discontinuity of the 802.11 interfering transmission, specifically when the 802.11 transmitter turns ON in the middle of a Bluetooth packet payload after it was initially OFF during transmission of the header, causing a false decision in the Bluetooth receiver that the received packet is mostly free of errors due to header error check success while considerable portions of the payload may be corrupt. Despite this, the ratio of residual BER to raw BER is still low. Coupled with the ability to remove most random payload bit errors by employing FEC for the payload, i.e., by using HV1 or HV2 packets, and noting that the decrease in ASR rate is almost in direct proportion to the FER, we conclude that the considerable decrease in ASR performance is mostly a result of frame losses rather than random bit errors. Hence, we focus our analysis on the effects of frame losses.

4. EFFECTS OF FRAME LOSSES

To identify the effects of frame losses, speech waveforms before and after Bluetooth transmission were analyzed. We also examined recognition results during the Viterbi process where the HMM state alignment of each analysis frame and the corresponding best tokens with their acoustic log-likelihoods were tracked, for various phonetically different segments. Fig. 1(b) shows a 0.95 s segment of an example speech file where several packet losses occur due to 802.11b interference, compared to the same "clean" segment before transmission, shown in Fig. 1(a). Based on this analysis, the effects of frame losses are identified as follows:

4.1. Speech waveform gaps

Gaps occurring in the waveform are a direct result of lost packets. Such gaps are shown in Fig. 1(b). Their effect on ASR performance depends on their number and location. The number of

lost packets within the duration of an analysis window adversely affects ASR performance. More lost packets lead to greater loss of information in an analysis window, consequently MFCC parameters experience larger offsets, and hence lower acoustic log-likelihoods. A few consecutive lost packets in the middle of a Hamming window have higher distorting effects on its MFCCs than if they were located at the edges. The phonetic significance of the region where the gaps occur also plays a role. Packet losses have less effect on recognition log-likelihoods for weak speech segments or silence.

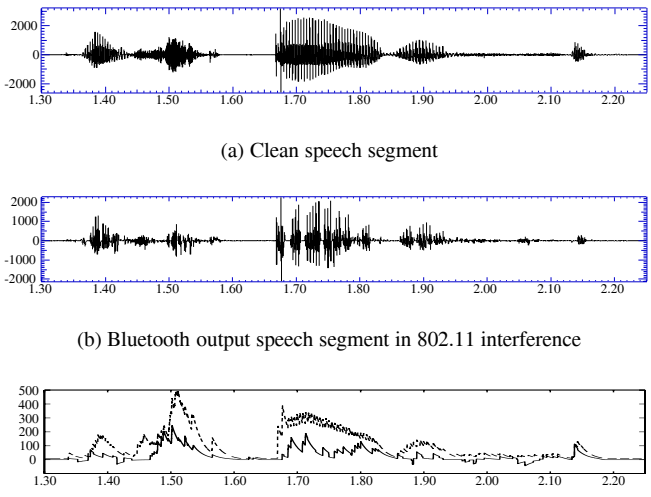
4.2. Step-size errors

As the CVSD decoder is disabled upon packet loss, its most recent pre-loss step-size value is held, using it as the starting value to decode new correctly received packets. However, the incoming bits were output from the CVSD encoder based on a step-size that is different from the pre-loss value. This deviation in the step-size value following a loss (as opposed to the value it would have taken if no previous packets were lost) manifests itself as erroneous scaling of post-loss packets. Fig. 1(c) shows the "clean" step-size values aligned to step-size errors resulting from packet losses for the 0.95 s speech segment. Upon analysis of these errors, we find that step-size error peaks occur at packet loss locations. Although the step-size error decays between peaks, this decay may continue for several packet lengths before reaching zero error, depending on the peak level it starts decaying from and the time of the next peak (packet loss). Consequently, step-size errors propagate along the waveform and thus erroneously scale even the regions where no packet losses occur. This incorrect scaling is evident by comparing the clean and distorted waveforms in Figs. 1(a) and 1(b) at 1.5 and 1.8 s. Moreover, as the number of consecutively lost packets increases, the step-size value experiences a greater offset, leading to a bigger incorrect scaling factor for the following correctly received packets. Step-size errors also depend directly on the average level of the waveform prior to the loss, where weak magnitudes are less affected by such deviations in CVSD step-size. We also found that silences or short pauses effectively "reset" the step-size value, even if immediately prior speech has been incorrectly scaled up to that pause, and consequently, speech following the pause would be correctly scaled until new packet losses occur causing new step-size errors. For regions where several closely separated losses occur such that the step-size error decay is smaller than the error increases caused by new packet losses, the error builds up, increasing the scaling error and causing larger waveform offsets. These larger offsets naturally lead to higher MFCC deviations, resulting in an increase of the acoustic log-likelihood degradation for the corresponding frames, i.e., more degraded ASR.

Hence, measures should be taken to correct the CVSD decoder step-size errors as much as possible, while simultaneously compensating for the lost speech.

5. ALTERNATIVE DECODING SCHEMES

The drawback of the silence substitution scheme used above is that waveform discontinuity occurs at replaced packet boundaries. Silence substitution can be alternatively implemented by substituting a 0, 1, . . . sequence for a lost packet's corrupt CVSD encoded bits, causing the CVSD decoder step-size to quickly decay to its minimum value since the Bluetooth CVSD codec uses 4-bit syllabic companding, where the adaptive step-size value is adjusted based



(a) Clean speech segment

(b) Bluetooth output speech segment in 802.11 interference

(c) Step-size values (dashed) and the corresponding errors (solid)

Fig. 1. Analysis over a 0.95 s segment (x-axis: time (s), y-axis: amplitude).

on whether the four most recent encoded bits are equal or not. Accordingly, the speech output decays to zero within a short period ($\approxeq$ 0.5 ms, the time constant of the accumulator decay factor [1, p. 140]), continuing to be zero until a new packet is correctly received where the step-size begins to increase following the new packet's bits. Although this scheme does not solve the lost speech or the incorrect scaling problem since it only "resets" the step-size during packet losses, it ensures waveform continuity at replaced packet boundaries.

In contrast, *Packet Repetition* replaces lost packets with copies of those that arrived immediately before the loss. It performs reasonably well both subjectively and in terms of intelligibility [17]. We propose performing packet repetition on the CVSD decoder speech output, as well as on its encoded payload input directly. Thus, lost packets and step-size errors can be simultaneously dealt with by the following schemes:

1. CVSD disabling + silence substitution: As above.

2. CVSD disabling + waveform repetition: The previous waveform output is repeated. Waveform continuity is achieved only at the boundary between the repeated waveform and the subsequent decoded speech.

3. Step-size resetting: The CVSD decoder output is used as the final output speech waveform, thus implicitly performing silence substitution. Waveform continuity is also ensured.

4. Step-size resetting + waveform repetition: The speech waveform output immediately before a lost packet is repeated in place of the CVSD decoder silence output for the loss duration. Upon successful reception of a packet following the loss, the output is switched back to the CVSD decoder output, using the minimum step-size as the initial value. Hence, waveform discontinuities are inevitable at replacement packet boundaries.

5. CVSD packet repetition + silence substitution: Silence can be substituted for the CVSD decoder output while repeating the CVSD encoded input upon packet less. The advantage of repeating CVSD input packets over disabling it or step-size resetting is that it incorporates the dynamic change of the step-size value immediately before a loss into its estimation since the CVSD decoder

Employed Schemes	Continuity	ASR
1. CVSD disabling + Silence subst.	No	76.42
2. CVSD disabling + Waveform rep.	1 side	80.11
3. Step-size resetting	2 sides	79.99
4. Step-size resetting + Waveform rep.	No	76.29
5. CVSD packet rep. + Silence subst.	No	65.77
6. CVSD packet rep. + Waveform rep.	No	73.74
7. CVSD packet rep.	2 sides	81.52

Table 2. Characteristics and ASR performance (%) of the alternative decoding and compensation schemes.

mimics its operation prior to the loss. Thus, subsequent speech would have better scaling. However, the output waveform will not be continuous at replacement packet boundaries due to silence substitution of the decoder's output.

6. CVSD packet repetition + waveform repetition: Waveform repetition can also be performed as an alternative to using the CVSD decoder output. Although the step-size value at the onset of speech after a loss would incorporate the step-size dynamic shape immediately before the loss, the output waveform boundaries around replacement waveforms would still not be continuous.

7. CVSD packet repetition: Waveform continuity is ensured by using the CVSD decoder output directly as the output waveform, without performing any processing on the waveform itself as in the two cases above. In this case, the replacement waveform would be a scaled repetition of the waveform immediately preceding a loss.

6. ASR RESULTS

Recognition tests were performed using the compensation and step-size correction schemes described above. Worst-case interference over the Bluetooth channel was assumed during model simulations. Table 2 summarizes the characteristics and ASR performance of these schemes. The ASR results of Table 2 show that ensuring waveform continuity on one or both sides of the replacement waveforms clearly improves ASR performance, as shown by the performance of schemes 2, 3 and 7. Comparing the performance of schemes 5 and 6 confirms that waveform repetition generally outperforms silence substitution. Finally, the best ASR performance improvement (6.67%) is obtained by incorporating the pre-loss dynamic characteristics of the CVSD decoder step-size into its post-loss estimation through CVSD packet repetition, and using the resulting continuous and scaled waveform as the output speech.

7. CONCLUSIONS

We present a thorough analysis of the effects of 802.11 interference on Bluetooth speech in general, and from an ASR perspective in particular. Our analysis shows that ASR performance degradation in the presence of 802.11 interference is not only due to missing speech segments caused by packet losses, but also due to propagating step-size errors which result in erroneous scaling of correctly received packets. Accordingly, we proposed several alternative decoding schemes employing insertion-based compensation techniques. ASR results show the superiority of simultaneously incorporating pre-loss step-size information in its post-loss estimation by performing CVSD packet repetition while replacing lost speech, compared to silence substitution and the other decoding schemes.

8. REFERENCES

[1] *Specification of the Bluetooth system — Core*, version 1.1, feb 2001. Available at http://bluetooth.com/dev/specifications.asp

[2] *ANSI/IEEE std 802.11 — WLAN MAC and physical layer specifications*, 1999 edition. Available at http://standards.ieee.org/getieee802/802.11.html

[3] I. Howitt, "Bluetooth performance in the presence of 802.11b WLAN", *IEEE Trans. Vehicular Tech.*, vol. 51, no. 6, pp. 1640–1651, 2002.

[4] I. Howitt, "IEEE 802.11 and Bluetooth coexistence analysis methodology", *Proc. 53rd IEEE Vehicular Technology Conf., VTC*, pp. 1114–1118, 2001.

[5] G. Ennis, "Impact of Bluetooth on 802.11 direct sequence", *IEEE 802.11-00/319*, 1998.

[6] J. Zyren, "Extension of Bluetooth and 802.11 direct sequence interference model", *IEEE 802.11-98/378*, 1998.

[7] J. C. Haartsen and S. Zürbes, "Bluetooth voice and data performance in 802.11 DS WLAN environment", Ericsson SIG Publication, 1999.

[8] C. F. Chiasserini and R. R. Rao, "Performance of IEEE 802.11 WLANs in a Bluetooth Environment", *Proc. IEEE Wireless Commun. and Networking Conf., WCNC*, pp. 94–99, 2000.

[9] A. Kamerman, "Coexistence between Bluetooth and IEEE 802.11 — CCK solutions to avoid mutual interference", *IEEE 802.11-00/162*, 2000.

[10] J. Zyren, "Reliability of IEEE 802.11 hi rate DSSS WLANs in high density Bluetooth environment", Bluetooth'99 SIG Publication, 1999.

[11] R. Shorey and B. A. Miller, "The Bluetooth technology: merits and limitations", *Proc. IEEE Internat. Conf. Personal Wireless Commun., ICPWC*, pp. 80–84, 2000.

[12] B. Jiang and O. Yang, "Performance evaluation of Bluetooth system in the presence of WLAN IEEE 802.11 system", *Proc. IEEE Canadian Conf. Electrical and Computer Eng., CCECE*, pp. 1633–1636, 2003.

[13] J. G. Gruber and L. Strawczynski, "Subjective effects of variable delay and clipping in dynamically managed voice systems", *IEEE Trans. Commun.*, vol. 33, no. 8, pp. 801–808, 1985.

[14] C. Perkins, O. Hodson and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio", *IEEE Network Mag.*, vol. 12, no. 5, 1998.

[15] B. W. Wah, X. Su and D. Lin., "A survey of error-concealment schemes for real-time audio and video transmissions over the internet", *Proc. IEEE Internat. Symp. Multimedia Software Eng.*, pp. 17–24, 2000.

[16] S. McGarrity, "Bluetooth voice Simulink model", *Matlab digest*, nov 2001. Available at http://www.mathworks.com/company/digest/nov01/bluetooth.shtml

[17] R. C. F. Tucker and J. E. Flood, "Optimizing the performance of packet-switched speech", *Proc. IEEE Conf. Digital Processing of Signals in Communications*, pp. 227–234, 1985.