Phone Impact Based Speech Transmission Technique for Reliable Speech Recognition in Poor Wireless Network Conditions

Azar Taufique[1,3], *Kumaran Vijayasankar*[1], *Wooil Kim*[2], *John H.L. Hansen*[2], *Marco Tacca*[1], *Andrea Fumagalli*[1]

[1]Open Networking Advanced Research Laboratory, [2]Center for Robust Speech Systems
The University of Texas at Dallas, Richardson, Texas, USA
{axt082000,kumaran,wikim,John.Hansen,mtacca,andreaf}@utdallas.edu

Abstract

This paper presents a preliminary study on an effective differentiable network service technique to achieve improved speech recognition under severely poor wireless channel conditions, by leveraging multiple priority levels applied to speech classes. Each speech class is assigned a different priority level based on its level of impact on speech recognition performance. Based on their priority level, frames of each speech class are given distinct levels of network quality of service (QoS) to satisfy the delay requirement and enable speech recognition at the receiver. This proposed Phone Impact (PI) based priority class is compared to the Voiced/Unvoiced (VU) based priority class in this study. The experimental results prove that the proposed scheme is effective at providing wireless network service for robust speech recognition under poor channel conditions, showing up to 2.67 dB and 5.93 dB lower Signal to Noise Ratio (SNR) operating regions compared to the VU based and plain protocols respectively. The PI based method also shows acceptable WERs at lower SNRs where VU and plain systems significantly degrade in speech recognition performance in case of retry limit of 6.

Index Terms: Phone Impact, Priority Class, Speech Recognition, IEEE 802.11, Differentiated Maximum Retry Limit.

1. Introduction

Speech recognition is increasingly gaining importance in automated response systems that provide customer interaction services such as emergency relief in disaster scenarios, etc. Many of such systems receive speech signal through wireless networks including cellular networks, wireless local area networks, etc. Wireless networks are prone to frame (i.e., packet carrying speech signal) loss, which can be caused due to collisions from the transmission of other nodes or due to frame errors caused by low SNR at the receiver. In order to provide reliable delivery of the signal frames, wireless link layer protocols make multiple attempts to deliver a frame till an acknowledgment from the destination is received or a maximum limit of attempts is reached. In the latter case notification of the failure in delivering the frame is given to higher layers.

Under poor channel conditions the probability of the frame not being received by the receiver is high, increasing the average number of transmission attempts made by the link layer protocols. The increase in the number of attempts made by a transmitter leads to a high collision probability with other transmission attempts, further increasing the average number of attempts per frame made by the link layer. Such an increase in the number of attempts results in more delayed time in service per a frame, thereby increasing the end-to-end frame delivery delay. Moreover, under such conditions, a higher portion of frames may reach the maximum limit of the retry attempt number, thus degrading the channel reliability.

Speech recognition as similar to other real time applications requires the end-to-end delay of frames to be within a tolerable limit in order to offer a real time service. It also requires a minimum amount of information to be received in order to achieve reliable recognition performance. The loss of frames can be compensated to an extent by the use of improved language models or by employing packet loss concealment techniques [1, 2]. However, the performance of such techniques are limited by the information content and timing of the received speech signal.

Since at a low SNR condition both the delay requirements and successful delivery of all frames cannot be satisfied, performance of speech recognition systems might be severely compromised. One way to meet the delay requirement at such a condition can be to drop a subset of frames from being retransmitted. However, this technique would significantly restrict the information content of the received signal. An effective way for this is to drop the frames in such a way that the portion of speech signals that has greater impact on speech recognition is successfully delivered. The presence of different classes of frames within an application having different levels of impact on the quality of the application has been studied in literature [3, 4]. Classification of voice signals is performed in these works and the overall quality of the received voice signal is shown to improve by providing a distinct level of QoS to each class. However, to the best of authors' knowledge, there has been no prior work that uses such a method to offer improved performance of speech recognition under poor channel conditions.

In this paper, the use of a *Phone Impact (PI) based priority class* in conjunction with a differentiated maximum retry (DMR) limit technique at the wireless link layer is proposed to improve speech recognition performance under poor channel conditions. Frames are assigned different number of retransmission attempts (i.e., QoS level) depending on the level of importance of their carried phonetic information. IEEE 802.11 is chosen as the wireless protocol for the study and the speech recognition system performance is measured in terms of word error rate (WER). The performance of the proposed scheme is compared against a previously published work [3], which in this paper is referred to as *Voiced/Unvoiced (VU) based priority class* method. The proposed Phone Impact based class method and the DMR technique are described in the following sections.

2. Impact of Speech Class on Speech Quality

In this section, the impact of speech phone class on speech quality is investigated by employing speech recognition as a performance measure. Here, the SPHINX3 [5] large vocabulary speech recognition system and the TIMIT [6] corpus are employed. The TIMIT corpus consists of 5.6 hours of speech data including a total of 630 speakers, where all data is pho-

[3] Affiliated with University of Engineering and Technology,Lahore,Pakistan

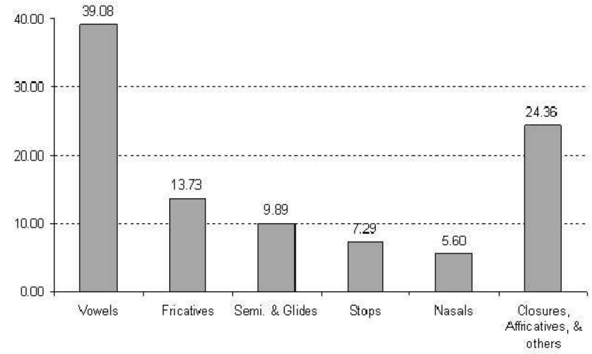

Figure 1: Distribution of frames for each speech phone.

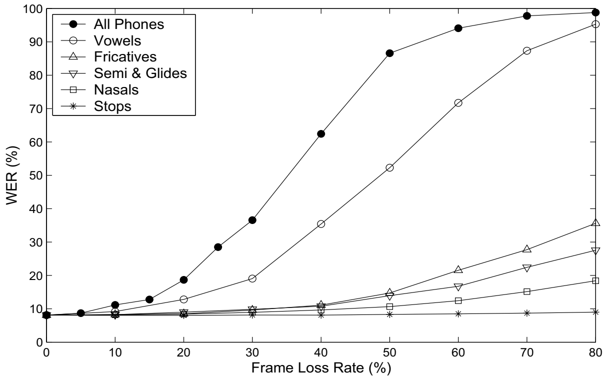

Figure 2: WER vs. Frame loss rate for speech classes.

netically balanced. Details on the speech recognition system and corpus will be presented in Sec. 4. Figure. 1 shows the distribution of speech frames for each speech phone class (i.e, vowels, fricatives, semi & glides, stops, nasals, and others) in the TIMIT corpus. Figure. 2 shows the impact of each phone class on the speech recognition performance in WER. In order to obtain these plots, frames of speech are randomly dropped corresponding to the target phone class on each plot according to the given frame loss rate. In this experiment, a speech frame consists of 5 msec length block and the dropped (i.e., lost) frame is replaced with a silence segment. From the plots in Figure. 2, it can be observed that each phone class has different impact on the speech recognition performance. In case of "Vowels", we obtained 52.29% WER for 50% of frame loss rate. Considering the occupancy of vowel sounds in speech (i.e., 39.08% from Figure. 1), same amount of frame loss rate for "All Phones" becomes 19.54%, which is expected to get 18.13%[1] in WER. This indicates that the vowel sounds have approximately 3 times the impact on speech recognition than the average of all phone classes (i.e., 52.29% vs. 18.13%). The "stop" class does not have any effect on speech recognition even dropping all of stop sounds. This is due to the fact that the linguistic models associated with the speech recognition engine is able to recover these lost phonemes based on probability distribution of the word sequences. This finding suggests that a different strategy of transmitting or processing each phone class will be effective to increase overall speech recognition performance or improve the intelligibility of the delivered speech signal, when communicating over the adverse network environment where the packet loss might easily happen. Figure. 3 presents the effects of Voiced/Unvoiced sounds on speech recognition obtained in a similar manner as the plots of Figure. 2, showing different impacts on speech recognition performance. The voiced portion of the signal constitutes 61% of the speech data.

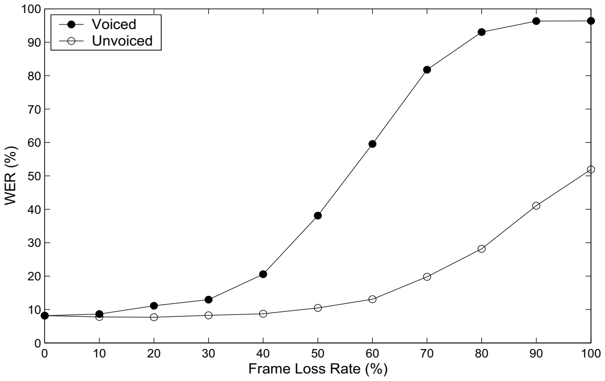

Figure 3: WER vs. Frame loss rate WER for voiced/unvoiced classes.

3. Differentiated Maximum Retry (DMR) Limit Technique

IEEE 802.11 is a carrier sense multiple access with collision avoidance (CSMA/CA) protocol developed for wireless local area networks (WLAN) [7]. The channel access procedure used in this study is the distributed coordination function (DCF) of the IEEE 802.11 protocol. Under DCF, a (active) node that has a frame to transmit initially will continue its attempts to deliver the frame till either the frame gets successfully delivered or the retry limit ($RETRYLIMIT$) is reached. In case the retransmission limit is reached, the current frame is discarded and the next awaiting frame is transmitted. Increasing the $RETRYLIMIT$ will increase the chances of delivering a frame successfully from source to destination but will however cause the node to transmit the frame many times which can lead to collision with other frames and also increase the channel utilization.

Under poor channel conditions (low SNR) the probability of success of delivering a frame becomes low, and a higher number of attempts is required to deliver a frame successfully. Higher number of attempts implies an increase in time to service a frame, which may push the network to saturation, increasing the end-to-end delay beyond the tolerable limit. In order to meet the delay requirement and still achieve the required performance, the DMR technique can be used when possible. In DMR, not all frames are transmitted with a high $RETRYLIMIT$. The application layer identifies the frame that is sent to the IEEE 802.11 layer with a priority level. In our study only two levels of priority are used. In the proposed PI based priority class, the vowels, semi and glide phone frames are classified as the higher priority, whereas all the other phones are considered as the lower priority. In the VU based priority class, voiced sound frames are considered to be of higher priority. The IEEE 802.11 protocol sends the higher priority frames with a higher $RETRYLIMIT$ and the lower priority frames with lower $RETRYLIMIT$.

As the lower priority frames take less processing time, the overall service time required for a frame also decreases. Thus at the cost of lower priority frames being more likely lost, the higher priority frames are sent with a higher reliability. Since the higher priority frames have more impact on speech recognition, increase of their reliable deliveries will help enhance the received speech quality. It should be noted, that under the PI based priority class only 50% (i.e., 39.08% + 9.89% from Figure. 1) of speech data will be classified as higher priority as opposed to 61% in the VU based priority class classification. This enables the PI based classification to be more robust to poor channel conditions. Though lesser number of phones are

[1] It is obtained by interpolation of 12.78% and 18.67% of WERs on 15% and 20% loss rates respectively for the "All Phones" case from Figure. 2.

Table 1: Parameter values used in simulation.

Path Loss Exponent β = 4	Fading is Flat Rayleigh
Average Transmitter Power = 100 mW	PHY Header = 192 bits
Transmission data rate = 1 Mbps	Speech Frame = 10 bytes
RTP Header = 20 bytes	UDP Header = 20 bytes
$SIFS$ = 10 μs	Vulnerable Period = 20 μs
$DIFS$ = 50 μs	Slot Time = 20 μs
MAC ACK = 14 bytes	CW_MIN = 63 slots
CW_MAX = 1023 slots	MAC Header = 34 bytes

protected when the PI based priority class is used, it provides better performance to the VU based priority class as shown in the following section.

4. Experimental Results

The TIMIT [6] speech corpus is used for performance evaluation of the proposed method. A total of 4.1 hours of speech (462 speakers, 4,620 utterances) are used for training the acoustic model of the speech recognizer, and 1.5 hours of data (168 speakers, 1,680 utterances) are used for test. The training and the test sets do not overlap each other in speakers and uttered sentences. We employ SPHINX3 [5] as Hidden Markov Model (HMM) based speech recognizer to obtain recognition accuracy in varying channels/transmission conditions. Each HMM represents a tri-phone which consists of 3 states with an 8-component Gaussian Mixture Model per state, which is tied with 1,138 states. The task has 6,233 words as the vocabulary, and the bigram language model is adapted on the TIMIT database using a Broadcast News language model as an initial model. A conventional Mel-Frequency Cepstral Coefficient (MFCC) feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [8]. An analysis window of 25 msec in duration is used with a 10 msec skip rate for 8 kHz speech data. To evaluate the ultimate performance gain that can be achieved by applying DMR to classified speech frame, the paper assumes that an "Oracle" knowledge of phone class (including voiced/unvoiced) for each speech frame is available at the source side. The test part of the TIMIT corpus is used at the application layer as the source of the speech signal. The data is converted into frames of 10 bytes and are generated at the rate of 100 frames per second. This is in accordance to the G.729 codec and is one of the commonly used encoding methods [9].

Five speech sources concurrently transmitting to a common speech recognition system are considered in the simulation. Real time transport protocol (RTP) and user datagram protocol (UDP) are assumed to be used at the transport layer. All sources can sense each other and are assumed to perceive the same channel condition (i.e., no hidden node). The SNR and the bit error rate (BER) are same for all the sources. A conventional repetition-based packet loss concealment (PLC) technique [1] is used at the receiver end to recover the lost frames. The presented results are obtained by taking an average over the five sources. The IEEE 802.11 was simulated using a custom built simulator and the parameters used are shown in Table 1.

First, the performance of the PI based priority class and the VU based priority class methods when 6 retry attempts (in practice this value or close to this is used) for the higher priority and 1 retry attempt for the lower priority are used, is compared against the plain IEEE 802.11 protocol operating with 6 retry attempts for all frames. Figure. 4 shows the logarithm of the end-to-end delay across SNR. The maximum tolerable delay is chosen as 1 sec in this study. It should be noted that most applications have a tolerable delay lower than this value. It is observed that the proposed PI based priority class when combined with DMR is more robust against poor channel conditions, which is able to meet the delay requirement till 11.14 dB as opposed to 12.98 dB by the VU based class and 16.02 dB by the plain protocol. The corresponding WER is shown in Figure. 5 for the SNR values at which the delay of at least one of the systems (Plain, VU based or PI based) is within the tolerable limit. The PI based priority class method achieves a WER comparable to other systems and also obtains low WER at SNR values where other systems show significantly high WER. It is worth to note that the proposed PI based priority class approach provides reasonable WERs at lower SNRs (i.e., 11-12 dB), however both the plain protocol and VU based method show severely degraded recognition performance at the SNRs, which cannot be employed for the speech recognition system in a real-life situation.

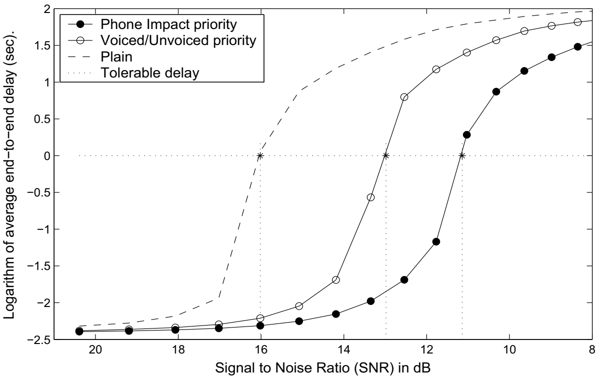

Figure 4: Logarithm of average end-to-end delay vs. SNR when the higher priority frames are given a retry limit of 6 and lower priority frames are given a retry limit of 1.

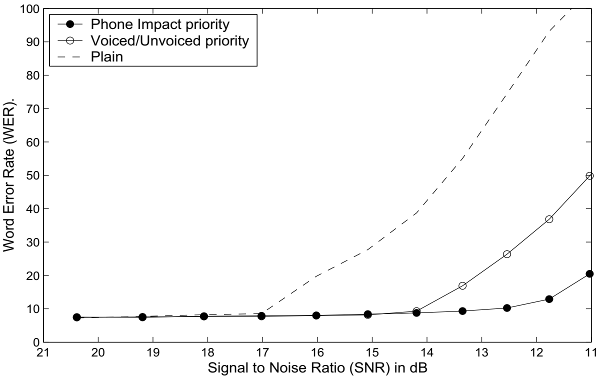

Figure 5: WER vs. SNR when the higher priority frames are given a retry limit of 6 and lower priority frames are given a retry limit of 1.

The corresponding delivery ratio of total number of frames delivered by PI based priority class is compared with VU based and plain protocol in Figure. 6. It is evident that the delivery ratio of PI based priority class is lower than plain and VU based protocols. Such an increase in frame loss rate is due to more number of the lost frames which are assigned a lower priority. In the PI based prority class method, 50 % of frames are considered as the lower priority, which is larger comapred to 39 % frames (i.e., unvoiced sounds) in the VU based method. It is able to protect those frames which are more important for speech recognition by giving higher retry attempts at the expense of giving lower retry attempts to other phones. Therefore, the proposed PI based technique is more effective to meet delay requirements, providing good QoS and better WER in poor SNR conditions whereas other protocols loose QoS, WER, and delay requirements trying to give high importance even to those frames which do not have much impact on speech recognition.

In the next experiment, 3 retry attempts (in order to see the

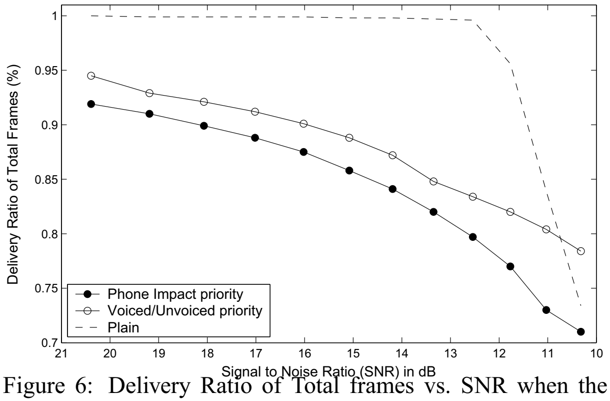

Figure 6: Delivery Ratio of Total frames vs. SNR when the higher priority frames are given a retry limit of 6 and lower priority frames are given a retry limit of 1.

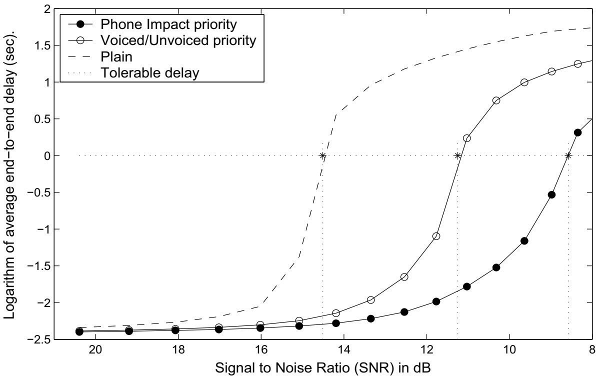

Figure 7: Logarithm of average end-to-end delay vs. SNR when the higher priority frames are given a retry limit of 3 and lower priority frames are given a retry limit of 1.

effect of half number of retry attempts as compared to previous case) are used for higher priority and 1 retry attempt is used for lower priority, offering a comparison against the plain protocol that uses 3 retry attempts for all frames. Figure. 7 shows the logarithm of end-to-end delay across SNR. Here again it can be observed that the PI based method is more robust and is able to meet the delay requirement till 8.58 dB as opposed to 11.25 dB by the VU based and 14.51 dB by the plain protocol. Figure. 8 and Figure. 9 shows the corresponding WER and delivery ratio of total frames, revealing a trend similar to the earlier experiment. Here, the PI based priority class method shows consistently improved WERs for the lower SNRs, compared to the VU based method. It can be observed that the WER values in this case is better than the one when 6 and 1 retry limits were used for higher and lower priority levels at lower SNR regions. This is because, since the retry limit for higher priority (3 in this case as compared to 6 in previous case) is lower, the delay is reduced thus helping in improving the WER. This suggests that different retry limit combinations would be more effective at different SNR conditions to satisfy delay and WER requirements and will be explained in future work.

5. Conclusion and Future Work

An effective technique for speech transmission yielding satisfactory performance in terms of WER over poor wireless channel conditions was presented. It was shown that the proposed Phone Impact based priority scheme can perform at up to 2.67 dB lower SNR regions compared to the Voice/Unvoiced based method, and up to 5.93 dB when compared to plain IEEE 802.11 implementation. The PI based method also showed consistently improved WERs for all SNR conditions compared to

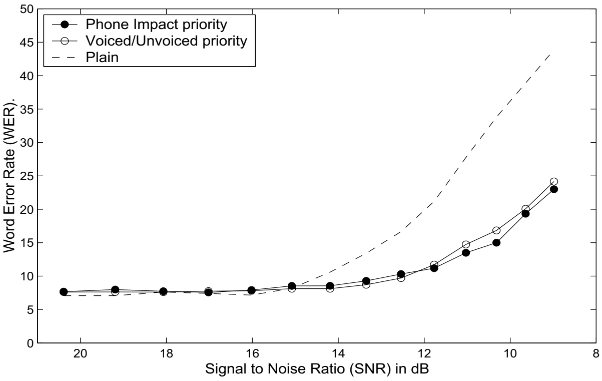

Figure 8: WER vs. SNR when the higher priority frames are given a retry limit of 3 and lower priority frames are given a retry limit of 1.

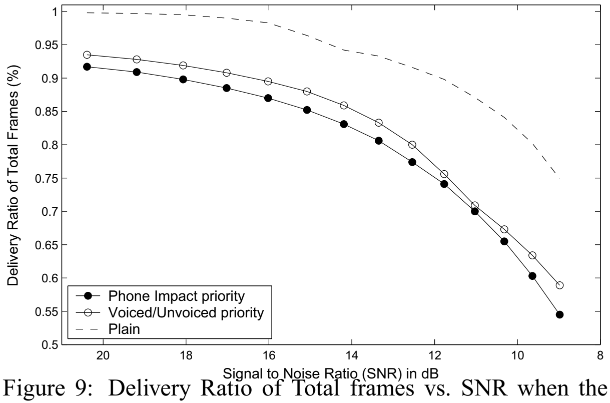

Figure 9: Delivery Ratio of Total frames vs. SNR when the higher priority frames are given a retry limit of 3 and lower priority frames are given a retry limit of 1.

VU based method. In case of the retry limit of 6 it provided acceptable WERs at lower SNRs (11-12dB), where both VU and plain systems show significantly high WERs. It has been shown that the obtained WER depends on the retry limits assigned to each priority class. As a future work, a technique to determine *a priori* the optimal retry limits for a given SNR will be investigated.

6. References

[1] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," in *IEEE Network Magzine*, Sep./Oct. 1998, pp. 40–48.

[2] C.A. Rodbro, M.N. Murthi, S.V. Anderson, and S.H. Jensen, "Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP," in *IEEE Trans. Audio, Spech, and Language Processing*, Sept. 2006, pp. 1609–1623.

[3] C. Hoene, I. Carreras, and A. Wolisz, "Voice Over IP: Improving the Quality Over Wireless LAN by Adopting a Booster Mechanism—An Experimental Apprach," in *Proc. of SPIE-Voice Over IP (VoIP) Technology*, August 2001, pp. 157–168.

[4] M. Petracca, J.C.De Martin, G. Litovsky, M. Tacca, and A. Fumagalli, "Low-complexity Perceptual Packet Marking for Speech Transmission over Tiny Mote Device ," in *Proc. of IEEE ICME*, June/July 2009, pp. 806–809.

[5] http://www.cmusphinx.sourceforge.net.

[6] http://www.ldc.upenn.edu.

[7] "Wireless LAN Medium Access Control (MAC) and physical layer (PHY) specification," 1997.

[8] ETSI standard document ETSI ES 201 108 v1.1.2 (2000-04), 2000.

[9] CISCO, *Voice Over IP-Per Call Bandwidth Consumption*, (http://www.cisco.com/application/pdf/paws/7934/bwidth_consume.pdf).