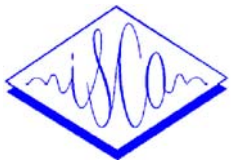

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Robust Speech Recognition over Packet Networks: An Overview

Kyu Jeong Han[†], *Naveen Srinivasamurthy*[‡], *Shrikanth Narayanan*[†]

[†]Department of Electrical Engineering – Systems
University of Southern California
Los Angeles, California 90089
kyuhan@usc.edu and shri@sipi.usc.edu

[‡]Qualcomm Incorporated
Standards Engineering
San Diego, California 92121
naveens@qualcomm.com

Abstract

Conventional circuit-switched networks are increasingly being replaced by packet-based networks for voice communication applications. Additionally, there has been an increased deployment of services supporting speech based interactions. These trends demand reliable transmission of speech data not just for playback but also to ensure acceptable automatic speech recognition (ASR) performance. In this paper, we present an overview of techniques that have been investigated to improve ASR performance against two major degradation factors in the context of packet networks: (1) information loss due to a low bit-rate codec and (2) packet loss due to channel (network) conditions. In addition, we highlight another key issue, packet loss rate, by showing ASR performance as a function of packet size and channel condition.

1. Introduction

As customer demand for convenient access to a variety of services in network-based applications increases, provisioning dependable ASR is becoming more important for service providers such as telephone/cellular phone or internet-related companies. A general ASR system that is currently being deployed in network-based applications uses a client/server architecture. Typically, the client is a low complexity front-end terminal which acquires speech data. The speech data are encoded and transmitted to a server over a lossy (packet) network. The server hosts a high complexity speech recognizer which uses the encoded speech data for recognition. This architecture not only reduces computational burden at the low complexity portable device, but also makes it more straightforward to update the recognizer and its application resources. In this context, reliable transmission of speech data (through wired and/or wireless networks) to ensure *minimal degradation in ASR performance* is strongly desired.

There have been two main network-based client/server ASR systems as shown in Fig. 1. The first is Network Speech Recognition (NSR), where the speech signal is encoded, transmitted and decoded and the speech features for ASR are extracted from the decoded speech at the server. The other is Distributed Speech Recognition (DSR) where speech features for ASR are extracted, encoded and transmitted by the client. The decoded speech features are directly used at the server for recognition. NSR has the advantage that the client does not have to be changed, i.e., we can use the same front-end terminal such as those used in Global System for Mobile Communications (GSM) or Voice over Internet Protocol (VoIP), for transmission of speech data. However, DSR is known to provide superior ASR performance [1] since the coding can be explicitly optimized to maximize speech recognition performance and not playback.

During the last decade, there have been a number of efforts that have focused on reliable transmission of speech data for network-based ASR. Lately, the research has increasingly focused on improving ASR performance for speech transmitted over packet networks. This trend is natural since conventional circuit-switched networks are currently being replaced by packet-based networks. This paper summarizes previous research on robust ASR over packet networks. Specifically, two major degradation factors in ASR performance in the context of packet networks are considered: (1) source coding - information loss due to a low bit-rate codec and (2) channel errors - packet loss due to channel (network) conditions. Section 2 addresses the source coding problem and presents techniques that have been adopted to improve ASR performance by improved source coding. Section 3 addresses the channel coding and error recovery mechanisms. Different error protection and mitigation schemes are summarized and analyzed. In addition, another key issue is addressed by analyzing ASR performance as a function of packet size and channel condition, i.e., packet loss rate (PLR).

The structure of rest of this paper is as follows: Sections 2 and 3 summarize prior techniques to improve ASR performance over packet networks in terms of information loss and packet loss, respectively. Section 4, shows the dependence of ASR performance on packet size and PLR. Finally, conclusions are given in Section 5.

2. Source Coding Techniques

Transmission of speech data over networks for ASR applications requires encoding speech data, much similar to conventional speech communication, in order to meet bandwidth requirements. As mentioned in Section 1, NSR and DSR encode different speech sources: while speech signals get encoded in NSR, speech features used for ASR are encoded in DSR. Both methods, however, will result in some information loss, unavoidable due to the use of low bit-rate codecs. In this section, we review research on speech (source) codecs for each case and comment on possible future research directions.

Early research in NSR focused on studying the influence of speech codecs used for voice communication on

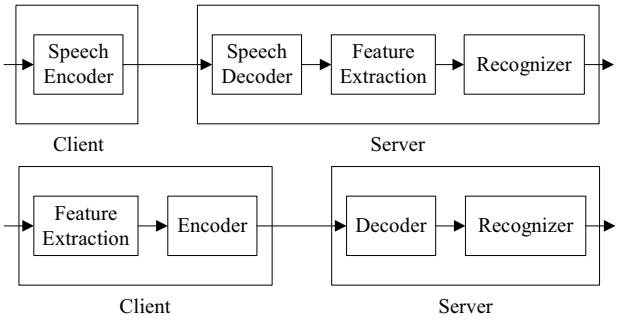

Fig.1. *Two main methods to transmit speech data for ASR through networks: Network Speech Recognition (top figure) and Distributed Speech Recognition (bottom figure).*

ASR performance [2-3]. These studies found that, although degradation in ASR performance was reduced under 'matched conditions' (when decoded speech signals were used for both training and testing a speech recognizer), there still remains a considerable performance gap to be overcome to support their use in practice. This performance degradation, which is NSR's most critical disadvantage, was attributed primarily to encoding that was perceptually-optimized for speech playback and not for machine speech recognition. It is important to note that speech processing required to satisfy perceptual criteria in speech codecs are distinct from those required to ensure maximal discriminability in speech pattern classification. Nevertheless, a number of efforts for NSR have been pursued primarily on the grounds that NSR has the advantage that the encoding at a client does not have to be modified for performing recognition at a server. Among these recent efforts in developing coding techniques for NSR, the most attractive one is the proposal of feature extraction from transmitted codec parameters (bit-stream) and not from decoded speech signals [4-7]. This has been shown to result in better ASR performance than the case when speech features were extracted from decoded speech signals, and has closely approached ASR performance achieved by DSR systems. (As mentioned in Section 1, DSR is known to be superior in ASR performance to NSR.) The reason for the good ASR performance of this method is that it extracts speech features directly from encoded spectral information and thus, avoids spectral degradation introduced during speech reconstruction. Further research should attempt an integrated analysis of the characteristics of codec parameters in conjunction with efficient feature extraction/selection methods used for recognition.

The ASR performance degradation in NSR naturally led to a new research direction that focused on transmitting speech features instead of speech signals. This method, referred to as DSR, can be viewed as an approach to enable better network-based ASR performance by reducing the information loss which arises due to inefficient representation of speech features in source coding. The superior ASR performance achievable by DSR has motivated a research trend in developing DSR encoders which are optimized for recognition, not for speech playback as is the case in conventional speech coding [8-9]. This is based on the idea that speech codecs optimized for perceptual quality are not suitable for recognition and the assumption that there exist (different) optimal codecs for recognition. The results obtained thus far are promising and have resulted in *Aurora,* a DSR system standardized by the European Telecommunications Standards Institute (ETSI) [10]. Despite these excellent research efforts, to ensure practical acceptance, the following critical disadvantages of DSR must still be overcome: (1) service providers have to release a new front-end terminal, which obviously incurs additional cost and (2) speech reconstructed using speech features that have been transmitted for recognition are not yet acceptable for playback [11,21]. Continuing progress however signals promise in this regards [23].

Future work should focus on development of source/channel coding techniques that not only are optimized for efficient data delivery but that are targeted toward the final application, i.e., recognition. An interesting question is whether the fact that the final target is a computer algorithm (pattern recognizer for speech recognition), not a human, can result in different source/channel coding algorithms enabling better ASR performance.

3. Packet Loss Recovery Techniques

Packet loss and delay jitter are both caused by network congestion and are two major issues in network-based applications. Fortunately, for network-based ASR delay jitter introduces limited problems because a recognizer can wait for the delayed data to arrive [12]. Retransmission, which is used in Transmission Control Protocol (TCP), is the best method to recover lost packets. However, it is not suitable for real time applications. Hence, User Datagram Protocol (UDP) which does not provide recovery services is generally used for transmission of speech data. For these reasons, packet loss in UDP is considered in this section as the only major degradation factor resulting in ASR performance due to undesirable network conditions.

There have been several research efforts on packet loss recovery techniques in order to improve ASR performance over packet networks, mainly based on methods outlined in [13]. The authors mention five schemes that can be applied for robust ASR over packet networks: forward error correction (FEC), insertion, interpolation, regeneration and interleaving. FEC is the error protection scheme, chosen by the ETSI DSR standard, *Aurora*, which is implemented by adding redundant bits to a packet. More gains in ASR performance can be obtained by applying unequal error protection schemes to FEC [14].

Insertion, interpolation and regeneration are basic schemes to replace lost packets according to different prespecified rules. In cases when packet size is small, reasonable improvements in ASR performance have been obtained [15-16]. Interleaving is an efficient scheme to enhance ASR performance from the effects of burst packet losses by shuffling several frames or packets. This scheme when combined with repetition or interpolation has been shown to provide additional gains in ASR performance [15].

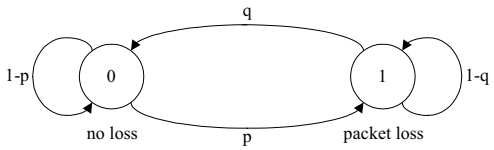

Fig. 2. *A 2-state Markov model (Gilbert model) to model network packet loss*.

Table 1. *Channel conditions according to p,q and PLR. PLR is calculated as PLR=p/(p+q)*

Condition	1	2	3	4
PLR	0.006	0.090	0.286	0.385
p	0.005	0.066	0.200	0.250
q	0.853	0.670	0.500	0.400

Unlike the aforementioned schemes, multiple description coding (MDC) [22] makes use of network diversity to enable robust ASR over packet networks [17]. Here, multiple descriptions are created for the same source data. Each individual description carries limited information about the source data, but, the combination of all the descriptions yields near optimal representation of the original source data. The MDC scheme relies on different descriptions being transmitted over different routes. By assuming that packet losses on different routes are independent, it is highly unlikely that all descriptions for a given source data are simultaneously lost. This ensures that for every given instant the server receives at least some information about the speech utterances acquired at the client. It can be expected that the MDC scheme might be an effective technique to protect against burst packet losses. A networked ASR system employing MDC scheme has shown around 15% enhancement in ASR performance compared to the case of using the G.729 standard at a PLR of 10%, with 90% of the burst packet losses being less than or equal to 11 frames [17].

Another trend has been in the use of soft decision at the decoding step instead of hard decision [18,19]. The soft decision algorithms incorporate information about the reliability of received data, provided by channel coding schemes, directly into the speech recognizer decoding step to improve ASR performance. In general, soft decoding is more relevant to mitigate bit errors and not packet errors. However, soft decision decoding has also been applied to mitigate performance degradation due to packet losses [18].

The general ASR performance over packet networks has been investigated by [12] and [20] for NSR and DSR, respectively. Similar results were obtained for both cases, i.e., ASR performance over packet networks is very sensitive to burst packet losses while it is tolerant to independent or random packet losses. In [12], the additional degradation caused by burst packet losses, when compared to random packet losses, was tested by fixing the PLR and gradually increasing the probability of burst packet loss. In [20], the mean length of burst packet losses was set to 4. Unfortunately, no research work so far has shown acceptable improvements in ASR performance against burst packet losses. Hence future work should concentrate on improving ASR performance against burst packet losses. In order to achieve better performance, more flexibility is required both at the encoding and at the packetization stages. Flexibility at the packetization stage can be achieved by changing the number of frames in a packet. In the next section we investigate ASR performance for different packet sizes.

4. ASR Performance and Packet Size

As mentioned in Section 3, the research investigations that have been done so far in order to alleviate the effect of packet losses in ASR performance have mostly handled short packets (Typically each packet is assumed to have less than five frames. The reason for choosing small number of frames is to minimize delay introduced due to packetization). Moreover, they have not shown reasonable results in overcoming burst packet losses. To enable more flexibility in the system design we explore the relationship in ASR performance between packet size and packet loss (PLR) through simulation experiments.

For the experiments, we consider connected digit recognition as the recognition task. The TIDIGITS database consisting of 8440 files from 55 male and 55 female adult speakers for training, and 4004 files from 52 male and 52 female adult speakers for testing, was used. The HMM for each digit contained 5 states and the observation in each state used 4 mixture Gaussian density functions. For modeling different packet loss environments (conditions 1 through 4 in Table 1), we used the Gilbert model, a 2-state Markov model shown in Fig. 2. p is the probability that the next packet is lost given the current packet is not lost and q is the probability that the next packet is not lost given that the current packet is lost. The parameter q controls the amount of burst packet losses [14]. Conditions 1 and 2 exhibit predominantly solitary losses and fairly insignificant number of burst packet losses. Approximately, 90% of the burst packet losses at conditions 1 and 2 consisted of three packets or less while, at conditions 3 and 4, 90% of the burst packet losses consisted of five packets or less [20]. We assumed that there was no data loss at the speech codec. The baseline ASR performance was 96.96% which is the case for no packet losses.

Table 2 shows ASR performance for 4 different packet sizes, where the number of frames per packet was 5, 10, 15 and 20, respectively. The experiments were conducted for 4 different channel conditions, i.e., percentage PLR of 0.6, 9, 28.6 and 38.5, respectively. First, we can confirm from this table that ASR performance is very sensitive to burst packet losses, by noticing that recognition performance becomes worse as the burst characteristics of channels increases, i.e., recognition performance in the last row in Table 1 is the lowest in each column. This is consistent with results obtained in previous research studies.

As mentioned in Section 3, there have been no acceptable approaches to overcome the effects of burst packet losses on ASR performance. Since packet loss distribution in real network environments mostly exhibits burst characteristics, the solution to ASR performance's sensitivity for burst packet losses is critically desired. Second, we can see by focusing our attention to the change in recognition performance from packet size 5 to 20 that the recognition

Table 2. *Recognition performance (%) for different frame sizes and channel conditions. The best recognition performance for each channel condition is indicated in bold.*

Channel Conditions	Packet Size (number of frames)			
	5	10	15	20
1	**96.94**	96.61	96.88	96.90
2	**94.78**	94.35	94.72	93.59
3	82.11	84.97	**85.91**	**85.91**
4	75.83	**80.35**	78.17	78.20

performance in each channel condition does not degrade proportionally as the size of a packet increases. The best recognition performance for different channel conditions is achieved for different packet sizes (indicated in bold in Table 2). This is not consistent with the fact that (perceptual) audio quality is known to degrade as packet size increases, which means that the optimal packet size for ASR over packet networks is dependent on the channel condition. From these results, based on a TIDIGITS connected digit recognition task in this paper, we can conclude that choice of packet size for ASR applications should consider channel conditions for reliable ASR performance over packet networks.

5. Conclusions

In this paper, we summarized and analyzed techniques that have been proposed to improve ASR performance when speech is sent over packet networks. The specific focus has been on two major degradation factors, namely data loss due to a codec and packet loss due to channel conditions. In addition, we addressed the role of packet size for reliable transmission of speech data over packet networks. We studied ASR performance as a function of packet size and packet loss rate and showed that the packet size that achieves minimal ASR error rate depends on the channel conditions. There are a number of open questions that remain to be tackled, especially in terms of dealing with burst packet losses, for achieving robust ASR that would enable reliable practical use of voice-based services over packet networks.

6. References

[1] H. Kelleher, D. Pearce, D. Ealey and L. Mauuary, "Speech recognition performance comparison between DSR and AMR transmission Speech," in *Proc. ICSLP 2002*, Denver, CO, U.S.A., pp. 1873-1876, Sept. 2002.

[2] S. Euler and J. Zinke, "The influence of speech coding algorithms on automatic speech recognition," in *Proc. ICASSP 1994*, Adelaide, Australia, pp. 621-624, April 1994.

[3] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. ICSLP 1996*, Philadelphia, PA, U.S.A., pp. 2344-2347, Oct. 1996.

[4] S. H. Choi, H. K. Kim, H. S. Lee and R. M. Gray, "Speech recognition method using quantised LSP parameters in CELP-type coders," *IEE Electronic Letters*, vol. 34, no. 2, pp. 156-157, Jan. 1998.

[5] H. K. Kim and R. Cox, "Bitstream-based feature extraction for wireless speech recognition," in *Proc. ICASSP 2000*, Istanbul, Turkey, pp. 1607-1610, June 2000.

[6] C. Palaez-Moreno, A. Gallardo-Antolin and F. Diaz-de-Maria, "Recognizing voice over IP: A robust front-end for speech recognition on the World Wide Web," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 209-218, Jun 2001.

[7] B. Raj, J. Migdal and R. Singh, "Distributed speech recognition with codec parameters," in *Proc. ASRU 2001*, Trento, Italy, pp. 127-130, Dec. 2001.

[8] V. V. Digalakis, L. G. Neumeyer and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the World Wide Web," *IEEE Journal on Select. Areas Comm.*, vol. 17, no. 1, pp. 82-90, Jan. 1999.

[9] N. Srinivasamurthy, A. Ortega and S. Narayanan, "Efficient scalable encoding for distributed speech recognition," submitted to *IEEE Trans. Speech and Audio Processing*, 2004.

[10] ETSI Standard, *ES 201 108 V1.1.2 Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms*, April 2000.

[11] D. Chazan, R. Hoory, G. Cohen and M. Zibulski, "Speech reconstruction from mel-frequency cepstral coefficients and pitch frequency," in *Proc. ICASSP 2000*, Istanbul, Turkey, pp. 1607-1610, June 2000.

[12] J. V. Sciver, J. Z. Ma, F. Vanpoucke and H. V. Hamme, "Investigation of speech recognition over IP channels," in Proc. *ICASSP 2002*, Orlando, FL, U.S.A., pp. 3812-3815, May 2002.

[13] C. Perkins, O. Hodson and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40-48, Sept./Oct. 1998.

[14] C. Boulis, M. Ostendorf, E. A. Riskin and S. Otterson, "Graceful degradation of speech recognition performance over packet-erasure networks," IEEE Trans. Speech and Audio Processing, vol. 10, no. 8, pp. 580-590, Nov. 2002.

[15] P. Mayorga, R. Lamy and L. Besacier, "Recovering of packet loss for distributed speech recognition," in Proc. Eusipco 2002, Toulouse, France, Sept. 2002

[16] B. Milner and S. Semnani, "Robust speech recognition over IP networks," in Proc. *ICASSP 2000*, Istanbul, Turkey, pp. 1791-1794, June 2000.

[17] X. Zhong, J. Arrowood, A. Moreno and M. Clements, "Multiple description coding for recognizing voice over IP," in Proc. *Digital Signal Processing Workshop 2002*, Callaway Gardens, GA, U.S.A., pp. 383-386, Oct. 2002.

[18] A. Bernard and A. Alwan, "Low-bitrate distributed speech recognition for packet-based and wireless communication," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, pp. 570-579, Nov. 2002.

[19] A. Potamianos and V. Weerackody, "Soft-feature decoding for speech recognition over wireless channels," in Proc. *ICASSP 2001*, Salt Lake City, UT, U.S.A., pp. 269-272, May 2001.

[20] D. Quercia, L. Docio-Fernandez, C. Garcia-Mateo, L. Farinetti and J. C. De Martin, "Performance analysis of distributed speech recognition over IP networks on the AURORA database," in Proc. *ICASSP 2002*, Orlando, FL, U.S.A., pp. 3820-3823, May 2002.

[21] T. Ramabadran, J. Meunier, M. Jasiuk, and B. Kushner, .Enhancing distributed speech recognition with back-end speech reconstruction,. in *Eurospeech 2001*, (Aalborg, Denmark), September 2001.

[22] Vivek K Goyal, "Multiple Description Coding: Compression Meets the Network," IEEE Signal Processing Magazine, vol. 18, no. 5, pp. 74-93, Sept. 2001.

[23] T. Ramabadran, A. Sorin, M. McLauglin, D. Chazan, D. Pearce and R. Hoory, "The ETSI Extended Distributed Speech Recognition (DSR) Standards: Server-Side Speech Reconstruction", Proc. ICASSP, (Montreal, Canada), May 2004