International Journal of

MODELLING & SIMULATION

VOLUME 21, Number 2, 2001

(Continued on back cover)

ISSN 0228-6203

Published by **ACTA PRESS** ANAHEIM • CALGARY • ZURICH

International Journal of Modelling and Simulation, Vol. 21, No. 2, 2001

SIMULATIVE ANALYSIS OF AN ADAPTIVE CONTROL MECHANISM FOR PACKETIZED VOICE ACROSS THE INTERNET

M. Roccetti,* V. Ghini,* and G. Pau*

Abstract

Proposed new control mechanisms for the transmission of packetized voice across the Internet utilize the presence of silent intervals in conversational speech in order to dynamically adapt the behaviour of the audio application to the network fluctuating traffic conditions so as to minimize the effect of packet loss and varying delays on the quality of audio delivered to the destinations. An accurate model of the on-off characteristics of the conversational speech is thus necessary to analyze the performance of those audio communication systems. In this paper, an eight-state Markov model of voice activity in conversational speech has been used in order to assess the adequacy of an Internet audio mechanism that dynamically sets the playout delay value of packet audio in Internet voice-based connections. Based on this model, several simulation experiments have been carried out that show that a sufficient number of silence periods (of sufficiently long duration) occur in a typical human conversation that permit an adequate application of the proposed audio mechanism. In addition, a number of simulative/experimental trials are reported that show that the proposed Internet audio mechanism strikes a favourable balance between the average playout delay and the packet loss percentage experienced during audio conversations over the Internet.

Key Words

Internet, multimedia systems and applications, packet voice, modelling and simulation of packet audio over the Internet

1. Introduction

The recent interest in multimedia conferencing is the result of the incorporation of cheap audio (and video) hardware in today's workstations, as well as the result of the development of a global infrastructure capable of supporting multimedia traffic — the Mbone/Internet. Unfortunately, Internet telephony is still often dismissed as an impractical application because of the poor quality experienced by many users of Internet audio tools. The problem with the today's Internet is that routers operate on a FIFO basis and statistically multiplex traffic from different sources. The impact of this behaviour on real-time traffic is to introduce delay variation (known as jitter) to the interpacket timing relationship and even high packet loss rates. In the absence of network support to provide guarantees of quality to users of Internet voice software, an interesting approach to cope with the problems caused by jitter and high packet loss rates is to use control mechanism [1]. Those mechanisms adapt the behaviour of the audio application to the network conditions so as to minimize the impact of packet loss and delay jitter. Jitter has to be removed from audio packet streams because it causes the speech to be unintelligible. Hence, in order to compensate for variable network delays, a smoothing playout buffer is used at the receiver. Received audio packets are first queued into the buffer and then the playout of each packet is delayed for some quantity of time beyond the reception of the first packet, as seen in Fig. 1. These control mechanisms are adaptive because jitter on the Internet may vary significantly with time. They adaptively adjust the playout delay in order to keep the buffering delay as small as possible, while minimizing the number of packets delayed past the point at which they are scheduled to be played out. Clearly, a critical trade-off exists between the amount

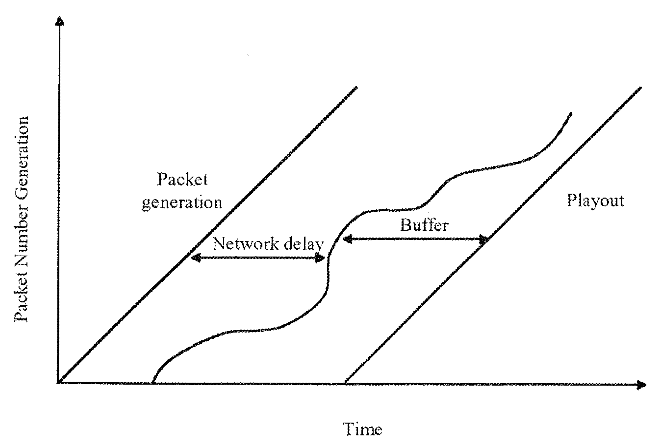

Figure 1. Smoothing out jitter delay at the receiver.

* Dipartimento di Scienze dell'Informazione, Università di Bologna, Via Mura A. Zamboni 7 - I-40127 Bologna, Italy; e-mail: roccetti@cs.unibo.it, ghini@cs.unibo.it, giovanni@csr.unibo.it

Recommended by Prof. Ramesh K. Karne
(paper no. 1999-046)

of delay that is introduced in the buffer and the percentage of late packets that are not received in time for playout: the longer the additional buffering delay, the more likely it is that a packet will arrive before its scheduled playout time. But if, on one side, a too large percentage of audio packet loss (over 5–10%) may impair the intelligibility of an audio transmission, on the other side, too large playout delays (e.g., more than 250-300 milliseconds) may disrupt the interactivity of an audio conversation [2, 3].

An audio segment may be considered as constituted of short bursts of energy (called talkspurts) separated by periods of silence (during which no audio packet is generated). NeVot [4], vat [5], rat [6], and FreePhone [7] are popular Internet audio tools that adopt a control mechanism for adaptively adjusting the playout delay during different talkspurts. However, extensive experiments have been carried out that have shown that, in some circumstances, the cited mechanisms may suffer from a number of problems, especially when they are deployed over wide-area networks. In particular, the following problems may be pointed out:

1. An external software-based mechanism (e.g., the IP/NTP protocol) is frequently used to synchronize the system clocks at both the sender and receiver of the audio communication, as low-frequency clock drift between the two hosts can cause receiver buffer overflow or underflow. Unfortunately, those synchronization mechanisms are not typically widespread over the Internet and, in addition, may be too inaccurate for coping with the real-time nature of the audio generation/playout process.
2. The packet transmission delays experienced over the Internet are assumed to follow a Gaussian distribution. This assumption, which is employed in all the cited tools to adjust the playout delay, seems to be a plausible conjecture only for those limited time intervals in which the overall load of the underlying network is quite light [2].
3. The computational complexity of existing playout delay control mechanisms is (on average) too large, and has a negative impact on the real-time audio playout process.

Thus, in order to adequately support real-time voice over packet-switched networks (such as the Internet), we have designed an adaptive mechanism for the control of the playout delay that ameliorates all the negative effects of the audio tools mentioned above, while maintaining satisfactory values in the trade-off between average playout delay and packet loss due to late arrivals [1]. The main characteristic of such an audio mechanism is that it dynamically adapts the playout delay of packet audio in Internet voice-based connections but relies on the existence of a sufficient amount of silent intervals (of sufficiently long duration) needed for performing the dynamical setting of the playout delay. Hence, because our scheme exploits the presence of silent intervals for adjusting the playout delay to fluctuating network conditions, an accurate modelling of the voice activity characteristics of human conversational speech is necessary to understand whether a sufficient number of (sufficiently long) silent intervals occur in human conversations.

In this paper, we report on the use of an eight-state Markov model for voice activity in conversational speech that we have exploited in order to assess the efficacy of our Internet audio mechanism. Based on this model, we have developed several simulation experiments that show the existence of a sufficient number of silent intervals during typical voice-based conversational speech. Those silent periods are both sufficiently long and frequent to adequately permit the application of our Internet audio mechanism. In particular, we present results of a simulative/experimental analysis that show that our Internet audio mechanism makes adequate use of those silent intervals for the dynamical adaptation of the playout delay, and strikes a favourable balance between the average playout delay and packet loss percentage. In essence, the simple, accurate, and comprehensive simulative results we have obtained regarding human voice activity in conversational speech have illustrated, prior to the development of a prototype implementation of the mechanism, the adequacy of the approach adopted in the design of our audio playout delay control scheme. Based on these results, it was then possible to implement and develop the complete mechanism without incurring additional costs due to late discovery of unexpected errors or inefficiency. The remainder of the paper is structured as follows. In Section 2 we briefly recall the main design features of our mechanism. Section 3 is devoted to presenting and discussing the adopted simulation model and the obtained simulation results. Section 4 provides some final remarks.

2. An Adaptive Internet Mechanism for Packet Voice

We have developed a novel playout delay control mechanism that is suitable for adjusting the talkspurt playout delays of unicast, voice-based audio communications across the Internet [1]. The mechanism was designed to dynamically adjust the talkspurt playout delays to the network traffic conditions without assuming either the existence of an external mechanism for maintaining an accurate clock synchronization between the sender and the receiver, or a specific distribution of the end-to-end transmission delays experienced by the audio packets during an audio communication.

Our technique for dynamically adjusting the talkspurt playout delay keeps the same playout delay constant throughout a given talkspurt but permits different playout delays in different talkspurts. It is based on obtaining, in periodic intervals (about 1 second), an estimation of the upper bound for the packet transmission delays experienced during an audio communication. Such an upper bound is periodically computed using round-trip time values obtained from packet exchanges of a three-way handshake protocol performed between the sender and the receiver of the audio communication. Prior to the beginning of the first talkspurt (and afterward in periodic intervals), the sender initiates a packet protocol exchange activity at the end of which the receiver is provided with the sender's

Table 1
State Transition Parameters for Modified Brady's Model [8]

$\underline{\alpha}'_{1,4}$	$\underline{\alpha}_{1,7}$	$\underline{\alpha}_{3,1}$	$\underline{\alpha}_{2,1}$	$\alpha_{7,1}$	$\underline{\alpha}_{4,1}$	$\underline{\alpha}_{5,1}$	$\underline{\alpha}_{1,2}\underline{\alpha}_{7,2}$
$\underline{\alpha}'_{6,5}$	$\underline{\alpha}_{6,8}$	$\underline{\alpha}_{2,6}$	$\underline{\alpha}_{3,6}$	$\alpha_{8,6}$	$\underline{\alpha}_{5,6}$	$\underline{\alpha}_{4,6}$	$\underline{\alpha}_{6,3}\underline{\alpha}_{8,3}$
0.83305	5.4890	2.1572	2.3245	27.62	2.2222*	1.0438*	0.27853
* During the first 200 ms after transition to States 4 or 5, $\underline{\alpha}_{4,1}$, $\underline{\alpha}_{5,6}$, $\underline{\alpha}_{5,1}$, and $\underline{\alpha}_{4,6}$ are set to 0.							

estimate of an upper bound for the transmission delay. This upped bound is then used to dynamically compute the talkspurt playout delay and to adjust the dimensions of the receiver's playout buffer. All the complex technical details concerning this three-way handshake protocol fall outside the scope of this paper and are not reported here for the sake of brevity. They may be found in [1]; instead it is more important to recall here that the proposed mechanism guarantees that the talkspurt playout delay may be dynamically set from one talkspurt to the next, provided that intervening silent periods of sufficiently long duration are exploited for effecting the adjustments.

In simple words, the computed variation of the playout delay is accommodated at the receiver with the introduction of artificially elongated or reduced silence periods of the human conversation. In particular, an improvement of the transmission delay experienced by audio packets (equal to (milliseconds) is accommodated by *installing* at the receiver an up-to-date playout delay value that causes an artificial contraction of the silent period chosen for the installation. Such a reduction of the silence period perceived by the receiver (w.r.t. the original duration of the corresponding silence period generated by the sender) amounts to exactly (milliseconds. Instead, a deterioration of the transmission delay experienced by audio packets (equal to (milliseconds) causes, at the receiver, a perception of a silent period whose original duration is elongated by δ milliseconds.

3. Simulation Model, Results, and Data Interpretation

The need for silent intervals to allow a playout delay control mechanism to adjust to the fluctuating network conditions renders the scheme proposed in [1] particularly relevant for voice-based applications with intervening silence periods between subsequent talkspurts. In order to assess the efficacy of such a mechanism, an accurate modelling of the talkspurt/silence characteristics of conversational speech is mandatory for understanding whether sufficient (and sufficiently long) silent intervals occur in typical human conversations that may permit the periodical activity of dynamically setting the playout delay.

To this aim, we adopted the modified eight-state Brady's model of conversational speech introduced in [8], which is able to describe the main on-off characteristics of human conversations. The motivations behind the choice of this particular model (represented in Fig. 2) are its accuracy and ease of implementation for carrying out simulations and analyses.

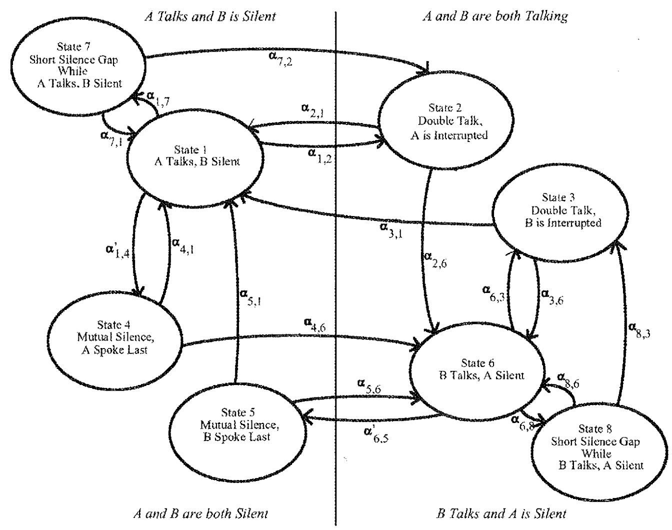

Figure 2. Modified 8-state Brady model [8].

Fig. 2 is divided into quadrants, with each quadrant representing a different state for parties A and B engaged in a conversation. In particular, such a model is able to reproduce all three of the following different types of silences occurring in human speech: (1) listening pauses, which occur when a party is silent and listening to the other party; (2) long speaking pauses, which occur between phrases or sentences while a party is speaking; (3) short speaking pauses, which occur between words or syllables while a party is speaking.

In Table 1, the values of all the state transition parameters of the Markovian Brady's model depicted in Fig. 3 are reported, as calculated in [8].

Based on this model, we have developed a set of simulative experiments in order to assess, respectively, the overall quantity, the frequency, and the duration of silence intervals within human conversational speech. The first question to be answered in order to understand whether the designed playout delay control mechanism could be profitably applied to the human voice was: "Is there a sufficient total amount of silent intervals in human conversational speech to permit an adequate application of the designed playout delay control mechanism?" We obtained a first positive response to this question. In fact, based on the use of the model, we observed that the percentage of silent intervals within a simulated packetized conversation amounts to about 63–66%, depending on the voice packetization interval that is typically chosen in the range of 10–30 milliseconds. This result is summarized in Fig. 3, where the total number of silent intervals (with relative duration) is shown, as obtained in a simulated one-hour-long two-party conversation. As seen from the figure, the smaller the packet size (i.e., 10 milliseconds), the larger the number of silent intervals (i.e., 5,075).

Then, since an important parameter that influences the efficacy of our control mechanism is the frequency of the intervening silence periods, we addressed the following question: "How frequent are those silent intervals within a human conversation?" Needless to say, the more frequent those silent intervals are, the more likely it is that our mechanism will be successful in dynamically adjusting the playout delay. Hence, we used the modified Brady's model to understand how large, on average, the talkspurts expected in a simulated packetized conversation are. From this simulative experiment, we observed the following important result: the smaller the chosen packet size, the more likely it is that our mechanism will have the possibility of dynamically setting the playout delay. In fact, the total number of silent periods increases and the average talkspurt length decreases (to about 244 millisecond) with small-sized packets. Instead, the larger the packet size, the larger the average duration of the talkspurt (i.e., 451 milliseconds). The main result concerning the quantity and the duration of the talkspurts is depicted in Fig. 4, where the percentage of talkspurts with length smaller than a fixed amount is reported.

The final question we asked prior to the development of our Internet audio mechanism was: "Given that, according to the policy adopted by our control mechanism, the duration of intervening silence periods is artificially reduced when an improvement of the audio packet transmission delays is experienced, how long on average are those silent intervals?" To this end, an important result obtained from the simulation experiments we carried out concerns the duration of the intervening silence periods in a human conversation. The average silence length of the silent periods obtained in a simulated two-party one-hour-long conversation was measured as ranging in the interval of 465–770 milliseconds, yet again depending on the packetization interval. In particular, the larger the packet size (i.e., 30 milliseconds), the larger the average silence duration (i.e., 770 milliseconds). The main results regarding the silence interval duration are summarized in Fig. 5 and Fig. 6, where, respectively, the probability distribution of

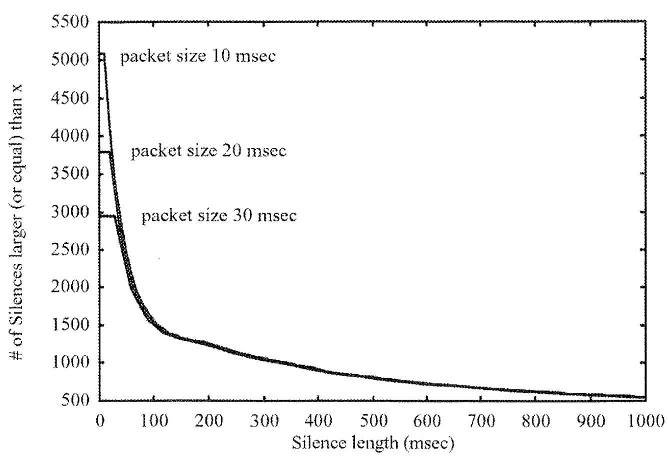

Figure 3. Total amount of silent periods.

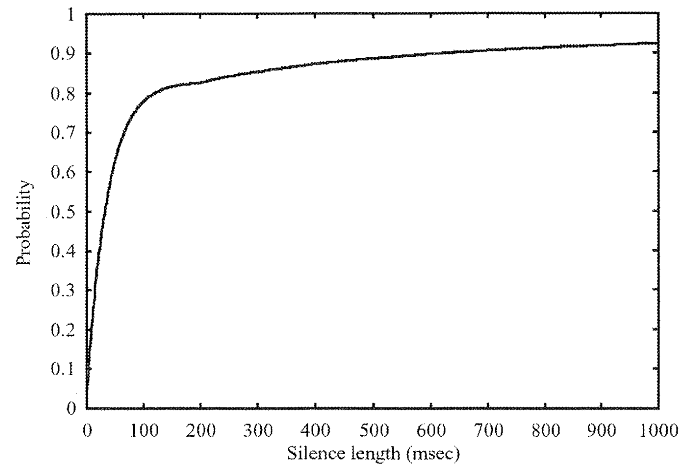

Figure 5. Silence duration.

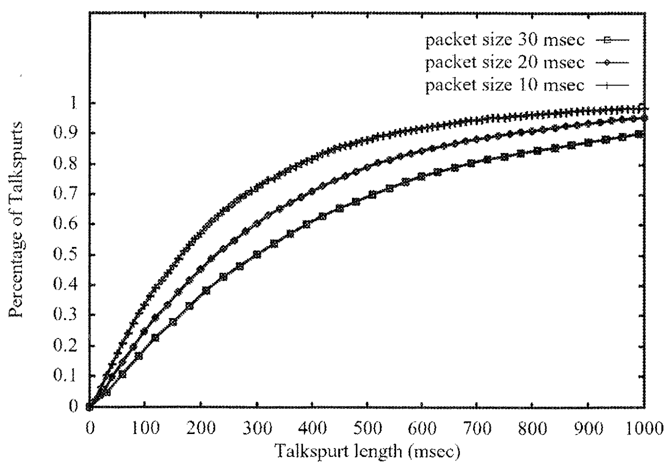

Figure 4. Percentage of talkspurts with duration smaller than a given value x.

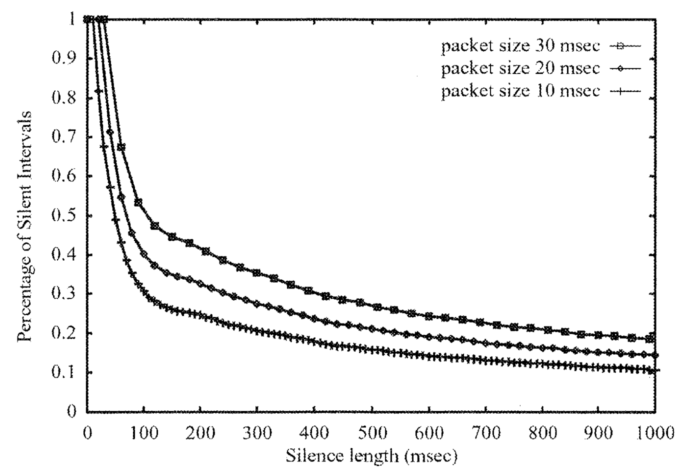

Figure 6. Percentage of silent intervals with duration larger than a given value x.

Table 2
Experimental Results: Average Playout Delay and Packet Loss

Experiment	Start Time	Playout Delay	Packet Loss
#1	08:20am 10 /15 / 97 (Th)	188 msec	7%
#2	02:00pm 10 / 9 / 97 (Fr)	238 msec	5%
#3	04:30pm 10 / 4 / 97 (Su)	202 msec	6%
#4	07:40pm 9 / 21 / 97 (Mo)	229 msec	5%
#5	04:15pm 9 / 16 / 97 (We)	207 msec	5%

the silence interval duration and the percentage of silent periods of length larger than a fixed amount are reported.

Based on the positive results provided by the adopted simulation model, we proceeded to the development of a prototype implementation of our Internet audio mechanism. Thus, a working prototype implementation of the playout control mechanism designed in [1] was performed. The UNIX socket interface and the datagram-based UDP protocol were used to transmit and receive the sampled audio packets. The coding scheme that was used to produce the audio packets used 8-kHz sampled speech with bit rates of 8-kbit/sec [9]. Based on the observed simulation results, we decided to use a 30-milliseconds packet size in order to permit an adequate reduction of the silent periods in the case of improvement of network traffic conditions.

Using such a prototype implementation, an extensive experimentation was carried out aiming at assessing the performance of the audio mechanism. Several audio delay traces (about 30) were obtained by transmitting over an Internet connection (from Cesena, Italy, to Geneva, Switzerland) about 15,000 audio packets generated from prerecorded 10-minute-long audio files. In Table 2 the values of the average playout delay and the packet loss percentage are reported for only 5 of the 30 experiments carried out. It is worth mentioning that, besides the results provided in Table 2, in all the other 25 cited experiments both an acceptable value of the average playout delay (ranging in the interval 180–250 milliseconds) and a tolerable loss percentage of up to 6–7% were experienced.

Finally, we developed a software simulator, using the C programming language, that is able to read in the transmission delay of each packet from a given real audio trace, detects if it has arrived before the playout time that is computed by our control mechanism, and executes the algorithm. The simulator is also able to calculate the average playout delay and the packet loss for each given trace. Thus, we run the simulator on each obtained audio delay trace using the receiver buffer size as the control parameter to be varied to achieve different loss percentages [10]. Running the simulator over all the 30 experimental traces of our experiments and then averaging the results, we obtained the plot of Fig. 7, where the playout delay is represented as a function of the loss percentage. The effect that various delays and loss rates have on the quality of the perceived audio is intuitively represented by means of three different quality ranges: good, potentially useful, and

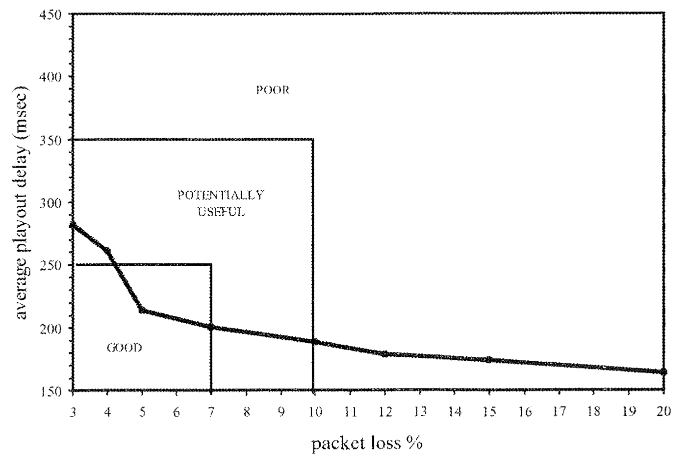

Figure 7. Performance of playout delay control mechanism.

poor [3]. As seen from the figure, and based on the consideration that audio of acceptable quality may be obtained only if lower delays are achieved while the loss percentage does not exceed the value of 10%, we deduce that our algorithm shows very good performance.

4. Conclusion

We have reported on the use of an eight-state Markov model for voice activity in conversational speech that has been exploited in order to assess the performance of an adaptive playout delay control mechanism recently proposed [1, 11]. One of the most critical characteristics of this audio mechanism is its reliance on the existence of silent intervals (of sufficiently long duration) for performing the dynamical setting of the playout delay. Based on the use of the model, we have conducted several simulation experiments that show that a sufficient number of silent periods occur during a human conversation (about 66%). In addition, those silent intervals turn out to be both sufficiently long and frequent so as to permit an adequate application of the proposed audio mechanism. Further simulation experiments jointly conducted with experimental measurements show that the proposed audio mechanism strikes a favourable balance between average playout delay and packet loss percentage experienced during audio conversations over the Internet. We conclude by pointing out that the simple, accurate, and comprehensive simulative results we have obtained regarding human voice activity in conversational speech, even prior to the development of a prototype implementation of our Internet audio mechanism, have illustrated in advance the adequacy of the approach adopted in the design of the audio control mechanism.

Acknowledgements

This research was partially funded by Italian MURST and Italian CNR under the Grant No. 98.00387.CT12. The first author is also indebted to Microsoft Research Europe (Cambridge, UK) for supporting his work under a grant.

References

[1] M. Roccetti, V. Ghini, G. Pau, P. Salomoni, & M.E. Bonfigli, Design and experimental evaluation of an adaptive playout delay control mechanism for packetized audio for use over the Internet, *Multimedia Tools and Application.* To appear.

[2] J. Bolot, H. Crepin, & A. Vega Garcia, Analysis of audio packet loss on the Internet, *Proc. Network and Operating System Support for Digital Audio and Video,* Durham, NC, 1995, 163–174.

[3] T.J. Kostas, M.S. Borella, I. Sidhu, G.M. Schuster, J. Grabiec, & J. Mahler, Real-time voice over packet-switched networks, *IEEE Network, 12,* 1998, 18–27.

[4] H. Schulzrinne, Voice communication across the Internet: A network voice terminal, *Tech. Rep., Dept. of ECE and CS University of Massachusetts,* Amherst, MA, 1992.

[5] V. Jacobson & S. McCanne, vat, ftp://ftp.ee.lbl.gov/conferencing/vat/.

[6] V. Hardman, M.A. Sasse, & I. Kouvelas, Successful multiparty audio communication over the Internet, *Comm. of the ACM, 41*(5), 1998, 74–80.

[7] J. Bolot & A. Vega Garcia, Control mechanism for packet audio in the Internet, *Proc. IEEE INFOCOM'96,* San Francisco, CA, 1996, 232–239.

[8] H.P. Stern, S.A. Mahmoud, & K.K Wong, A comprehensive model for voice activity in conversational speech: Development and application to performance analysis of new-generation wireless communication systems, *Wireless Networks, 2*(4), 1996, 359–367.

[9] ITU-T Recommendation G.729, Coding of Speech at 8 kb/s using Conjugate Structure Algebraic Code Excited Linear Prediction, 1996.

[10] S.B. Moon, J. Kurose, & D. Towsley, Packet audio playout delay adjustment: Performance bounds and algorithms, *ACM Multimedia Systems, 6*(1), 1998, 17–28.

[11] M. Roccetti, V. Ghini, & G. Pau, Modelling and simulation of an adaptive playout delay control mechanism for packet audio across the Internet, *Proc. 1999 IASTED Int. Conf. on Modelling and Simulation,* Philadelphia, PA, 1999, 153–157.

Biographies

Marco Roccetti holds an Italian Laurea degree in electronic engineering from the University of Bologna, Italy. He is currently an Associate Professor of Computer Science at the Department of Computer Science, University of Bologna. His research interests include design, implementation, and evaluation of distributed multimedia systems.

Vittorio Ghini has an Italian Laurea degree in computer science from the University of Bologna, Italy. He is a Ph.D. candidate at the Department of Computer Science of that university. His research interests are in the field of QoS management mechanisms for IP networks.

Giovanni Pau has an Italian Laurea degree in computer science from the University of Bologna, Italy. Currently, he is a Ph.D. student at the Department of Computer Engineering of that university. His research interests lie in the area of network communication protocols.