Efficient Overlay Audio Conferencing

Norbert Egi, Nick Blundell, and Laurent Mathy

Computing Department, Lancaster University, Lancaster, LA1 4WA, UK
{egi, n.blundell, laurent}@comp.lancs.ac.uk

Abstract. In this paper, we present a thorough and realistic analysis of audio conferencing over application-level multicast (ALM).

Through flexibility and ease-of-deployment, ALM is a compelling alternative group-communication technique to IP Multicast — which has yet to see wide-scale deployment in the Internet. However, proposed ALM techniques suffer from inherent latency inefficiencies, which we show, through realistic simulation and exploration of perceived quality in multi-party conversation, to be greatly problematic for the realisation of truly-scalable audio-conferencing systems over ALM.

In this work, we propose to adapt dynamically the application-level distribution structure to the conversational pattern of the audio conference. The contribution of this paper is threefold: we develop a novel perceptual quality model for multi-party audio conversations; we provide dynamic adaptation via a simple next-speaker prediction technique and we validate the proposed approach by using a large and detailed corpus of real multi-party conversations.

Key words: ALM, VoIP, Conversation Analysis, Adaptive Conferencing

1 Introduction

It is well known that the mouth-to-ear latency of an echo-less voice-communication channel should not exceed 300 ms in order to allow natural conversation [2] — audible echo can reduce this threshold by two orders of magnitude.

This limit is especially important for Internet VoIP applications, where the communication channel comprises non-trivial application and network latency components. With typical one-way application latencies of 60–400 ms [6] and Internet round-trip latencies of 150–200 ms [4], such VoIP applications operate with communication-channel latencies that are at or above the upper threshold of human tolerance.

In particular, this poses a problem for application-level group communication techniques, which are inherently less latency- or cost-efficient than their scarcely-deployed network-level counterpart (*i.e.* IP multicast), for example: multiple unicasting between participants cannot scale to support even modestly sized groups; standard overlay-tree flooding (*i.e.* that is performed by proposed application-level multicast (ALM) techniques) results in highly-varied node-pair latencies; and centralised reflector servers do not accommodate well groups of widely distributed membership (*i.e.* since there is no obvious place to put the reflector). A specific solution is therefore required to support group audio applications.

In [1] we proposed ALNAC, a dynamic application-level multicast (ALM) routing protocol especially designed for audio-conferencing applications, and we argued that perceptual quality of multi-party conversation could be improved by exploiting the patterns in natural conversation that allow for prediction — with a high accuracy — of who will speak next in conversation.

In this paper, we develop the preliminary work in [1] into a thorough investigation of the problem and make the following contributions. In Section 2, we give an in-depth exploration of the specific effects of communication-channel latency on multi-party conversation, leading to a novel model of perceptual quality. In Section 4, we propose, and conduct a thorough analysis of, several approaches to next-speaker prediction algorithms, using a large corpus of highly-detailed talkspurt data from actual multi-party conversation. Finally, in Section 5, we evaluate, by simulation, our ALM-based audio-conferencing proposal under conditions of realistic network latency and through using a realistic model of multi-party conversation.

2   Issues of Latency in Multi-Party Conversation

In interactive scenarios, latency is a problem usually because it *cannot* be perceived: in fact, only when a source's sound is reflected (echoed) back can latency be gauged; otherwise, the listener's brain interprets what is heard or what is not heard as events that happen in real-time, for example: we would quite-happily perceive a live radio show as such despite that it may in fact have a two-minute censorship delay.

Thus, when engaged in conversation we are constantly (subconsciously) projecting times at which *responses* to our spoken *cues* should arrive; if they do not arrive within our expected time range (a range bounded by well-studied latency-tolerance threshold), we perceive that they will never arrive and repeat cues unnecessarily in an attempt to repair the conversation.

With these two considerations in mind, we can argue, therefore, that perceived quality is not simply dependent upon a communication channel's latency but upon the delay with which specific responses are heard after their cues. This observation is particularly relevant to multi-party conversation, since not only will a participant hear responses to their own cues — if they choose to speak — but they will also hear responses to the cues of other participants, and so perceptual quality of those responses will be dependent not upon the *absolute latencies* with which they are heard but upon the *difference* in absolute latencies of those responses and their cues.

We remark that not all speech acts cue a response (for example at the end of the discussion on a topic). On the other hand, one should keep in mind that conversational turns (i.e. change of speaker) are typically delimited by silence gaps of no more than 1 second [11].

2.1 Issues of Stream Synchronisation

Due to the inherent latency inefficiency of ALM techniques, there is a potential that participants of a multi-party communication system will observe highly-varied network delays between streams, which, by affecting the synchronisation of responses and their cues, will impact upon perceived quality.

To the best of our knowledge, there has been no study on the effects of stream desynchronisation on the perceived quality of multi-party conversation (*i.e.* from a listener's perspective). We therefore performed a simple listening experiment in which we desynchronised a participant channel of a recorded meeting from the ICSI meeting corpus [5]: in the experiment, we took one recorded audio channel (a channel of a participant who was engaged in conversation for a large proportion of the particular meeting segment) and shifted it by various time constants, before mixing all of the separate channels into a single audio file; a set of mixes were thus created (including the original mix without shifting). Volunteers, who had no insight into the particular transformations that was performed, were asked to categorise between those mixes that sounded "strange" and those that sounded normal.

Interestingly, none of the listeners in the experiment could perceive a difference between mixes that were desynchronised by less than or equal to 1,000 ms, which indicates that we have a higher tolerance to the lateness of responses when we listen to a conversation than when we are actively engaged in it (in which case the maximum mouth-to-ear round-trip tolerance is about 600 ms).

2.2 Quality Model for Multi-Party Conversation

In line with our observations on quality-perception in multi-party conversation, we propose a perceptual-quality model that is not based on channel latency, as has so far been considered in the literature, but rather on the 'lateness' of individual spoken responses with respect to their cues.

In [2], the authors proposed a simple utility function for describing the perceived quality of mouth-to-ear channel latency in which two score-levels are defined: a high score level to reflect 'very good' latency perception, and a low score level to reflect 'bad' latency perception. We base our own utility function on a similar principle but instead consider tolerance to round-trip mouth-to-ear latency, since conversation — as a two-way process — is affected by round-trip latency and since asymmetric latencies are highly-likely in ALM. In addition to the original utility function, we extend the function to distinguish between the tolerance thresholds to response lateness for a participant's own cues and for the cues of other participants.

The resulting utility function for perceptual quality is depicted in Figure 1(a). Note that, in the model, since a non-cued talkspurt cannot be perceived as being late, it is automatically awarded a score of 1.
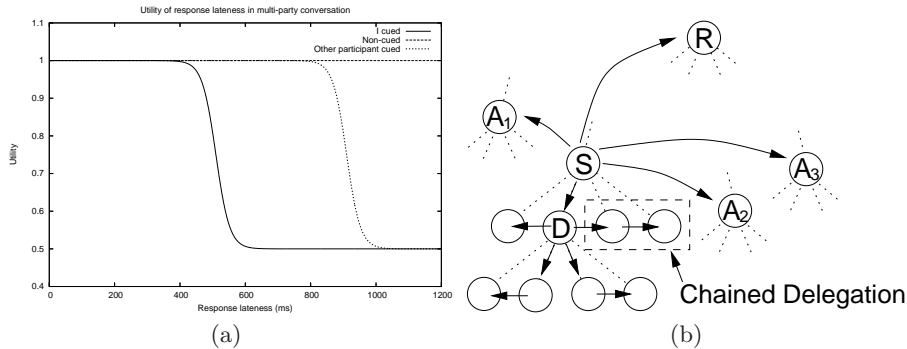
Fig. 1. (a) Utility function for the perceptual quality of response lateness in multi-party conversation; (b) Dynamic routing of ALNAC through the process of delegation.

3 ALNAC: A Dynamic Overlay Routing Protocol

ALNAC (Application-Level Network Audio-Conferencing routing protocol) [1] is a light-weight ALM routing protocol, designed especially to optimise audio-packet delivery for those audio-conference participants who are most sensitive to communication-channel latency (*i.e.* those who are currently engaged in conversation), whilst minimising the impact of such optimisation on members that are least sensitive to communication-channel latency (*i.e.* those members that take only a listening role in the current conversation).

More precisely, ALNAC operates over an ALM tree structure. ALNAC adapts a basic flooding technique whereby a speaker sends audio samples to the tree root and to its children (for forwarding in their respective sub-trees). The adaptation is that a speaker will send audio samples, in addition to the root, directly to a set of predicted next speakers who are identified as highly likely active participants in the current conversation by a prediction algorithm. On the other hand, because the out-degree of a node (i.e. the maximum number of forwarding the node will do) can be limited due to bandwidth constraints, some of the speaker's children on the ALM tree may have to be *deprived* from receiving the audio samples from the speaker directly. To ensure that all samples are eventually flooded to all nodes in the tree, a speaker will *delegate* the responsibility for supplying the deprived nodes among the nodes to whom it *is* sending directly. Note however, that delegation can be recursive (i.e. a supplier can further delegate). Figure 1(b) illustrates the audio sample distribution process in ALNAC.

We therefore see that, in essence, ALNAC builds a dynamic overlay over an ALM tree (as opposed to adapting the tree to conversation changes).

Note that in [1], a very primitive, static prediction algorithm was proposed. A more efficient and adaptive algorithm is described in the following section.

4 Next-Speaker Prediction

From Section 3, it should be clear that the ability of ALNAC to identify the participants that are actively engaged in conversation is critical for the effectiveness of the protocol.

In [1], through the analysis of a limited number of textual transcripts of actual conversation and packet-trace files of an audio-conferencing application, we showed that in natural, multi-party conversation there is a high correlation between those participants who spoke recently and those who will speak next; the explanation for this result lies in the relationships between conversational turns, such as *adjacency pairs* (*e.g.* questions and answers, exclamations and responses, *etc.*), which have been well-documented in the study of conversation analysis [10]. We also devised a rudimentary next-speaker prediction algorithm.

In the context of audio conferencing over application-level multicast, we define the problem of next-speaker prediction as a problem of maximising the probability that one participant of a constrained set of recent-speaking participants, which set we refer to as the *backlog*, will speak next. Thus, the role of a next-speaker prediction algorithm essentially is to create a prioritised list of participants, ranking them by their level of 'activeness' in the conversation, such that a minimum backlog may be extrapolated from the priority list to perform optimised overlay routing.

In this section, we extend our previous work on next-speaker prediction into a more-complete analysis through the incorporation of corpus data collected and processed by the ICSI meeting project [5]. The corpus comprises the data of over seventy full-length meetings of natural, multi-party conversation, featuring interactions among wide varieties of participants (*i.e.* gender, age, ethnicity, *etc.*), and was produced primarily to aid linguistical research on group conversation and interaction. The data for each meeting comprises recordings of per-participant audio and highly-detailed transcripts, painstakingly annotated per-talkspurt with timing and semantic information. Figure 2(b) shows a sample of talkspurt patterns plotted from corpus meeting.

Using only timing information of talkspurts, as is readily available with little processing overhead to participants of an audio conference, we present, due to lack of space, only our most sophisticated next-speaker prediction algorithm, that gives the best overall efficiency under the most circumstances and is heuristically derived through talkspurt analysis of the ICSI meeting corpus data; furthermore we give an evaluation of the algorithm as well.

4.1 Prediction Algorithm

The next-speaker prediction algorithm presented in this section follows a strategy of associating with each group member (*i.e.* conference participant) a *priority* which quantifies the recent conversational contribution of the participant (and thus, the participant's likely immediate future contribution). The algorithm takes as input (a description of) audio samples/packets and produces a list of participants (ordered according to their computed priorities) on detection of
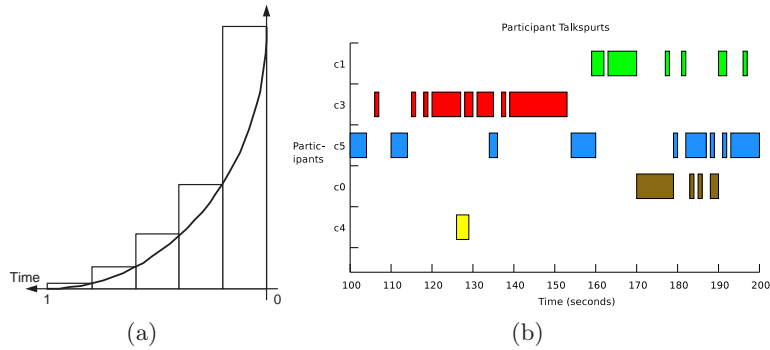
Fig. 2. (a) Calculation of sub-window weights. (b) A sample of talkspurt activity among participants of the ICSI meeting corpus.

turn boundaries (*i.e.* at points of speaker-change) to reflect changing levels of participation throughout the course of the meeting.

Analysis of the ICSI corpus shows that disregarding turns shorter than 800 ms alleviate the problem of falsely indicating that a particular participant is currently engaged in conversation, that is caused by unintentional talkspurts (*e.g.* environmental noises) or intentional back-channel talkspurts (*e.g.* 'hmmm').

The reader should note that each participant runs an independent instance of the next-speaker prediction algorithm and that a description of all audio packets (both received and produced by a participant) are used as algorithm input.

Our next-speaker prediction algorithm analyses both the recent turn-activity of participants and the *occupancy* (*i.e.* the cumulative duration) of their talkspurts within a recent time window: participants who produced talkspurts within the bounds of the window are awarded a priority based upon an occupancy calculation of their talkspurts, and those participants who have no talkspurts within the window are given a priority of zero (*i.e.* they may be considered not to be engaged in the current conversation).

To give weight to those talkspurts that occurred more-recently in the window (*i.e.* so that participants with more-recent talkspurts are favoured), the time window is decomposed into n sub-windows of equal length and a weight is calculated for each sub-window based on some decay function. Figure 2(a) depicts this decomposition, where sub-window weights are calculated as the area of rectangles delimited by the decay function.

4.2 Practical considerations

Our next-speaker prediction algorithm is conceptually simple and easy to implement. However, in order to achieve effective ALNAC routing in all circumstances, we propose some simple extensions.

The next-speaker prediction algorithm simply returns a priority-ordered list of participants. Although always using the maximum allowed backlog increases

ALNAC's effectiveness, but it also results in a higher protocol overhead caused by more delegation (see Section 3). In order to keep actual backlog sizes (and therefore delegation) to a minimum, while still achieving high prediction accuracy, we introduce the concept of *backlog priority threshold*, whereby prediction algorithms will only return a priority-ordered list of participants whose priority is higher than the backlog priority threshold. Obviously, a higher threshold forces the algorithms to 'hide' participants whose recent conversational contribution is 'minor'.

Since in some applications, such as tele-teaching, the turn of a participant with, say, a teaching role can be expected to last for a very long duration of time (*i.e.* with only sparse interruptions by students throughout the duration of the session) we propose that, in order to avoid unnecessary overhead of the ALNAC routing process (*i.e.* by having an unnecessary-large backlog), algorithm priorities should be re-computed periodically intra-turn with a period of M ms (*i.e.* in addition to being computed on turn boundaries), such that the previous turn of, say, a student becomes discounted after a period of time *during* the teacher's very long turn.
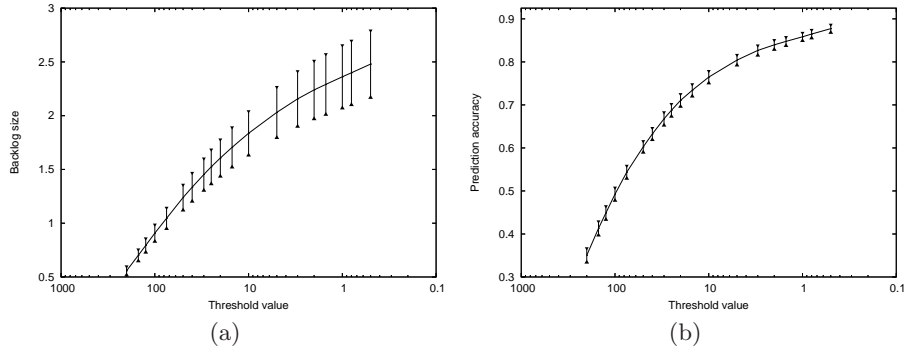
4.3 Evaluation of the Algorithm

We have defined our new next-speaker prediction algorithm that, through the observation only of recent talkspurt patterns among participants of multi-party conversation, tries to maximise the accuracy of next-speaker prediction. In this section, we evaluate the efficiency of this algorithm (*i.e.* its ability to predict the next speaker with a minimised backlog).

To understand the relationship between the backlog-priority threshold, the backlog size and, of course, prediction accuracy, we performed an analysis of the algorithm over talkspurt data in all 75 ICSI meetings of the corpus, computing the average and 95% confidence intervals for backlog size and prediction accuracy for a range of threshold values that were chosen sensibly to suit the priority mechanism of the algorithm (see Figure 3).

In Figure 4 the results of threshold analysis for our new and 'Initial' algorithms have been combined to show the prediction accuracy of each algorithm against average backlog size. Note that, the 'Initial algorithm' represents the prediction accuracy of our rudimentary algorithm from [1], in which prioritization of participants for prediction is based upon only the order of recent talkspurts.

In Figure 4, we see that our new algorithm gives improved performance over the initial algorithm; this occurs as a result of this algorithm being capable of making intelligent judgements as to whether a talkspurt is significant in prediction or not, for example: whether a talkspurt is a short burst of noise or back-channel speech (*i.e.* 'mmm-hmm', 'yeah', *etc.*) from participants who have no intent to become engaged in the current conversation. In summary, we see that a high prediction accuracy may be achieved in multi-party conversation by considering only a small backlog of previous speakers (≤ 3); this result confirms the results of our less-extensive analysis of textual transcript turn patterns and packet traces in [1].

Fig. 3. Backlog-priority threshold characteristics of the new algorithm
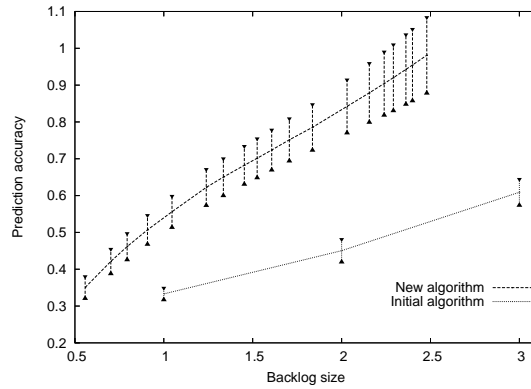
Fig. 4. Comparison of the algorithms' prediction efficiency

5 Simulations

Since existing topology-based simulators cannot simulate the realistic dynamics of Internet node-pair latencies (*i.e.* due to complexities of traffic patterns and network structuring), we implemented an event-based network simulator that uses latency matrices, populated by actual Internet latency measurements. The latency matrices were obtained for 1740 arbitrary Internet hosts from [3] and for PlanetLab [9] nodes from [12].

The goal of ALNAC is to improve the perceptual quality experienced by not *some* but *all* participants of an Internet audio conferencing system over application-level multicast, and so here a performance comparison is made among ALNAC and two other approaches, Narada and TBCP, that are representative of conventional (*i.e.* non-adaptive) per-source–tree and shared-tree application-level multicast, respectively:

– the performance of Narada [13] is important as a benchmark since, through a technique of exhaustive probing to construct a highly-optimised overlay network — albeit doing so with a high protocol overhead — the protocol is designed especially to provide low-latency, any-source group communication;
– the performance of TBCP [8] is important as a benchmark since the protocol is able quickly to build good quality shared trees with a low protocol overhead, and is thus representative of more scalable, alternative approaches to Narada.

Two additional ALM protocols used in the comparison are multiple-unicast, representative of the latency performance of network-level multicast, and a variation on multiple-unicast which is introduced here as *magic multicast*: through an exhaustive off-line search of the simulator's latency matrix, the magic-multicast ALM protocol exploits all of the best shortcut paths that improve latency over those of unicast, and is therefore representative of an ALM protocol with the best possible latency performance.

Recall that ALNAC defines only the way in which (audio) data should be flooded over a given overlay network and not the structure of the overlay network itself, and so ALNAC may run on top of any overlay network from which distribution trees may be inferred. In this case ALNAC is run on top of a (shared) TBCP tree to demonstrate how such shared-tree overlay networks can benefit from dynamic flooding whilst maintaining properties of good scalability — indeed, ALNAC may be run over, say, Narada, but the benefit of doing so would be less.

In the simulations ALNAC was run using our new next-speaker prediction algorithm introduced in Section 4.

5.1 Results

Figure 5 plot conversational response lateness performance of the ALM protocols for simulations of an audio conference among 40 nodes. The simulations were run using reconstructed conversation among participants of four of the largest (participant-wise) transcripts from the ICSI meeting corpus, with ten repetitions of each simulation (*i.e.* with random selection of overlay nodes from the King latency matrix [3] and random placement of ICSI meeting participants on overlay nodes). Note that, where applicable, a low maximum out-degree constraint of 4 is enforced on the overlay construction protocols (*e.g.* `narada-4`, `tbcp-4`) to be representative of a low and fair forwarding overhead for each node.

Note that, Narada was simulated both with and without a maximum out-degree constraint, since given the freedom of being unbound, the protocol is better able to optimise the overlay for latency.

Figure 5(a) shows the distribution of response lateness for conversational responses that were cued by participants themselves (*i.e.* those participants that were engaged in conversation) and Figure 5(b) shows response lateness for responses cued by other participants (*i.e.* those heard by participants who were not at that time engaged in conversation).
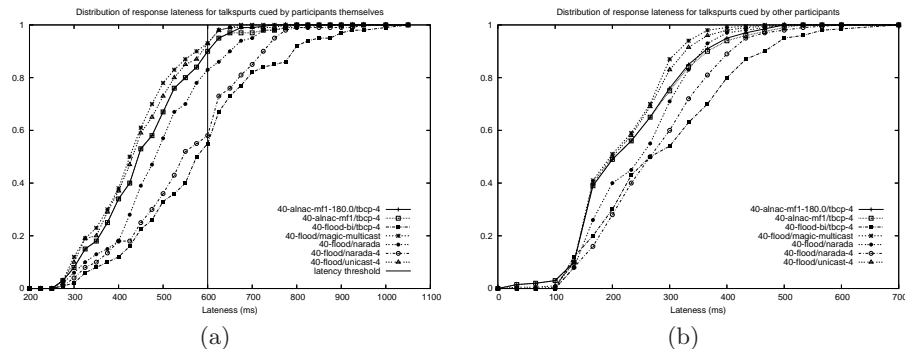
Fig. 5. Distributions of conversational response lateness.

Figure 5(a) captures the essence of ALNAC, since despite in the experiment that ALNAC was run over a shared TBCP overlay tree and with a maximum node out-degree of only 4, perceptual lateness for participants that became engaged in conversation throughout the simulated audio conference is significantly improved over that of the other conventional ALM protocols; even improving over that of Narada with an unbound maximum out-degree (`flood/narada`), which had a tendency to burden centrally-located nodes with a large out-degree of 9–11; in fact, the performance of ALNAC is very close to that of the ideal protocols `flood/unicast` and `flood/magic-multicast`.

This is actually as expected, since ALNAC exploits the predictability of conversational patterns in order to deliver — through a process of semi–tree-circumvention — audio data with minimal latency to those participants who are currently engaged in conversation, whereas the other ALM protocols, during the overlay-network construction phase, try to strike a balance of optimisation for the whole group with the result that, when the time comes for them to communicate, a significant proportion of participant-pairs will be unable to tolerate the excessive communication-channel latency.

In Figure 5(a), the only reason that the performance of ALNAC does not mirror exactly that of multiple-unicast and magic-multicast is that, although highly accurate due to the natural ordering of turn-taking that occurs in multiparty conversation, next-speaker prediction is not infallible and nor is the identification of cues and responses that are used to measure performance of the protocols: next-speaker prediction often falters during the occasional 'hot spots' that occur in multiparty conversation.

Figure 5(b) is very interesting, since by capturing the perceptual quality experienced by participants that listened to the conversation of other participants without themselves being engaged in conversation at that particular time, it is complementary to Figure 5(a), and thus completes the picture of perceptual quality experienced by *all* participants of the simulated audio conferences. As

argued in Section 2, a non-engaged participant (*i.e.* as opposed to a participant who their self is issuing cues) will perceive a response to be late not by the absolute latency with which it is received but by its relative lateness with respect to the time that the cue was received; this feature of perception is captured in Figure 5(b) by the fact that, despite there being some variation in the degrees of lateness experienced by participants under the various protocols, none of the responses were actually significantly late as to impair perceptual quality of the participants, for any of the protocols; the reason that these distributions have consistently smaller ranges than those of Figure 5(a) is because, from the perspective of a non-engaged participant, cues and responses undergo the same or similar application delays (*i.e.* packetisation, decoding, and buffering), which are effectively cancelled between two incoming streams, such that, and unlike for participants who are engaged in conversation, perceptual lateness of non-engaged participants is actually a reflection only of the difference in the respective network delays incurred by incoming streams. The curves of this figure thus reflect the distributions of path lengths of the overlay networks build by the various ALM protocols.

An interesting observation to make of Figure 5(b) is that for non-engaged participants ALNAC actually improves perceptual lateness over TBCP with conventional flooding yet both use the same shared overlay tree; this can be explained by the fact that ALNAC nodes will make full use of their maximum forwarding capacity to disseminate data on the tree, for example: with conventional shared-tree flooding, a leaf node will begin dissemination only through a single node (*i.e.* its parent or the tree's root); whereas an ALNAC node will also or alternatively begin dissemination through a number of nodes that host those participants who are predicted to be engaged in conversation at that particular time.

6   Conclusions

In this paper, we have presented a novel and thorough investigation of two properties of multi-party conversation that are highly important in the realisation of VoIP applications over ALM: (i) the effects of communication-channel latency on quality perception in multi-party conversation; and (ii) the problem of next-speaker prediction in multi-party conversation (*i.e.* which participants are at current most-sensitive to communication-channel latency). The two main contributions, namely our the quality model for multi-party conversation and our efficient next-speaker prediction algorithm, although developed in the context of our work on ALNAC, are readily applicable in the wider context of audio-conferencing systems. Indeed, they can, for instance, be used to evaluate and guide the operation of related proposals such as ACTIVE[7] (a proposal based on the strategy of shaping the ALM tree so that active speakers are near the root).

We have also presented the ALNAC protocol and conducted simulations that model, realistically, characteristics both of the network and of multi-party conversation.

Based on our analysis and simulations, we conclude that in order to support truly-scalable audio conferencing over ALM, an ALM routing protocol *must* be reactive to the conversational patterns of participants, such that perceived quality may be improved for not just *some* of the participants (*i.e.* by fortune of their location in the overlay tree(s)) but for *all* participants. The ALNAC protocol, including its next-speaker prediction algorithm, was shown to be a scalable, elegant and general solution to this problem, capable of efficiently supporting both meeting-type and orator-type audio conference applications.

References

1. Blundell, N., Mathy, L.: Minimising *Perceived* Latency in Audio-Conferencing Systems over Application-Level Multicast. In Proceedings of MIPS 2004, Grenoble, France, Nov 2004.
2. Boutremans, C.,Le Boudec, J.-Y.: Adaptive Delay Aware Error Control For Internet Telephony. In Proceedings of the 2nd IP-Telephony Workshop, New York, April 2001.
3. Gil, T. M., Kaashoek, F., Li, J., Morris, R., Stribling, J.: King Dataset. http://www.pdos.lcs.mit.edu/p2psim/kingdata/, August 2004.
4. Gummadi, K.P., Gummadi, R., Gribble, S. D., et al.: The Impact of DHT Routing Geometry on Resilience and Proximity. In Proceedings of the ACM SIGCOMM 2003, Karlsruhe, Germany, August 2003.
5. Janin, A., Ang, J., Bhagat, S., et al.: The ICSI Meeting Project: Resources and Research. Proceedings of NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, May 2004.
6. Jiang, W., Koguchi, K., Schulzrinne, H.: QoS Evaluation of VoIP End-Points. Proceedings of IEEE International Conference on Communications (ICC 2003), Anchorage, Alaska, May 2003.
7. Liu, L., Zimmermann, R.: ACTIVE: Adaptive Low-Latency Peer-to-Peer Streaming. In Proceedings of the Twelfth Annual Multimedia Computing and Networking (MMCN '05), San Jose, California, January 2005.
8. Mathy, L., Canonico, R., Hutchison, D.: An Overlay Tree Building Control Protocol. Proc. of Intl. workshop on Networked Group Communication (NGC), Nov 2001. 76–87
9. PlanetLab. http://www.planet-lab.org.
10. Lectures on Conversation. Blackwell, Oxford, UK, 1992.
11. Schmitz, U.: Eloquent Silence. Linguistik-Server Essen (LINSE), 1994.
12. Stribling J.: PlanetLab all pairs ping data. http://pdos.csail.mit.edu/~strib/pl_app/.
13. Chu Y.-H., Rao S. G.,Zhang H.: A Case for End-System Multicast. In Proceedings of ACM SIGMETRICS 2000, Santa Clara, California, US, June 2000.