

Tandem-Free VoIP Conferencing: A Bridge to Next-Generation Networks

Paxton J. Smith, Peter Kabal, and Maier L. Blostein, McGill University

Rafi Rabipour, Nortel Networks

ABSTRACT

This article surveys approaches to teleconferencing in voice over IP networks. The considerations for conferencing include perceived quality, scalability, control, and compatibility. Architectures used for conferencing range from centralized bridges to full mesh. Centralized conference bridges used with compressed speech degrade speech quality when multiple talkers are mixed and subjected to tandem coding operations. Full mesh and multicast solutions (mixing at the endpoints) are inappropriate when the number of conferees is large. This article discusses a hybrid solution that incorporates tandem-free bridging (the bridge selects and forwards packets) and endpoint mixing.

INTRODUCTION

Conferencing capability is an essential part of a voice communication network. Wide-area conferencing facilitates group collaboration for geographically dispersed organizations, such as business, military, government, and educational institutions, and permits three-way calling between subscribers. For the Internet to evolve into a global telecommunications network, enhanced conferencing services must be provided. In fact, the high penetration of IP telephony networks will require voice conferencing to have a quality of service on par with that offered by the public switched telephone network (PSTN).

Today, commercial voice over IP (VoIP) conferencing services are provided with centralized conference bridges. This approach requires that the voice signals pass through two speech codecs in tandem, resulting in poor speech quality if low-bit-rate (high compression) speech codecs are used. Good quality of service requires high bit rates and hence more bandwidth.

If compressed speech must be used, then bridgeless, peer-to-peer conferencing arrangements such as full mesh and multicast conferencing eliminate the tandeming altogether. However, in these arrangements the worst case endpoint bandwidth grows with the size of the conference. What is more, wide-area multicast conferencing is difficult to achieve since native

support for multicasting is not widespread and multisender sessions are not yet well understood.

The viability of cost-effective VoIP conferencing is contingent on new approaches to the problem. A novel approach, described later, is a hybrid system that uses centralized speaker selection and decentralized mixing.

This article surveys both popular and unconventional conferencing architectures used in packet voice networks. The article is organized as follows. First, the basic design considerations of voice conferencing systems are discussed. Then, various conferencing architectures are described together with their advantages and disadvantages. A promising new architecture is then presented. The article ends with a comparison between the surveyed models, followed by conclusions.

DESIGN ISSUES FOR VOICE CONFERENCING SYSTEMS

Conferencing system architectures can be distinguished by the location of their audio mixing functions and their connection topologies. This leads to two general classes of architectures, namely centralized and decentralized, as well as a third *hybrid* class. Systems from these families can be evaluated in terms of perceived quality, scalability, controllability, and compatibility with existing standards and practices.

PERCEIVED QUALITY

Speech quality is the strongest factor that affects the perceived quality of a conferencing system [1, 2]. Quality is affected by the performance of the speech coding algorithm in clean and frame-erased channel environments, the number of transcodings, end-to-end delay, perceived echo, and so on. The topology of the conferencing system influences the number of transcodings and end-to-end delay. With improved speech quality and intelligibility, listener fatigue is decreased, and there are fewer misunderstood words, helping to maintain the pace and productivity of the conference.

Perceived quality is also affected by the level of participation of the conferees (i.e., the ability

of the conferees to converse naturally). For instance, some systems allow only one conferee to be heard at a time, which reduces interactivity and forces the conferees to compete for talking privileges. Empirical evidence has shown that allowing two or three simultaneous talkers is adequate to maintain a good level of participation [2].

Finally, perceived quality can be affected by the terminal equipment or transmission facilities used to participate in the conference. This is mainly a concern for nonstandardized desktop conferencing systems where the quality and configuration of microphones, headphones or speakers, and sound cards can vary between stations.

SYSTEM SCALABILITY

The two main issues affecting scalability of a conferencing system are the computational complexity and bandwidth requirements of the bridge and endpoint. Computational complexity is dominated by the number of speech processing operations (e.g., encoding, decoding, and mixing) at any one node. Architectures that distribute the processing load over many bridges (or endpoints) are more scalable.

Bandwidth scalability depends on the connection model. Large bandwidth requirements may create bottlenecks over low-speed links. The problem can be remedied by using resource reservations, although only partially, since reservations for the worst case may be unpractical if the conference membership is large.

CONFERENCE CONTROL

A conferencing system must evolve to support new services and standards. For example, a carrier-grade conferencing system should be supplemented with features such as subconferencing and muting. The system's architecture may complicate its ability to meet current and future requirements.

SYSTEM COMPATIBILITY

Conformance to current standards used in VoIP networks facilitates compatibility with systems operating in different domains or built by different vendors. Sometimes proprietary schemes introduce nonstandard media processing or signaling in order to provide key differentiators from competitors' products. If these modifications are simple, their use may become widespread.

CENTRALIZED CONFERENCE ARCHITECTURES

Traditional "meet me" teleconferences have been provided by centralized conference bridges, to which conferees dial in at a prearranged time. The endpoints establish one-to-one media and signaling connections with the bridge. The bridge establishes voice paths between endpoints by summing the input signals together and returning the summed signal(s) to the conferees. To prevent howling and direct talker echo, the conferees receive a tailored audio signal comprising the sum of all conferees' voices except their own. Customarily, the bridge reduces background

noise and the probability of hybrid echo by including only M out of N active talkers in the conference sum(s). This use of *speaker selection* implies that $M + 1$ sums are formed: one sum for each of the M talkers plus one sum for the $N - M$ unselected conferees (the listeners).

A speaker selection algorithm typically chooses one to three signals for output. The signals are divided into frames, compared, and then selected based on the order in which the conferees begin talking or their relative "loudness." These selection criteria are used by the First-Come-First-Served (FCFS) or Loudest Talker (LT) algorithms, respectively. FCFS uses a voice activity detector (VAD) decision as input, while LT requires a measure of signal power or average absolute level. LT was meant to select over 0.125–3 ms intervals and results in voice break-ups when used with 10–30 ms frames commonly used by most speech coders. In contrast, FCFS provides smoother switching and is usually preferred for packet-based systems. FCFS treats naturally low or loud talkers equally, but prevents interruptions and produces annoying speech level contrasts when a new talker is selected mid-talkspurt [1].

CONVENTIONAL VOIP CONFERENCE BRIDGES

A generic VoIP conference bridge works as follows. Compressed speech arrives at the bridge encapsulated in Real Time Protocol (RTP) packets, which are then disassembled and added to a jitter (playout) buffer. At their scheduled playout times, the speech data are decoded and added to a mix buffer, from which $M + 1$ sums are formed. These $M + 1$ signals are encoded, encapsulated in RTP packets, and then distributed to the conference endpoints. Copies of the $(M + 1)$ th signal (i.e., the sum of all M talkers) are sent to the $N - M$ unselected conferees.

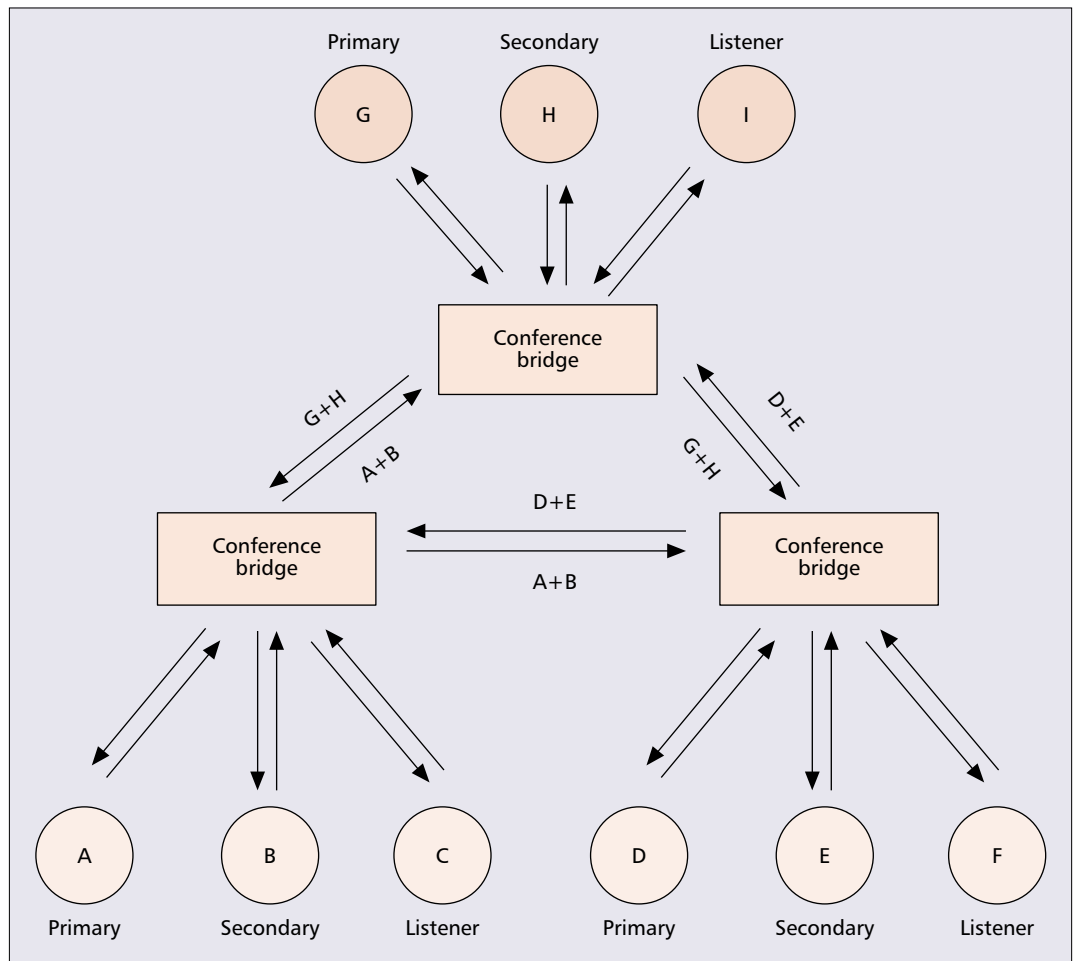
Conference bridges are often built on stand-alone platforms that provide auxiliary services to the network, such as recording, announcements, and lawful interception. Commercial VoIP bridges are available on two kinds of platforms: software-based bridges hosted on dedicated general-purpose computers, or as applications running on digital signal processor (DSP)-based media servers. The former embodiment is intended for smaller LAN conferences and achieves scalability by sharing the speech processing functions over multiple servers.

DSP-based bridges are meant for large-scale carrier-grade applications. Scalability is achieved by adding more DSP cards to the server chassis. The easiest way to build such a bridge is to place a packet interface in front of existing time-division multiplexing (TDM)-based audio bridging circuitry. However, pure packet-based audio bridges yield higher port densities per DSP since there is no delay in clocking the signals between the packet and TDM interfaces. Major vendors such as Cisco Systems, Lucent Technologies, and Nortel Networks offer DSP-based bridges.

Speaker selection is used to limit the number of output sums to $M + 1$; hence, only as many encodings are required (although N decoding operations are required in the worst case). Network topology permitting, a bandwidth optimization can be achieved if the VoIP bridge

Sometimes proprietary schemes introduce nonstandard media processing or signaling in order to provide key differentiators from competitor's products. If these modifications are simple, their use may become widespread.

Speech distortions produced by centralized conference bridges can be reduced by using speaker (signal) selection and forwarding instead of mixing. The idea is to select and forward the compressed speech signal(s) to the endpoints without undergoing the usual decoding, mixing, and re-encoding process.



■ **Figure 1.** A standard multiple bridge configuration.

distributes the stream common to the $N - M$ listeners via multicast [3]. In this case, the selected speakers receive their custom streams on a unicast port; otherwise, they receive the listener stream on a multicast port.

Multiple bridges are required when the conference membership is large, or conferees are dispersed over a large geographical area. For example, both transoceanic and intercampus conferences often comprise small clusters of closely situated conferees; the groups could be connected via satellite, leased lines, or the Internet. Conferees connect to their local bridge, which forms and distributes the local conference sum to its peer bridges (Fig. 1). The local bridge receives the sums of the peer bridges in return, and then mixes them with the sum(s) heard by the local conferees.

The use of centralized conference bridges can result in significant reduction of speech quality due to the tandem arrangement of low-bit-rate speech codecs and vocoding of the multispeaker signal. Overall quality of service is reduced by the additional delay imposed by the jitter buffers and codec processing, resulting in near double the total end-to-end delay. The problem is worse for multiple-bridge operation since the signal undergoes additional transcodings. The speech processing operations are computationally demanding, limiting the scalability of the bridge.

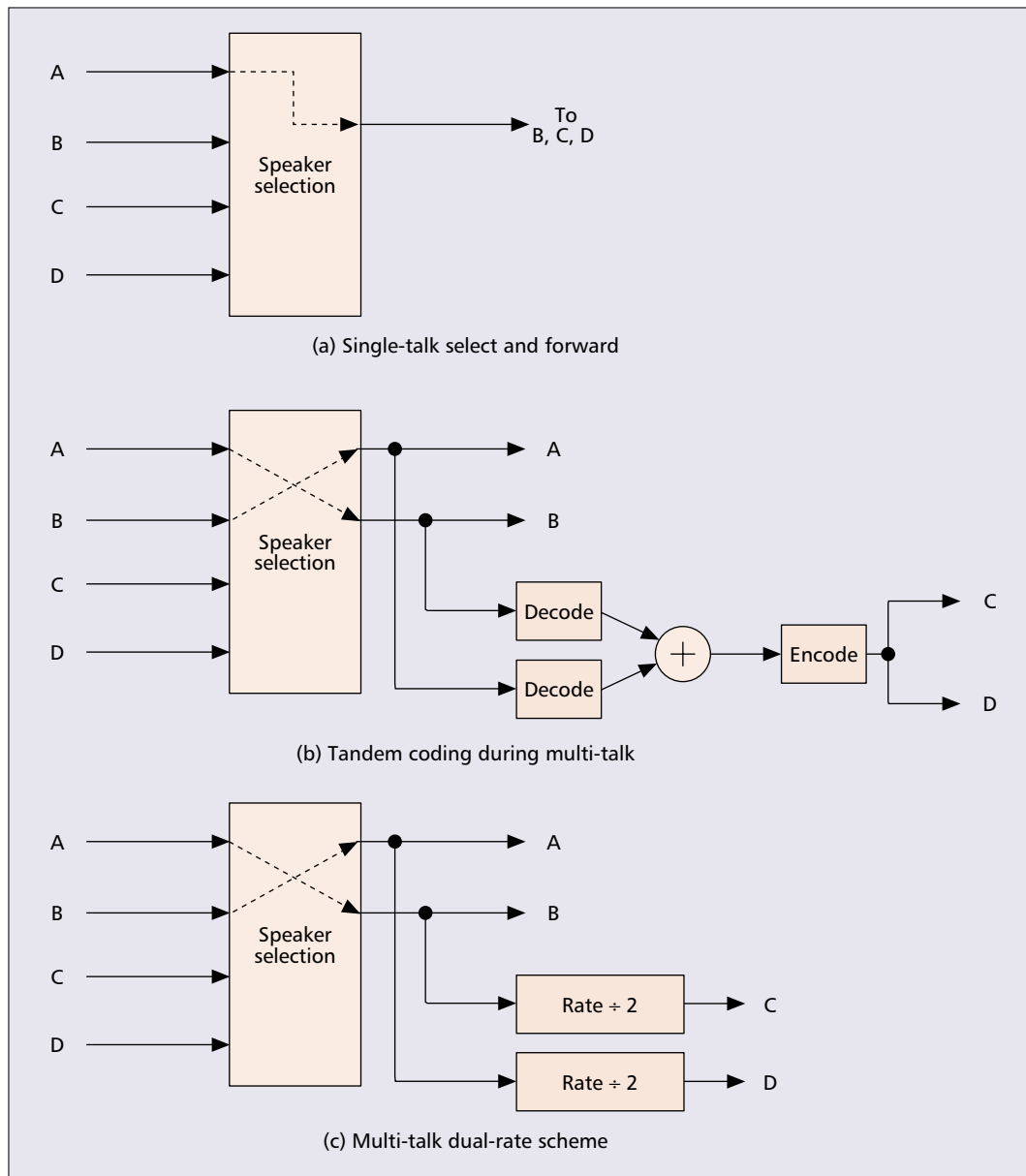
Conference bridge designers can mitigate these problems using the following techniques.

SELECT-AND-FORWARD CONFERENCE BRIDGES

Speech distortions produced by centralized conference bridges can be reduced by using speaker (signal) selection and forwarding instead of mixing. The idea is to select and forward the compressed speech signal(s) to the endpoints without undergoing the usual decoding, mixing, and re-encoding process. Such bridging techniques can be characterized by:

- The number of simultaneous talkers allowed
- The use of either a partial or full decoding process for feature extraction
- The amount of tandeming/transcoding
- Codec dependencies

The first tandem-free conference bridge was built by Forgie (Lincoln Labs) when it was observed that tandeming the multispeaker signal with Linear Predictive Coding (LPC) at 2.4 kb/s resulted in extremely poor quality speech [1]. Forgie's solution was to select and return the compressed signal of the primary speaker to the other $N - 1$ conferees (Fig. 2a). Meanwhile, the signal of the primary interrupter (i.e., the conferee with second highest priority) was sent to the primary speaker. Speaker selection was accomplished using FCFS, effectively reducing the conference to a series of monologues.



If the select-and-forward process is used during single-talk, while normal mixing — thus tandeming — is restricted to periods of multi-talk, then the speech quality is improved most of the time, and the computational complexity of the bridge is reduced.

■ **Figure 2.** Select-and-forward bridges.

If the select-and-forward process is used during single-talk, while normal mixing — thus tandeming — is restricted to periods of multi-talk, speech quality is improved *most* of the time, and the computational complexity of the bridge is reduced. Such a system was proposed by Nahumi of AT&T [4] (Fig. 2b), using FCFS to select M speakers. A partial decoding process was used to monitor gain and spectral parameters in the bitstream, which in turn were used to derive a VAD decision. Since a maximum of M conferees were selected, only M full decoders and $M + 1$ encoders were allocated to the conference. The technique introduced audible pops into the synthesized speech on transitions between single- and multi-talk (resolvable with additional processing) since the algorithmic delay between these two modes was different.

The two modified bridges described above emit only one output stream; hence, they fit nicely with the one-to-one connection model

used in conventional centralized conferencing. At the endpoint, the same decoder channel is used to synthesize the received frames, even though the correct decoder state is not known when a switch of talkers occurs. Nonetheless, these designs leverage the fact that the audible distortion due to decoder state loss persists for only a very short period of time, since the decoder and encoder states rapidly resynchronize.

Another approach is possible if the terminal has the ability to receive and mix multiple streams. Champion (COMSEC) proposed such an approach where the bridge forwarded one signal during single-talk, but selected and forwarded the signals of the primary and secondary speakers during multi-talk [5]. In order to preserve the downstream channel bandwidth, the bridge transcoded the two selected streams to half-rate before returning them to the $N - 2$ listeners (Fig 2c). Speaker selection was accom-

*Interdomain
multisender
multicasting is
still a topic of
research. It is
expected that in
the near term, the
use of multicast
will be restricted
to single-sender
noninteractive
conferences, such
as streaming
media and file
transfers.*

plished using FCFS, but the VAD decision was computed at the source and included in the upstream packets. This meant that the bridge did not need to know the semantics of the bit-stream in order to perform speaker selection.

The three bridging techniques surveyed above provide inconsistent speech quality during multi-talk, by either selecting only one speaker or transcoding. The first two systems can be deployed without upgrading the conference terminals; the third cannot. However, decentralized decoding and mixing are key to tandem-free operation in VoIP conferencing. The following section deals with the decentralized class of conferencing models, which also avoids tandeming.

DECENTRALIZED CONFERENCING ARCHITECTURES

In a decentralized conference, media are exchanged between endpoints *without* using a centralized bridge. Improved speech quality is inherent since the absence of the bridge eliminates tandeming. The endpoints, however, must have the ability to receive and mix multiple streams. Distributing the speech processing functions across the endpoints implies that no single entity requires as much computing power as a conventional VoIP bridge. Decentralized conferencing is represented by the full mesh and multicast conferencing models.

FULL MESH CONFERENCING

In this type of conference, a full duplex media connection is set up between every pair of participants, resulting in a “mesh” of connections. Each endpoint transmits a copy of its stream to the $N - 1$ other endpoints, and receives $N - 1$ streams in return, each on its own port. Each pair of endpoints can communicate with any mutually supported codec type. Typically, signaling control is centralized at a server so that a consistent view of the conference state is maintained [3, 6], wherein the conference state could comprise the conference membership or requests for supplementary audio services.

In the worst case, $N^2 - N$ streams will flow through the network, while at each endpoint there must be bandwidth for $N - 1$ full-duplex connections. If silence suppression is used, the worst case bandwidth requirement only occurs when all conferees talk at the same time. Furthermore, the endpoints themselves are burdened by the task of decoding and mixing up to $N - 1$ inbound streams. Due to these constraints, this architecture is suitable for small LAN or campus conferences where large amounts of bandwidth are available and endpoints are powerful desktop workstations [6].

Pseudo wide-area full mesh conferences are possible using a hybrid approach analogous to multiple-bridge operation used in centralized conferencing (conventional bridging as shown in Fig. 1). Here, bridges are used to connect two or more full mesh conferences together [3]. Each bridge forms a local conference sum and transmits it to its peer bridges (Fig. 3). The local bridge receives composite signals in return, and distributes these signals to the local conferees.

Note that in this arrangement, the signals traveling between bridges undergo three transcodings and pass through two jitter buffers. However, if the signals are transmitted to the local bridges uncompressed, the conference sum exchanged between bridges undergoes only one encoding.

MULTICAST CONFERENCING

Multicast conferencing is synonymous with wide-area conferences over the Multicast Backbone (Mbone). In a multicast conference, each endpoint transmits a single copy of its stream to the conference multicast address, and receives $N - 1$ streams in return. From a receiver perspective, nothing changes from the full mesh scenario except that the streams arrive on one port.

Multicast conferencing is another form of a “meet me” conference. Instead of connecting to a conference bridge, endpoints join the conference by subscribing to the conference multicast address. This address could be advertised by one of the endpoints or by a central server, or distributed to the conferees prior to the conference. Once the address is known, popular software such as the Video Audio Tool (VAT) or the Robust Audio Tool (RAT), can be used to participate in the multicast conference from a desktop workstation.

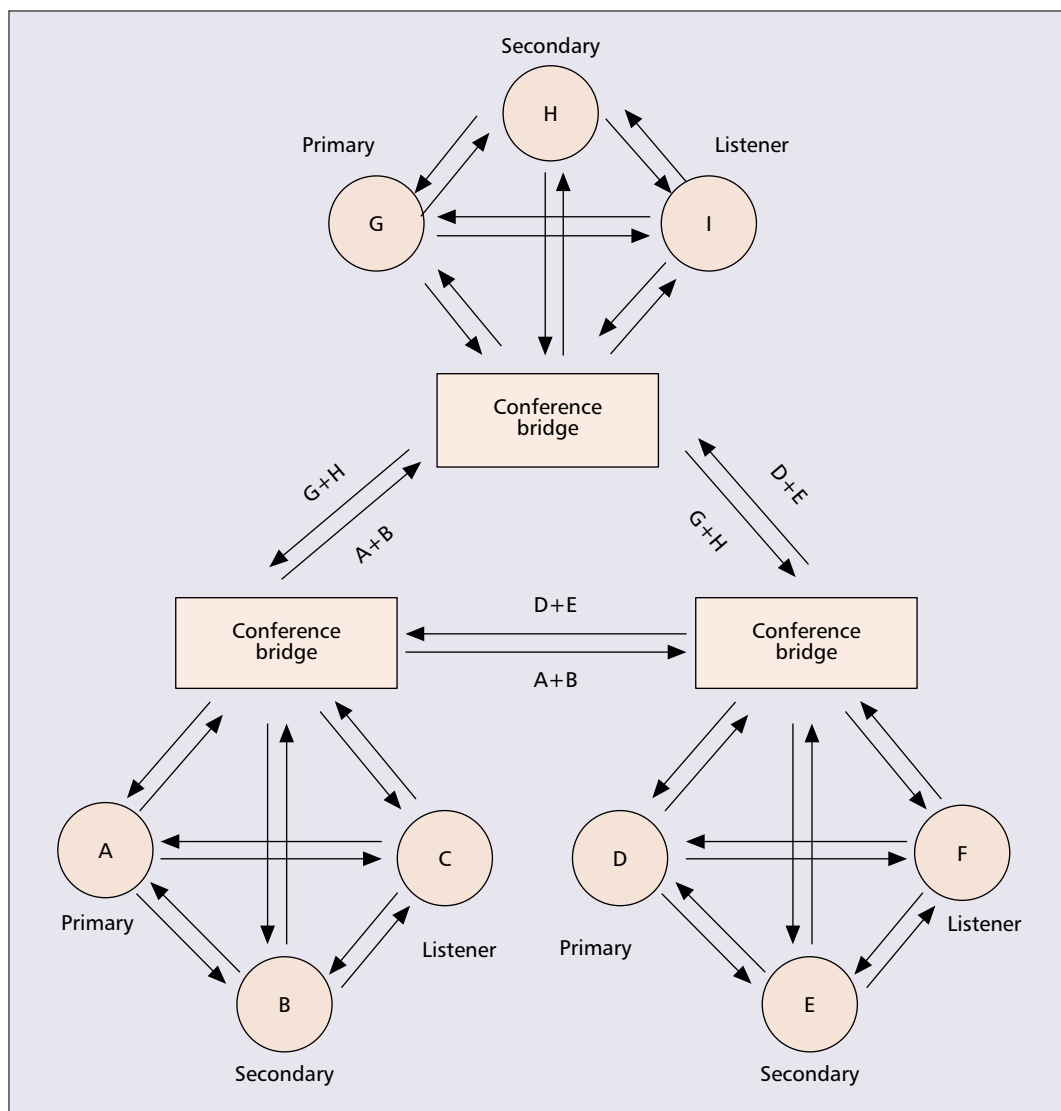
Multicast makes efficient use of network bandwidth, although endpoints require bandwidth for a total of N streams: $N - 1$ inbound (same as full mesh) and one outbound. The speech processing requirements of multicast and full mesh endpoints are the same. Similar to full mesh, bridges can be used to connect multiple multicast conferences together over a wide area. A commercial embodiment of this architecture is the Hoot and Holler conferencing system by Cisco Systems.

Since the scope of multicast packets cannot be explicitly controlled, the media streams must be encrypted to provide privacy. Interdomain multisender multicasting is still a topic of research. It is expected that in the near term, the use of multicast will be restricted to single-sender noninteractive conferences, such as streaming media and file transfers.

The conferencing systems reviewed thus far have several shortcomings: both conventional and select-and-forward conference bridges provide poor or inconsistent speech quality, respectively, while full mesh and multicast architectures impose large bandwidth constraints at the endpoints, and the latter requires network layer support. The following section presents another conferencing architecture with improved performance.

TANDEM-FREE CONFERENCING ARCHITECTURE

The Tandem-Free Conferencing (TFC) architecture, recently proposed by Burns *et al.* [7] (Nortel Networks) and Rabipour and Coverdale [8], is a hybrid between traditional centralized and decentralized approaches. The model uses a tandem-free bridge (TFB), which is a multi-talker select-and-forward conference bridge. The TFB selects M current speakers and forwards their



■ **Figure 3.** Full mesh conferences connected by bridges.

Unlike prior systems, no tandeming or transcoding occurs during multi-talk, thereby providing consistent speech quality over the course of the conference. In addition, the sources can compute and encapsulate the parameters used for speaker selection into TFC data frames and add them to the upstream packets.

compressed signals back to the $N - M$ endpoints, where they are decoded and mixed. If $M = 2$, the primary speaker receives the signal of the secondary speaker, and vice versa.

Unlike prior systems [4, 5], no tandeming or transcoding occurs during multitalk, thereby providing consistent speech quality over the course of the conference. In addition, the sources can compute and encapsulate the parameters used for speaker selection into TFC data frames and add them to the upstream packets. The TFB reads these parameters and uses them to perform speaker selection. Since a partial or full decoding process is not required, the TFB is free of any codec dependencies.

In the case of bundling k codec frames in one packet, k TFC data frames are laid out following the RTP header in the same order in which the codec data appears. Separating the TFC data and RTP payload avoids problems delineating packets that carry multiple frames of variable rate codec data. An *exemplary* 1-byte TFC data frame could be a 1-bit further frame indicator, a 1-bit VAD field (optionally populated), and a 6-bit power field. The use of the TFC data can be

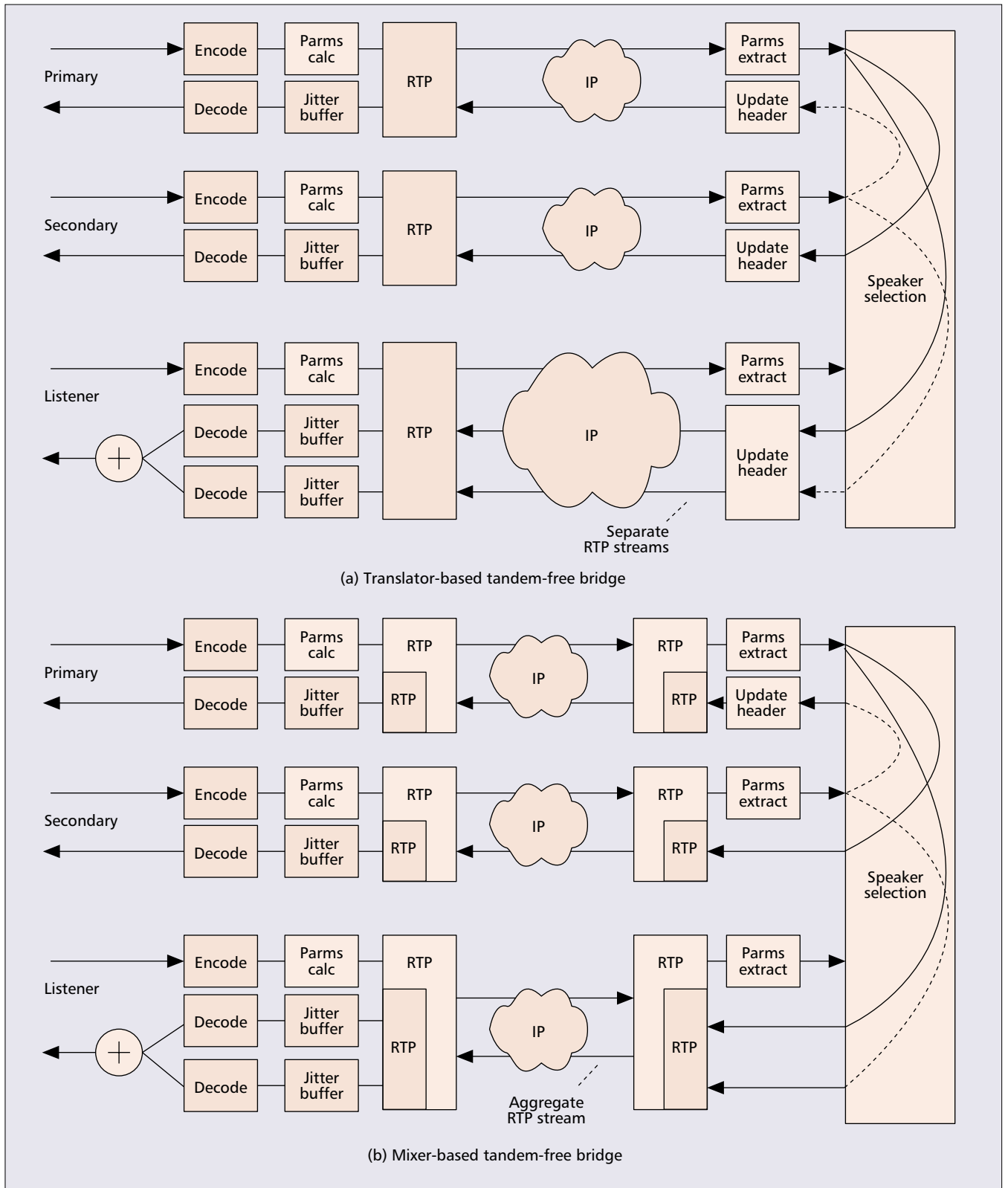
signaled in the session description at conference initiation.

Since the TFB is an intermediate RTP system, it can be modeled as an RTP *translator* or *mixer* [9]. A translator-based TFB selects and forwards whole RTP packets to the endpoints (Fig. 4a). This implies that up to M packets are received at the endpoints per packet interval. The forwarded packets retain their original timestamp and synchronization source (SSRC) information. However, each packet's marker bit and sequence number fields are updated such that the endpoints do not mistake intentionally discarded (i.e., nonselected) packets as lost ones.

At the endpoints, packets are received from the TFB through a single RTP session and are demultiplexed by their SSRCs. Note that the same receiver model is used by multicast conference endpoints. As such, the translator-based TFB can distribute the selected streams via multicast rather than multi-unicast. It follows that this model is suitable for LAN or campus conferencing systems, and is compatible with existing multicast conferencing tools.

Today, some carrier-grade IP telephony networks still enforce the PSTN's one-to-one media connection model. In these cases, terminating multiple streams on the TFC endpoints may cause problems for existing service logic; for instance, the call waiting feature would need to place multiple inbound streams on hold while

the second call is answered. A solution is to combine the selected packets into one *aggregate* RTP packet, complete with its own RTP header (Fig. 4b). In this way, the handling of multiple streams is confined to the media processing layer, allowing call processing to function in the normal way. Specifically, the TFB is modeled as



■ **Figure 4.** Connection models for tandem-free conferencing with two selected talkers.

an RTP mixer, and the aggregate packets are carried in a single RTP session. After IP/UDP/RTP processing of the aggregate packet, the endpoints extract the individual RTP packets and send them through another RTP layer, followed by the dejittering, decoding, and mixing process.

The TFC architecture provides an attractive scheme for wide-area conferences carried by multiple bridges. The configuration is analogous to normal multiple bridge operation, except the mixing is replaced with a select-and-forward process (Fig. 5) [7]. A master TFB is necessary to synchronize the currently selected speakers with the other TFBs. Conferees connect to their local TFB, which selects and forwards M streams to the master TFB. The master reselects M speakers, and returns these streams to the slaves. When a slave receives its own stream(s) back from the master, it knows it hosts one of the M selected talkers. Therefore, the slave forwards the selected stream(s) to all other slave TFBs (as well as the master). The TFBs that do not host a selected talker forward the received stream(s) to all other TFBs except the source TFB. Note that this approach avoids tandeming altogether, yielding a great improvement over the two previously described wide-area conferencing arrangements.

COMPARISON OF ARCHITECTURES

Of the conferencing systems surveyed in the previous sections, the TFC architecture has some interesting properties. The system eliminates tandeming, operates independent of the speech codec, and reduces the computational demands of the bridge. The disadvantages are that protocol extensions are necessary for carrying the TFC data, and endpoints must support multiple stream termination and mixing. This section explores these issues further, and makes comparisons to the conventional VoIP bridge, as well as full mesh and multicast models, in terms of the performance criteria outlined earlier.

PERCEIVED QUALITY

Since speech quality is the strongest factor influencing the performance of a conferencing system, decentralized and TFC architectures should greatly outperform conventional VoIP and known select-and-forward techniques. This was confirmed by live subjective comparisons carried out at McGill University, which solicited conferees' opinions of different conferencing systems [10]. In particular, for systems using the G.729A speech coder, conferees unanimously preferred a decentralized or TFC arrangement rather than one using a conventional VoIP conference bridge. Conferees reported the speech quality produced by the conventional VoIP bridge to be poor and muffled.

Unlike conventional VoIP and select-and-forward bridging techniques, the full mesh and multicast models do not degrade periods of multitalk. In principle, they allow up to N simultaneous talkers without distortion. However, when many conferees speak at the same time, congestion may occur at the endpoints, leading

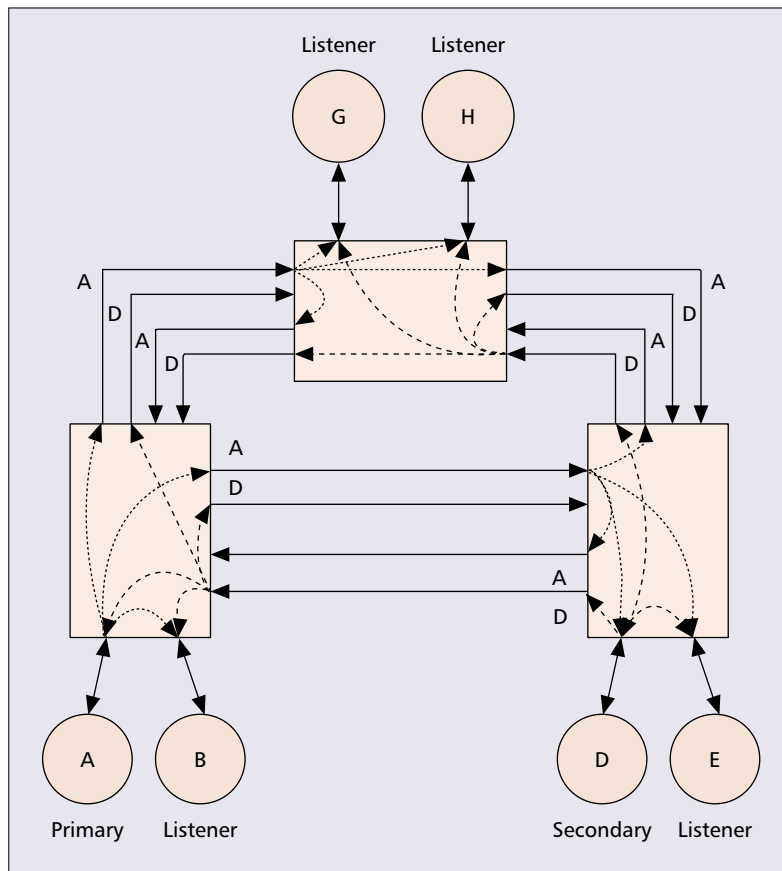


Figure 5. Multiple tandem-free bridge operation with two selected talkers.

to unpredictable packet dropping. In addition, the ability to distinguish between different voices decreases as more signals are added to the conference sum.

TFC provides consistent speech quality by limiting the number of simultaneous talkers to M . Using speaker selection will improve intelligibility during multitalk since a smaller number of voices will be heard. If the speaker selection algorithm is designed to allow interruptions (unlike FCFS) without resulting in voice breakups (like LT), greater interactivity is provided [10].

The full mesh and multicast conferencing models yield the best end-to-end delay performance, while conventional VoIP conference bridges yield the worst. TFC fits somewhere between. Packets must traverse both the upstream and downstream links, but bypass the jitter buffers normally encountered at a conventional VoIP bridge in favor of an inter-stream synchronization algorithm. Since the streams are resynchronized and mixed at the endpoints, the TFC outputs need not be exactly synchronized, allowing for a reduction in algorithmic delay.

SYSTEM SCALABILITY

Network planning and maintenance is simplified by conventional VoIP bridges due to the well understood and scalable one-to-one connection model. Per link bandwidth requirements for the multicast and full mesh endpoints can quickly become unmanageable as N becomes large. In

Since TFC and decentralized conference endpoints receive the audio streams separately, some auxiliary user audio controls can be implemented entirely on the endpoint. For instance, per stream gain and recording controls, and 3-D audio rendering can be implemented without any signaling.

contrast, the TFC model is bandwidth scalable to any size of conference, since it restricts the maximum downstream bit rate to that of M streams. This is a manageable increase for today's terrestrial IP networks, although RTP header compression would be required for operation over 56 kb/s modems and cellular wireless channels. If the TFB distributes the selected streams via multicast, congestion is relieved on links that serve multiple endpoints.

The upstream link is increased by the bit rate of the side information contained in the TFC data frames. Considering the previously described 1-byte TFC data frame and RTP packing scheme, this yields an increase of 0.8 kb/s for codecs with 10 ms frame durations (i.e., G.729), half this for codecs with 20 ms frame durations, and so on.

Computational complexity is a limiting factor for conventional VoIP bridges. By contrast, the TFB is of very low complexity, since only the packet I/O and speaker selection overhead is necessary. In other words, a TFB can carry the same number of conferees at *far* less cost. The TFB could be deployed as an application running on a general-purpose processor in a router or a network server.

Unless silence suppression is used, full mesh or multicast conference endpoints must perform $N - 1$ decoding operations in the worst case. However, the complexity of the TFC endpoint is raised by only $M - 1$ decoding and $M - 1$ mixing operations. This is not so much a problem for high powered media gateways and desktop workstations, but it could pose a problem for tightly engineered wireless handsets and IP phones. Furthermore, since only M streams are active simultaneously, only M speech decoders (which are typically much less complex than speech encoders) need to be assigned to a single conference [4].

CONFERENCE CONTROL

Other services generally provided in conferencing, such as chairperson control or absolute talking privileges, are easily supported by arrangements with centralized media. For instance, a conventional VoIP bridge can provide a subconference to a subset of conferees by excluding their signals from the main sum and adding them to a new one. The two sums are then distributed to their respective conferees. A TFB provides the same service by replacing mixing with selection. On the other hand, with full mesh, the main conferees need to stop transmitting to the subconferees, and vice versa, while the subconferees of a multicast conference drop their membership on the main conference multicast address and join a new one. In the two decentralized cases, the required actions would need to be coordinated by a controller.

Since TFC and decentralized conference endpoints receive the audio streams separately, some auxiliary user audio controls can be implemented entirely on the endpoint. For instance, per stream gain and recording controls, and 3D audio rendering can be implemented without any signaling. In a typical centralized conference, these services can only be provided by the bridge.

SYSTEM COMPATIBILITY

Conventional VoIP bridges, as well as the select-and-forward conference bridges of Forgie and Nahumi, provide the best options in terms of compatibility with existing carrier-grade practices, since endpoints used in two-party calls can be used in multiparty calls. The full mesh, multicast, and TFC endpoints do not have this ability, primarily due to the required mixing duties. The former two also require special call signaling. Another drawback of both multicast and TFC conferences is that the endpoints must share a common speech codec. This is a minor concern for carrier-grade scenarios since network access is provided by gateways, and most gateways will support the same speech coding standards.

If the mixer-based TFB is used, conference initiation and maintenance is the same as used in conventional centralized conferences. This makes deploying TFC a far less onerous task than deploying the full mesh or multicast models. Then the main issues are endpoint support for adding TFC data to outbound packets, receiving and delineating the aggregate packets, and mixing. The first two issues require new RTP payload types to be defined by the Internet Engineering Task Force. The ability to receive (and mix) multiple streams can be added to soft-phones with little trouble, and is already supported by multicast conferencing tools. Multiple stream termination is inherent to the Megaco/H.248 standards. An overall TFC specification could be developed under the auspices of International Telecommunication Union — Telecommunication Standardization Sector Study Group 16.

CONCLUSION

The TFC architecture improves the quality of VoIP conferences that use compressed speech to reduce bandwidth requirements, and is a good solution to the problem of providing large-scale voice conferencing services over IP. Conventional conference bridges are acceptable if high-bit-rate voice coding is used. Multicast conferencing is an attractive approach that also eliminates tandeming and reduces delay, but is limited in scope since native support for multicast is not widespread, and it requires large bandwidths at the endpoints.

Since vendors of VoIP equipment have only begun to roll out their new products, design traditions have not yet been established. The time is appropriate to build in support for new methods such as tandem-free conferencing.

REFERENCES

- [1] J. Forgie, C. Feehrer, and P. Weene, "Voice Conferencing Technology Final Report," Tech. rep. DDC AD-A074498, MIT Lincoln Lab., Mar. 1979.
- [2] J. D. Tardelli et al., "The Benefits of Multi-Speaker Conferencing and the Design of Conference Bridge Control Algorithms," *Proc. IEEE Int'l. Conf. Acoustics, Speech, Sig. Processing*, Minneapolis, MN, vol. 2, Apr. 1993, pp. 435-38.
- [3] ITU-T Rec. H.323, "Packet-Based Multimedia Communication Systems," Nov. 2000.
- [4] D. Nahumi, "Conferencing Arrangement for Compressed Information Signals," U.S. Patent 5,390,177, Feb. 1995.

- [5] T. G. Champion, "Multi-speaker Conferencing Over Narrowband Channels," *Proc. IEEE MILCOM*, Washington, DC., Nov. 1991, pp. 1220–23.
- [6] J. Rosenberg and H. Schulzrinne, "Models for Multi-party Conferencing in SIP," Internet draft, IETF, Nov. 2000, work in progress.
- [7] N. K. Burns, P. K. Edholm, and F. F. Simard, "Apparatus and Method for Packet-based Media Communications," Canadian Patent App. 2,319,655, June 2001, U.S. Patent App. 09/475,047, Dec. 1999.
- [8] R. Rabipour and P. Coverdale, "Tandem-Free VoX Conferencing," Internal memo, Nortel Networks, Montreal, Canada, Aug. 1999.
- [9] H. Schulzrinne *et al.*, "RTP: A Transport Protocol for Real-Time Applications," RFC 1889, IETF, Jan. 1996.
- [10] P. J. Smith, P. Kabal, and R. Rabipour, "Speaker Selection for Tandem-Free Operation VoIP Conference Bridges," *Proc. IEEE Wksp. Speech Coding*, Tsukuba, Japan, Oct. 2002.

BIOGRAPHIES

PAXTON SMITH (paxtons@tsp.ece.mcgill.ca) received his B.Sc. in computer engineering from the University of Manitoba, Winnipeg, Canada, and his M.Eng. in electrical engineering from McGill University, Montreal, Canada, in 1999 and 2002, respectively. He holds a position at McGill's Telecommunication and Signal Processing (TSP) Laboratory, where

he researches speech processing and architecture issues for packet-voice conferencing systems.

PETER KABAL received his Ph.D. degree in electrical engineering from the University of Toronto, Canada, in 1975. He is a professor of electrical and computer engineering at McGill University, Montreal, Quebec, and holds the NSERC/Nortel Industrial Research Chair. His current research interests focus on DSP as applied to speech and audio processing, adaptive filtering, and data transmission.

MAIER L. BLOSTEIN received his B.Eng. and M.Eng. from McGill University in 1954 and 1959, respectively, and a Ph.D. from the University of Illinois, Urbana, in 1963. He has been on the faculty at McGill since 1963. He was director of INRS-Telecommunications and director of the Systems Research Laboratory at Bell Northern Research (BNR), Montreal, Canada, 1975–1985, and served as founding president of the Canadian Institute of Telecommunications Research, 1989–1997. He is currently Professor Emeritus of the Department of Electrical and Computer Engineering at McGill.

RAFI RABIPOUR received his M.Eng. in electrical engineering from McGill University in 1982. He started his professional career at the BNR Laboratory in Montreal, developing DSP techniques for telephony applications. He is now engaged in research and development of voice quality enhancement features and packet-voice signal processing applications for Nortel Networks' line of wireless products.

Since vendors of VoIP equipment have only begun to roll out their new products, design traditions have not yet been established.