Proc. IEEE Int. Conf. Commun. (Anchorage, AK), pp. 794-798, May 2003

Tandem-Free Operation for VoIP Conference Bridges

Paxton J. Smith[†], Peter Kabal[†], Maier Blostein[†], and Rafi Rabipour[‡]
[†]Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, H3A 2A7
[‡]Wireless Speech and Data Processing, Nortel Networks, Montreal, Canada, H4S 2A9

***Abstract*— Traditional telephone conferencing has been accomplished by way of a centralized conference bridge. The tandem arrangement of high compression speech codecs in conventional VoIP conference bridges lead to speech distortions and require a substantial number of computations. Decentralized architectures avoid the speech degradations and delay, but lack strong control and depend on silence suppression to make the endpoint bandwidth and processing requirements scalable. One solution is to use centralized speaker selection and forwarding, and decentralized decoding and mixing. This approach eliminates the problem of tandem encodings but maintains centralized control, thereby improving the speech quality and scalability of the conference. This paper considers design options and solutions for this model in the context of modern IP telephony networks. Performance was evaluated with real conferees over live conferences using a PC-based conferencing test bed, built using a custom software-based bridge and a third-party endpoint. Conferees strongly preferred the speech quality of the new arrangement to that of a conventional VoIP conference bridge.**

I. Introduction

Currently, circuit-switched telephone networks are migrating to packet-based Internet Protocol (IP) networks. In addition to providing new services, IP telephony networks must preserve the rich feature set of the trusted Public Switched Telephone Network (PSTN). One such essential service is voice conferencing, which has historically been accomplished by having users dial in to a digital conference bridge. The purpose of the bridge is to sum up the input signals of each conferee and subsequently supply the summed signal(s) back to each conferee. The conferees hear the sum of all other conferees' signals except their own.

Deploying voice conferencing services with centralized bridges is a seemingly convenient choice since the design principles and caveats of such arrangements are well-known. However, various problems arise when the model is applied to Voice over IP (VoIP). Packet networks inject variable delays into the arrival process, requiring the bridge receiver to absorb this variable delay by means of a jitter buffer. The speech decoder receives compressed speech frames from the jitter buffer at regular intervals such that a continuous stream of PCM samples is fed into the audio bridge. The bridge performs its regular summation duties and returns the composite signal(s) to the network interface where the signal is again compressed, encapsulated in Real Time Protocol (RTP) packets, and sent into the network.

This architecture results in two well-known problems that reduce the speech quality of the conference: tandem encodings and the encoding of a multi-speaker signal. These problems have previously been identified and partially solved in [1–3]. However, speech quality is also reduced due to increased delay, which stems from the per stream jitter buffer and codec processing at the conference bridge.

Another disadvantage of VoIP conference bridges is that they are subject to heavy processing demands when used with compressed speech. This is largely due to the fact that Code-Excited Linear Prediction (CELP)-based codecs are of high computational complexity. Additional processing such as RTP framing, packet loss concealment, buffer management, and clock skew compensation make DSP platforms an attractive—although expensive—choice for implementing such bridges.

This paper presents an architecture for a centralized conferencing arrangement that eliminates the problem of tandem encodings at VoIP conference bridges. This work is based on a generic model described by Burns *et al.* [4], and Rabipour and Coverdale [5], in which the codec data of the primary and secondary speakers are selected and forwarded to the endpoints, where they are decoded and mixed. The upstream packets include an additional field containing the signal power such that speaker selection can be performed without decoding. The model will be known as the Tandem-Free Operation (TFO) conferencing model herein.

The current work expands the ideas of Burns *et al.* and Rabipour and Coverdale to include specific techniques which can be used to realize the system in the context of modern VoIP networks. Performance is evaluated relative to a conventional VoIP conference bridge. Earlier work in [6] considered various speaker selection algorithms for TFO conference bridges, and proposed an improved approach, i.e., the Multi-Speaker/Iinterrupter (MS/I) algorithm.

II. Alternatives to the Conventional VoIP Bridge

The tandem encoding problem can be eliminated or mitigated in several ways, with the most obvious being the use of a bridgeless, decentralized conferencing arrangement. The problems with decentralized architectures are well-known; hence, industrial research has tended to focus on bridging techniques which reduce the impact of tandem encodings. Prior solutions which are also based on speaker selection and forwarding are reviewed here.

A. Full Mesh and Multicast Conferencing

In full mesh and multicast conferences, media is exchanged directly between the endpoints, and the endpoints must have the ability to receive and mix multiple streams.

In a full mesh conference, each endpoint establishes a one-to-one media connection with the $N-1$ other endpoints; each pair of endpoints must share a common codec. The source speech signal is coded once and then copies are distributed via multi-unicast. If signalling control is not centralized at a server, each endpoint must manage $N-1$ signalling connections. These conferences may only be scalable to a few participants [7].

In a multicast conference, each endpoint transmits a single copy of its audio to the conference multicast address and receives $N-1$ streams in return. This implies that the endpoints must share a common codec. The model makes efficient use of network bandwidth and is often associated with wide-area MBONE conferences. However, silence suppression or Discontinuous Transmission (DTX) must be used for large conferences to prevent bandwidth bottlenecks at the endpoints.

Regardless of the speech quality advantages, these two scenarios are not widely used in carrier-grade applications. Full mesh conferencing violates the one-to-one signalling and media connection model of the PSTN and current IP telephony networks, while multicast conferencing is hindered by the lack of widespread support for network-layer multicast [7]. Due to these problems, centralized bridges remain the conferencing architecture of choice for public carrier networks.

B. Single-Talker Select-and-Forward

Forgie first identified the problem of signal summation in centralized conferencing arrangements, and recommended that signal selection be used instead [1]. His bridge selected a primary speaker from N input streams, then replicated and forwarded his/her compressed signal back to the $N-1$ other endpoints without undergoing the usual mixing and re-encoding process. Single-talker systems eliminate the tandem encodings, yet severely limit the conferees' ability to interact.

C. Multi-Talker Select-and-Forward

Reduced interactivity is a concern since it is expected that multi-talk accounts for 6–11% of the total conference time [1,8], while empirical evidence gathered during earlier work demonstrated this can be as high as 40% [6]. Nahumi proposed that the codec data be passed through the bridge during single-talk, but undergo the usual decoding, mixing, and re-encoding process during multi-talk [3]. A two-talker system by Champion improved upon this by selecting a primary and secondary talker (during multi-talk), transcoding their streams to half-rate, and then returning *both* streams to the endpoints where they were decoded and mixed [2].

One problem with the aforementioned centralized models is that speech is degraded during periods of multi-talk. Another is that both Forgie's and Nahumi's systems obtain the speaker selection parameters following a partial or full decoding process. Finally, Champion's system requires use of a coder which supports full- and half-rate encoding modes.

III. TFO Conferencing Architecture

The TFO architecture improves on many of the problems introduced by conventional centralized bridges and the variants described above. The TFO model uses centralized control (speaker selection), and decentralized decoding and mixing, eliminating the tandem encodings with high compression codecs. For a conference with N participants, each endpoint transmits its stream to the Tandem-Free Bridge (TFB), but receives M-out-of-N streams in return. The bridge selectively forwards the codec data of the M selected talkers to the endpoints and discards the rest. The elimination of the codec processing instantly yields a reduction in the end-to-end delay (e.g., 15–25 ms for G.729 and 37.5–67.5 ms for G.723.1).

An option of the TFO architecture is that the features used for speaker selection can be computed at the source and added to each upstream RTP packet. In this way, the TFB can extract said parameters from the bitstream *without* a partial or full decoding process, allowing the TFB to operate independently of the type of speech codec.

As in a decentralized conference, each TFO conferencing endpoint must be capable of simultaneously receiving, decoding, and mixing multiple streams of speech. Typically, M is set to two or three, depending on the desired level of transparency of the speaker selection, per link bandwidth, or endpoint processing power.

A. RTP Connections

The most natural way to model the TFB is as an RTP translator [9] which accepts N streams, but only emits M. Streams can be carried in a single session, or $N-1$ individual sessions. In the former case, the endpoint terminates streams analogously to a multicast endpoint. Then, the TFB can multicast the selected packets back to the endpoints instead of multi-unicasting. In the latter case, there are $N-1$ individual sessions, although only M are active at any one time. This model has applications where the call control layer is flexible enough to allow endpoints to terminate multiple

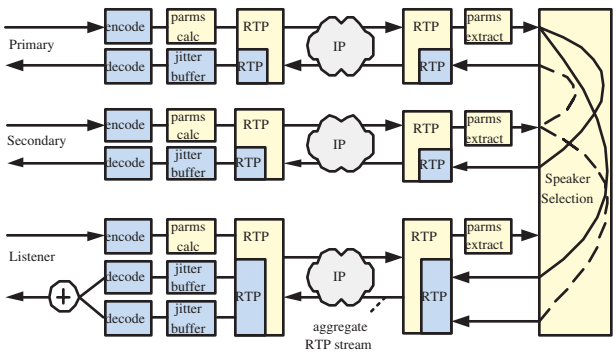

Fig. 1. TFB using aggregate RTP streams.

RTP streams (e.g., LAN or Enterprise conferencing systems). The translator approach leverages the inherent capabilities of multicast conferencing tools, such as the Robust Audio Tool (RAT) [10], which are designed to receive and mix multiple RTP streams.

However, the translator model is not compatible with carrier-grade voice networks where one-to-one connection models are strictly enforced. For these cases, a conventional VoIP bridge which emits one composite output stream for each conferee possesses a more suitable connection model—such VoIP bridges are essentially RTP mixers [9]. The TFB can emulate such behavior by bundling the M selected packets into one aggregate RTP packet, as shown in Fig. 1. This allows the higher level call processing functions to treat the bundled stream in the usual way (e.g., call transfer). Suitable RTP payload types for aggregate streams have been presented in [11]. Essentially, the selected codec data travels through two RTP layers: the first packetizes the codec data as usual, while the second forms the aggregate. Receivers must have the ability to disassemble the composite packets and route the individual streams to their own decoding paths.

B. Speaker Selection

The speaker selection unit assigns priority to the conferees according to some heuristic—the signals of the first M priority positions are selected for output. The typical approach for packet networks is to assign priority based on the order the conferees become active (i.e., when VAD $= 1$), while digital PCM bridges often base priority on the conferees' average speech level or power. The former approach, known as First-Come-First-Served (FCFS), limits positive conversational events such as reinforcement, overlap, and interruptions [1]. The latter approach, known as Loudest Talker (LT), was designed for use with short frames (e.g., 0.125–3 ms) and results in disturbing voice break-ups when used with the longer frames (e.g., 10–30 ms) of popular VoIP speech coders.

A better approach for speaker selection is one which allows barge-in, but controls spurious switching. This is accomplished by the Multi-Speaker/Interrupter (MS/I) algorithm [6]. The algorithm considers three measures: speech activity, smoothed signal power, and a barge-in threshold. The signal power of the ith frame belonging to the jth stream, $E_j(i)$, is filtered to yield a metric, $\hat{E}_j(i)$, possessing a fast-rise slow-decay characteristic. Conferees are assigned priority positions which are maintained as state information. The active conferees are identified using a VAD, and then promoted in priority only if their $\hat{E}_j(i)$ exceeds that of higher priority conferee by a barge-in threshold, B_{th}. This threshold adds hysteresis to the system, helping to eliminate spurious switching.

C. Stream Synchronization

Speaker selection is performed following intra-stream and inter-stream synchronization. The most straightforward method is to perform intra-stream synchronization over all streams, followed by a select-and-forward operation scheduled at periodic intervals [12]. However, this approach does not reconstruct the temporal relationships *between* streams. Alternatively, intra-stream synchronization is performed only on the primary talker's stream, i.e., the master stream, followed by inter-stream synchronization across the slave streams [13]. That is, the slave streams are synchronized relative to the primary talker stream.

Approaches such as the above are convenient when the *periodic* output of packets is required—for instance, when a receiver is feeding a sound device. However, the output process of the TFB need not be periodic, allowing some reductions in queuing delay. For example, the TFB can select-and-forward packets for a given time interval as soon as all packets for that interval have arrived—there is no need to wait. Further reduction in delay can be achieved by performing the select-and-forward operation once per packet arrival. In this case, the packets are mapped to the appropriate time interval and speaker selection is performed using the other conferees' *last* known state. For instance, the metric $\hat{E}_j(i)$ can be derived by setting $E_j(i) = E_j(i-1)$. This works well since $E_j(i)$ is strongly correlated over lags $\leq$ 20–50 ms, hence the general trend in $\hat{E}_j(i)$ is preserved. Wrong decisions will cause a premature or delayed change of speakers, although only by 1–2 frames.

D. Side Information

Recall that the speaker selection parameters can be carried as additional fields in the upstream RTP packets. The most natural method for transporting these parameters is to introduce a new payload type for RTP packets. The use of TFO payload is signalled at session initiation. The actual choice of side information depends on the method of speaker selection used at the TFB, although typical features are a VAD decision, the signal power, or the talkspurt state. An exemplary TFO data frame is shown in Fig. 2. Here, the MSB of the 1-byte frame is a Further Frame Indication (FFI), while the VAD and power can be coded with the remaining 7-bits (e.g., 1-bit for VAD, 6-bits for power). Note that if silence suppression/DTX is available, then speech activity can be determined by monitoring SID frames in the transmission, and the VAD bit can be omitted. Otherwise, the signal power can be used to derive a VAD decision at the TFB.

It is beneficial to match the TFO data frame rate to that of the speech codec being used in the session. This simplifies selecting individual frames for output in the case where k codec frames are carried in one RTP packet. Here, the TFO

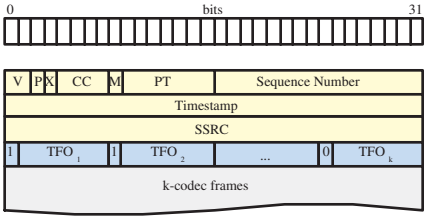

Fig. 2. RTP packet with TFO payload.

data frames are laid out one after another following the RTP header. This has the advantage of separating the TFO data from the speech data, which avoids parsing problems when the packet carries bundled variable rate codec data. This scheme increases the upstream bitrate by $\epsilon = 8/T_f$ kbps, where T_f is the speech codec frame duration in ms.

IV. Complexity Analysis

The computational and bandwidth requirements of the TFO bridge and endpoint were compared against their conventional VoIP bridge, multicast, and full mesh conference counterparts. The following analysis considers the worst-case, i.e., those times when all conferees talk at the same time, or when silence suppression is not used. The number of selected speakers is M, the total number of conferees is N.

A. Computational Requirements

Table I breaks down the conference algorithms into four speech processing operations: (1) speech encoding, (2) speech decoding, (3) speaker selection, and (4) mixing. Here, one mixing operation is assumed to be the vector addition of two decoded speech frames. The table shows the number of these operations for each conferencing system per packet interval. The effect of operating system scheduler, packet or audio I/O, and list/queue management is not considered.

Using speaker selection in the VoIP bridge limits the number of mix operations to $M + 1$, so scalability is limited by the decode operations. Both the conventional and TFO endpoints are scalable since their required number of operations is independent of N, although the TFO endpoint must maintain state for all $N - 1$ conferees. Scalability of the multicast and full mesh clients is limited by the $N - 1$ decodes.

The actual CPU% of the four operations components was measured on a 800 MHz Windows 2000 PC and used to scale the operations of Table I. For G.711 and G.729A, the TFO bridge yielded a reduction in complexity of 5-fold and 300-fold over the VoIP bridge, respectively, for a four-member conference. This reduction is primarily due to the absence of speech encoding and decoding operations at the bridge. In the worst case, it was found that a TFO endpoint requires up to 20% more CPU than a typical VoIP endpoint when using the G.729A codec.

B. Bandwidth Utilization

The worst-case bandwidth utilization is an important design constraint. The analysis is straightforward and summarized in Table II. The translator model of the TFO bridge is used here, since individually forwarding the selected packets means packet overhead is increased by a factor M. Note that B is the bitrate of the IP/UDP/RTP header plus speech data, and ϵ is the bitrate of the extra TFO data field. For G.711 and G.729A, and B is 80 kbps and 24 kbps for 20 ms RTP payloads, and ϵ is 0.4 kbps and 0.8 kbps, respectively.

TABLE I

Number of Operations.

Model	Bridge				Endpoint			
	Enc.	Dec.	Select	Mix	Enc.	Dec.	Select	Mix
TFO	—	—	N	—	1	M	—	1
VoIP Bridge	$M+1$	N	N	$M^2 - M - 1$	1	1	—	—
Multicast	—	—	—	—	1	$N-1$	$N-1$	$M-1$
Full Mesh								

TABLE II

Worst-case bandwidth requirements (kbps).

Model	Bridge		Endpoint	
	In	Out	In	Out
TFO	$N(B+\epsilon)$	$M(N-1)B$	MB	$B+\epsilon$
VoIP Bridge	NB	NB	B	B
Multicast	—	—	$(N-1)B$	B
Full Mesh	—	—	$(N-1)B$	$(N-1)B$

Consider the bandwidth required on a link to a single endpoint in a conferencing arrangement. If $M = 2$ and G.729A is used, then a TFO endpoint will use 24.8 kbps more bandwidth than one connected to a conventional VoIP bridge; however, the peak rate over a single link is independent of the size of the conference, N. This is not the case for multicast and full mesh systems. For a smaller four-member conference, a TFO endpoint requires approximately 25% less bandwidth than a multicast or full mesh endpoint

V. Subjective System Evaluation

The performance evaluation of teleconferencing systems has historically been subjective in nature [1,8]. To this end, real conferees were recruited and used to evaluate the speech quality provided by a conventional VoIP bridge and the proposed TFO conferencing architectures. It was desired to see if the gain in speech quality due to the elimination of tandeming would outweigh the effect of artifacts introduced by speaker selection and encoder-decoder state desynchronization.

A. Platform

The series of live, four-member conferences, were configured over the LAN at the TSP Lab of McGill University. The conference endpoints consisted of four Linux PCs running RAT version 4.10. RAT was modified to support G.729A, and the signal power was carried in the RTP header extension. RTP packets carried 20 ms of speech from either the G.711 or G.729A codec, and RAT's unsophisticated silence suppression was turned off so as not to mask the effects of speaker selection. The number of selected speakers, M, was two.

A TFB was constructed in software according to the design principles outlined in Section III. The flexibility of the LAN allowed the TFB to use the RTP translator model. By avoiding the bundling on the downstream link, the instantaneous select-and-forward could be performed as described in Section III-C. Overall, the system could emulate the TFO, conventional conference bridge, and multicast conferencing through a combination of TFB and (custom) RAT parameters.

B. Method

To encourage interaction between the conferees, a simple game was developed based on the TV game show Family Feud [8]. Each round lasted approximately 10 minutes during which several questions were solved. Conferees' opinions were solicited through interviews following each comparison. Conferees were asked to choose which system they preferred and why. In all, 12 subjects used the system.

C. Summary of Conferees' Opinions

A notable outcome of the experiments is that the conferees did not detect the presence of the MS/I speaker selection algorithm, even though the number of output streams was limited to two. Speech quality was perceived to be high even in the face of minor clicks and pops due to occasional switching errors and G.729A state loss. In general, TFO conferencing was found to provide equivalent or better speech quality than the VoIP bridge or a multicast conference when G.729A was used.

In contrast, VoIP bridge was consistently rated the worst by the conferees. The system was unpleasant and difficult to use due to severe distortion caused by tandeming a multi-speaker signal with G.729A.

VI. Discussion

This work demonstrates the feasibility and advantages of the Tandem-Free Operation (TFO) conferencing model. The arrangement yields improved speech quality and scalability over a conventional VoIP bridge. Subjective testing revealed that tandem encodings with G.729A resulted in speech quality that was significantly degraded. Hence the traditional conferencing service is compromised when G.729A or equivalent codecs are used. Conventional centralized conferencing should use coders with better resiliency to tandem encodings in order to approach the quality of modern PCM conference circuits. Alternatively, the TFO conferencing system allows the use of low bitrate speech coders without sacrificing speech quality.

Deploying such TFO conferencing services is inexpensive since the TFB requires limited computational power. It can easily run on a modern multi-purpose workstation, or could be implemented as additional service logic within a router. Multiple bridges may work together to support a conference by exchanging the streams of the two selected talkers and then reselecting. In the conventional case, multiple bridges exchange partial sums and then remix the total conference sum, resulting in at least three encodings (e.g., in a transoceanic conference).

The TFO conferencing model raises the bandwidth requirements over that of a traditional centralized conference by the bitrate of the upstream TFO payload and the extra stream in the return path. However, this is manageable for high-bandwidth terrestrial VoIP networks. The worst-case bandwidth requirement per link remains fixed at $M + 1$ streams. This is in contrast to multicast conferences which have a worst-case requirement of N streams. Multicast conferences are subject to unpredictable packet dropping since the worst-case bandwidth may not be available.

A drawback of both TFO and multicast conferences is that they require all endpoints to share a common speech codec. This is a minor concern since access to the IP telephony network is provided by a gateway, and most gateways will support the same speech coding standards. In addition, low-level layers of the endpoints need to be changed to support reception of multiple streams. Further, a new RTP payload needs to be standardized by the IETF for transporting the speech features in upstream RTP packets. Another is necessary for the bundled downstream payload, if this mode of operation is used.

As VoIP networks mature and evolve, wide-area multicast conferences may, in some cases, supplant conference bridges. Nevertheless, the TFO conferencing model is well-suited as an alternative technology. Carriers could reduce costs by implementing their three-way calling line-option with the TFO model, thus freeing-up expensive audio bridge resources for larger "business" class conferences that demand G.711 quality.

References

[1] J. Forgie, C. Feehrer, and P. Weene, "Voice Conferencing Technology Final Report," M.I.T. Lincoln Lab., Lexington, MA, Tech. Rep. DDC AD-A074498, Mar. 1979.

[2] T. G. Champion, "Multi-speaker conferencing over narrowband channels," in *Proc. IEEE Military Communications Conf.*, Washington, D.C., Nov. 1991, pp. 1220–1223.

[3] D. Nahumi, "Conferencing arrangement for compressed information signals," United States Patent 5,390,177, AT&T Corporation, Murray Hill, NJ, Feb. 1995.

[4] N. K. Burns, P. K. Edholm, and F. F. Simard, "Apparatus and method for packet-based media communications," Canadian Patent Application 2,319,655, opened June 2001, U.S. Patent Application 09/475,047, Dec. 1999, Nortel Networks Corporation, Ottawa, Canada.

[5] R. Rabipour and P. Coverdale, "Tandem-free VoX conferencing," Internal memo, Nortel Networks, Montreal, Canada, Aug. 1999.

[6] P. J. Smith, P. Kabal, and R. Rabipour, "Speaker selection for Tandem-Free Operation VoIP conference bridges," in *Proc. IEEE Workshop on Speech Coding*, Tsukuba, Ibaraki, Japan, Oct. 2002.

[7] J. Rosenberg and H. Schulzrinne, "Models for multi-party conferencing in SIP," Internet Draft, Internet Engineering Task Force—Work in Progress, Nov. 2000.

[8] J. D. Tardelli, P. D. Gatewood, E. W. Kreamer, and P. A. La Follette, "The benefits of multi-speaker conferencing and the design of conference bridge control algorithms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, Minneapolis,USA, Apr. 1993, pp. 435–438.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," RFC 1889, Internet Engineering Task Force, Jan. 1996.

[10] O. Hodson and C. Perkins, "Robust Audio Tool (RAT) version 4," [online] available at: http://www-mice.cs.ucl.ac.uk/multimedia/software/rat, Nov. 2000.

[11] J. Rosenberg and H. Schulzrinne, "Issues and options for an aggregation service within RTP," Expired Internet Draft, Internet Engineering Task Force, Nov. 1996.

[12] K. Singh, G. Nair, and H. Schulzrinne, "Centralized conferencing using SIP," in *Proc. 2nd IP-Telephony Workshop (IPTel2001)*, New York, NY, Apr. 2001.

[13] Y. Ishibashi, S. Tasaka, and Y. Tachibana, "Adaptive causality and media synchronization control for networked multimedia applications," in *Proc. IEEE Int. Conf. on Communications*, vol. 3, Helsinki, Finland, June 2001, pp. 952–958.