Proc. IEEE Workshop Speech Coding (Tsukuba, Japan), pp. 120-122, Oct. 2002

SPEAKER SELECTION FOR TANDEM-FREE OPERATION VOIP CONFERENCE BRIDGES

Paxton J. Smith [†], *Peter Kabal* [†], *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

A conventional Voice-over-Internet Protocol (VoIP) conference bridge reduces the speech quality due to tandeming the mixed multi-speaker signal with high compression speech codecs. One solution is to select and forward the compressed signal(s) to the endpoints, where they are decoded and mixed. In such arrangements, speaker selection is usually accomplished with an order-based approach which prevents listeners from interrupting the current speaker(s). This paper presents an alternative in which talking privileges are assigned based on order of activity and signal power. Subjective evaluations indicate that speaker switching is smooth, nearly transparent, and unanimously preferred over a VoIP conference with tandemed connections.

1. INTRODUCTION

Traditional disparate voice and data networks are converging into one network, interconnected by an IP packet core. Speech coders are used to achieve efficient transmission of speech signals over the network, and to enable communication over bitrate-constrained links. However, the use of speech compression in conference connections creates various problems for VoIP conference bridges.

Speech quality is reduced by a conventional VoIP conference bridge due to the tandem arrangement, i.e., the serial connection, of high compression speech codecs, and the encoding of the mixed multi-speaker signal. Another problem is the increased end-to-end delay due to jitter buffer and codec processing, and the reduced ability to scale due to the computational demands of the speech codecs. Regardless of the tandem encoding problem, centralized bridges are preferred by carriers because they fit well with traditional approaches.

Solutions have been discussed in the literature which help to improve the speech quality in such arrangements. The basic approach is to use speaker (signal) selection instead of summation [1]. In other words, the bridge selects and forwards the compressed stream(s) of the M dominant speaker(s) ($M = 1$ or 2) back to the endpoints without undergoing the usual decoding, mixing, and re-encoding process. Previous solutions are characterized by one or more of the following features:

1. Partial or full decoding is required for feature extraction,
2. The system is codec-dependent,
3. Tandeming only occurs during multi-talk,
4. Only one talker can be heard at any time.

A full decoding process is required when speaker selection parameters (e.g., a Voice Activity Detector (VAD) decision) must be determined from the decoded signals [1]. Alternatively, a partial decoding process can be used to monitor gain or spectral parameters in the bitstream [2], although this implies a codec-dependency. Stronger codec-dependencies do exist, for instance, in [3], the system uses a dual-rate Sinusoidal Transform Coding (STC) algorithm. During multi-talk, two streams are selected and transcoded to half-rate such that the bandwidth used by the downstream channel remains constant.

If the bridge limits the mixing and re-encoding process to periods of multi-talk, and relays the compressed signals during single-talk, then the speech quality is improved *most* of the time. This method has problems during transitions from single-talk to multi-talk, resulting in audible pops in the synthesized speech [4]. In addition, multi-talk is thought to account for, on average, 6–11% of total conference time [1, 5], but this can vary as shown in Fig. 1 (data collected during a four person conference [6]).

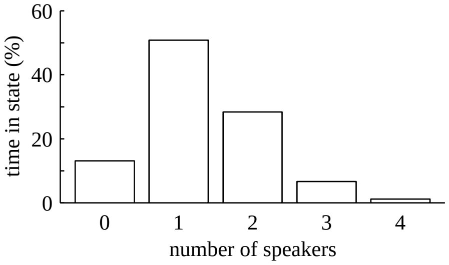

Fig. 1. State distribution of an active four-way conversation.

One feature common to these approaches is that speaker selection is accomplished based on the order that the conferees talk, i.e., the First-Come-First-Served (FCFS) criterion is used. Since FCFS has no notion of preemption, an unselected conferee cannot enter the conversation if M talkers are active. Another feature is that these systems somehow degrade periods of multi-talk, either by transcoding/tandeming or by selecting only one speaker.

The remainder of the paper presents solutions for both of the above shortcomings of prior systems. First, the Tandem-Free Operation (TFO) conferencing architecture proposed by Burns *et al.* [7] and Rabipour and Coverdale [8] is briefly described. Then, the paper focusses on a new speaker selection algorithm suitable for the TFO conferencing architecture. The algorithm improves interactivity relative to FCFS by allowing interruptions, and results in less frequent switching relative to Loudest Talker (LT) algorithms. Performance is evaluated in terms of speech clipping, and subjectively through live conferences.

This work was supported by the Nortel Networks Wireless Speech and Data Processing group in Montreal, Canada.

0-7803-7549-1/02/$17.00 ©2002 IEEE.

2. TFO CONFERENCING

The TFO conferencing architecture [7, 8] uses centralized speaker selection, and decentralized decoding and mixing. The Tandem-Free Bridge (TFB) selects a primary and secondary talker, i.e., $M = 2$, from N input streams and forwards their compressed signals back to the other conferees. The primary speaker is sent the signal of the secondary speaker and vice versa, while the $N-2$ listeners receive both streams. Fig. 2 illustrates the approach. Deferring the decoding and mixing processes to the endpoints eliminates tandeming, thereby improving the speech quality.

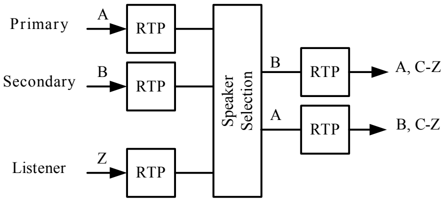

Fig. 2. Generic TFB. The selected signals can be sent in separate packets or can be bundled in one aggregate packet.

A novel feature of the TFO architecture is that the endpoints transmit speech activity features as additional fields in the upstream packets. This allows speaker selection to be performed *without* decoding. Thus, the bridge does not require speech codecs.

In terms of speaker selection, Burns *et al.* alludes to the use of FCFS, while Rabipour suggests using signal power and a hysteresis. The following section describes the latter approach.

3. PROPOSED ALGORITHM

The Multi-Speaker/Interrupter (MS/I) algorithm assigns talking privileges according to order of activity, the power signal envelope, $\hat{E}_i$, and a "barge-in" threshold, B_{th}. The input to the algorithm is the signal power, $\bar{E}_i$, and a VAD decision for the ith frame. $\bar{E}_i$ is carried as side information in the upstream packets (see Section 2). The VAD decision can be derived from either $\bar{E}_i$ or by the arrival of Silence Insertion Descriptor (SID) frames.

The hypothesis is that the envelope of the power signal will track the trend in conferees' activity better than the instantaneous frame power, resulting in less frequent switching. However, large increases in power—for instance, at speech onsets—should be followed closely, while decreases should be decayed slowly such that the conferee remains enabled during short fluctuations in energy (e.g., articulation pauses). Hence, $\hat{E}_i$ is calculated as:

$$\hat{E}_{i+1} = \max(\hat{E}_i, \beta\hat{E}_i + (1-\beta)\bar{E}_i), \qquad (1)$$

where $\bar{E}_i$ is the signal power of the ith speech frame, and β is the weight of the exponential average. A barge-in threshold is used to control spurious switching when the $\hat{E}_i$'s of two or more conferees' are close. A conferee of priority m can only preempt a conferee of priority $m-k$ if

$$10\log(\hat{E}_i^{(m)}/\hat{E}_i^{(m-l)}) > B_{th}, \quad \forall l = 1\ldots k, \quad k \leq m, \qquad (2)$$

where B_{th} is a "barge-in" threshold in dB. Note that interactivity decreases with increasing B_{th} and/or β.

The algorithm uses its own hangover mechanism in addition to the underlying VAD decision. In effect, the VAD's hangover is extended and the conferee's $\hat{E}_i$ is decayed exponentially for T_h s, after which time $\hat{E}_i = 0$. This allows the most recently active conferees a greater chance of being selected [9].

3.1. Algorithm Comparison

The performance of FCFS, LT, and MS/I were evaluated by driving the algorithms with actual speech traces recorded from four-person conferences. Speech clipping metrics were used as a basis for comparison. Two out of four conferees were selected for output, the switching interval was set to 20 ms to coincide with default RTP payload durations, $T_h = 1.5$ s, $B_{th} = 3.3$ dB, and $\beta = e^{t/\tau}$, where $t = 20$ ms, and $\tau = 50$ ms. Note that τ and B_{th} were manually tuned so as to minimize audible clipping/switching. Simulation results are summarized in Table 1, while Fig. 3 plots the selected speech of the primary talker (of a four person conference) for the three algorithms.

Since front-end clipping (FEC) is less noticeable than mid-speech clipping (MSC) (and back-end clipping (BEC)) [10], it is desirable for a selection algorithm to reduce MSC *without* preventing barge-ins altogether. Table 1 shows that MSC and BEC are greatly reduced over the LT algorithm, and FEC is greatly reduced relative to FCFS. Switching was, at times, clearly audible with FCFS, but was smooth with MS/I. In general, MS/I keeps a selected conferee enabled until the end of their talkspurt or until a loud, interrupting conferee breaks into the conference. The higher rate of switching for BEC is expected since new talkers tend to start talking when they anticipate that one of the current talkers might stop.

Table 1. Front-end, back-end, and mid-speech clipping due to speaker selection when selecting one, two, and three out of four talkers. Note that mid-speech and back-end clipping do not occur when using FCFS.

Type	Method	Speech Clipping Duration L (ms)			Percentage of Speech Clipped P (%)			Frequency of Clip Occurrence F (clips/min)		
No. Talkers		1	2	3	1	2	3	1	2	3
FEC	FCFS	410	300	218	14.6	5.4	0.9	7.7	3.8	0.9
	LT	108	64	31	3.8	0.5	0.0	7.6	1.8	0.2
	MS/I	152	98	78	5.7	1.3	0.1	8.1	3.0	0.3
MSC	LT	111	76	60	14.1	5.2	2.1	27.2	15.4	7.8
	MS/I	272	188	129	6.9	1.4	0.2	5.3	1.7	0.3
BEC	LT	246	167	153	19.4	10.9	5.8	16.8	14.5	9.1
	MS/I	304	193	188	14.1	3.7	0.6	10.1	4.1	0.9

4. SUBJECTIVE EVALUATION

The goal of the subjective comparisons was two-fold: (1) evaluate the transparency of the MS/I algorithm, and (2) compare the speech quality of a conventional VoIP conference bridge (with tandem connections) to that of a TFO conference (using MS/I). To this end, a PC-based conferencing test-bed was arranged on the LAN of the TSP Lab of McGill University. The Robust Audio Tool (RAT) [11] was the endpoint, and a TFB was built in software. The system could emulate a conventional VoIP conference bridge, a TFO conference, or a multicast conference.

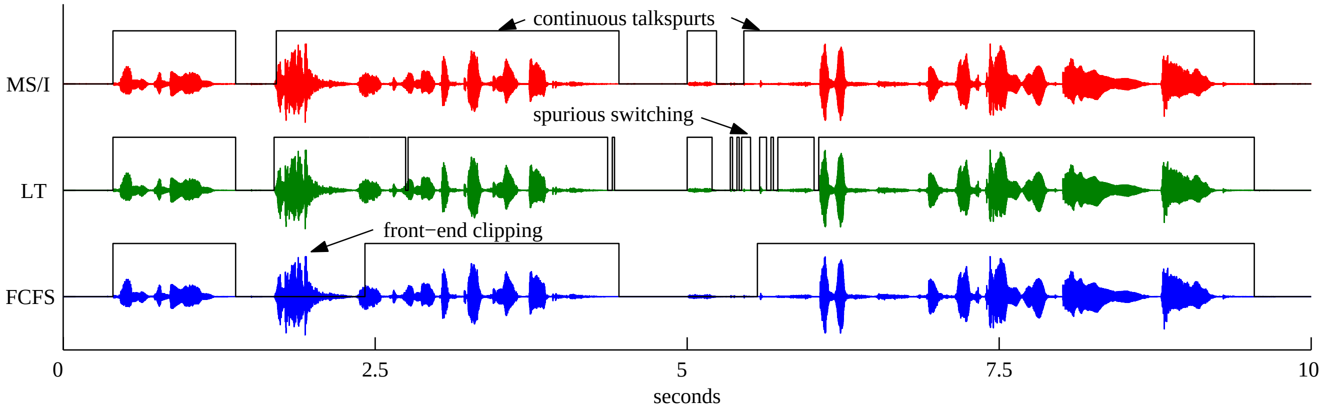

Fig. 3. Selected speech of primary talker using Multi-Speaker/Interrupter, Loudest Talker, and First-Come-First-Served.

G.729A was added to RAT and the experiments were run *without* silence suppression so that the effect of tandeming or speaker selection was not masked. Parameters used in the live conferences were the same as in the simulations, except two out of four speakers were selected when using MS/I. As in [1], the conferees' opinions were gathered through postconference interviews. A total of 12 listeners participated in the experiments (some more than once). Further details of the experiments are described in [6]. Table 2 shows a summary of the results.

Table 2. Summary of conferees' opinions and rankings.

Rank	System	M-of-N	Speech Codec	Speech Quality	Comments
1	VoIP Bridge Multicast	(4-of-4)	G.711	Good	
2	TFO-MS/I	(2-of-4)	G.729A	Good	Few Pops/Clicks
3	Multicast	(4-of-4)	G.729A	Good	More Noise
4	VoIP Bridge	(4-of-4)	G.729A	Poor	Less Intelligible

The MS/I algorithm was evaluated by comparing it to a multicast conference in which no tandeming or speaker selection was performed. This scenario was evaluated with three groups of four conferees using G.729A. Some experienced listeners could detect occasional pops when MS/I was used, which was likely due to codec state de-synchronization of the unselected streams. Overall, most conferees did not detect the presence of the MS/I algorithm.

Next, the quality of a TFO-MS/I conferencing arrangement was compared to a VoIP bridge. The test was carried out with two groups of four conferees using G.729A. As expected, TFO-MS/I was unanimously preferred. Conferees strongly felt that the speech quality of the VoIP bridge was poor and muffled-sounding.

5. CONCLUSION

A new method for speaker selection has been proposed for the TFO conferencing architecture. Preliminary testing suggests that the algorithm is nearly transparent to listeners, and that the TFO-MS/I system provides a significant improvement in speech quality over conventional VoIP bridges. The algorithm does not result in spurious switching, but does allow for interruptions. The former feature helps maintain synchronization between the encoder and decoder, and the latter improves interactivity. The algorithm can be extended to work in conferences with heterogeneous endpoints.

References

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

[2] D. Nahumi, "Conferencing arrangement for compressed information signals." United States Patent 5,390,177, Feb. 1995.

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

[4] D. Nahumi, "Delay synchronization in compressed audio systems." United States Patent 5,754,534, May 1998.

[5] 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," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Minneapolis,USA), vol. 2, pp. 435–438, Apr. 1993.

[6] P. J. Smith, "Voice conferencing over IP networks," Master's thesis, McGill University, Montreal, Canada, available online at http://www.tsp.ece.mcgill.ca, Jan. 2002.

[7] 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.

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

[9] M. A. Marouf and P. W. Vancil, "Method and apparatus for controlling signal level in a digital conference arrangement." United States Patent 4,499,578, Feb. 1985.

[10] J. G. Gruber and N. H. Le, "Performance requirements for integrated voice/data networks," *IEEE J. Selected Areas Communications*, vol. SAC-1, pp. 981–1005, Dec. 1983.

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