THE BENEFITS OF MULTI-SPEAKER CONFERENCING AND THE DESIGN OF CONFERENCE BRIDGE CONTROL ALGORITHMS

J.D. Tardelli, P.D. Gatewood, E.W. Kreamer, P.A. La Follette
ARCON Corporation, 260 Bear Hill Road, Waltham, MA 02154 USA

ABSTRACT

Conferencing over narrowband digital voice communication systems is limited to simplex broadcast and other limited modes because of system bandwidth, tandem limitations, and the inability of some algorithms to represent multiple speakers. Advances in narrowband algorithms show promise in representing multiple speakers. This research looked at the benefits to be gained with a multi-speaker narrowband conference capability and sets design parameters for such a system. It has provided subjective opinions of 18 conferencing systems by 16 subjects, audio recordings of input and output channels for each subject and system, and data files that detail the speech activity and channel switch settings for each subject and system every 50ms. An important conferencing objective or conference type effect has been found. That is, for different types of discussion, the speech dynamics and system requirements are substantially different. Various Speech/Bridge Dynamic states are evaluated. The percent time that the system is in these states is provided along with interval distributions.

1. INTRODUCTION

Speech conferences conducted over a telecommunications network can take several forms. Conferences conducted over narrowband digital systems normally use the simplex broadcast mode where only one conferee is in control at any time and no interruptors are heard. Thus the conference becomes a series of monologues. The limitations of simplex broadcast are imposed by the system's bandwidth, requirements to tandem speakers at the bridge, and the inability of some algorithms to represent multiple speakers. A solution to these limitations has been addressed by Champion[1].

The primary issues regarding narrowband speech conferences that are addressed by this work are as follows. Will the additional capabilities provide for a more efficient conference? How many additional speakers are needed for a significant improvement? What are the design parameters to be used by the conference bridge control algorithm? The answers will be dependent on the dynamics of conversational speech as affected by the conference mode used.

Research on the dynamics of two party conversational dynamics by Brady and others [2, 3] provides initial design parameters for speech activity detection but may not be applicable to the conferencing environment. Research into teleconferencing [4, 5] addressed various conferencing modes and used subjective questionnaires to measure users preferences. This work combines subjective measures with extensive data collection and analysis to provide an objective representation of conference dynamics. The study was performed on an analog testbed with speech detection at each source and dynamic channel switching at a bridge. No attempt was made to simulate the reduced bandwidth of a narrowband digital multi-speaker conferencing system.

2. DEFINITIONS

Conferencing modes are indicated throughout this paper as:

- Mode 1 SB Simplex Broadcast
- Mode 2 SI Simplex Broadcast w/ Speaker Interrupt
- Mode 3 MS2 2-Speaker Multi-Speaker Conference
- Mode 4 MS3 3-Speaker Multi-Speaker Conference
- Mode 5 MS4 Open Conference w/ Active Switching
- Mode 6 OC Open Conference w/ No Switching

A voice conference can also be classified by the objectives of the conference, such as direct one-way information exchange, problem solving, general discussion or directed discussion. These objectives can be represented by various types of conference tasks. Two types of tasks were chosen: problem solving and leader-led discussion. For the discussion, the amount of control exerted by the leader varied with the individual conferees.

The problem solving tasks (FF) were loosely based on the television game show, "Family Feud." They were designed so that the conference participants would be presented with a statement or question that had been presented to "...100 men and women in the studio audience." The conferees then were required to determine a specified number of alternative responses to the statement or question, rank the alternatives according to how the "studio audience" might have answered, and agree on a final listing.

Role Play scenarios (RP) were developed for the discussion task type. Each topic was chosen for its ability to generate discussion and to allow 4 discussion viewpoints or "roles". Each scenario included a general issue and/or discussion that was seen by all 4 members of the conference. After that, each topic included 4 opposing viewpoints that the conference members were required to adopt.

Three channel conditions were used with each mode and task type. There was a high quality (H) transparent channel, a medium quality channel (M) that introduced bandpass filtering from 100Hz to 3400Hz, and a low quality channel (L) that combined -12dBm white noise on the receive channels with the bandpass degradation of the medium state. A "system condition" is considered as a specific conferencing mode and channel condition.

3. EXPERIMENTAL DESIGN & TESTBED

The experimental design was based on 4 groups of 4 conferees each performing 2 conference task types for each of 6 conference modes. For each conference mode, the 3 channel

conditions were used. Each system condition was presented twice per conferee group.

The Factors important to the variance of the experiment were: Mode; Task Type (Task Version); Channel Condition; Presentation Order; Original/Repeat; Group (Subject (Room), (Sex)).

A conferencing testbed, Figure 1, was constructed that allowed 4 conferees to communicate from controlled environments over high quality transducers through a controllable switch matrix. Various transmission line deficiencies could be introduced at either side of the switch network. Synchronized digital audio recordings were made on DAT recorders at several locations in the network. A monitoring station was provided for the administration and control of the experiment. Signal activity was monitored on the individual transmit channels and provided to the bridge controller. The bridge dynamically set the switching matrix depending on the conference mode, switching history and signal activity on the various channels.

The testbed was calibrated using a 1KHz/114dB tone generator to eliminate clipping by the recorders and to provide the Equivalent Peak Level (EPL) of speech to be at +22dBm. The additive white noise of the Low channel condition was set 34dBm below the speech peaks.

4. ALGORITHMS

Speech Activity - Each conferee's station was connected to a spike discriminator that was set to count the crossings of the time domain signal above a threshold (+0.1V) with a 5ms sensitivity factor. This threshold is 38dB below the EPL of speech. This "# of firings" was read and reset to zero by the controlling PC every 50ms. The PC calculated a speech activity weight for each conferee that consisted of a single pole filter on the scaled # of firings. Separate thresholds were used for speech onset and offset detection. When offset was detected, the activity weight was reset to zero. The algorithm is:

```
W(n) = A[W(n-1)] + B[#firings],
where  A = .8 and B = 500.
Speech Onset for W(n) >= 2500
Speech Offset for W(n) < 25
```

Speech onset would be declared with 5 firings during the 50ms polling window. The offset threshold accommodated conversational pauses while recognizing final speech offsets. For a single onset followed by no firings, a speech offset would be declared with a delay of 1.15 seconds. The thresholds were fixed during the experiment. The logging of all activity data allowed post analysis with other thresholds.

Bridge Switching - The bridge switching algorithm, Figure 2, was the same for all conference modes except for simplex broadcast with interrupt. This was possible since simplex broadcast is multi-speaker conferencing with 1 channel, and for our case, open conferencing is multi-speaker conferencing with 4 channels.

The first come, first served criterion was used in the selection algorithm. When there were conflicts in contention for a channel, the speaker with the greatest activity weight was chosen. Once a conferee was assigned a channel, the channel was not lost until a speech offset was detected. For the SI mode, any channel set as interruptor had the highest priority to be switched to prime.

5. EXPERIMENTAL PROCEDURES

Subjects - Participants for this project were recruited to fulfill two primary requirements: reliability and speaking ability. Sixteen individuals were accepted as participants. Three groups were composed of 2 males and 2 females, while the fourth had 3 females and 1 male.

Training - A one day training session was held with each group to introduce them to the conferencing task types (FF & RP); the equipment; the communication modes and channel conditions. Details of the various system conditions were not discussed with the subjects.

The subjective rating sheet to be used by the participants for each system condition was discussed. This sheet consisted of 9 statements concerning the system just exercised. Each could be completed from a set of 5 ratings from which the conferee had to choose. The order of these ratings ranked the system from natural to unnatural, excellent to bad, relaxed to fatiguing, seldom to full time, etc. The topics covered in these statements were:

1. The level of effort required to participate.
2. How naturally were others talking.
3. Your level of participation.
4. System interference with the task (FF or RP).
5. Reasons for others to repeat themselves.
6. Overall conditions.
7. Reasons for loss of information.
8. Ability to talk naturally.
9. Overall conference quality.

Testing - Four test days were held with each group. Breaks were distributed to reduce fatigue. Task versions were randomly assigned to mode/conditions for each group. The presentation order of mode/condition was randomized within the sets of FF and RP tasks for each group.

6. RESULTS

Subjective - A correlation of the 9 individual subjective ratings made by each participant indicated that they were highly interrelated. Further, factor analysis suggested that only a single factor was being tapped. Closer examination showed 3 ratings clusters; overall system quality (Items 6,9), individual participation and perception of conferences (Items 1,2,3,8), and system impact on conference (Items 4,5,7). Each cluster was transformed to a composite variable (Quality; Participation; and System).

The subjective data were examined using Analysis of Variance (ANOVA) prior to collapsing across variables. The effects of Presentation Order, Group, Room, Task Type, and Condition for ratings were examined. ANOVAs indicated that there were no effects for Group or Room. There was a Presentation Order effect. This indicated a training effect. The ratings from the second presentation were felt to be more representative of the conferees true opinions, therefore, further analyses were conducted using only these data. Significance was determined using $p \leq .05$. A 2 (Task Type) $\times$ 6 (Mode) $\times$ 3 (Condition) ANOVA was conducted on each of the 3 composite variables.

For Quality, there was a significant interaction between Task Type and Mode, an indication that the two Task Types produced different ratings across Modes, reflecting reported difficulties completing the Family Feud task while in Modes 1 and 2. Main effects were indicated for Task Type, Mode and Condition. The Task Type main effect reflects a preference most participants had for Role Play. The Mode main effect reflects the differences in communicating with each Mode. The Condition main effect reflects the different degraded Channel Conditions which were imposed.

Results for System showed the same interaction and main effects minus the main effect for Condition, demonstrating

that the conferees realized that Channel Conditions were not system related, but felt they impacted overall quality.

Results for Participation showed a single significant main effect on Mode, suggesting that the only determinate for participation in the conference was the Mode. These relationships are presented in Figures 3 and 4 where 1 indicates a high rating and 5 indicates a low rating.

Objective - Speech activity for each conferee and the bridge switch settings were logged every 50ms allowing various states of Speech and Bridge Dynamics to be studied. The overall percent time and the distributions of time intervals in a state are good indicators of what was happening. The speech dynamic states that were studied include Single Speech, Multiple Speech, and Silence for the combined conference independent of mode. A mean of the individual conferees speech activity (Mean Speech) was also studied. The bridge dynamic states were studied but are not presented here. The interaction of speech and bridge dynamics is presented as the mean of the "Serviced" state for each conferee or its converse the "Ignored" state.

An analysis of the speech activity used to control the bridge switch would provide an inflated value for activity because of the required hangover time at the end of each speech segment. Postprocessing of the logged speech activity weights for each conferee allowed us to adjust the reported speech dynamics results, to allow for direct comparison with previous work [2, 3]. This did not change the switching dynamics of the actual conference experiments or the subjective data in any way. It did provide a realistic look at whether a conferee was or was not being serviced when speaking in a particular conference mode.

ANOVA was used to analyze the objective data prior to making decisions regarding collapsing across variables, particularly, to determine the effect of Presentation Order, Group, Room, Task Type, and Condition. There were no effects for Group or Room, therefore, data were collapsed across these variables. A 2 (Task Type) $\times$ 6 (Mode) $\times$ 3 (Condition) Repeated Measures (2) ANOVA was conducted on each of the variables using $p \leq .05$. With the exception of Mean Ignore, Task Type was the only significant effect for each of the variables. Mean Ignore results showed a significant interaction for Task Type and Mode. There were significant main effects for Task Type and Mode. Condition was not significant. These results demonstrated that the amount of communications did vary by task type, but neither Mode nor Condition had an effect. Figures 5 and 6 show the Speech Dynamics and the Speech/Bridge interaction as a function of conference Mode for the FF and RP.

The most interesting descriptor of the various states is the length of time that the state lasts or its interval. The cumulative distributions of these intervals for the various states are provided in Figures 7 - 12. They are plotted on log time axes to allow for comparison with previous results [2, 3]. The FF and RP task types are separated along with Mode. Onlymodes 1,3 and 5 are provided.

7. DISCUSSION

The subjective results indicate that users prefer a multi-speaker system. The interrupted simplex broadcast mode (M2) is not well received, especially for RPs. A 2-speaker multi-speaker conference (M3) does improve the user acceptance of the system for FFs and is equivalent to Modes 4-6 for RPs. The objective results indicate that the conferee's speech dynamics are not affected by Mode. Rather than adapting to the specific system, the conferee continues to try to be heard, even when the system condition does not allow it. This indicates that the design parameters used at the bridge must account for the Multiple Speech interval distributions presented in Figures 7 - 12 rather than the Mean Speech distributions. Of most importance is the Task Type effect. Since conference systems are multi-purpose, they must be designed for the worst-case application. This research has shown that a problem solving task is considerably more demanding on a conferencing system than the conversational material traditionally used to test and design them. An M3 conference provides improvement in the % time a conferee is serviced for both Task Types and does represent a good choice for the number of channels in a narrowband multi-speaker conferencing system.

Sponsored by RL/ERT Contract F19628-92-C-0042.

REFERENCES

[1] T. Champion, "Multi-speaker Conferencing Over Narrowband Channels," MILCOM 91 Conference Record, 1991 IEEE Military Communications Conference, pp. 51.6.1 - 51.6.4.

[2] P.T. Brady, "A Technique for Investigating On-Off Patterns of Speech," BSTJ, 44, Jan. 1965, pp. 1-22.

[3] P.T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," BSTJ, 47, Jan. 1968, pp. 73-91.

[4] J.W. Forgie, C.E. Feehrer, & P.L. Weene, "Voice Conferencing Technology Program: Final Report," Lincoln Laboratory, MIT, Lexington, MA, August 1977.

[5] D. Busson, N. Irizarry, & C. Stengel, "Secure Conferencing HF Communications: Final Report," RADC-TR-86-55, RADC, Air Force Systems Command, Griffiss AFB, Rome New York, April 1986.

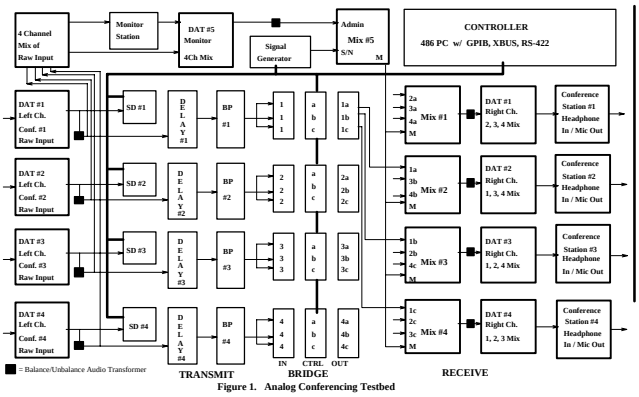

Figure 1. Analog Conferencing Testbed

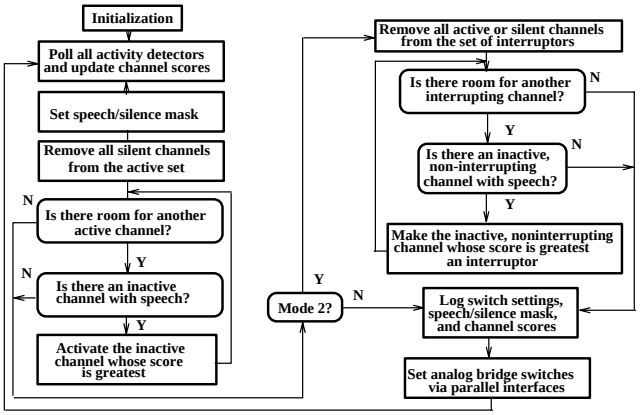

Figure 2. Conference Bridge Algorithm