A METHOD FOR COMPUTER PROCESSING OF SMALL GROUP VERBAL BEHAVIOR

STEVE R. HOWELL, KUTZTOWN UNIVERSITY, CHRISTOPHER KARPOWITZ, BRIGHAM YOUNG UNIVERSITY, TALI MENDELBERG, PRINCETON UNIVERSITY

ABSTRACT

We present a novel methodology and software tool for recording small group interactions, such as during deliberation or discussion, and automatically processing the resulting audio recordings to provide statistics on individual speaker and whole-group verbal behaviors. We discuss data from a large study using this technique.

INTRODUCTION

Many researchers are interested in phenomena that occur in groups, particularly in groups of conversing or debating participants (e.g. work teams in meetings, study groups, focus groups, legislators, juries). One logical approach to this work is to attempt to record the participants, using unobtrusive table or room microphones. However, while this kind of recording is useful as a record of the event, any kind of processing of the recording into usable data must rely on laborious human coding of the audio (due to the difficulty of source separation). This kind of human transcription can be done, and commercial services for this do exist. However, it is costly and time-consuming, putting it out of the reach of most investigators who are processing any significant number of meetings or deliberations. This is especially true if the researchers would like to record timing information of when utterances were spoken (which costs significantly more and takes significantly more time), or worse yet, if the researchers require identification of which actual speaker issued which utterance. With more than a few voices, it is not possible to reliably determine the speaker from audio alone. Even male versus female distinctions, which most transcription services will do, can be inaccurate, since the pitch of male and female voices can overlap. Distinguishing the speaker uniquely generally requires a video recording rather than an audio recording for processing (frequently even video recording from two angles to cover all potential speakers). Most researchers are not equipped to do this, and most importantly transcription companies are not generally equipped to process video but only audio.

We have developed an alternative methodology, which uses low-profile headset microphones to record individual audio tracks from each of 5 closely-seated participants in a round-table group deliberation environment. The audio tracks were recorded via inexpensive and easily available commercial multi-track recording equipment and software attached to and running on a nearby office PC. In the first experiment using this methodology over 155 groups were processed, which would have made human processing impractical. Hence, a user-friendly software application was created (Verbal Behavior Analysis Software v1.3) to analyze the individual audio tracks (.wav files), determine when that person was speaking, convert that information to human-readable output, and merge that information with data from the other speakers. This provides an integrated, time-coded data set for the verbal behavior of that group by individual speaker and utterance time. Furthermore, the software provides canned statistics for each group on the verbal behavior of the group and its participants, such as total utterance duration, mean duration of utterance, number of interruptions made, number of interruptions received, proportion of groups' total talk time contributed by this speaker, etc. This data can then be imported to statistical software for analysis alongside other variables that might have been recorded about the deliberative or discussion environment (e.g. experimental condition, gender, age, etc.) This software (and the accompanying recording configuration) was designed for use beyond this particular experiment, and is publicly available for research purposes from **steverhowell.com**

METHOD

Our method for recording small group interaction was first used for a joint research project (Karpowitz, Mendelberg, and Howell, 2009) between Princeton University (PU) and Brigham Young University (BYU), and data collection was performed at both sites.

Subjects met 5 at a time and engaged in small group deliberation about social justice issues around a round table. Each group of five deliberators was recorded using a total of 6 microphones and two separate digital video cameras. Five individual Shure low profile headset microphones were worn by the participants. The unidirectional cardiod pattern of these microphones helped eliminate any contamination of each speaker's audio by background noise and other participants' speech. The sixth microphone was an omnidirectional flat tabletop model. The microphones were connected to a MOTU 8PRE 8-channel microphone preamplifier. This preamplifier connected via a Firewire cable to a standard Microsoft Windows lab PC running Adobe Audition multi-track recording software.

A simple Microsoft Visual Basic 6.0 application was written (using the 'sendkeys' function) to automate the operation of the Audition software to ensure that recording was started on all channels at the same time, to name the channels according to experimental naming standards to ease data archiving and post-processing, and to copy the final files to a large network server disk drive for storage. The audio files are so large (often over a GB per group) that they would rapidly fill the hard drive of the recording PC.

Once the individual participants' audio channels were recorded, they were processed using Verbal Behavior Analysis Software V1.3 (See Figure 1) created by Dr. Steve R. Howell of Kutztown University. This software application first performed voice activity detection (VAD) on each channel. Each participant's audio was converted from an audio file (.wav file) to an amplitude data file (.amp) of average speaking amplitudes, by calculating the average amplitude of the speaker's voice during every .25 second interval of the recording. These averaged amplitudes for each speaker were then converted to binary on-off Voice Activity files (.vad). That is, if the amplitude for a .25 second interval for this speaker was greater than some minimum threshold, then their speaking status was set to 1 or ON for that interval, otherwise it was set to 0. [1]

This process yielded data files (.vad) for each subject with their speaking turns (utterances) identified. This data was then post-processed to ensure that slight pauses during utterances were bridged if they were less than 1 second in duration (to avoid have long single utterances broken into two shorter utterances by brief pauses). Then to avoid spurious short utterances due to microphone noise, etc., any of these utterances that did not contain at least one .25 second interval of some minimum high amplitude during the utterance were eliminated. [2]

Once all individual .vad files were processed, the software integrated them into a single group data file (.grp) for each deliberative group. Verbal behavior statistics were then run on this data, including such measures as total amount of speaking time for the group, % of time for this speaker, total number of interruptions of this speaker, total number of times this speaker interrupted someone, etc. These statistics were then integrated into the data set for the rest of the experiment, and included in analyses of the behavior of the groups under different conditions, gender balances, etc.

[1] The minimum amplitude threshold is a key parameter that must be set by inspection of the audio that has been recorded. In the present experiment, this threshold was set to 7. This was the value which produced the best VAD results based on comparisons to laborious human coding for speech activity (directly from the audio) for three test groups. See the Results section.

[2] For the present experiment, the ' minimum maximum' for an utterance was set to 10, since it was determined that 99% of test utterance sequences rose in amplitude at least once to the 10 level during the rise and fall vocal amplitude during speech. See the Results section.

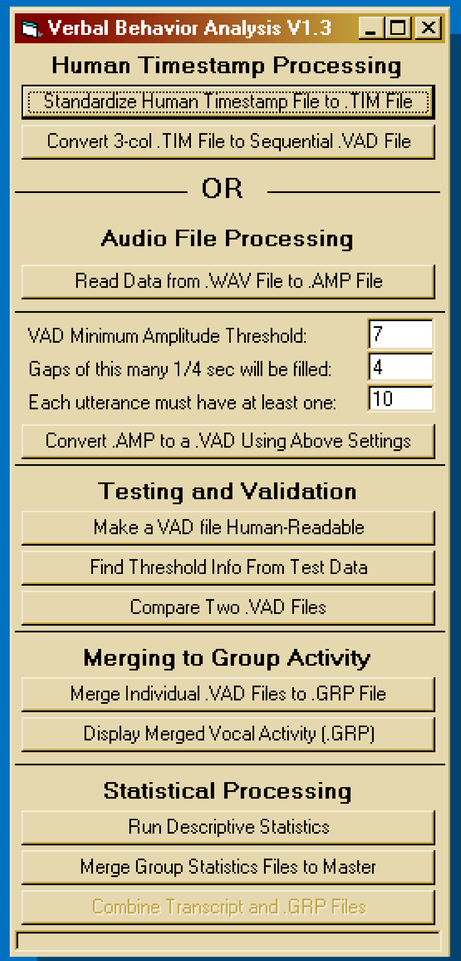

Figure 1: Verbal Behavior Analysis Software V1.3 Control Panel

Note top to bottom workflow design with one button per step. Also note dual points of entry to the process, either from audio files or from human-generated 'timestamp' files of when each speaker spoke and for how long. A utility to combine the verbal behavior analysis of the audio, including timestamps and speaker identity, with the human- transcribed actual text is presently under development and is disabled in this version.

RESULTS

Once the audio files were processed to amplitude files, Voice Activity Detection was performed using various test settings for the minimum amplitude thresholds. After each VAD, the resulting start-stop data was compared to a human-processed benchmark group for comparison. Results varied with different minimum amplitude thresholds, but the best overall threshold was an amplitude of 7. See Table 1 for accuracy results, over all quarter-second intervals, of the 15 minute to 45 minute long sessions. These results underreport the true accuracy, as qualitative analysis showed that most utterances were recognized via both manual and software VAD processes, but the start and stop times differed slightly.

Table 1: Accuracy for Match of Software VAD to Human VAD (by ¼ sec)

Group	Matches %	False Alarm %	Correlation
Group 1-005	0.732	0.1814	0.6914
Group 1-010	0.803	0.3205	0.6313
Group 1-036	0.774	0.3434	0.6554

DISCUSSION

• The Shure headset microphones produced sufficiently good recording quality that the isolated speaker was clearly distinguishable from background noise and from other speakers around the table. However, recording volume must be kept low enough that the primary speaker is picked up, but not the others in the group.

•The best quality, hypercardiod headset microphones that can be obtained should be used with this method, as the greater the elimination of background noise in each recording, the more accurate the automated VBA processing can be and the lower the false alarm rate in voice detection. Other than that, any multitrack recording software should work.

•It is essential that a test set of groups be manually processed for voice start-stop times on each audio track. This benchmark information is used when calibrating the settings of the VBA software ("Find Threshold Information From Test Data" button) to find the most accurate threshold values. In the initial data collection, recordings from PU were much more accurately processed than those from BYU. This was due to the fact that the microphone recording volume settings differed between the labs, and that the human-processed benchmark groups were all from PU. 3 groups were manually processed out of 155 total for that data.

•Since some inaccuracy remains due some groups being louder or quieter than others, we recommend looking only at relative differences between experimental conditions, not absolute results, for automatic verbal behavior analysis.

REFERENCES

•Karpowitz, C., Mendelberg, T., and Howell, S. R.(2009). Deliberation, Gender, and Speaking Behavior. *Proceedings of the 2009 Conference of the American Political Science Association, Toronto, Ontario.*

Presented at the 22nd Annual Convention of the Association For Psychological Science, Boston, MA, May 27 – May 30, 2010.