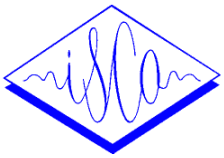
ISCA Archive
http://www.isca-speech.org/archive



7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

THE ISL MEETING CORPUS: THE IMPACT OF MEETING TYPE ON SPEECH STYLE

Susanne Burger, Victoria MacLaren, Hua Yu

Interactive Systems Laboratories
Carnegie Mellon University, USA
sburger@cs.cmu.edu

ABSTRACT

Speech research is becoming very interested in new application domains, such as meeting summarization and automatic transcription, and has thus begun to work with recorded meeting data. The following paper gives an overview of the meeting data collection at Interactive Systems Laboratories. There are currently over 100 meetings of different types recorded. An experiment is described which aimed at testing the possibility of controlling issues of speaking style by meeting type. Results show that depending on the meeting type, speaking style varies in terms of turn length, speed and disfluencies.

10.21437/ICSLP.2002-140

1. Introduction

Databases containing meetings are becoming more attractive to speech research and technology. Automatic speech recognition is finding a new challenge in the recognition of real spontaneous speech in multi-party conversation, with new tasks such as automatic transcription and summarization of meetings in mind [1].

Collecting meetings gives also a new task to speech data collection: first, to generate many topical types, which will facilitate a range of speech styles, domains and technical settings in a controlled environment; the second, to gather many meetings within these topical archetypes.

The Interactive System Labs of CMU, Pittsburgh has collected meetings since 1999. The database currently consists of more than 100 diverse meetings[1].

In the following, we will give an overview of how meetings are recorded (scenario, participants, environment, equipment) and prepared (transcription) at our lab. We will briefly describe the data we have collected to date. Finally, we will use statistical results to demonstrate how speaking styles change by choosing a variety of scenarios and situations.

2. Meeting Data Collection at ISL

We define a meeting as a minimum of three individuals speaking to one another. A meeting recorded at ISL results in a maximum of eight mono audio files in WAV format, so-called speaker and recording protocol files containing information about the participants, equipment, environment and scenario, three video tapes, one transcription file of the entire meeting, a so-called marker file containing begin and end time stamps for conversation contributions, and a list of the meeting's vocabulary.

2.1. Recording

Different aspects can influence the type and quality of a meeting. Since we are trying to collect meetings in a controlled manner, we record according to the following variables: meeting scenario or topic, meeting participants, recording environment and recording equipment.

2.1.1. Meeting Scenario

A meeting type can be controlled by a given scenario or topic, or meeting parties are invited to discuss a prearranged topic.

Creating the scenario provides the opportunity to satisfy specific research needs. For instance, to satisfy the needs of emotion recognition, we can create a scenario that may invoke anger or excitement, e.g. a controversial discussion about a political issue. Furnishing the scenario limits the spoken vocabulary to a specific domain. Giving a military strategy situation to military personnel provides specific military vocabulary, acronyms and jargon.

We experimented with the following meeting types: *Project/Work Planning, Military Block Parties, Games, Chatting, and Topic Discussion*.

Project/Work-Planning: The participants either planned a project or discussed work. Projects led to a vocabulary oriented to that project; work meetings had a more varied vocabulary because different projects were planned. We recorded research groups, research project meetings, student projects, product development groups and our own data collection work meetings. Participants were generally colleagues.

Military Block Parties: Block parties are strategic exercises. Military personnel pretend to be in a real combat situation and have to solve special tasks communicating via radio or in a special room. Participants did not always know each other beforehand and communicated in a military formality. The environment often included more than one room. Sound quality, therefore, was often poor.

Games: Game-like tasks that had to be completed within certain time constraints were assigned to a group. Tasks have included building an object, designing an advertisement, and making an unanimous purchase decision. Conversation during more conventional games was also recorded, such as board games and cards. The participants were colleagues, but often knew one another casually. The used vocabulary was dependant on the game's task.

[1] The authors would like to thank the ISL data collection team, especially Robert Isenberg, Denise Hill, Debra Vlasak, Raina Jones and all the meeting participants.

Chatting: Participants were seated together and left alone. The group gossiped, discussed a mutual interest, or swapped stories. *Chatting* differs from the other meeting categories in that the group was not provided with a planned scenario or task, but was rather asked to develop one. Participants were friendly with one another. The vocabulary was unpredictable.

Discussion: For *discussion* meetings, the group was provided with materials to promote a group discussion. Materials have included journal and news articles, video documentaries, and erotic advertisements. Often, a mediator is present or questions provided. Participants did not always know each other beforehand. They where chosen because of a special affiliation (e.g. nationality, gender) or opinion. The vocabulary was determined by the topic but expanded often into a more emotional vocabulary.

2.1.2. Participants

Researchers and students from our lab, project partners, and staff were recorded during their own meetings or acted as participants in scenario meetings. An on-campus flier campaign was also very successful in finding participants. In return for free food and a room, groups began to hold their meetings in the ISL lab and returned regularly.

We recorded student meetings, staff meetings, and block parties with active and retired military personnel. The age range, therefore, of the participants is between 18 and 70, with a focus on 20-30.

The participants have primarily been native speakers of American English, but have also included non-native speakers from various countries, such as Germany, France and China. All meetings were recorded in American English, but due to the origin of some speakers, some meetings have a larger portion of accented speech; even foreign speech has occasionally appeared.

To eliminate the issue of security and privacy, meeting participants were asked to sign a waiver. Participants were given the opportunity to review the transcript and audio and revoke this waiver at any time. Material deemed sensitive and private, as well as anything which may reveal the identity of a speaker, was made unrecognizable in the audio and eliminated from the transcription. A meeting can be shared with other scientific parties only if all meeting participants sign this waiver.

2.1.3. Environment

Almost all meetings were recorded in the ISL lab. Two carpeted, free standing walls and a Smart Board separated a section of the lab. The separated room was equipped with projectors, a TV and VCR, which could be used during the meetings. A round table with space for 10 individuals stood in the middle. The background noises in this environment are very similar to a quiet cubicle office environment: PCs running, fluorescent lighting, air conditioning noise, and sometimes people having background conversations. Recording personnel and equipment were hidden behind a screen throughout recording for increased privacy.

2.1.4. Equipment

As many as 8 microphones were fed through an Alesis mix board connected to a PC equipped with an ECHO Layla card. This Layla card allows 8-channel recording directly onto hard disk, eliminating the need for later synchronization or down sampling. Audio was recorded at 16kHz, 16Bit.

We recorded primarily with wired lavaliere microphones, but also experimented with table microphones and wireless microphones. Each speaker wore one lavaliere microphone and was recorded on an individual channel. Up to three video cameras, depending on the size of the group, recorded the meeting from multiple angles.

2.2. Additional Information

A database of demographic information was maintained regarding each meeting participant. Using an online database interface, each speaker was asked to provide mandatory information, such as gender, occupation and birth date, and optional information, such as height and weight. Additionally, participants provided data on the origin of their accent, specifically the city, state and country where they were born, grew up, and have lived the longest. Comments were sometimes made regarding the quality of a speaker's voice.

In addition to the demographic information, information regarding each meeting was collected through the same online database interface. Among the collected data for each recording were the scenario, recording date, medium, and location.

2.3. Data Preparation

Each meeting is transcribed. Transcriptions are completed in accordance with the VERBMOBIL conventions, which offer an established method, labeling set and the tools necessary for transcription and turn segmentation.

The transcription tool is called TransEdit, which was developed for the transcription requirements of the ISL lab. Aside from its pension for easy editing, it has the ability to display multiple audio files in parallel, which was indispensable in following meeting content and marking the beginnings and ends of turns on all channels.

Transcriptions are initially completed on a first pass level and then checked by another transcriber. In the second stage, transcription undergoes a so-called privacy check in order to eliminate any sensitive or personal information from the transcription.

2.4. Status

104 meetings have been collected so far, generating a combined total of 103 hours (4.3 days). Each meeting lasted an average of 60 minutes. The recorded audio (since every speaker had his/her own channel) is 588.5 hours in 552 wav files, 77430.5 MB of data. The meetings have an average of 6.4 participants.

45% of the meetings have been completely transcribed and checked. For 18 of these meetings, all participants signed the waiver and privacy checks have been completed. 19 meetings are in progress, 10 of which will also become shareable. The remaining 38% of meetings are currently scheduled for transcription, of which 24 have a complete set of signed waivers. Eventually, almost 50% of the meetings will become shareable data.

3. Experiment

Future meeting data collection can profit from knowing which kind of meeting type provides which features of speaking style. In an experiment, we define speaking style as a combination of features, such as length of speaker contribution, the number of word tokens per contribution, the sum of used sentence type (question or non-question) and the amount of disfluency. The experiment aimed at testing the possibility of controlling issues of speaking style by meeting type.

3.1. Data

At the time of the experiment, transcriptions for 42 meetings were completed and checked; these 42 transcriptions were included in the experiment. We counted the following variables in the transcription per speaker:

- *General*: word tokens, turns, duration of overall talk time per speaker
- *Short/long turns*: turns containing more than thirty words (wlon), turns containing less than one word, turns longer than 10 seconds, turns shorter than 0.7 seconds
- *Question/Non-question/Turn breaks*: turn breaks (interrupted turns), question marks, and periods
- *Disfluency*:
  Non-grammatical events: false starts, repetitions / corrections
  Pauses in Speech: human noise, empty pauses, breaths, filled pauses, laughs
  Articulatorical breaks: interrupted words

Short and long turns and turn breaks were calculated in accordance to the entire number of turns. The rate of questions and non-questions and the percentage of non-grammatical phrases were calculated according to the sum of counted question marks, periods and turn breaks. All other disfluencies are percentages of the total number of counted word tokens per dialogue.

3.2. Categories

According to meeting type, the meetings were categorized into: *Chatting* (4 meetings), *discussion* (11 meetings), *game* (7 meetings), *project planning* (7 meetings) and *work planning* (14 meetings). While project planning and work team meetings may appear synonymous, the difference in vocabulary and participant relationships was significant enough to categorize them separately.

3.3. Results

Speaker contributions (turns): The highest number of turns per minute was expressed in *discussion* (4.7 turns/minute); the lowest number of turns was in *project planning* (3 turns/minute). The most word tokens per minute were also expressed in *discussion* (42), the lowest number of words per minute in *game* (29) (see fig 1).

Project had the highest percentage of turns longer than 30 words or 10 seconds (12.8% word number, 10.5% duration). The lowest percentage of long turns was found in *chatting* (6.3% word number, 5.2% duration). *Work planning* had the highest occurrence of turns either containing just one word or shorter than 0.7 seconds (40.2% word number, 33.8% duration). The lowest percentage of short turns was found in *game,* in case of word amount (28.1%), and in *chatting,* in case of duration (16.8%). (See fig. 2).

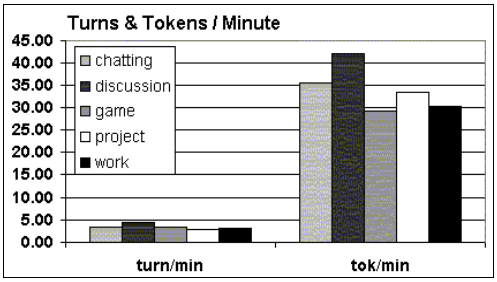


Fig 1: Turns and tokens (tok) per minute

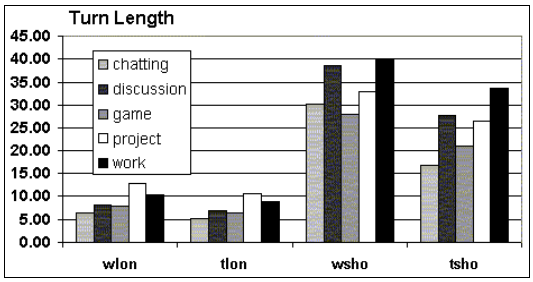


Fig 2: Turn length: wlon: more than 30 tokens, tlon: over 10 sec, wsho: 1 token, tsho: 0.7 sec or less

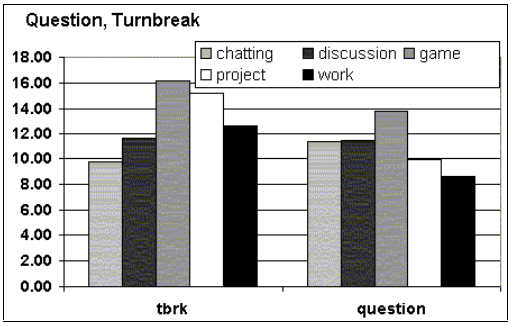


Fig 3: Questions and Turn breaks

The highest percentage of interrupted turns was found in *game* (16.1% of all turns), the lowest percentage in *chatting* (9.7% of all turns). The distribution of questions and non-questions showed most questions in *game* (13.8%), and fewest in *work* (8.6%). Non-questions (ended with a period in the transcription) expressed the opposite, i.e. the highest percentage in *work* (82.2) and the lowest percentage in *game* (74.7). (Fig. 3)

Summarizing all disfluencies (non-grammatical phrases, pauses in speech and broken words), the most disfluencies were found in *game* (22.4% per token), the fewest in *project* (16.6%). After examining only the non-grammatical phrases (false starts, repetition or correction), *game* again demonstrated the highest percentage of occurrences (7.6%), while *project* again had the lowest (4.3%). The particular results for pauses in speech of any kind (empty pauses, filled pauses, breathing,

human noise) showed the highest number of pauses for *discussion* (12.7%) and the lowest for *project* (10.5%). An especially interesting element was annotated laughing, which occurred most often in *discussion* (3%) and the least in *project* (1%). (See Fig. 4)

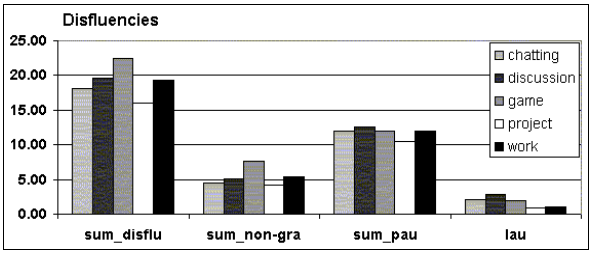


Fig 4: Disfluencies: disflu: all types, non-gra: repetitions, corrections, false starts; pau: empty and filled pauses, breath, human noise; lau: laughing

4. DISCUSSION

The results of the experiment support that there are differences in some aspects of speaking style between the meeting types. In Fig. 5, we ranked the speech style features (5 being the highest frequency of occurrence, 1 the lowest). This broad ranking allows for highlighting patterns of distinct features in different meeting types:

Chatting has remarkable "fast-talking", measured by the number of turns and the-word-per-minute rate, even despite the significant amount of pausing and laughing. Turn length fluctuates between 1 and 30 sec; meeting participants are of the same level, as there is neither a main speaker who has the longest talking time, nor is there an abundance of "uh-huh' affirmations. But one should not expect *chatting* in general to demonstrate very much controversy, as it is more a general exchange of experiences or stories.

Discussion shows an even higher ranking for turns and words per minutes than *chatting,* despite of a large amount of pauses and laughing. We found an increase of short turns, where we observed more "uh-huh" contributions. In a discussion, people may want to show agreement with other opinions. Longer turns ranked third (see table 5). A reason for that may be the involvement of mediators, who gave statements for discussion. Mediators also caused the high rank for questions by putting questions into the discussion. Contrary to our expectations, we discovered fewer turn breaks where participants interrupted each other.

Game demonstrated the lowest ratio of words per minute. In *Game*, we found similar results for long and short turns as in *Chatting*: more turns consisted of a medium length. Here we found the highest number of turn breaks and questions. These questions cannot be attributed to the rules of the game itself, as most of the games involved in the experiment were of a role-playing or task-solution genre. People often made their suggestions in the form of questions. The large number of turn breaks reflects the given time limit. Therefore, participants seemed to try save time by not waiting until other participants finished their contributions.

For the category *work,* we found a low ratio of words or turns per minute. The reason for the turn rate is an inflated number of longer turns. *Work* also demonstrated the largest number of short turns. All work meetings included one or two supervisors. These participants had long turns, accompanied by a lot of short affirmations from their respective groups. We found a significant number of non-grammatical phrases and few episodes of laughter.

The *project* meetings also combined a low rate of turns per minute with a high incident of long turns. Also here, members of the project management gave instructions and explanations while the group confirmed by short affirmative turns. Turn breaks occurred frequently. Similar to *game,* most project meetings had a time limit. Additionally, project meetings were the ones with the most participants. Therefore, participants interrupted each other more frequently to save time.

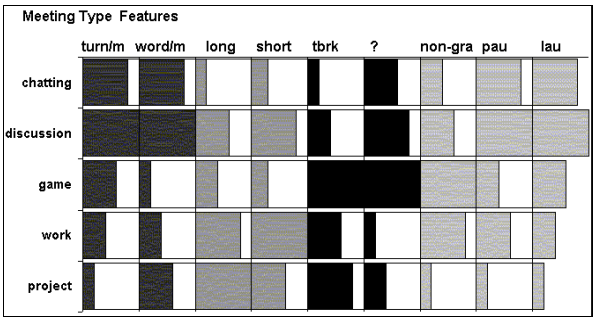


Fig 5: Feature ranking for turns/min, word/min long and short turns; turn break (tbrk), question, non-grammatical, pauses and laughing

The broadest variance in all variables was found for *project*. *Chatting* and *work* showed the highest consistency. The reason is that the category *project* contained meetings conducted by different groups, while *chatting* and *work* almost always consisted of the same group of people. Despite the interesting point that lots of pauses in speech and laughing did not decrease the high ranks of speed in chatting and discussion, disfluencies seem not to be a real feature of meeting type. This feature is too dependent on individual speaker behavior. Additionally, there was not a significant difference in overlapping turns between the meeting types.

5. Conclusions

We asked whether it is possible to describe what speaking style features will be present depending on a recorded meeting type, but the question cannot be answered by categorization alone. While there are some aspects of speaking style that are clearly indigenous to special meeting types, a categorization of meeting participants would also be necessary (age, relationship to each other). Future work would entail an evaluation of the found results, which would automatically recognize the meeting type upon recognition of special features. Other variables, such as participant type, or meeting structure will also be subject to future research.

6. References

[1] Alex Waibel at al: Advances in Automatic Meeting Record Creation and Access ICASSP 2001, Salt Lake City, May 2001.