Prosody and Spoken Dialog Systems

ELIZABETH NIELSEN

LING 575

What is prosody?

Phonetic level:

Pitch

Length

Volume

Phonemic level:

Tone

Stress

Phrasing

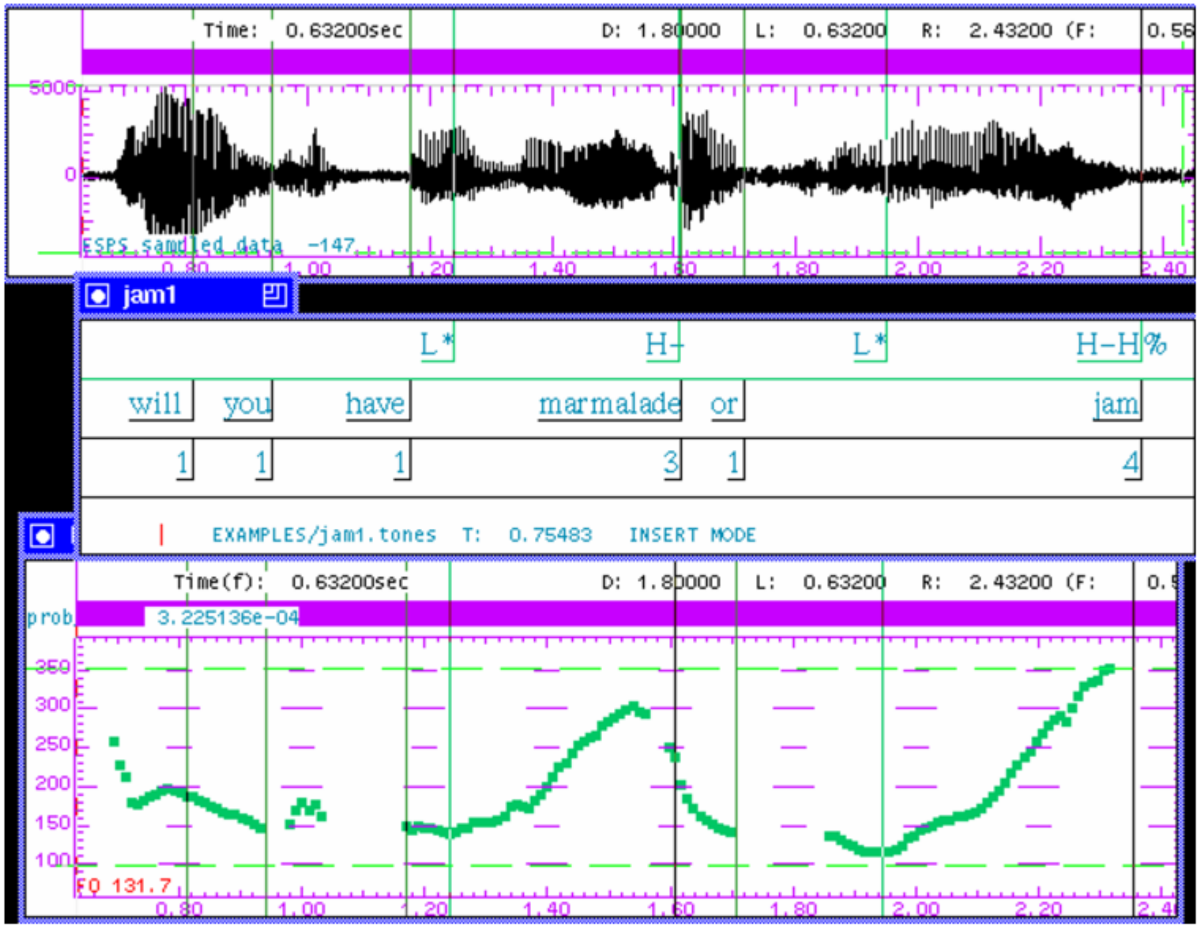
Examples:

English stress + pitch accent

Mandarin lexical tone

Japanese LPA

Annotating prosody



ttp://www.speech.cs.cmu.edu/tobi/ToBI.1.html

… or not

Shriberg and Stolcke (2002):

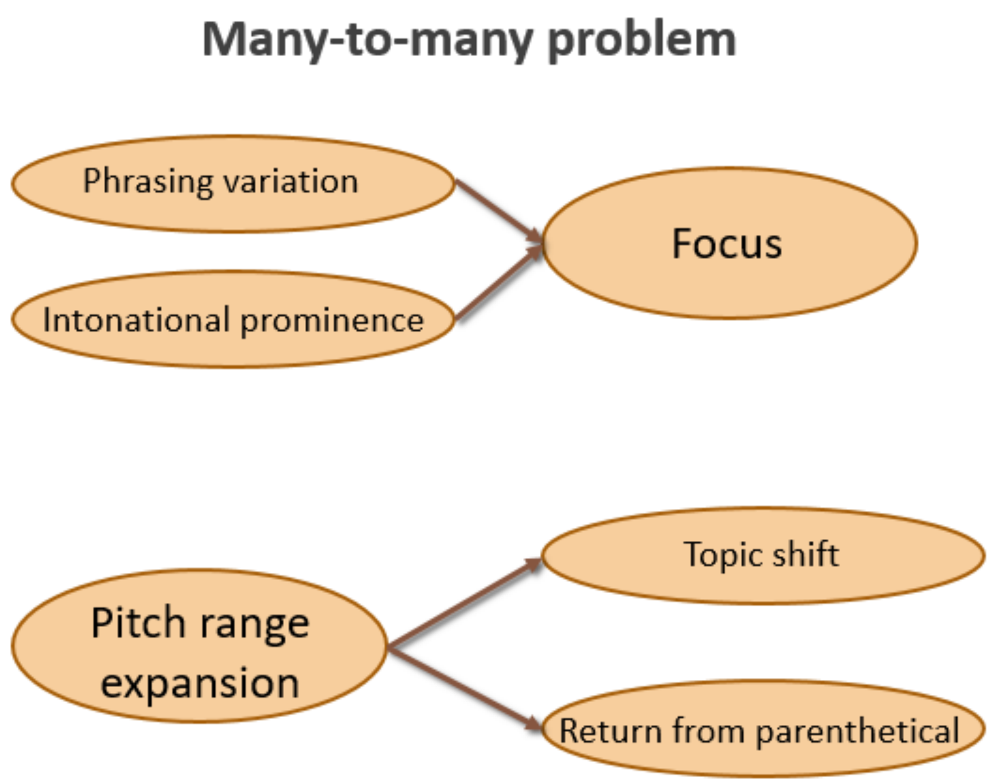
Problems with hand-annotation

- Interannotator unreliability (e.g. L*+H vs L+H*)
- Cost
- Must guess the correct level of granularity

Instead:

- Force align a transcribed text
- Extract features:
  - F0, pauses, segment duration, rate
- Learn which features matter for the task

Functions of prosodic variation



Functions of prosodic variation

Contour variation
- Syntactic mood
- Speaker attitude and beliefs
- Turn taking

Pitch accent location/type variation
- Focus
- Pronoun resolution
  - “Joe laughed at Bill and then he hit him.”

Phrasing variation
- Scope disambiguation
  - “I don’t travel by ship because I’m too cheap.”

Timing and pitch range variation
- Can result in phonemically different contours
  - Rise-fall-rise (L*+H L-H%)

Applications

Dialog act recognition – Kornel Laskowski and Elizabeth Shriberg (2010)

- Can't always have access to text
- Prosody + information about who is speaking when = almost as good as textual information

Improving ASR – Elizabeth Shriberg and Andreas Stolcke (2002)

- Sentence segmentation (better than LM alone)
- Dialog act recognition
- Topic segmentation
- Disfluency detection
- Word recognition

Improving TTS – Sridhar et al.

- Automatic dialog act tagging

Potential applications

TTS:

- Improve paragraph-length production (rate, pitch range variation)
- Use better parsers – get better phrasing
- Improve systems that try to model givenness
- Apply research about prosodic correlates of emotion
- Improve confirmation, turn taking in SDS
- Concept-to-speech applications?

ASR:

- Identifying salient information for NLU systems

Discussion questions

Why don’t more deployed systems make use of prosodic information?

Karen’s question about Shriberg and Stolcke (2002).

- “What happens when you use other types of classifiers?”

What about variation by dialect?

What's a real-life application scenario in which topic segmentation is done? Can it be analyzed as a special case of sentence segmentation?

References

Syrdal, A. and McGory, J., “Inter-transcriber reliability of ToBI prosodic labeling,” Proc. of the Intl. Conf. on Spoken Lang. Proc., Beijing: China, 235-238, 2000.

EMOTION DETECTION IN SPOKEN DIALOG SYSTEMS

Anna Gale

Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog

- Ang, et al.
- Which features have the most influence on detecting annoyance and frustration?
- Focus on prosody because it's not just what people say but how they say it that often indicates emotion of the speaker
- Use naturally-occurring instances of annoyance and frustration and fully automatic system

Approach

- Corpus: DARPA Communicator project (simulated travel plans)
  - 21,899 utterances
  - Labelled by 5 students with 7 possible emotion labels
  - Also labelled for speaking style, repeated errors/corrections, and data quality
- Study looks at Annoyance+Frustration vs. Else and Frustration vs. Else
  - Frustration = "extreme cases of anger"
  - Critique: poor terminology
- Look at prosodic features and language model features

Features

- Prosodic Features
  - Duration: max and average durations of phones
  - Speaking rate: # of vowels / duration of utterance
  - Pause: ratio of speech to pause time, duration of longest pause, number of long pauses
  - Pitch: min and max pitch (F0)
  - Energy: max or average RMS energy
  - Position of utterance in the dialog
  - Repeats and Corrections
- Language Model Features
  - Difference of log likelihoods of the two classes (Ineffective)
  - Sign of the log likelihood difference between the two classes

Results

	ANNOY.+FRUST. vs. ELSE				FRUST. vs. ELSE			
	True words		ASR words		True words		ASR words	
	Acc	Eff	Acc	Eff	Acc	Eff	Acc	Eff
Each human with other human, overall	72.6				68.8			
Human with human “Consensus” (biased)	83.9				77.3			
Consensus version, [All Features]	80.2	32.7			93.2	67.2		
Originally agreed, [All Features]	85.4	47.2			91.8	63.3		
Consensus version, [no STYLE] (“Baseline”)	75.2	21.2	75.1	21.9	86.4	46.5	87.0	49.5
Originally agreed, [no STYLE]	80.0	32.0	78.5	28.2	86.4	44.6	85.7	46.9
Consensus version, [no STYLE, no REP]	71.1	14.6	70.7	14.8	84.2	39.7	86.7	47.9
Originally agreed, [no STYLE, no REP]	77.1	23.0	74.5	18.6	80.4	31.8	83.6	39.6
Consensus version, [REP *only*]	69.8	12.8			76.6	21.1		
Originally agreed, [REP *only*]	74.7	18.5			85.4	14.3		
Consensus version, [LM *only*]	65.6	3.8						
Originally agreed, [LM *only*]	64.5	-0.9						

Results

- Most valuable features
  - Duration and speaking rate
  - Pitch
  - Repeats/corrections
- System does better with frustration vs. else than frustration+annoyance vs. else
  - However, small sample size so cannot draw firm conclusion

On NoMatchs, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection?

- Schmitt, et al.
- How do you detect anger with more than just acoustic and prosodic features?
- Acoustic features can be misconstrued
  - Ex. Loudness variation can indicate anger or it can be caused by technical problems
- Requires thinking about what indicators of anger exist in a conversation

Approach

- Corpus from an automated agent for internet-related problems
  - 1,911 calls
  - 22,724 utterances
  - Labelled angry, annoyed, non-angry
  - 22.4% of calls contained an angry or annoyed utterance
- Look at angry vs. non-angry and angry/annoyed vs. non-angry
- Look at both acoustic and non-acoustic features

Features

Two sets of features:

- **Acoustic**
  - Power
  - Mean
  - Rms
  - Mean harmonicity
  - Pitch
  - Voice Pitch
  - Intensity
  - Jitter Points
  - Formants 1-5
  - MFCC 1-12
- **Non-acoustic**
  - *ASR*
    - Utterance (unigram bag-of-words)
    - ASR confidence of transcription
    - Barged in: caller began speaking before prompt finished
    - Successful/Unsuccessful (NoInput or NoMatch)
  - *NLU*
    - Semantic parse
  - *Dialog Manager*
    - Last automated agent prompt
    - Number of tries to elicit desired response
  - *Context*
    - Number of help requests by the user
    - Number of operator requests by the user
    - Number of NoInput/NoMatch/BargeIn events

Results

Test A: Angry/Annoyed vs. Non-angry	only Acoustic	only Non-Acoustic	both
Accuracy	70.29 (+-2.94) %	61.43 (+-2.75) %	72.57 (+-2.37) %
Precision/Recall Class 'Ang./Ann.'	71.51% / 61.57%	68.35% / 42.57%	73.67% / 70.14%
Precision/Recall Class 'Non-angry'	69.19% / 73.00%	58.30% / 80.29%	71.57% / 75.00%
Test B: Angry vs. Non-angry	only Acoustic	only Non-Acoustic	both
Accuracy	87.06 (+-3.76) %	64.29 (+-1.32) %	87.23 (+-3.72) %
Precision/Recall Class 'Angry'	87.13% / 86.55%	66.0% / 58.9%	86.88% / 87.11%
Precision/Recall Class 'Non-angry'	86.97% / 87.53%	62.9% 69.9%	87.55% / 87.33%

Results

- 2.3% improvement in accuracy when including non-acoustic features
- Most relevant feature: audio duration
- Including Emotional History did not improve all test results

Discussion

- Are these ideas feasible to implement?
  - Difficulty of getting a labelled corpus
  - Schmitt, et al. uses already existing pieces of the system (ASR, Dialog Manager)
- Effect of implementing these techniques in dialog systems
  - System that detects anger can improve customer experience
- Are these approaches motivated by modeling human conversation or engineering considerations?
  - Emphasis on which features are most effective statistically

Turn-Taking and Backchanneling

Travis Nguyen
Prof. Gina-Anne Levow
LING 575
May 2, 2017

Agenda

- Turn-taking
- Backchanneling
- Proposed systems
- Discussion

Turn-Taking (1 of 2)

- Manner of conversing in which two or more participants speak one at a time
- Includes how to:
  - Contribute
  - Respond to previous utterances
  - Transition to another participant
- Linguistic and non-linguistic cues

Turn-Taking (2 of 2)

- Highly variable
  - Dependent on factors such as gender, culture, modality, etc.
  - Examples
    - Californian English
      - Non-final sentence in an utterance has rising tone
      - Final sentence in an utterance has falling tone + creaky voice
    - Eye gaze in Deaf community (United States)
      - Eye gaze towards participant indicates that it is participant’s turn to speak
    - Hand placement in Deaf-Blind community (United States)

Violation in Turn-Taking Rules

- Forcibly ending participant's turn
- Linguistic and non-linguistic cues
  - Linguistic
    - Interruption
  - Non-linguistic
    - Eye gaze (in spoken languages)
- Also dependent on factors such as gender, culture, modality, etc.
  - Examples
    - Male vs. female discourse
    - Italian-American culture
      - Overlap not intended as interruption

Backchanneling

- Receiver indicates to speaker that they are listening
- Linguistic and non-linguistic cues
  - Linguistic
    - Continuers (“uh-huh”)
    - Assessments (“No way!”)
  - Non-linguistic
- Also dependent on factors such as gender, culture, modality, etc.

Proposed Systems (1 of 2)

- Modeling turn-taking phenomena as taxonomy [Khouzaimi et al. (2015)]
  - Modeled on turn-taking phenomena (TTP) that humans employ
  - Each TTP modeled on two criteria:
    - Quantity of information Giver (speaker) has injected
    - Quantity of information that Taker (receiver) tries to add by taking the stage
  - Examples of taxonomy labels
    - Complete, incomplete, incoherent, insufficient information

Proposed Systems (2 of 2)

- Modeling backchanneling using a regression-based approach [Terrell et al. (2012)]
    - Experiment
        - Employed 48 people and coupled them
        - Assigned each person in dyad one role (narrator, listener)
        - Recorded videos of interactions
    - Results
        - Speech and eye gaze significant predictors of addressee backchannels
        - Pitch variability more significant than previously thought

Discussion

- How can spoken dialog systems account for the high variability in turn-taking and backchanneling rules?
- How can spoken dialog systems account for interruptions from the user?
- If a violation in turn-taking occurs in a spoken dialog system, who should get the stage?

Multi-Party Systems

Austin Almond

Multi-Party Systems

- Traum, 2004.
  - *Issues in Multiparty Dialogues*
- Purver et al., 2007.
  - *Detecting and Summarizing Action Items in Multi-Party Dialogue.*

Issues in Multiparty Dialogues

Traum, 200

ssues in Multiparty Dialogues

Issues

.Participant Roles

2.Interaction Management

3.Grounding and Obligations

Themes

- Two-Party Systems are much simpler
- Multi-Party Systems have unique challenges that make them complex
- Examples use the *Mission Rehearsal Exercise* (MRE) which is a military simulation

Participant Roles: Conversational Roles

Two-Party

- Speaker
- Addressee = Listener

Multi-Party

- Speaker
- Listener(s)
  - Addressee?
  - Ratified?
  - Known to be listening?
  - In-context?

Participant Roles: Speaker Identification

Two-Party

- **If** Speaker != A (me),
  Then Speaker := B (you)

Multi-Party

- Variety of cues
  - Style
  - Self-identification
  - Stereo mic
  - Visual cues/gestures
  - Metadata (computer-computer)

Participant Roles: Addressee Identification

Two-Party

- **If** Addressee != A (me), **Then** Addressee := B (you)

Multi-Party

- Distinguish *addressees* from *hearers*
- Cues:
  - Vocatives (such as names
  - Content of utterance
  - Context

Participant Roles: Addressee Identification

If utterance specifies addressee (e.g., a vocative or utterance of just a name when not expecting a short answer or clarification of type person)
then Addressee = specified addressee
else if speaker of current utterance is the same as the speaker of the immediately previous utterance
then Addressee = previous addressee
else if previous speaker is different from current speaker
then Addressee = previous speaker
else if unique other conversational participant
then Addressee = participant
else Addressee unknown

- Algorithm for addressee identification
- Critique:
  - Ignores participants entering or leaving
  - Ignores pauses / changes in topic

Participant Roles: Other Participant Roles

Two-Party

- **If** Role != Performer
  Then Role = Requester

Multi-Party

- Requester
- Performer (of primitive task)
- Delegator
- Authority
- Guard

- Social roles
- Institutional roles

nteraction Management

- Turn management
- Channel management
- Thread / conversation management
- Initiative management
- Attention management (no detail)

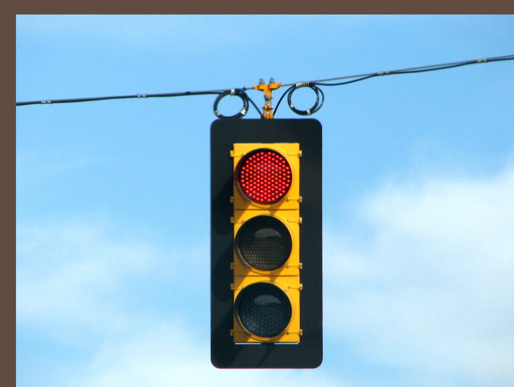
nteraction Management: Turn Management

Two-Party

- When to speak
- When to stop speaking
- <prompt bargein=“true”>
- Cues:
  - Prosody
  - Filled pauses

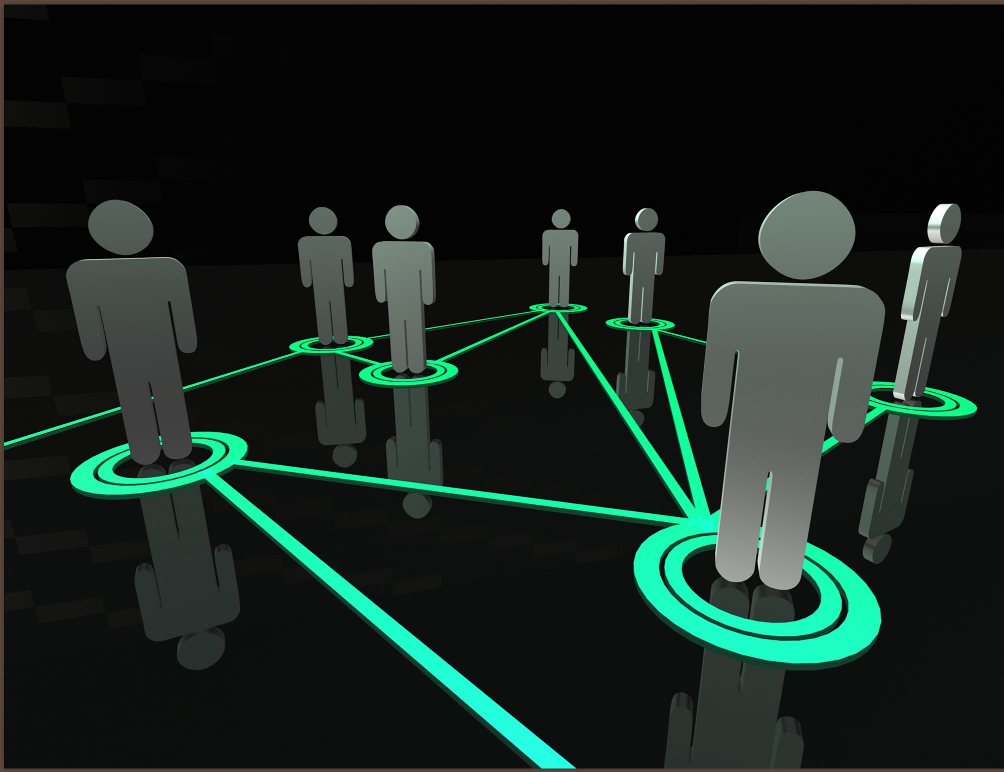
Multi-Party

- Assign turns to speakers?
- Release the floor to all?
- Require request to speak?



Channel, Thread, Conversation Management

- Multiple channels: speech, visual, text
- Stack-based topic organization
  - Fails to handle overlapping topics
- Sometimes channels map to conversation topics
- Formal situations follow a main conversation
  - Information situations all over the place
- Conversations are not always independent



nteraction Management: Initiative Management

Two-Party

- System-initiative
- User-initiative
- Mixed-initiative

Multi-Party

- Team leaders have more initiative
- Cross-initiative
  - Redirect to third party

Grounding and Obligations

- ***Grounding*** – the process of adding to the common ground between participants in conversation
- **Obligation** – Requiring a party to respond
  - Do we care more about **who** answers, or **what** the answer is?

By Frits Ahlefel

Action Items in Multi-Party Dialogue

Purver et al., 200

Action Items in Multi-Party Dialogue

- Research problem: Identify ***action items*** discussed in a meeting
  - ***Action item***: A public commitment to perform a given task
    - Consists of the following information:
      - Owner (who?)
      - Description (what?)
      - Timeframe (when?)

Action Items in Multi-Party Dialogue

- **Input**: Transcript of a meeting
- **Output**: List of action items

- Hierarchical classifier
  - Sub-classifiers
  - Super-classifier

Action Items in Multi-Party Dialogue

- **Sub-classifiers** tag utterances by type
  - Task description (what?)
  - Timeframe (when?)
  - Ownership (who?)
  - Agreement (yes/no)
- **Super-classifier** extracts phrases and summaries from utterances
  - Information can be spread across many utterances
  - **Speaker** and **addressee identification** are needed to determine ownership.

Subdialogue Detection

(2) A: Well maybe by uh Tuesday you could
B: Uh-huh
A: revise the uh
C: proposal
B: Mmm Tuesday let's see
A: and send it around
B: OK sure sounds good

Speaker identification allows tagging utterances by speaker

Action Item:

- **Description**: *revise the proposal*
  - "revise the uh"
  - "proposal"
- **Ownership**: *B*
  - "you could"
  - "OK sure sounds good"
- **Timeframe**: *by Tuesday*
  - "maybe by uh Tuesday"
- **Agreement**:
  - "Uh-huh"
  - "OK sure sounds good"

Parsing and Summarization

- **Timeframe** and **task** descriptions detected using syntactic and semantic features
  - COMLEX, VerbNet, WordNet, NOMLEX, KnowItAll, WSJ
- Spoken grammar is "ungrammatical, disfluent and errorful"
  - Only a few structures are detected – S, VP, NP, PP

Results

Subdialogue Detection

- Discourse-structural approach improves significantly over a flat classifier
- Running on manual transcripts beats error-prone ASR-produced transcripts

Summarization and Parsing

- No improvement over baseline
- Fails to account for summaries across multiple utterances
- Inaccurate sentence segmentation
  - Single-word excerpts of timeframe and ta often detected

Conclusions

- Multi-party systems are easier to model if they have:
  - Formal participant roles
  - Clear topics of discussion
  - Consistently present participants
  - Few simultaneous conversation threads
  - Few channels of communication
- Extracting information from a multi-party dialogue situation is difficult because:
  - Information is fragmented across multiple utterances
  - Multiple speakers may interject or add to an utterance
  - Multi-party dialog lacks formal grammatical structure

Discussion

- How feasible are these ideas to implement in a dialog system?
- What would be the effect of implementing these sorts of techniques in dialog systems?
- Are the approaches more motivated by modeling human conversational behavior or engineering considerations?