Recognition and Understanding of Prosody

Varsha Ganesh
What’s Prosody?

- Rhythm and Melody of Speech
- Characterized (for Modeling) by Features such as
  - Pitch
  - Pitch Contours
  - Intensity
  - Fundamental Frequency
  - Energy
What is it Good For?

- The usual “Make Dialogue Systems more natural and human-like” – but how exactly?
  - Exclusive to spoken language
  - Redundant source of information
  - Detect/Generate Emotion
  - Labeling Dialogue Acts
  - End of Utterance Detection (For Turn-taking)
  - Detect and Respond with Back-channels
The Prosody of Back-channels

What’s a back-channel?

- Listener responses to the speaker
- The listener doesn’t wish to take the floor – only shows that he is following what the speaker is saying
- Examples – uhuh, yeah, okay, I see, etc.

We shall discuss

- Analysis of the prosodic structure of back-channels (Benus et all)
- Analysis of the prosodic cues that “trigger” back-channels in the listener (Nigel et all)
Analysis of the Prosodic Structure of Back-channels

- To Answer
  - Do back-channels have a distinctive prosodic structure? If they do what are the features that characterize them?
- Language for which the analysis was done
  - American English
- Why do we care?
  - More natural speech-synthesis
  - Turn-taking (To understand that the speaker doesn’t want to take the turn, he is only responding to what you are saying)
- What did they find?
  - Back-channels do have prosody structure that allows for them to be distinguished from other discourse elements.
How did they find out?

- Corpus Collection
  - Made subject play one of two games
    - CARDS
      - One subject described a card on his screen and the other subject had to find it on his own screen.
    - OBJECTS
      - One subject described the position of an object on his screen while the other tried to move the object to match the position being described.
  - Active discussion – so conversation is natural
• Annotation

  ◦ Three labelers labeled and classified the occurrences of words that are commonly used as back-channels into several discourse functions (Because these words often have functions other than as back-channels)
    • Words classified – alright, mmhm, okay, right, uhhuh, yeah, yep, yes, yup
    • Discourse Functions – Acknowledgement, backchannel, beginning discourse segment, etc.
• Feature Extraction for Analysis
  ◦ Prosody features extracted from the words themselves and from the segment that precedes them.
    • Acoustic and Durational features
    • Minimum, maximum, mean and slope of
      • Pitch
      • Intensity
    • A few other features
Results

- Lexical Choice
  - *Mhm, uhuh, okay* among the most frequently used back-channels
  - Choice of back-channel depends on speaker
    - For instance, tokens of “alright”, “right” in the corpus came from a single speaker

- Prosodic Characteristics of Back-channels
  - Strongly marked prosodically
    - Higher pitch, intensity and pitch slope (as compared to agreements and other discourse functions)
  - However, the prosodic structure was compared only with occurrences of the words in functions other than as back-channels and not against the rest of the corpus.

- Context of Back-channels
  - Follow intonational phrases with rising pitch
  - Preceding segments end in rising pitch starting with a low tone around 0.5s before the back-channel
Study of Prosodic Features Which Cue Back-Channel Responses

- **To Answer**
  - Can the occurrence of back-channels be predicted by analyzing the prosodic structure of the preceding segments?

- **Language**
  - American English and Japanese

- **Why do we care?**
  - In Dialogue Systems that play roles as “listeners” (Tutor, Psychologist systems) we may wish to have them respond with back-channels to make them more realistic

- **What did they come up with?**
  - A bunch of rules which can predict back-channels to a certain extent – not a great accuracy, but a good start.
Method

- Define Precisely a Back-channel to set stage for proper analysis
  - Back-channel feedback
    - Responds directly to the content of an utterance of the other
    - Is optional
    - Does not require acknowledgement by the other

- Analyzed the corpus of several one-one conversations to come up with a set of rules that could predict the occurrence of a back-channel
  - A region of pitch less than the 26th percentile pitch level
  - Continuing for at least 110 milliseconds
  - Coming after at least 700 milliseconds of speech
  - Providing you have not output back-channel feedback within the preceding 800 milliseconds
  - After 700 milliseconds wait
The rules were tested by running the two tracks of each one-one conversation together –

- One track was the speaker, and the other was listener.
- The rules predicted back-channel occurrences in the listener track – only the fact that a back-channel would be produced – not which back-channel.
- The predictions were compared with actual occurrences of back-channels for evaluation of the rules.
- A prediction was considered correct if it a back-channel occurred in the track within a certain time-interval of the prediction (+- 500 milliseconds)
Results

- Each rule was evaluated for Coverage and Accuracy
  - A rule is said to have a larger coverage if it is likely to predict a larger number of back-channels
  - A rule is said to be accurate if its predictions concurred with observed back-channels in the actual track.
- Best predictor was the “low-pitch” region rule – it had a large coverage and accuracy.
Prosody and Emotions

- Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialogue - Ang et all
Frustration and Annoyance Detection

To Answer

- Can we detect whether a user of a dialogue system is frustrated or annoyed from the prosodic structure of his speech?

What did they find?

- Prediction possible with 60 – 70 % accuracy with both true words and ASR words
- Prediction accuracy close to inter-labeler agreement with human annotators for emotion
Method

- **Corpus Used**
  - Users calling automated systems to make air-travel arrangements
  - Staged calls – not to actually book tickets, so the overall levels of frustration detected by the system was rather low compared to expectation

- **Annotation**
  - Utterances labeled and classified into several emotion classes – Neutral, Annoyed, Frustrated, Tired, Amused, Other, Not-Applicable
  - Inter-labeler agreement only 71% (Kappa 0.47)
  - So second pass of labeling done – this time only the two most experienced labelers re-labeled utterances that the original labelers did not agree on
• **Prosodic Model**
  - **Prosodic Features Extracted**
    - Duration and speaking rate
    - Pause features
    - Pitch features
    - Energy features
  - **Non Prosodic Features Used**
    - Position of utterance in the dialog
    - Repeated attempts and explicit corrections
Results

- ANNOYANCE + FRUSTRATION vs. Else
  - Better results when run with just the corpus labels on which all labelers agreed.
  - Possibility for improvement might come from considering features such as hyper-articulation, pauses and raised-voice
  - “Feature-usage”
    - Temporal Features
      - Duration
      - Speaking-rate
      - (Longer duration and slower speaking-rate associated with frustration)
    - Pitch Features
Pragmatic Functions of Prosodic Features of Non-Lexical Utterances

- Non-lexical Utterances?
  - Utterances that are not words (Example – uhuh, mmhmm, etc.)
  - Not necessarily back-channels

- To Achieve
  - Labeling of all occurrences of non-lexical utterances for 8 dimensions of pragmatic functions. (Recap: Pragmatics is the study of how context contributes to meaning)
Method

- Prosodic features analyzed
  - Syllabification
  - Duration
  - Loudness
  - Pitch-height
  - Pitch-slope
  - Creaky Voice

- Task at hand – Find pragmatic functions to map to these prosodic features
- Hypotheses generated for each of the Prosodic Features for what pragmatic Features they might stand for (This was done by listening to the corpus repeatedly)
- Assumption – the meaning of the prosodic feature was orthogonal to the functional aspect of the utterance
- Hypotheses “tested” against the corpus – not sure what they mean by this… But let’s talk about their Hypothesis for each prosodic feature
• Syllabification
  ◦ Two-syllable items signal intention to take a listening role (Ex. Yeah-yeah, uhuh, etc.) – not considering those non-lexical utterances that are of two-syllables to begin with.
  ◦ (Summary) – two syllable non-lexical utterances are typically back-channels
• Duration
  ◦ Filler duration correlates to uncertainty of response
• Height
  ◦ Pitch height correlates with degree of interest
• Loudness
  ◦ Confidence, Importance
• Pitch Slope
  ◦ Degree of Understanding
Prosody for End of Utterance Detection (Ferrer et al)

To Answer
- Can prosodic features of utterances be used to detect an end of utterance better than the baseline does? (Baseline = Threshold of silence ranging from 0.5 s to 1 s)

Why?
- Better turn-taking capabilities for the system
  - So the system doesn’t barge in if the user is in the middle of an utterance but has paused to think
  - So the system doesn’t wait too long if the speaker is done but the silence hasn’t hit threshold

What did they find?
- 45% improvement over baseline
Method

- System decides whether speaker has reached end of utterance after a pause of any length (i.e., no threshold for minimum pause duration) using prosodic features
- A maximum pause duration threshold is still used
- Main prosodic features used
  - Pitch
- Different evaluations of prosodic features done for different pause durations
  - If pause duration $x$ is being considered, then the features analyzed for $x$ are a superset of features analyzed for $x-a$.
- Combined with a Language Model
  - Because some words are unlikely to occur at the end of a sentence
  - Example in Air Travel domain on which this system was tested, “to” was almost never the end of the utterance because the user is looking for information about flights “from somewhere TO elsewhere”
- Results
  - 45% better than baseline
Questions?