Prosody for Conversational Discourse

Joseph Tyler
Sensely.com
Background

- Voice interfaces are booming.
- Most interactions are short:
  - 1-2 turns
  - 1-2 sentences per turn (or less)

You have had more symptoms than usual in the past 24 hours.
Background

• Longer speech has its own challenges.

You have had more symptoms than usual in the past 24 hours. These symptoms could be caused by a dormant bacterial or viral infection. Take your medications as prescribed and spread your energy out across the day. Avoid circumstances that could potentially worsen your symptoms, and apply the breathing and relaxation techniques that you've learned. To follow up on your symptoms you will receive a daily check-in from this day forward till your health is back to normal.
Let’s look at an example.
A single sentence sounds ok

“You have had more symptoms than usual in the past 24 hours. These symptoms could be caused by a dormant bacterial or viral infection. Take your medications as prescribed and spread your energy out across the day. Avoid circumstances that could potentially worsen your symptoms, and apply the breathing and relaxation techniques that you've learned. To follow up on your symptoms you will receive a daily check-in from this day forward till your health is back to normal.”
“You have had more symptoms than usual in the past 24 hours. These symptoms could be caused by a dormant bacterial or viral infection. Take your medications as prescribed and spread your energy out across the day. Avoid circumstances that could potentially worsen your symptoms, and apply the breathing and relaxation techniques that you've learned. To follow up on your symptoms you will receive a daily check-in from this day forward till your health is back to normal.”
What sounds weird here?
What sounds weird here?

• Each sentence sounds the same.
What sounds weird here?

• A lack of prosodic variation
  – Every sentence ends at ~155Hz
  – Intersentential pauses are all ~.72 seconds
  – Every sentence’s average pitch: 215-230Hz
  – Every sentence ends with the same high-low pitch contour (H* L-L%).
Why is lack of variability a problem?

- Hard to pay attention
- Boring
- Uncanny, not like human speech
- Unclear how to incorporate new sentences into the discourse.
  - More difficult to process
  - Potential ambiguity of meaning
What’s the solution?
What’s the solution?

- Add in (prosodic) variation.
What’s the solution?

• Add in (prosodic) variation.

Variation that indicates relationships between sentences, i.e. discourse structure.
Before showing how, let’s define...

- Discourse
- Prosody
- Discourse prosody
Discourse = many sentences
Discourse = many sentences

• Discourse is a level of linguistic structure above the sentence.
• Discourse is structured.
  – Stories, topics with elaboration, paragraphs
A feature of discourse

• Discourse has boundaries.
  – Not all sentences are equally related to previous.
  – Intuitive examples:
    • Topic changes
    • Narrative episodes
    • Paragraphs
Prosody

Acoustic features of prosody:

- Intonation (f0, measured in Hz, perceived as pitch)
- Intensity/amplitude (perceived as loudness)
- Pausing
- Speech rate
- Voice quality (voicing, whispering, breathy voice, creaky/vocal fry, more)
- Many more...
Discourse Prosody

• Relationships between prosody and discourse information.
  – In speech: a speaker includes prosodic variation that correlates with discourse.
  – In comprehension: listeners use this variation to understand the discourse.
Human speech vs. TTS

- Human speech has discourse prosody
- TTS (often) does not
In human speech...

- Discourse boundaries have prosodic correlates:
  - Larger boundaries have longer pauses
  - Pre-boundary: lower pitch, final lengthening, quieter, often vocal fry (creaky voice)
  - Post-boundary: higher onset pitch, louder
In human speech...

• Discourse boundaries have prosodic correlates:
  – Larger boundaries have longer pauses
  – Pre-boundary: lower pitch, final lengthening, quieter, often vocal fry (creaky voice)
  – Post-boundary: **higher onset pitch**, louder
High onset pitch in BBC news segments

Figure 2.3  The height of each ‘sentence beginning’ (major tone group onset) in a news report\(^4\) (SEC B01). Sections 1 and 12 contain the opening and closing headlines, preceded and followed respectively by metatextual links.
The bigger the boundary...
the more the prosodic contrast.
Adding discourse prosody to our example

**S1:** You have had more symptoms than usual in the past 24 hours.

**S2:** These symptoms could be caused by a dormant bacterial or viral infection.

**S3:** Take your medications as prescribed and spread your energy out across the day.

**S4:** Avoid circumstances that could potentially worsen your symptoms, and apply the breathing and relaxation techniques that you've learned.

**S5:** To follow up on your symptoms you will receive a daily check-in from this day forward till your health is back to normal.
Adding discourse prosody to our example

**S1:** You have had more symptoms than usual in the past 24 hours.

**S2:** These symptoms could be caused by a dormant bacterial or viral infection.

**S3:** Take your medications as prescribed and spread your energy out across the day.

**S4:** Avoid circumstances that could potentially worsen your symptoms, and apply the breathing and relaxation techniques that you've learned.

**S5:** To follow up on your symptoms you will receive a daily check-in from this day forward till your health is back to normal.
What I changed

• Intersentential pauses:
  — They were all ~0.72. Now: .6, .4, .4, 1 secs

• Pitch:
  — Overall pitch lowering of S2-4
  — Lower onset pitch for S2-4, higher onset pitch for S5
  — Final lowering especially for S4

• Intensity (loudness): S3-4 lowered 2 dB
How I made the changes

- Praat: standard software for phonetics
How I made the changes

- Praat: standard software for phonetics

SSML could work too!
In conclusion...

- Human speech has discourse prosody.
- TTS (often) does not.
- If we use TTS for multiple sentences, we should include discourse prosodic variation.