

Human-Robot Interaction

Module 4: Interaction

Lecture 3: Human-human verbal interaction; Speech recognition

David Vernon
Carnegie Mellon University Africa

www.vernon.eu

Verbal Interaction

Understanding human speech is much harder than synthesizing human speech

This creates an imbalance in people's expectations about what the robot can do and what the robot can in fact do

Verbal Interaction

Intonation and emphasis can change the meaning of a sentence substantial

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

She said she did not take his money.

Topics

- Human-human verbal interaction
 - Components of speech
 - Written text versus spoken text
- Speech recognition
 - Basic principles of speech recognition
 - Limitations
 - Practice in HRI
 - Voice-activity detection
 - Language understanding in HRI

Human-human verbal interaction

- Speech is used to
 - Convey information
 - Create joint attention
 - Create a shared reality through communication
- Speech is open to multiple interpretations (remember Kismet)
- Verbal communication is enriched by paralinguistic information
 - Prosody (rhythm, timing, ...)
 - Non-verbal behavior
 - Gaze
 - Gestures
 - Facial Expressions

Components of Speech

Speech comprises **utterances**: the smallest unit in spoken language

- Often not grammatically perfect
- **Pauses**
- **Words**: the smallest units of sound we can utter to convey meaning
- **Phonemes**: small unites of sound that make up words
 - Three phonemes: [p], [a], [t]
 - Changing [p] to [b] changes the meaning of the word

Components of Speech

Speech also comprises **conversational fillers**

- They **keep the conversation going**
 - Active listening; they indicate you are paying attention
- They serve to stop others taking turns and **keep control of the conversation**
 - By leaving no pauses where the other person can interject and take a turn
 - "You know, I was, like, yeah, going to buy her, you know, something, but then I had, like, uhm, what the heck."
- They can increase the efficiency of verbal communications (but not always DV)
 - Enhances the experience of a shared reality

Written Text vs. Spoken Language

- Written text and spoken utterance are vastly different (listen to a transcript)
- People usually and quickly detect when the interaction partner does not understand the message
 - They change the sentence on the fly

Written Text vs. Spoken Language

- Natural, humanlike communication is crucial for HRI
- Technical prerequisites
 - Transcribe speech to words
 - Demonstrate that it understands the words by coming up with appropriate responses
 - Generate spoken language
 - On the basis of disfluent, **imperfectly phrased verbal speech**

Speech Recognition

- **ASR**: Automated Speech Recognition
- **SST**: Speech to Text
- Often make **unrealistic assumptions**
 - Recorded with a high-quality microphone
 - Positioned close to the speaker
 - In a relatively noise-free environment
- Digital signal processing (**DSP**) and microphone arrays can help
- Often, the speaker wears a head-mic or lapel-mic (close miking) ... not natural but pragmatic

Speech Recognition

Significant variety in the performance of speech recognition engines

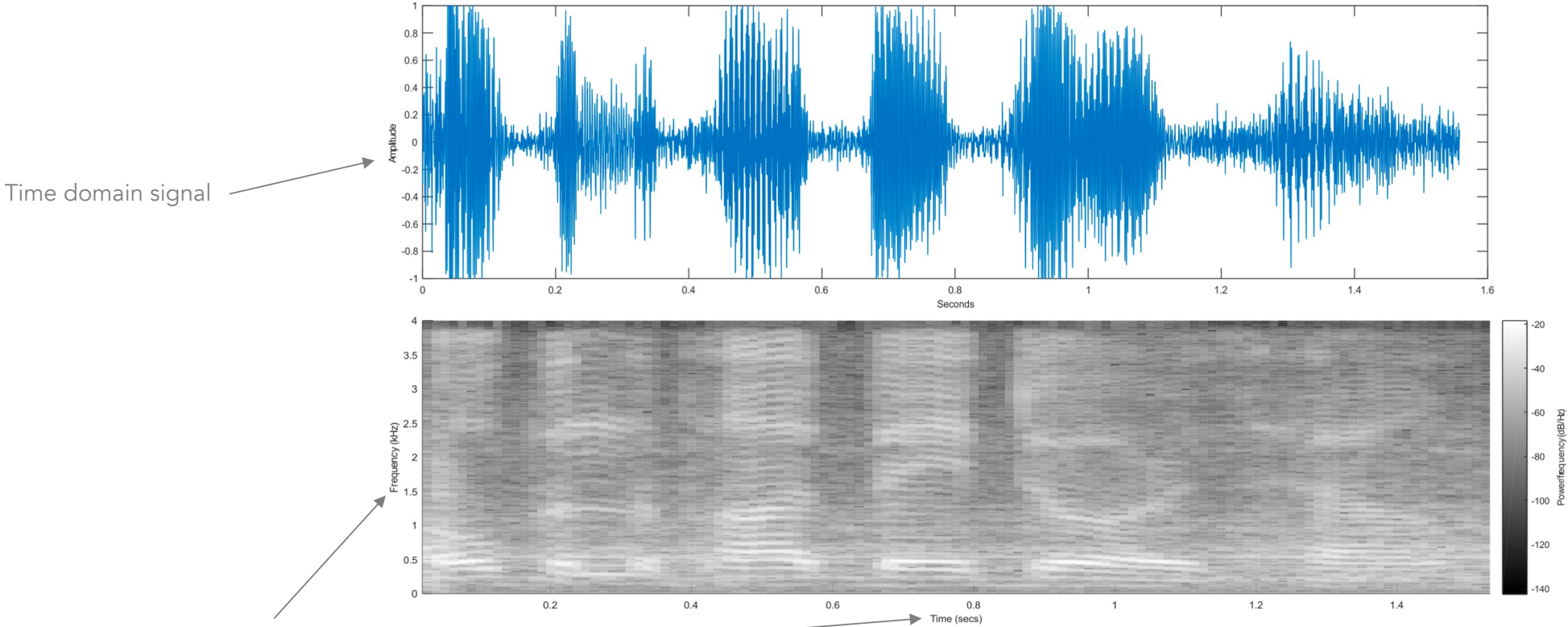
- Open
- Commercial

Basic Principles of Speech Recognition

The speech recognition process

- Digital recording of speech
 - In the time domain
 - Sample the amplitude of the signal over time
- Convert to the frequency domain
 - Make explicit the strength of certain frequencies
- Phonemes have different characteristics in the frequency domain
- Makes recognition easier

Basic Principles of Speech Recognition



Time domain signal

This is actually a time-frequency distribution

"Open the pod bay doors, HAL"

Basic Principles of Speech Recognition

Classic approach: extract phonemes, words, sentences

- Using Gaussian Mixture Models (GMM)
 - Hidden Markov models
- Use them to generate a **probabilistic model** how phonemes are strung together to form valid sentence
- "**robot**" is more likely than "**lobot**"
- "**the robot served the man**" is more likely than "**the robot swerved the man**"

Basic Principles of Speech Recognition

Modern approach

- Use deep neural networks to perform the interpretation
- Can deal with
 - Background noise
 - Crowded environments
 - Ill-formed speech
 - Speaker independent
 - Works with many different speakers
 - Of different genders

Limitations

All speech-recognition engines still struggle with **atypical speech**

Insufficient training on

- Young speakers
- Elderly speakers
- Local dialects
- Nonnative speakers
- Noise
- Reverberation
- Proper nouns (e.g. Kigali)
- Different backgrounds

Beyond the state of the art



Practice in HRI

Speech recognition using deep neural networks

- Typically, available as a pay-per-use remote (cloud) service
 - 1 cent / word
 - Frees up robot's computing resources
- Some on-board engines are available using first-gen DNN but not as good as a cloud service
 - Open source software:
e.g. Mozilla Foundation's **Common Voice** initiative & **DeepSpeech** recognition engine
- Simple API, sometimes with
 - Confidence rating
 - Alternative texts

Voice-activity Detection

- VAD is used to give the impression that the robot is listening
- Can be used to implement turn-taking without actually understanding the speech
 - Even better when using sound localization with a microphone array
 - Look at the speaker
- Deep learning has improved performance
- Free OpenSmile software package is currently (2020) leading in terms of performance

Language Understanding in HRI

Speech recognition is not the same as speech understanding

- Semantic content (what the utterance means) is difficult
- Sentiment analysis is used to extract the affect in the utterance
 - Prosody (rhythm, beat, ...)
 - Amplitude

Help understand whether the speaker is happy or agitated

- Natural language understanding systems can extract keywords

Language Understanding in HRI

- Natural language understanding systems can extract keywords
 - Command
 - Location
 - Person
 - Event
 - Date
- Allowing the system to respond appropriately
 - "Remind me to pick up the kids at 7 pm"
 - Sets a reminder for 7 pm for the "pick up the kids" event (whatever that is)
- Genuine understanding of natural language has not yet been achieved

Language Understanding in HRI

Consider the following dialogue where the robot was programmed to take everything literally (Williams et al., 2018)

Participant: Um, can I have one water?

Robot: Yes, that is permissible.

Participant: Great.

Robot: Please tell me your order.

Participant: Can I have one water?

Robot: Yes, that is permissible.

Participant: Great, I'll take one water.

Robot: Thank you for sharing that interesting prediction. Please tell me your order.

Language Understanding in HRI

Participant: Can I order one water?

Robot: Yes, that is permissible.

Participant: Okay, can you bring me one water?

Robot: Yes, I am able to do that.

Participant: May you bring me one water?

Robot: Yes, I am able to do that.

Participant: I would like to have one water.

Robot: Thank you for sharing that interesting fact. Please tell me your order.

[16-second pause]

Language Understanding in HRI

Participant: Can you bring me one water?

Robot: Yes, I am able to do that. Please tell me your order.

Participant: May you please bring me one water?

Robot: Yes, I am able to do that.

Participant: So can you do it?

Robot: Yes. Please tell me your order.

Participant: Can you go inside and get the water for me?

Robot: Yes, I am able to do that. Please tell me your order.

Language Understanding in HRI

Participant: My order is water.

Robot: Okay.

[Robot retrieves order for participant]

Reading

Bartneck, C., Belpaeme, T., Eyssele, F., Kanda, T., Keijsers, M., Sabanovic, S. (2020). Human-Robot Interaction - An Introduction, Cambridge University Press.

Chapter 6 – Verbal Interaction, pp. 98-106.