Speed Separation and Recognition Challenge: PASCAL CHiME

Juho Hirvonen

University of Helsinki

8.2.2012

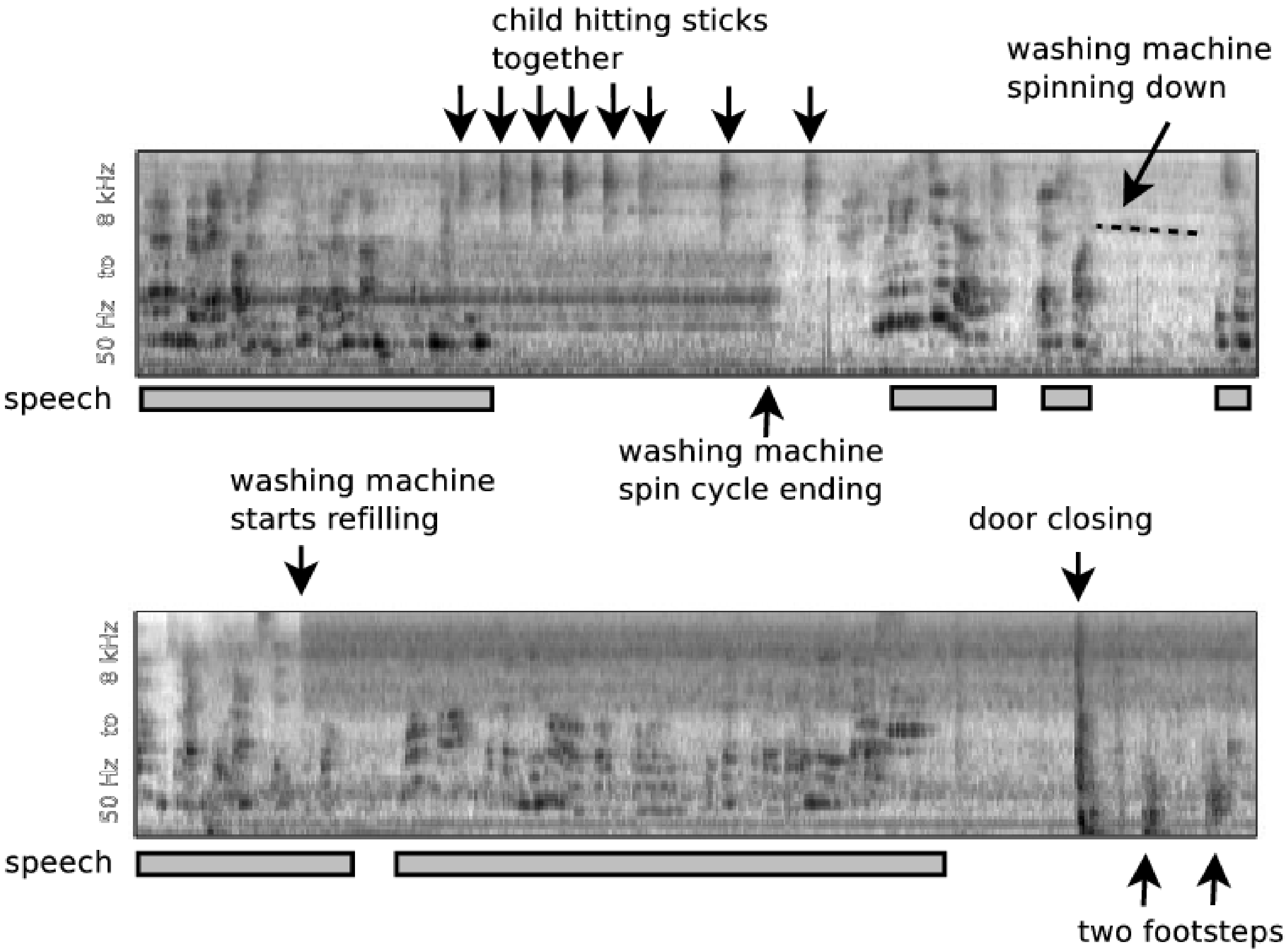
This Talk

- The Challenge: PASCAL CHiME
- Automated Speech Recognition (ASR)
  - Why?
- What is sound?
  - How does a computer process sound?
- Why is this challenge interesting?
  - Details of the challenge

The Challenge

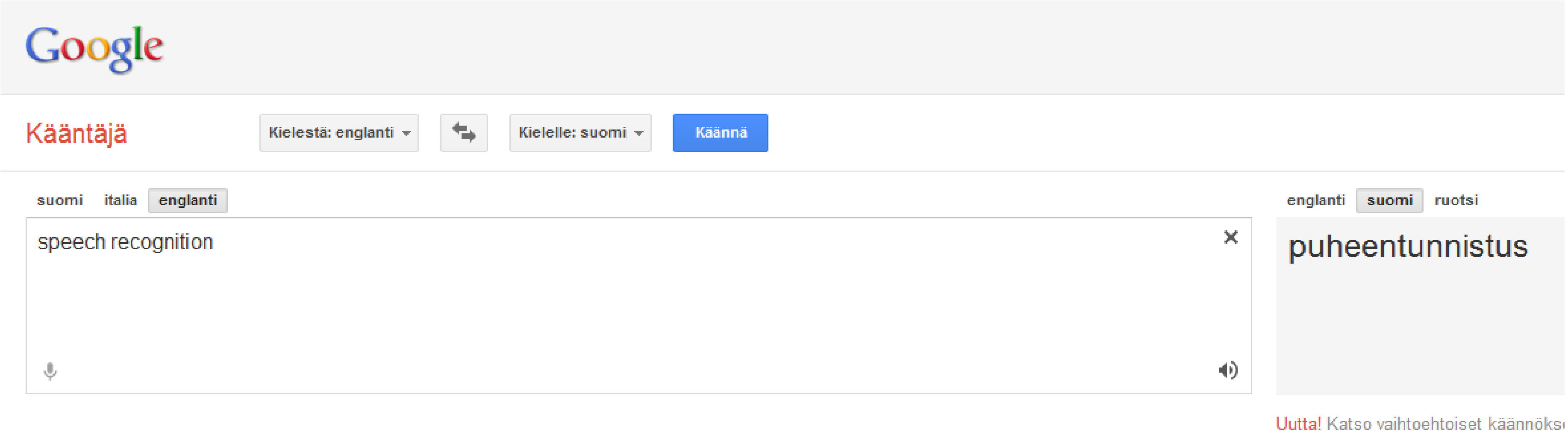
- Audio signal containing household sounds
- Task: recognise specific commands
  - Separate speech signal
  - Recognise speech
- Audio contains noise
  - People talking
  - Doors slamming

The Challenge: Audio



Applications

- Applications for speech recognition
  - Human-computer interaction
  - Speech to text
  - Translation
  - Mobile devices in general



speech recognition

puheentunnistus

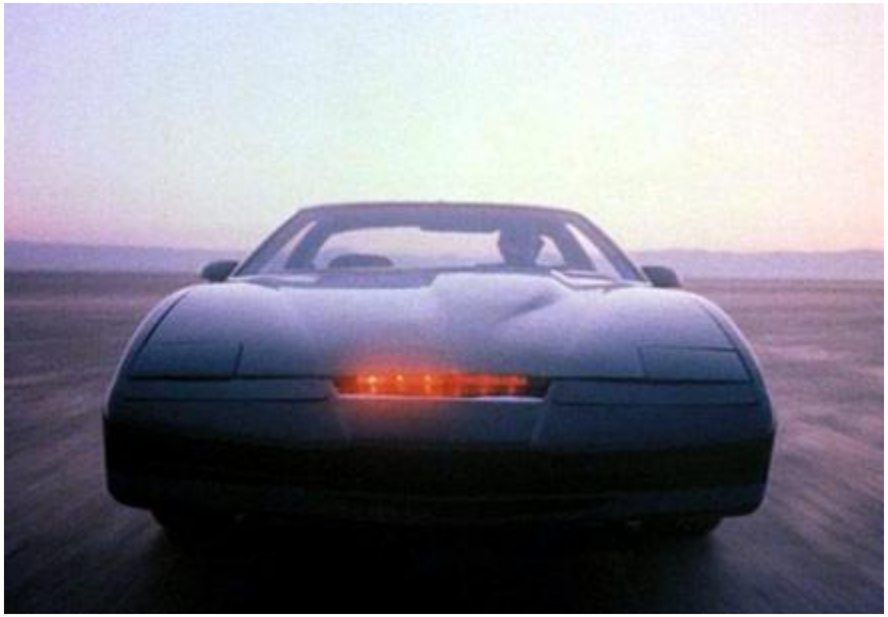
Uutta! Katso vaihtoehtoiset käännöks

Applications

3G
9:41 AM
Here's your email to Zach Friedman:
To: Zach Friedman
Subject: Sales numbers
Can you please send me the sales numbers from last quarter?
Cancel
Send
3G
9:41 AM
“Meet with Jamie for coffee at two today”
Ok, I scheduled your meeting with Jamie Roberts today.
7
Friday, October 2011
Coffee with Jamie...
2 PM
3G
9:41 AM
Tap the mic to say things like:
Call Jason
Phone
Play The Light of the Sun
iPod / Music
Tell Susan I'll be right there
Messages
Set up a meeting at 9
Calendar
Remind me to call mom
Reminders
How do I get home?
Maps and Directions (U.S. Only)
Email Lisa about the trip

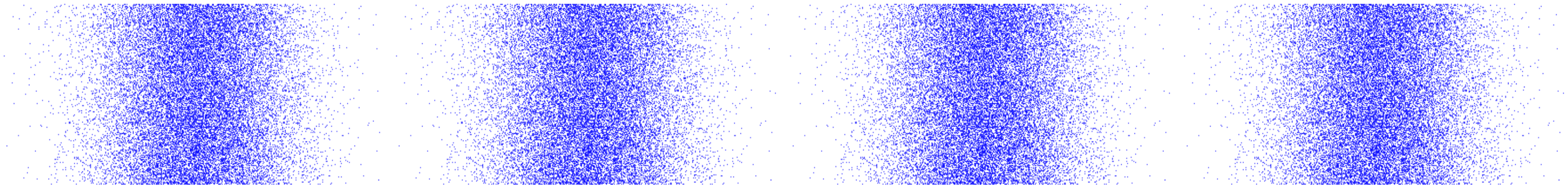
Applications

Applications



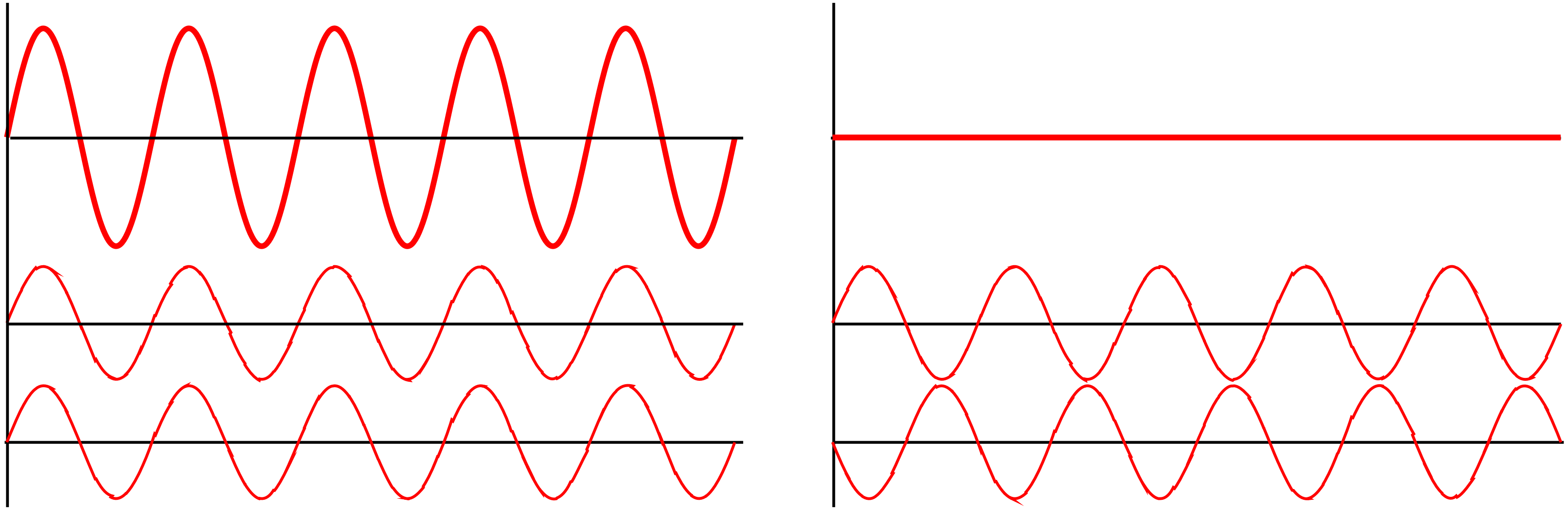
Speech

- Speech is sound, what is sound?
  - Pressure waves in a medium
  - Displacement of air molecules
- Physics formalism: a wave



Speech

- Usually presented in waveform

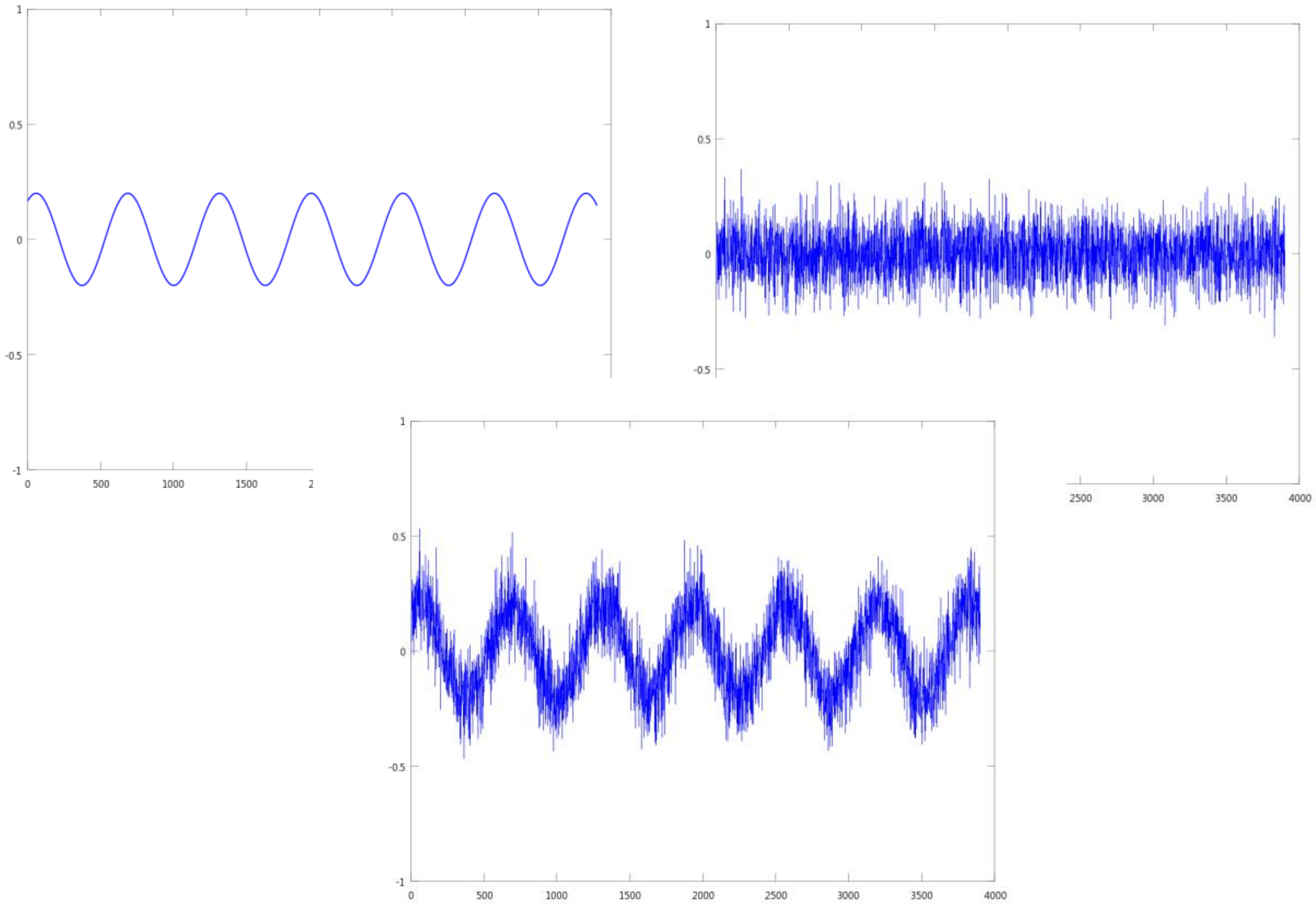


- Sound waves are additive

Image by Haade; available under Creative Commons Attribution-ShareAlike 3.0 Unported license

Speech

- Noise

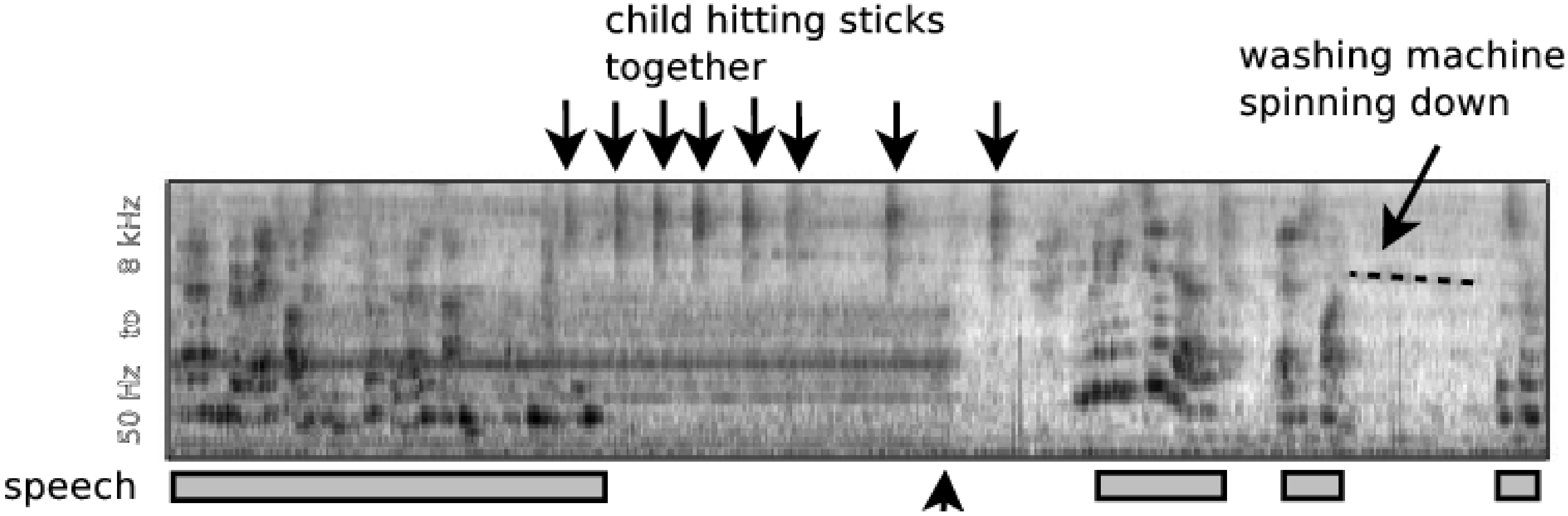


Speech

- Discrete representation
  - A set of (time, pressure) pairs
  - Sample frequency
- Problems if too few samples

Speech

- Alternative representation
  - Frequency vs. time



Speech

- Extracting signal information is a computational task
  - Basis in physics
  - Acoustics
  - Language
- For example: What is the frequency domain of the signal?

CHiME Challenge

- Speech recognition in an acoustically cluttered environment
- Recorded in an actual household
- Target voice commands mixed in

CHiME Challenge

- Why?
  - Realistic setting for speech recognition
  - Actual task: voice commands
  - Binaural hearing
- Different (possible) recognition subtasks
  - signal separation
  - feature extraction
  - speech recognition

Target

- Target voice commands of the following form

  <command:4><color:4><preposition:4><letter:25>

  <number:10><adverb:4>

  - For example: “place white at L 3 now”
  - In total 64 000 combinations
  - Phonetically similar vocabulary: C, D, E, G, P, T, …

Target

- Voice commands from the *Grid corpus*
  - Mixed into the background noise
- 34 speakers
- 600 different utterances
- Speaker location fixed
  - 2 meters from the microphone

Data Sets

- Test set, development test set and final test set
  - isolated utterances
  - background noise
  - utterances in noise
- Utterances in segmented form
- Utterances in continous audio with time infromation

Issues

- Signal-to-noise ratio
  - power of the signal : power of the noise
- Measure of how clear the signal is
- Varied in the data
  - Problem difficulty
  - Not done artificially, but by choosing the noise segment

Issues

- Different kinds of noise
  - Speech
  - Relatively high energy noise
  - Continous noise for a short time
  - Unpredictable

Available information

- Speaker identity in the development sets
  - Can be used for speaker-dependent models
- Continous background audio for acoustic modeling
  - 6 hours
- Speaker location fixed
  - If the speaker was moving, new problems

Concluding Remarks

- Quite realistic setting for speech recognition
  - Clear voice commands
  - Unpredictable, loud noise
- Multidisciplinary challenge
  - Signal processing
  - Machine learning
- Connections with research on human hearing