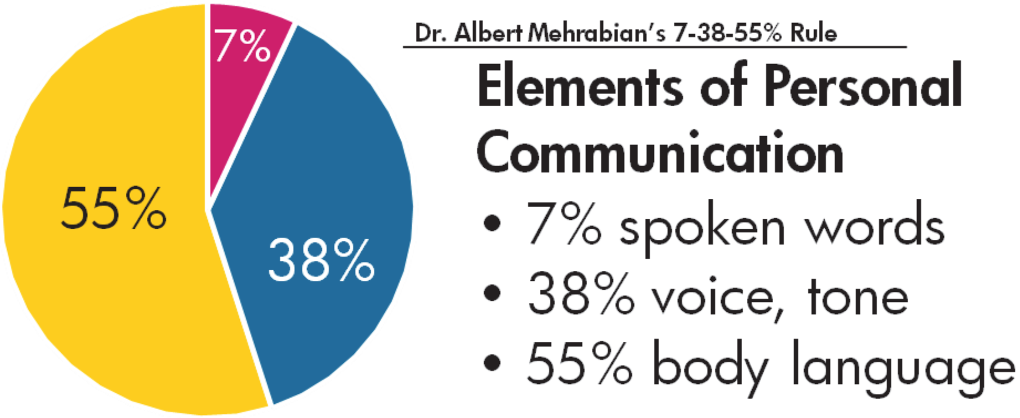
End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim

ETRI Electronics and Telecommunications Research Institute

Motivation



Dr. Albert Mehrabian's 7-38-55% Rule

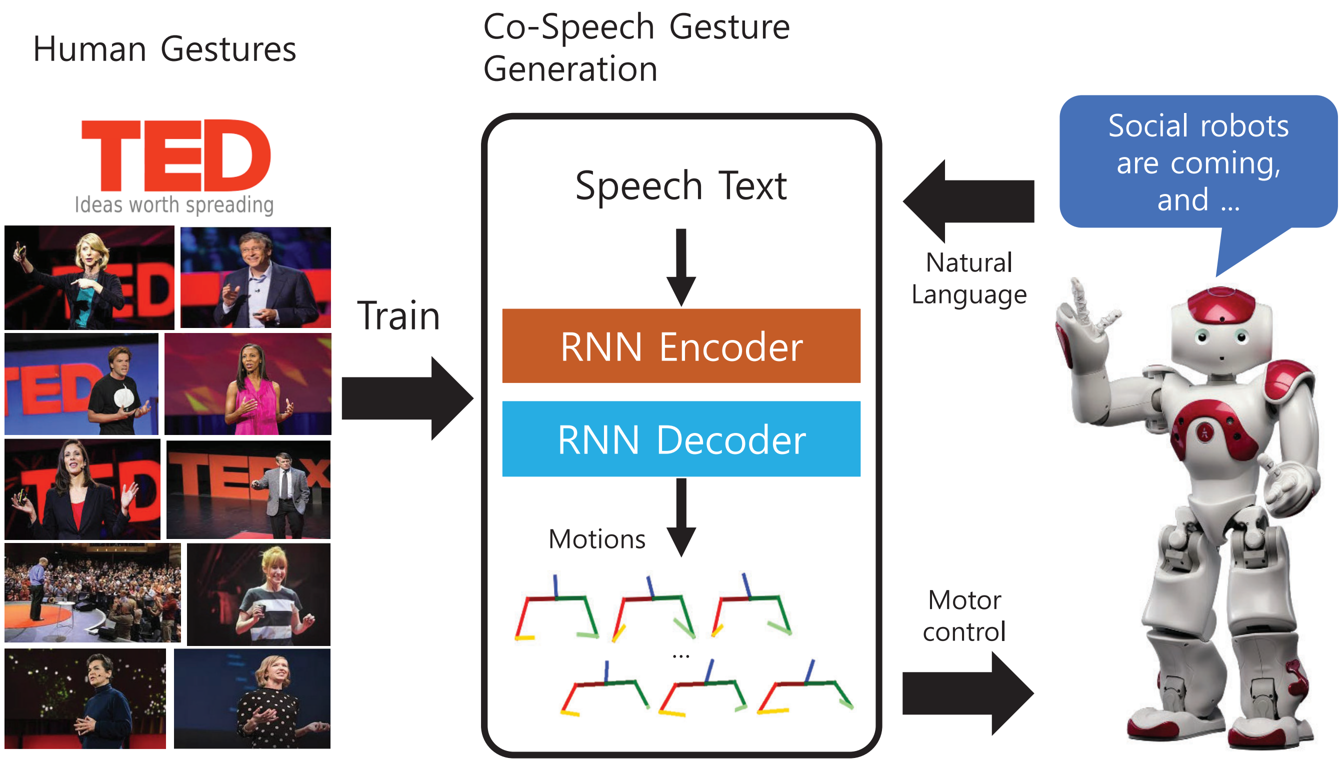
Elements of Personal Communication

- 7% spoken words
- 38% voice, tone
- 55% body language

- Co-speech gesture
  - Generating upper-body motion from speech text

Image credit: http://www.rightattitudes.com/2008/10/04/7-38-55-rule-personal-communication/

Overview



Human Gestures

Co-Speech Gesture Generation

Train

Speech Text

RNN Encoder

RNN Decoder

Motions

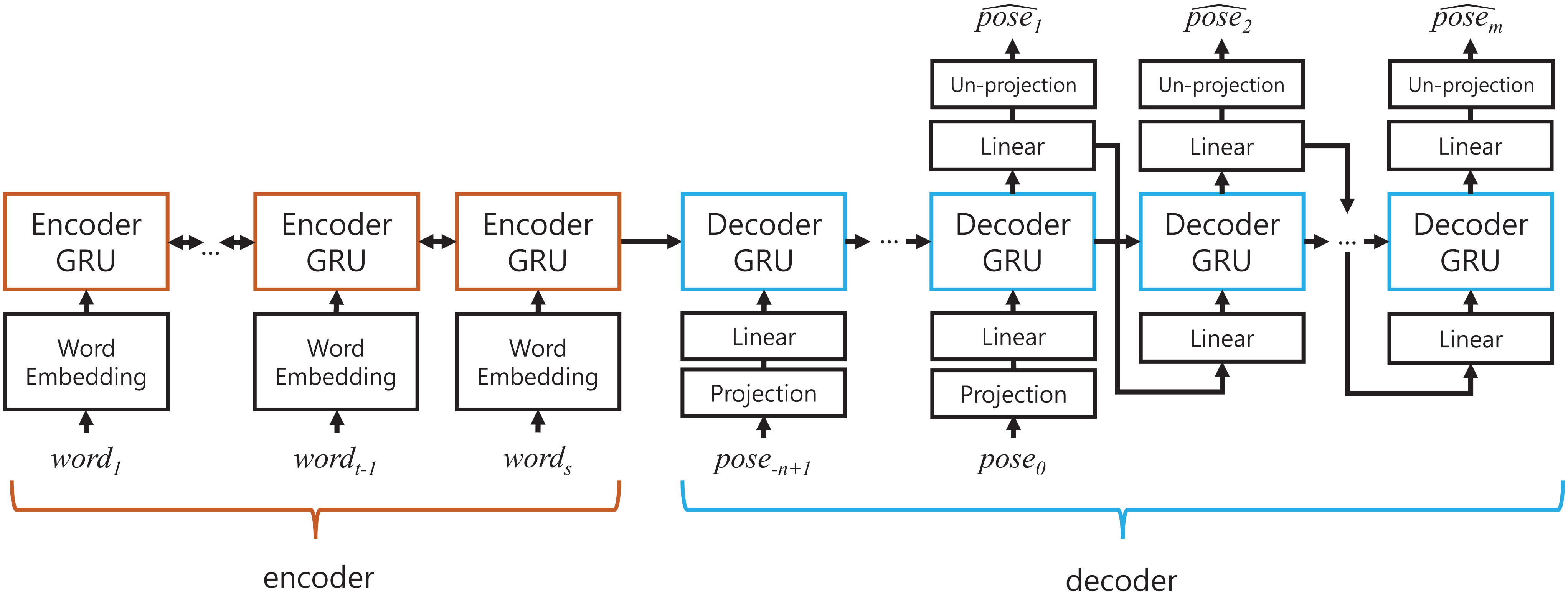
...

Natural Language

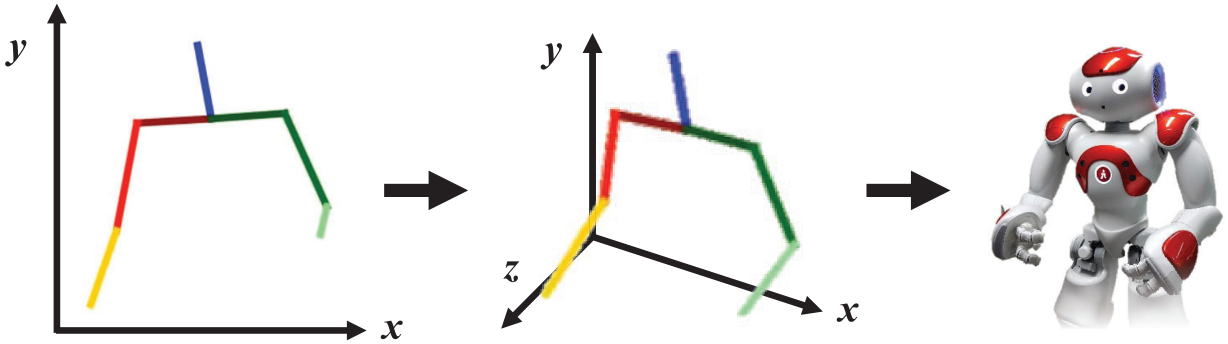
Motor control

Social robots are coming, and ...

End-to-end Architecture



Robot Prototype



3D Pose Estimation

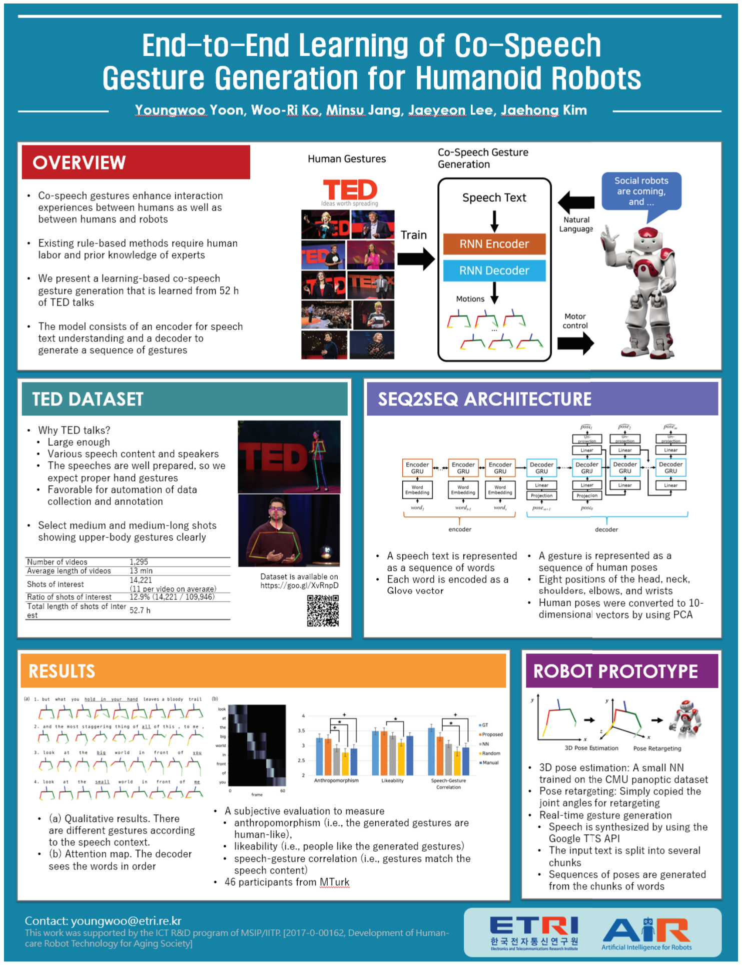
Pose Retargeting

Demo Video

Thanks you

TED Dataset is available on

https://sites.google.com/view/youngwoo-yoon/projects/co-speech-gesture-generation



End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim

OVERVIEW

- Co-speech gestures enhance interaction experiences between humans as well as between humans and robots
- Existing rule-based methods require human labor and prior knowledge of experts
- We present a learning-based co-speech gesture generation that is learned from 52 h of TED talks
- The model consists of an encoder for speech text understanding and a decoder to generate a sequence of gestures

TED DATASET

- Why TED talks?
  - Large enough
  - Various speech content and speakers
  - The speeches are well prepared, so we expect proper hand gestures
  - Favorable for automation of data collection and annotation
- Select medium and medium-long shots showing upper-body gestures clearly

Number of videos	1,295
Average length of videos	13 min
Shots of interest	14,221 (11 per video on average)
Ratio of shots of interest	12.9% (14,221 / 109,946)
Total length of shots of interest	52.7 h

Dataset is available on https://goo.gl/XvRnpD

SEQ2SEQ ARCHITECTURE

- A speech text is represented as a sequence of words
- Each word is encoded as a Glove vector
- A gesture is represented as a sequence of human poses
- Eight positions of the head, neck, shoulders, elbows, and wrists
- Human poses were converted to 10-dimensional vectors by using PCA

RESULTS

(a)
1. but what you hold in your hand leaves a bloody trail
2. and the most staggering thing of all of this , to me ,
3. look at the big world in front of you
4. look at the small world in front of me
(b)
look
at
the
big
world
in
front
of
you
0
frame
60
4
3.5
3
2.5
2
Anthropomorphism
Likeability
Speech-Gesture Correlation
GT
Proposed
NN
Random
Manual

- (a) Qualitative results. There are different gestures according to the speech context.
- (b) Attention map. The decoder sees the words in order
- A subjective evaluation to measure
  - anthropomorphism (i.e., the generated gestures are human-like),
  - likeability (i.e., people like the generated gestures)
  - speech-gesture correlation (i.e., gestures match the speech content)
- 46 participants from MTurk

ROBOT PROTOTYPE

- 3D pose estimation: A small NN trained on the CMU panoptic dataset
- Pose retargeting: Simply copied the joint angles for retargeting
- Real-time gesture generation
  - Speech is synthesized by using the Google TTS API
  - The input text is split into several chunks
  - Sequences of poses are generated from the chunks of words

Contact: youngwoo@etri.re.kr

This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00162, Development of Human-care Robot Technology for Aging Society]

ETRI
한국전자통신연구원
Electronics and Telecommunications Research Institute
AiR
Artificial Intelligence for Robots