A COMPLETE END-TO-END SPEAKER VERIFICATION SYSTEM USING DEEP NEURAL NETWORKS: FROM RAW SIGNALS TO VERIFICATION RESULT

Jee-Weon Jung, Hee-Soo Heo, IL-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu
School of Computer Science, University of Seoul

Overview

- Proposed an end-to-end DNN (text-dependent)
- Input : raw waveforms
- Trainable pre-emphasis embedding layer
- For processing raw waveform
- Parallel CNN-LSTM architecture
- Train CNN → add LSTM layer in the middle
- b-vector scheme* used for back-end classification

Main features

- Directly process raw waveform
- Learn pre-emphasis filter coefficients
- 3 step training
- Model is sequentially trained, expanding
  RWCNN → RWCNN-LSTM → RWE2E
- Evaluation
  - Input two raw waveforms → speaker verification result
  - Utter-level speaker feature extracted with
    LSTM sequential processing
  - Can handle variable length raw waveforms at evaluation
  - Fixed input length when training (duplicated or truncated)

Proposed system

- Dataset : RSR 2015 dataset part 1 / short utterances (3.2 s average) / text-dependent
- 197 speakers for DNN training / 103 speakers for evaluation
- 30 utterances × 9 sessions per speaker
- DNN configuration
  - RWCNN : pre-emp + 9 conv + 2 fully connected
  - RWCNN-LSTM : pre-emp + 6 conv + 1 LSTM + 2 fully connected
  - RWE2E : RWCNN-LSTM + b-vector + 5 fully connected

Experiments & results

- System (backend) EER (%)
  - RWCNN (CSS) 7.7
  - RWCNN + pre-emphasis embedding (CSS) (proposed) 5.3

- Input feature System (backend) EER (%)
  - mel-filterbank energies d-vector(CSS) 4.9
  - raw waveform (proposed) RWCNN (CSS) 5.3
  - raw waveform RWCNN (E2E) 4.0
  - RWCNN-LSTM (CSS) 3.9
  - RWCNN-LSTM (E2E) 3.7

Additional experiments

- Dataset : VoxCeleb* / text-independent
  (submitted for Interspeech 2018 with other proposals)
- Compared with other reported state-of-the-art systems


*H. S. Lee, et al. | Speaker verification using kernel based binary classifiers with binary operation derived features in ICASSP 2014