What is Cued Speech?

Cued Speech (CS) is a hand coding system in which the hand information complements the lip-reading. Consonants are coded by hand shapes and vowels by hand positions.

Asynchrony of lips and hand movements in Cued Speech

French word “Petit”. Top: lips and hand zoomed from the middle whole image. Bottom: the speech signal. Red vertical lines indicate where images are taken corresponding to the audio speech.

Observations:
- Common behavior at the end of each sentence;
- Existence of a turning-point at about 1s before the end;
- From the beginning to the turning-point, the same distribution;
- Decrease of $\Delta$, at the end of the sentences;
- Similar phonemnon for the two subjects.

Two parts in the model:
- a mean value $(0.139)$ before the turning-point;
- a linear regression after (with slope $-0.213$ for the first subject, with slope $-0.228$ for the second subject);
- This turning-point is the intersection of two lines $(0.84s)$;

A common hand preceding model

Evaluation and results

1. Multi-Gaussian

On the subset of database (138 sentences), with both the ground truth and the automatic tracked hand positions.

Based on the both hand positions:
- Ground truth > Predicted > Audio based

2. Multi-Gaussian and LSTM

On the whole database (476 sentences), only with the automatic tracked hand positions.

Temporal information captured by LSTM.

Based on the automatic hand positions:
- LSTM > Multi-Gaussian.

LSTM architecture:
- two hidden layers of 500 cells;
- 200 epochs are used;
- backpropagation through time (RPTT);
- cross-entropy cost function;
- using max-voting;
- Keras toolkit;
- GPU-accelerated library.

References
