AUTOMATIC TEMPORAL SEGMENTATION OF HAND MOVEMENTS FOR HAND POSITIONS RECOGNITION IN FRENCH CUED SPEECH

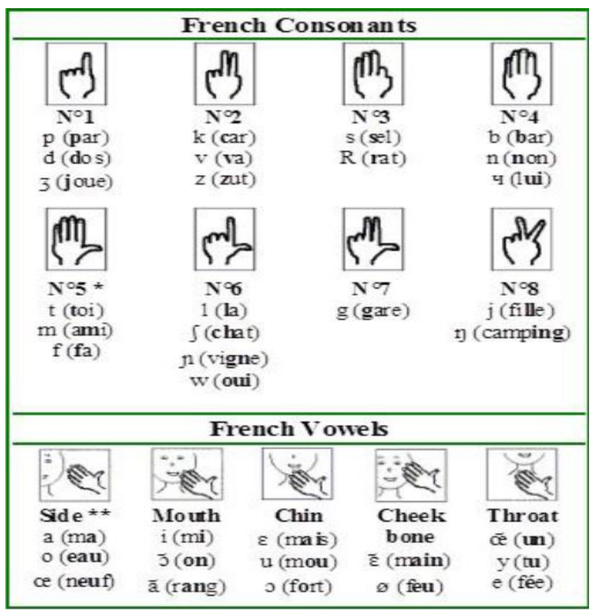
LI,LIU, GIPSA-lab, li.liu@gipsa-lab.fr
FENG,GANG, GIPSA-lab, gang.feng@gipsa-lab.fr
DENIS,BEAUTEMPS, GIPSA-lab, denis.beautemps@gipsa-lab.fr

séphora berrebi association

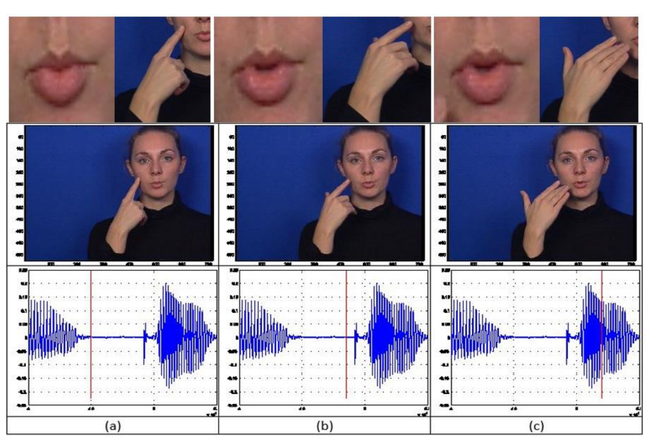
What is Cued Speech?

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.

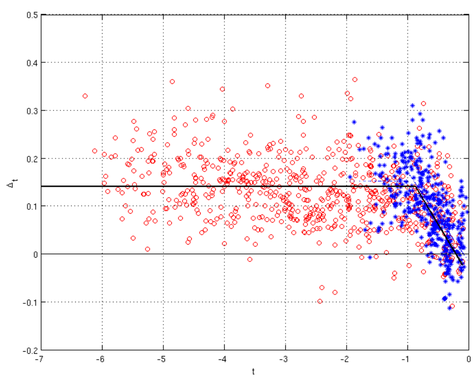
$$\Delta_t = A_t - H_t \qquad (1)$$

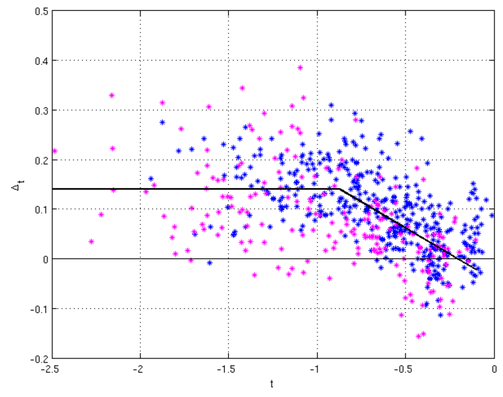
- A_t: target instant of the phoneme relization in audio speech
- H_t: target instant in hand relization (vowels)
- Δ_t: hand preceding time for vowels

Database

- Two professional Cued Speech speakers;
- For speaker 1: 88 short sentences + 50 long sentences (totally 1068 vowels) ;
- For speaker 2: 44 short sentence (196 vowels) ;
- GMM foreground extraction method is used for automatic hand position extraction.

Proposed hand preceding model





The X-axis is the vowel instant in a sentence. Y-axis: the preceding time Δ_t. In (a), the red circles shows the distribution of the long sentences, and the blue start the short sentences. The black curve shows the hand preceding model. In (b), the blue stars show the distribution of short sentences for the first subject and the magenta stars the second subject.

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 Δ_t at the end of the sentences;
- Similar phonemenon 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**

From hand preceding model to temporal segmentation

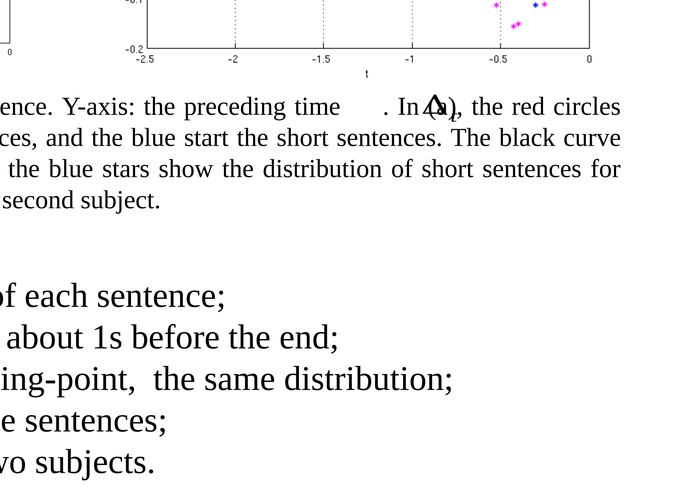
Audio based temporal segmentation

↓

Shift Δ_t

↓

Predicted temporal segmentation



The orange intervals: Δ_t

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

Spatial position / Temp. seg.	Auto hand pos	Manual hand pos
Audio based	45.41%	59.63%
Predicted	54.40%	71.33%
Ground truth	62.26%	86.57%

2. Multi-Gaussian and LSTM

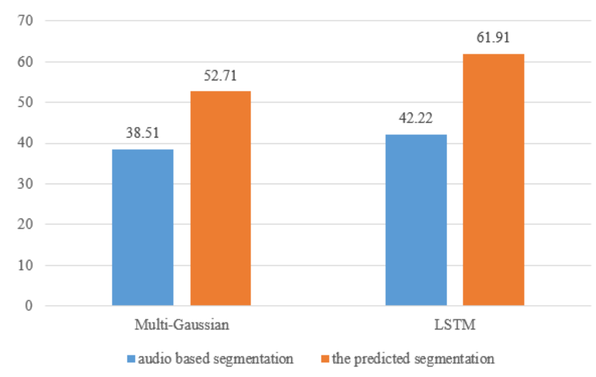
On the **whole** database (476 sentences), **only** with the automatic tracked hand positions. **Temporal information** captured by LSTM.

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

LSTM architecture:

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

Hand recognition using different hand position and temporal segmentations.



Hand positions recognition using the multi-Gaussian and LSTM

References

[1] Noureddine Aboutabit, Denis Beautemps, and Laurent Besacier, "Hand and lip desynchronization analysis in french cued speech: Automatic temporal segmentation of hand flow," in Proc. IEEE-ICASSP, 2006, vol. 1, pp. I–I .

[2] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE-ICASSP, 2013, pp. 6645–6649.

Grenoble Images Parole Signal Automatique
UMR CNRS 5216 – Grenoble Campus
38400 Saint Martin d'Hères - FRANCE