

Fusion of lip shape and hand gestures for vowel and consonant recognition in French Cued Speech

Panikos Heracleous, Nouredine Aboutabit & Denis Beautemps

{Panikos.Heracleous, Nouredine.Aboutabit, Denis.Beautemps}
@gipsa-lab.inpg.fr

To date, in many studies dealing with speech perception or speech recognition, visual information is also used to complement the audio information (lipreading). In automatic speech recognition, visual information is also used to improve the performance of a speech recognition system, especially under noisy conditions. The visual pattern, however is ambiguous, and speech cannot be recognized completely with lipreading alone. On the other hand, for the orally educated deaf or hearing-impaired people, lipreading remains the main modality to perceive speech. Because of this fact, in 1967 Cornett developed the Cued Speech system as a supplement to the lipreading (Cornett, 1967). Cued Speech improves speech perception for hearing-impaired people and offers a complete representation of the phonological system for hearing-impaired people, and therefore has a positive impact on the language development.

Cued Speech is a visual communication system that uses handshapes placed in different positions near the face in combination with natural speech lipreading to enhance speech perception from visual input. A manual cue in this system contains two components: the handshapes and the hand position relative to the face. Handshapes are designed to distinguish consonant phonemes whereas hand positions are used to distinguish vowel phonemes. In French Cued Speech eight handshapes are used in five positions.

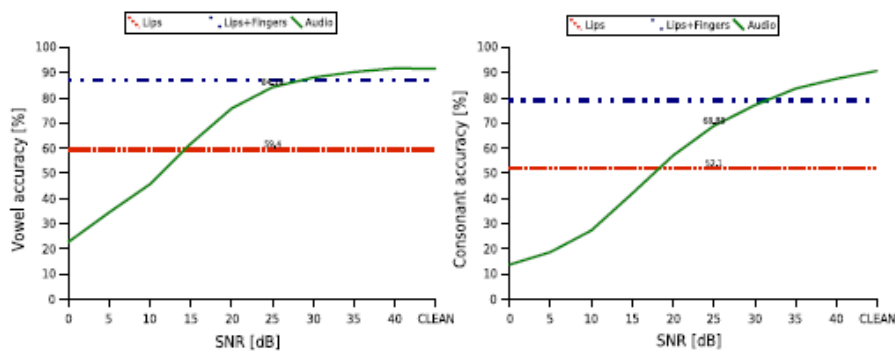


Figure 1 - Cued Speech vowel and consonant recognition using only lip and hand parameters.

This study focuses on recognition of French Cued Speech based on hidden Markov models (HMMs). To our knowledge, this is the first time that automatic recognition of French Cued Speech based on HMMs is introduced. In a first attempt for vowel recognition in Cued Speech, in Aboutabit *et al.*, (2008) a method based on separated identification was used and 75% vowel accuracy was obtained. In this study, however the proposed method is based on HMMs and it uses concatenative feature fusion and multistream HMM decision fusion to integrate the two modalities, *i.e.*, hand and lip modality into a combined one and then perform conventional automatic recognition. Fig. 1 shows the results obtained in a noisy environment in the function of several SNR (signal to noise ratio) levels. When hand position and lip shape modalities were fused, 87.6% vowel accuracy was obtained, showing a 70% relative improvement compared with using only lip parameters. In the case of consonant recognition, the relative improvement was 56%. Results also showed, that although only lip and hand parameters were used in the experiments, the obtained accuracies were comparable to those obtained using the acoustic signal. It should be also noted, that accuracy of Cued Speech recognition does not depend on noise.

REFERENCES

- Aboutabit, N., Beautemps, D., & Besacier, L., 2008 (accepted with revision). Lips and Hand Modeling for Recognition of the Cued Speech Gestures: The French Vowel Case. *Speech Communication*.
- Cornett, R. O., 1967. *Cued Speech*, American Annals of the Deaf, 112, pp. 3-13.