ESR 10. MULTISTREAM NEURAL ARCHITECTURES FOR CUED SPEECH RECOGNITION USINGA PRE-TRAINED VISUAL FEATURE EXTRACTOR AND CONSTRAINED CTC DECODING

Sanjana Sankar, Denis Beautemps, & Thomas Hueber

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

sanjana.sankar@gipsa-lab.grenoble-inp.fr

Cued Speech (CS) is a visual communication tool developedby Cornett [1] in 1967 to help people with hearing impair-ment to better understand the spoken language. It encodes speech as a combination of visible hand shapes (for consonants) and hand positions (for vowels) to highlight the uttered phoneme and complement lip-reading [1]. The purpose of this study is to automate the process of transcripting Cued Speech. We propose a simple and effective approach for automatic recognition of CS. The proposed approach is based on a pre-trained hand and lips tracker used for visual feature extraction and a phonetic decoder based on a multistream recurrent neural network trained with connectionist temporal classification loss and combined with a pronunciation lexicon. The proposed system is evaluated on an updated version of the French CS dataset CSF18 [2] for which the phonetic transcription has been manually checked and corrected. The use of an efficient pre-trained feature extractorallowed to reduce the complexity of the (visual) phonetic decoder. Also, cleaning the CSF18 dataset to account for differences in cueing and pronunciation of certain phonemes improves the performance significantly. With a decoding accuracy at the phonetic level of 70.88\%, the proposed system outperforms the previous CNN-HMM decoder developed by our lab [2] and competes with more complex baselines [3][4].

[1] R. O. Cornett, “Cued speech,” vol. 112, no. 1, pp. 3–13,1967.

[2] Li Liu, Thomas Hueber, Gang Feng, and DenisBeautemps, “Visual Recognition of Continuous CuedSpeech Using a Tandem CNN-HMM Approach,” inProc. of Interspeech, 2018, pp. 2643–2647.

[3] Katerina Papadimitriou and Gerasimos Potamianos, “Afully convolutional sequence learning approach for cuedspeech recognition from videos,” inProc. of EUSIPCO,2021, pp. 326–330.

[4] Jianrong Wang, Ziyue Tang, Xuewei Li, Mei Yu, QiangFang, and Li Liu, “Cross-Modal Knowledge DistillationMethod for Automatic Cued Speech Recognition,” inProc. of Interspeech 2021, 2021, pp. 2986–2990.