Multi Domain Cantonese Corpus (MDCC) Corpus for Automatic Speech Recognition for Cantonese Language

License agreement

The Multi Domain Cantonese Corpus (MDCC) is made available to the scientific community by the Center of Artificial Intelligence Research (CAiRE) of the Hong Kong University of Science and Technology (HKUST). The database is meant as a tool for studying new techniques and algorithms for the analysis of low resource languages such as Cantonese.

CAiRE maintains the copyright of the data, and it is its sole distributor. The data is made available for research purposes upon signature of this document. The requestor agrees to the following terms:

• Research only: The database is made available for research purposes only, any commercial use of the data is forbidden.

• Redistribution: The database will not be distributed, full or in part, to any third party without prior written approval from the CAiRE of HKUST.

• Publication: The use of the data is allowed for illustrative purposes in scientific publications only. In no case should the original subjects be caused embarrassment.

• Citation: All documents reporting on any research which uses the database has to acknowledge the Computer Vision Laboratory by citing the following article:

Yu, T., Frieske, R., Xu, P., Cahyawijaya, S., Yiu, C.T., Lovenia, H., Dai, W., Barezi, E.J., Chen, Q., Ma, X., Shi, B.E., & Fung, P. (2022). Automatic Speech Recognition Datasets in Cantonese: A Survey and New Dataset. ArXiv, abs/2201.02419.

•Warranty: THE DATA AND THE SOFTWARE COMING WITH IT IS PROVIDED “AS IS” AND THE PROVIDER GIVES NO EXPRESS OR IMPLIED WARRANTIES OF ANY KIND, INCLUDING WITHOUT LIMITATION THE WARRANTIES OF FITNESS FOR ANY PARTICULAR PURPOSE AND NON-INFRINGEMENT. IN NO EVENT SHALL THE PROVIDER BE HELD RESPONSIBLE FOR LOSS

OR DAMAGE CAUSED BY THE USE OF THE DATA.

Name Position

Date Signature

Organization

Email to send the signed document:

E-Mail: caire@ust.hk

www.caire.ust.hk