DNN-Based Missing-Data Mask Estimation for Noise-Robust ASR in Dual-Microphone Smartphones

Iván López-Espejo[*], *José A. González*[†], *Angel M. Gomez*[*], *and Antonio M. Peinado*[*]

[*]Dpt. of Signal Theory, Telematics and Com., University of Granada, Spain
[†]Dpt. of Computer Science, University of Sheffield, UK
{iloes,amgg,amp}@ugr.es, j.gonzalez@sheffield.ac.uk

Automatic speech recognition (ASR) technology is experiencing a new upswing in recent times thanks to the latest portable electronic devices (e.g. smartphones or tablets). In addition, these devices are beginning to integrate small microphone arrays (i.e. microphone arrays composed by a few number of sensors) especially intended to perform noise reduction on the speech signal. While this small microphone array feature is especially being exploited for speech enhancement purposes, few benefit is being taken for noise-robust ASR. Thus, we wish to present some of the advances recently achieved in our research group on the topic of noise-robust ASR with small microphone arrays. In particular, we show that the dual-channel information provided by a smartphone with a dual-microphone can be exploited to easily estimate accurate missing-data masks to perform noise-robust ASR. The followed approach is based on deep neural networks (DNNs), which have demonstrated to be a powerful tool in the field of signal processing in many different ways. Once the missing-data mask is estimated, its quality is evaluated both in terms of the percentage of wrongly estimated mask bins and the word accuracy obtained when used by a spectral reconstruction method. The considered spectral reconstruction method is called truncated-Gaussian based imputation (TGI). Moreover, such experiments are performed on the AURORA2-2C-CT (Aurora-2 - 2 Channels - Close-Talk) database, also developed in our research group. The AURORA2-2C-CT is based on the well-known Aurora-2 database and emulates the acquisition of noisy speech signals with a dual-microphone smartphone used in close-talk conditions (i.e. when the loudspeaker of the smartphone is placed at the ear of the user). Our experimental results show that the DNN is able to exploit the dual-channel information in a simple and efficient way outperforming state-of-the-art single-channel noise-robust approaches.