Chapter 7

Automatic Speech Recognition Using Missing Data Techniques: Handling of Real-World Data

Jort F. Gemmeke, Maarten Van Segbroeck, Yujun Wang, Bert Cranen, Hugo Van hamme

Abstract In this chapter, we investigate the performance of a missing data recognizer on real-world speech from the SPEEC ON and SpeechDat-Car databases. In previous work we hypothesized that in real-world speech, which is corrupted not only by environmental noise, but also by speaker, reverberation and channel effects, the ‘reliable’ features do not match an acoustic model trained on clean speech. In a series of experiments, we investigate the validity of this hypothesis and explore to what extent performance can be improved by combining MDT with three conventional techniques, viz. multi-condition training, dereverberation and feature enhancement. Our results confirm our hypothesis and show that the mismatch can be reduced by multi-condition training of the acoustic models and feature enhancement, and that these effects combine to some degree. Our experiments with dereverberation reveal that reverberation can have a major impact on recognition performance, but that MDT with a suitable missing data mask is capable of compensating both the environmental noise as well as the reverberation at once.

7.1 Introduction

Automatic speech recognition (ASR) performance drops rapidly when speech is corrupted with increasing levels of unfamiliar background noise (i.e., noise not seen during training) since the observed acoustic features no longer match the acoustic models. One of the most effective approaches to improving the noise robustness of a speech recognizer is to perform multi-condition training [15]: Rather than acoustic
models being trained on speech from a quiet environment only, they are trained directly on noisy speech signals. By carefully selecting the training speech to reflect the multiple acoustic conditions under which the system must operate, it is possible to minimize the mismatch between training and test/usage conditions. While often effective, recognition accuracies obtained with multi-condition training quickly deteriorate when the noisy environment deviates from the one that was used for training. Another disadvantage of multi-condition training is that the performance for truly clean speech tends to degrade.

Missing Data Techniques (MDTs) [25] are a very different approach to improving noise robustness that ideally overcomes the problems of multi-condition training. MDTs, first proposed in [6], build on two assumptions: The first assumption is that it is possible to estimate, prior to decoding, which spectro-temporal elements in the acoustic representation of noisy speech are reliable (i.e., dominated by speech energy) and which are unreliable (i.e., dominated by background noise). These reliability estimates are referred to as missing data masks. The second assumption is that the statistics of the features which are considered as dominated by speech energy match with the statistics of clean speech training data. This assumption implies that the acoustic models of MDT recognizers can be trained using clean speech.

In the unreliable elements, the speech information is considered missing, and the challenge is then to do speech recognition with partially observed data. In this work, we focus on the so-called imputation approach [24], which handles the missing elements by replacing them with clean speech estimates. Classic imputation methods include correlation and cluster-based reconstruction [23, 25], state-dependent imputation [17], which combines front-end imputation and classifier modification, and the Gaussian-dependent method [32], which additionally allows for reconstruction in the cepstral and PROSPECT domains. The latter method is employed in this chapter.

While imputation has proven effective for increasing noise robustness in the presence of both stationary and non-stationary noise, most of the existing knowledge about the effectiveness of MDT has been acquired using databases with noisy speech that has been constructed by artificially adding noise of various types and intensities to clean speech (see, e.g., [7, 23]). Using artificially corrupted data is attractive as it allows creating a missing data mask based on exact knowledge of the speech and noise power in each time-frequency cell. This facilitates comparison of different MDT approaches and allows for analysis of the influence of errors in reliability estimation.

Real-world recordings, however, are generally not only corrupted by background noise, but can also be affected by room acoustics. Moreover, real-world recordings are more likely to introduce a mismatch between the observed speech and the speech on which the recognizer is trained, due to microphone characteristics and speaker-specific behavior such as lip noises and the Lombard effect. Very few reports exist that describe the effectiveness of single-channel MDT recognition on real-world recordings (notable exceptions are [13, 19, 27]). In previous research we have used the SPEECON[16] and SpeechDat-Car[30] databases for that purpose. The SPEECON and SpeechDat-Car databases are recorded in realistic