Chapter 10

Summary

Recently, significant research advancements have transformed the field of hidden-Markov-model (HMM)-based automatic pattern recognition. This has allowed for a wide use of automatic pattern recognition systems (PRSs) like automatic speech recognition and automatic fault diagnosis in many modern applications such as navigation systems and dictation software.

Unfortunately, in the presence of noise, the recorded signals undergo massive corruptions, which degrade the performance of automatic pattern recognition systems rapidly. The main focus of this thesis is to develop application-specific frameworks for achieving a high level of robustness of PRSs against different noise types and levels.

In the realm of automatic speech recognition (ASR), a range of so-called uncertainty-of-observation decoding approaches have recently been proposed to achieve the intended robustness. The main idea of these approaches is to reduce the mismatch between the corrupted acoustic data and its underlying statistical model by considering the observed distorted acoustic observations not as deterministic, but rather as random variables with time-varying levels of uncertainty. The larger the uncertainty of a certain acoustic feature is, the smaller its contribution to the overall decoding (recognition) procedure becomes.

In this thesis, a new algorithm for estimating the acoustic feature uncertainties has been developed for Gaussian-mixture-model (GMM)-based acoustic models. This new algorithm avoids the shortcomings of the conventional estimation approaches, while achieving significant performance improvements in adverse acoustic conditions.
How to deploy the observation uncertainties in the recognition process of GMM/HMM-based ASR systems is another important question that has been addressed in this thesis. Investigating the uncertainty-of-observation decoding techniques and the relationships among them has allowed the development of a new approach, termed significance decoding, that inherits some of their properties and gives satisfactory improvements in real scenarios.

Deep-neural-network (DNN)-based acoustic models have recently made a breakthrough in ASR. Therefore, different aspects of speech recognition using uncertain observations, e.g., uncertainty estimation, uncertainty propagation and uncertainty-based modification of ASR scores, have also been investigated for DNN systems.

Another strategy for immunizing the ASR performance against noise is to use, in conjunction with the conventional acoustic features, another information source that is independent of the acoustic noise. Recently, the video signal of the speaker’s lip movements has been considered as a good candidate for this purpose. However, fusion of audio and video data is still a critical question. In this work, it has been investigated, how to improve the accuracy of audio-visual ASR (AV-ASR) systems by optimally combining the audio speech signals and video images. A novel framework has been proposed to optimize the fusion between the audio and video modality by using so-called stream weights. The stream weights help the AV-ASR systems to dynamically assess which modality (audio or video) is more reliable and, hence, should contribute more to the recognition decision. The new framework introduced in this thesis has led to significant improvements over the conventional approaches in terms of recognition accuracy.

The improvement gained using the aforementioned approaches was the motivation to also exploit them in the field of speech enhancement. Therefore, a new framework, called twin-HMM-based audio-visual speech enhancement, has been proposed to denoise and enhance the quality of distorted speech in multimedia signals, making use of the available video information. The quality of the acoustic signals has been greatly improved by this approach even under exceedingly difficult acoustical conditions.

The structure of electromechanical devices has recently become very complex so that manual diagnosis of their possible faults has become quite difficult. In the last part of this thesis, a noise-robust multimodal automatic fault diagnosis (AFD) system has been developed to detect faults in different components of elevator doors. The system has
been designed so that faulty parts can be detected before the operation of the elevator is affected. The robustness against noise has been achieved using multi-modal features extracted from vibration sensors and microphones. Such a system can significantly reduce the operation and maintenance cost by replacing periodic regular maintenance by predictive maintenance based on the AFD system output while still ensuring reliable operation.

Overall, in this thesis, novel approaches based on observation uncertainties and fusion of multimodal features have been proposed to increase the robustness of HMM-based PRSs against acoustic noise. Using these approaches, significant performance improvements have been achieved in applications like automatic speech recognition, speech enhancement, and automatic fault diagnosis. A potential extension of this study lies in exploiting the proposed approaches in other machine learning applications like machine translation and Bayesian-network-based speaker localization. DNNs have recently achieved a breakthrough in the realm of pattern recognition. Therefore, a wider use of DNNs in the different parts of the proposed frameworks is another natural extension of this work.