Chapter 2
Uncertainty Decoding and Conditional Bayesian Estimation

Reinhold Haeb-Umbach

Abstract In this contribution classification rules for HMM-based speech recognition in the presence of a mismatch between training and test data are presented. The observed feature vectors are regarded as corrupted versions of underlying and unobservable clean feature vectors, which have the same statistics as the training data. Optimal classification then consists of two steps. First, the posterior density of the clean feature vector, given the observed feature vectors, has to be determined, and second, this posterior is employed in a modified classification rule, which accounts for imperfect estimates. We discuss different variants of the classification rule and further elaborate on the estimation of the clean speech feature posterior, using conditional Bayesian estimation. It is shown that this concept is fairly general and can be applied to different scenarios, such as noisy or reverberant speech recognition.

2.1 Introduction

Improving the robustness of state-of-the-art automatic speech recognition (ASR) continues to be an important research area. Current hidden Markov model (HMM) based speech recognition systems are notorious for performing well in matched training and test conditions while quickly degrading in the presence of a mismatch. While such a mismatch may be caused by many factors, probably one of the most studied problems is improving the robustness of a recognizer trained on clean training data to test data being corrupted by acoustic environmental noise.

Huge research efforts have been devoted to overcoming this lack of robustness, and a wealth of methods has been proposed. These can be categorized into either methods that try to compensate the effect of distortions on the features (so-called
front-end methods) or approaches that modify the models used in the recognizer to better match the incoming distorted feature stream (back-end methods).

Traditionally, front-end methods aim at obtaining point estimates of the uncorrupted, clean features. Likewise, back-end methods usually try to obtain point estimates of parameters, such as the mean vectors of the observation probabilities. These estimates are then “plugged” into the Bayesian decision rule as if they were perfect estimates. However, more recently the focus has shifted to estimating the features or parameters together with a measure of reliability of the estimate and propagating the uncertainty to the decision rule [4, 9, 17, 20, 26–28, 37, 46, 47]. The underlying rationale is that an estimate is never perfect and that the recognizer can benefit from knowing the estimation error variance by de-emphasizing the contributions of unreliable estimates to the overall decision on the word sequence.

The use of such optimal decision rules is by no means new. How to modify the Bayesian decision rule in the presence of missing or noisy features can be found in many textbooks on pattern recognition; see, e.g., [22]. Two developments, however, are fairly recent: how to modify the decision rule for HMM classifiers and how to obtain reliability information for a given class of distortions.

In this contribution we first show how a suboptimal decision rule for HMM-based speech recognition in the presence of corrupted feature vectors can be derived from the optimal, though infeasible, decoding rule. It will be seen that this decoding rule exploits the temporal correlation between the feature vectors, which is otherwise not used in an HMM-based recognizer due to the predominant conditional independence assumption. Several other decoding rules proposed in the literature, including missing-feature techniques, can be viewed as approximations to this rule.

A key element of the classification rule is the posterior density of the clean feature vector, given the observed and corrupted feature vectors. We show how this posterior can be estimated in principle by using well-known results from the field of optimal filtering. While the principle approach is applicable to various kinds of signal degradation, a crucial element, the observation model, has to be developed specifically for each kind of degradation, such as noise or reverberation.

This contribution is organized as follows: starting from the optimal Bayesian decision rule, we first derive a classification rule for an HMM-based speech recognizer in the presence of corrupted feature vectors, and follow it with a discussion of related uncertainty decoding rules, including missing feature theory. While uncertainty can be treated in either the feature or the model domain, we concentrate here on the feature domain and motivate this in Section 2.3. Next, the estimation of the clean speech feature posterior is discussed, comprising the components a priori model, observation model and inference algorithm. Since an in-depth treatment for a specific problem, such as noisy or reverberant speech recognition, is beyond the scope of this contribution, we refer to other chapters of this book for details. Finally, we briefly discuss the limitations of this approach and thus share perspectives for future research. The chapter ends with some conclusions in Section 2.6.