Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 874–894

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Uncertainty-based learning of acoustic models from noisy data☆,☆☆

Alexey Ozerov[a], Mathieu Lagrange[b], Emmanuel Vincent[c,*]

[a] *Technicolor Research & Innovation, France*
[b] *STMS – IRCAM – CNRS – UPMC, France*
[c] *INRIA, Centre de Rennes – Bretagne Atlantique, France*

Received 14 January 2012; received in revised form 29 May 2012; accepted 5 July 2012
Available online 24 July 2012

Abstract

We consider the problem of acoustic modeling of noisy speech data, where the uncertainty over the data is given by a Gaussian distribution. While this uncertainty has been exploited at the decoding stage via uncertainty decoding, its usage at the training stage remains limited to static model adaptation. We introduce a new expectation maximization (EM) based technique, which we call *uncertainty training*, that allows us to train Gaussian mixture models (GMMs) or hidden Markov models (HMMs) directly from noisy data with dynamic uncertainty. We evaluate the potential of this technique for a GMM-based speaker recognition task on speech data corrupted by real-world domestic background noise, using a state-of-the-art signal enhancement technique and various uncertainty estimation techniques as a front-end. Compared to conventional training, the proposed training algorithm results in 3–4% absolute improvement in speaker recognition accuracy by training from either matched, unmatched or multi-condition noisy data. This algorithm is also applicable with minor modifications to maximum a posteriori (MAP) or maximum likelihood linear regression (MLLR) acoustic model adaptation from noisy data and to other data than audio.

Keywords: Noisy data; Training; Uncertainty; Classification; Acoustic model; Gaussian mixture model; Hidden Markov model; Expectation–maximization

1. Introduction

Classification and detection systems often face a variety of distortions (e.g., additive or convolutive) resulting in noisy data. In order to achieve noise robustness, at least three complementary approaches can be taken. At the signal level, one can apply enhancement techniques such as noise suppression (Ephraim, 1992), source separation (Vincent et al., 2012) or dereverberation (Delcroix et al., 2009). At the feature level, one can define features that are robust to the considered type of noise or to the residual noise after enhancement (Nadeu et al., 1997). Finally, at the classifier (or decoder) level, one can account for possible distortion of the features within the classifier itself. In this paper, we focus

☆ This paper has been recommended for acceptance by Jon Barker.
☆☆ This work was performed while A. Ozerov was with INRIA and partly supported by OSEO, the French State agency for innovation, under the Quaero program.
* Corresponding author. Tel.: +33 299 842 269.
E-mail addresses: alexey.ozerov@technicolor.com (A. Ozerov), mathieu.lagrange@ircam.fr (M. Lagrange), emmanuel.vincent@inria.fr (E. Vincent).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.07.002

on the latter approach by considering the problem of acoustic modeling of noisy speech data using Gaussian mixture models (GMMs) or hidden Markov models (HMMs).

The most straightforward approach to increasing the accuracy of the classifier is to train the models over *matched* training data exhibiting the same type and amount of noise as the test data (Droppo and Acero, 2008). Unfortunately, such data are not always available and one may be constrained to use *clean*, *multi-condition* or even *unmatched* training data whose noise characteristics do not match those of the test data. This is an example of the general problem known as *concept shift* in the machine learning community whereby the noise contribution varies between the training and test datasets (Moreno-Torres et al., 2012). One approach to this problem consists of clustering the model components and adapting their means and covariances within each cluster via a static (time-invariant) transform (Deng et al., 2000; Gales, 2011). This approach accounts for the *uncertainty* over the data induced by noise, but it does not exploit estimates of this uncertainty that may be available from the signal enhancement front-end and it is restricted to rather stationary noise environments by design. More recently, several approaches have been proposed to dynamically adapt the model parameters in each time frame in response to nonstationary noise. A separate signal enhancement front-end is employed that allows the use of harmonicity cues and spatial cues, which are essential for signal enhancement but not modeled by feature-domain GMMs or HMMs. The uncertainty over the data is then typically encoded either by a set of binary flags indicating whether each data dimension is "observed" or "missing" (Cooke, 2001) or by a Gaussian distribution whose mean and covariance matrix represent, respectively, the estimated underlying clean data and noise covariance (Deng et al., 2005). This last approach is the most flexible, since it allows to the amount of noise to be quantified along with the noise correlation between different data dimensions in each time frame. In the following, we focus on this approach, which has been successfully employed by the best scoring system (Delcroix et al., 2011) of the 2011 The PASCAL CHiME Speech Separation and Recognition Challenge (Barker et al., 2013).

While several algorithms have been derived that exploit uncertainty over the test data (Cooke, 2001; Barker et al., 2005; Deng et al., 2005; Srinivasan and Wang, 2007; Delcroix et al., 2009; Shao et al., 2010; Kolossa et al., 2010), uncertainty over the training data has not been fully exploited so far. Most approaches (Cooke, 2001; Barker et al., 2005; Deng et al., 2005; Srinivasan and Wang, 2007; Shao et al., 2010) assume *conventional training from clean data*. This training strategy is not always applicable in the case of, e.g., field recording or mobile recording where the whole recording might be corrupted by noise. Also, even when sufficient clean data are available for training, the uncertainty over the test data is never perfectly estimated in practice such that some noise may remain that is not accounted for. Recently, Delcroix et al. (2011) and Kolossa et al. (2011) achieved better results by *conventional training from noisy data*. Nevertheless, this heuristic strategy remains sensitive to mismatched training and test noise conditions and, even in matched conditions, the noise variance is overestimated by a factor of two. Indeed, the noise is taken into account both at the training stage within the model parameters and at the decoding stage within the uncertainty and these two contributions add up. Liao and Gales (2007) proposed a more principled training algorithm for use with static model adaptation, but the exploitation of dynamic Gaussian uncertainty over the training data remains an open issue.

In order to address this issue, we introduce a new EM based technique that allows us to train GMMs and HMMs directly from noisy data with dynamic Gaussian uncertainty. By analogy with the *uncertainty decoding* algorithm of Deng et al. (2005), we refer to this training strategy as *uncertainty training*. The proposed algorithm generalizes both the algorithm of Ghahramani and Jordan (1994) for binary uncertainty and the algorithm of Arberet et al. (2012) for Gaussian uncertainty with diagonal covariance and zero-mean GMMs with diagonal covariances, which were applied in different contexts. Furthermore, it is also applicable with minor modifications to maximum a posteriori (MAP) (Gauvain and Lee, 1994) or maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) model adaptation and to other noise-corrupted data, e.g., microarray data for which different genes and different conditions have different levels of experimental and biological noise whose variance can be estimated (Sanguinetti et al., 2005). This article expands our preliminary paper (Ozerov et al., 2011) by providing more insight about the proposed GMM training algorithm, by extending it to HMMs, and by extensively evaluating it for a speaker recognition task with real-world data and uncertainty estimates as opposed to synthetic data and oracle (i.e., ideal) uncertainty. For the sake of conciseness, we focus on GMMs in most of the paper and in the experimental study, and we present the algorithm for HMMs in Appendix B.

As a by-product, we also introduce the following two new uncertainty estimators. For the particular task and signal enhancement algorithm employed, we show that the best uncertainty estimator among the variety of estimators considered here is obtained by computing the uncertainty resulting from multichannel Wiener filtering (Fischer and Kammeyer, 1997) in the time-frequency domain and propagating it to the Mel Frequency Ceptral Coefficient (MFCC)