Joint Bayesian Predictive Classification and Parallel Model Combination for Robust Speech Recognition

Svein G. Pettersen, Magne H. Johnsen, Tor A. Myrvoll

Department of Electronics and Telecommunications
Norwegian University of Science and Technology
sveingun@iet.ntnu.no, mhj@iet.ntnu.no, myrvoll@iet.ntnu.no

Abstract

In this paper we present an approach that makes use of both Bayesian predictive classification (BPC) and parallel model combination (PMC) to achieve increased robustness towards noise. PMC provides a method for finding parameter estimates for speech corrupted by noise, while BPC is a method that compensates for uncertainty of parameter estimates. Thus, these methods can be combined in order to obtain knowledge about the mismatch situation and simultaneously account for uncertainty in this knowledge. We apply this technique in an unsupervised approach on the Aurora2 database and show that good performance is obtained.

1. Introduction

Robustness is essential for practical automatic speech recognition (ASR) systems in order to avoid severe degradation of performance due to mismatch between training and application conditions. Ambient background noise, channel and microphone variations, as well as speaker variations like dialect, age and sex, are examples of factors that cause mismatch.

A variety of schemes have been proposed to deal with the robustness issue, such as robust frontends, SNR-enhancement and model based compensation. Further, within a specific scheme, alternative strategies are chosen for different mismatch situations. Methods for unsupervised model compensation due to background noise include strategies such as histogram equalization (HE) [1], PMC [2], minimum mean square error filtering of noisy features [6] and BPC-based model compensation [5]. Some of the methods, like HE, require access to the whole utterance and thus introduce a delay. PMC is originally based on a separate noise model estimate trained from a representative noise database. Online BPC-based techniques can be applied either by updating parameters in frame-based Viterbi-mode (VBPC) or utterance-based model compensation mode (BP-MC).

This paper addresses unsupervised methods using one pass recognition where we only have access to clean MFCC-based models. Under this restriction we investigate the use of PMC and BPC to obtain better model estimates and at the same time compensate for uncertainty. The first step in the proposed approach is to use PMC to obtain model estimates for noisy speech. Then, instead of using the PMC estimates directly as a new model for noisy speech, we incorporate this knowledge into a set of prior distributions for use with BPC. This enables us to take our uncertainty about the noisy model into account. An alternative interpretation of the procedure is that we first obtain improved model estimates through PMC and then use BPC with "standard" priors to account for uncertainty.

2. Parallel model combination

The objective of the PMC technique is to estimate corrupted speech parameters for a Hidden Markov Model (HMM) given information about both clean speech and noise. Clean speech models are obtained from a set of HMMs trained using noise-free data, while noise models can either be obtained from separate recordings in noisy environments or from noise-only segments from the utterance to be recognized. Assuming additive noise in the linear domain, an observation in the log spectral domain is given by

$$X_i^l(t) = \log\Big(\exp(S_i^l(t)) + \exp(N_i^l(t))\Big) \tag{1}$$

where $S_i^l(t)$ and $N_i^l(t)$ denote the clean speech and noise at time t respectively, indexed by the element number in the feature vector i. The superscript l is used to indicate that the values are in the log spectral domain. In order to calculate the corrupted HMM parameters one has to find the expectation of (1). Unfortunately, this has no simple closed-form solution. Several approximations have been proposed to be able to solve this problem, such as the log-normal approximation [3] and the log-add approximation [3].

2.1. Log-normal approximation

When using the log-normal approximation, speech and noise parameters are transformed back to the linear domain before they are combined. For static parameters, the corrupted speech parameters can then be found as follows. After mapping speech and noise parameters from the cepstral domain to the log spectral domain, the parameters in the linear domain are given by

$$\mu_i = \exp(\mu_i^l + \Sigma_{ii}^l/2) \tag{2}$$

$$\Sigma_{ij} = \mu_i\mu_j\Big[\exp(\Sigma_{ij}^l) - 1\Big] \tag{3}$$

where μ_i and Σ_{ij} denote elements of the mean vector and covariance matrix respectively. The corresponding parameters in the log spectral domain are denoted by μ_i^l and Σ_{ij}^l. Let $\{\tilde{\mu}, \tilde{\Sigma}\}$ denote the noise parameters. By assuming that the sum of two lognormally distributed variables is itself approximately lognormally distributed, the corrupted speech parameters in the linear domain are given by

$$\hat{\mu} \approx \mu + \tilde{\mu} \tag{4}$$

$$\hat{\Sigma} \approx \Sigma + \tilde{\Sigma}. \tag{5}$$

These parameters can then be transformed back to the log spectral domain by inversion of (2) and (3).

$$\hat{\mu}_i^l = \log(\hat{\mu}_i) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \quad (6)$$

$$\hat{\Sigma}_{ij}^l = \log\left(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\hat{\mu}_j} + 1\right) \quad (7)$$

A compensation scheme for delta parameters has also been proposed [2]. The corrupted delta parameters are calculated as weighted sums of the clean speech and noise delta parameters. The corrupted delta means are given by

$$\Delta\hat{\mu}_i = \bar{\gamma}_i\Delta\mu_i + \bar{\eta}_i\Delta\tilde{\mu}_i \quad (8)$$

where $\Delta\mu_i$ and $\Delta\tilde{\mu}_i$ are the clean speech and noise delta means respectively, and $\bar{\gamma}_i$ and $\bar{\eta}_i$ are given by

$$\bar{\gamma}_i = \frac{\frac{\mu_i}{\tilde{\mu}_i}}{\frac{\mu_i}{\tilde{\mu}_i} + 1} \quad (9)$$

$$\bar{\eta}_i = \frac{1}{\frac{\mu_i}{\tilde{\mu}_i} + 1}. \quad (10)$$

2.2. Log-add approximation

When using the log-add approximation, speech and noise models are combined in the log spectral domain. The effect of the variance on the corrupted speech mean is ignored. The corrupted mean values in the log spectral domain are approximated by

$$\hat{\mu}_i^l \approx \log\left(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)\right). \quad (11)$$

This is a first order approximation of the expectation of the corrupted speech mean. Using a similar first order approximation, corrupted variances (diagonal covariance matrix) can be found as

$$\hat{\Sigma}_{ii}^l \approx \Sigma_{ii}^l + (\mu_i^l)^2 + 2\mu_i^l\log\left(1 + \exp(\tilde{\mu}_i^l - \mu_i^l)\right) + \left[\log\left(1 + \exp(\tilde{\mu}_i^l - \mu_i^l)\right)\right]^2 - (\hat{\mu}_i^l)^2. \quad (12)$$

By assuming that simple differences are used for delta parameters, eq. (1) can be rewritten as [4]

$$\begin{aligned}\Delta X_i^l(t) = &\log\left(\exp(\Delta S_i^l(t) + S_i^l(t-w))\right. \quad (13)\\ &\left. + \exp(\Delta N_i(t) + N_i^l(t-w))\right)\\ &- \log\left(\exp(S_i^l(t-w)) + \exp(N_i^l(t-w))\right)\end{aligned}$$

where w is the difference offset. Further, by assuming stationarity, the corrupted delta means can be approximated by

$$\begin{aligned}\Delta\hat{\mu}_i^l = &\log\left(\exp(\Delta\mu_i^l + \mu_i^l) + \exp(\Delta\tilde{\mu}_i + \tilde{\mu}_i^l)\right) \quad (14)\\ &- \log\left(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)\right).\end{aligned}$$

3. Bayesian predictive classification for ASR

3.1. Bayesian predictive classification

Let W denote a word sequence, and let $\mathbf{X}$ denote the associated acoustic observation (i.e. a sequence of feature vectors). Since the true joint distribution of $(W, \mathbf{X})$ is not known, a plug-in maximum a posteriori (MAP) decoder is normally used for ASR. This decoder is given by

$$\hat{W} = \arg\max_W p(W|\mathbf{X}) = \arg\max_W p_\Lambda(\mathbf{X}|W)P_\Gamma(W) \quad (15)$$

where $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$ are parametric models, with parameters Λ and Γ that must be estimated from training data.

Due to uncertainty in the assumption about the parametric models and in the parameter estimates, the plug-in MAP decoder is not optimal. Further, if there is mismatch between training conditions and testing conditions, the performance of the decoder can be significantly reduced. BPC can to some extent compensate for such uncertainty [5].

In the following, only uncertainty in the acoustic model will be considered. In the BPC approach a prior distribution is used for describing the uncertainty of the parameter estimates. The probability density function (pdf) $p_\Lambda(\mathbf{X}|W)$ in (15) is then replaced by a predictive pdf given by

$$\tilde{p}(\mathbf{X}|W) = \int_\Omega p(\mathbf{X}|\Lambda, W)p(\Lambda|\phi, W)d\Lambda \quad (16)$$

where $p(\Lambda|\phi, W)$ is the prior distribution with hyperparameters ϕ, and Ω is the admissible region of the parameter space.

3.2. Bayesian predictive density based model compensation

Bayesian predictive density based model compensation (BP-MC) [5] is a simple approach to BPC for speech recognition. Let the observation pdf for a state i have the form of a mixture of multivariate Gaussian pdfs $\{f(\boldsymbol{x}|\theta_{ik})\}$, where θ_{ik} are the parameters of the pdf for mixture k. To apply the BP-MC approach, prior distributions $\{p(\theta_{ik}|\phi_{ik})\}$ with hyperparameters ϕ_{ik} are needed for all mixtures. Each mixture pdf is replaced by a predictive density of the form

$$\tilde{f}_{ik}(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta_{ik})p(\theta_{ik}|\phi_{ik})d\theta_{ik}. \quad (17)$$

Although this approach can be used to compensate for mismatch in both mean and variance of the observation pdf, only the mean will be considered in this paper.

3.3. Prior specification

To make the BPC approach efficient, appropriate prior pdfs must be found. Knowledge about the mismatch situation is important to be able to specify suitable priors. Priors based on the influence of spectral mismatch on cepstral coefficients have been commonly used [5]. Symmetric uniform priors for the mean values then have the following form.

$$|\mu_{ikd} - \mu_{ikd}^*| \leqslant Cd^{-1}\rho^d \quad \text{for } d \geqslant 1 \quad (18)$$

Here μ_{ikd}^* is the original mean for state i, mixture k and element d of the feature vector, μ_{ikd} is the corresponding corrupted mean and C and ρ are constants ($C > 0$; $0 \leqslant \rho < 1$).

The approach presented in this paper adds further knowledge to the prior by using PMC to estimate the prior mean.

4. Experiments

4.1. The AURORA2 task

Experiments are performed on set A of the Aurora2 database, which consists of US-English digits in presence of artificially

added noise. Four different types of noise are found in this set: subway, babble, car, and exhibition noise. Each of these noise types are added at different signal to noise ratios (SNRs), from 20 dB down to -5 dB in steps of 5 dB.

4.2. Experimental setup

The Hidden Markov model Toolkit (HTK) frontend was used for feature extraction. Standard Aurora2 scripts have been used for training of the recognizer, but the 0th order cepstral coefficient was used instead of log-energy.

Recognition was run using static, delta and delta-delta parameters. PMC was used to estimate new means for static and delta, while delta-delta parameters were left unchanged. In some experiments, PMC was also used to estimate new variances for static parameters. Noise parameters were estimated using the first 10 frames of each test utterance.

BPC was used to compensate for uncertainty in all parameters. Since preliminary BPC tests gave approximately equal performance for Gaussian and uniform priors with the same variance, the former was chosen for the experiments.

4.3. Results

Plots showing recognition performance for different techniques, including baseline (○), and different noise types are given in Figures 1-8. In these plots, recognition using only PMC with updated mean values has been given the label PMC (△). When variances are also compensated for, the label PMC-v (□) is used. The joint BPC and PMC technique is labeled BPC-PMC (◇). For each noise type there are two plots, one showing results when using the log-add approximation for PMC and one showing results for the log-normal approximation.

The results show that BPC-PMC obtains improved performance compared to PMC for subway, car, and exhibition noise at low SNRs. In addition, the performance at high SNRs is comparable to PMC. However, for babble noise performance drops when using BPC-PMC compared to PMC. PMC-v only seems to work better than PMC in a few cases.

The PMC technique obtains significant improvement of recognition performance compared to baseline. For SNRs of 10 dB and higher, word accuracies of at least 80% are obtained for all noise types. This indicates that good estimates of noisy models are obtained. However, for low SNRs there is a higher degree of uncertainty in the compensated models. By using the noise estimates in prior specification for BPC, we are, to some extent, able to compensate for this uncertainty. This, however, does not seem to apply for babble noise where a drop in performance was observed. The use of BPC results in increased variances. Apparently, this causes problems in the case of babble noise.

5. Conclusion

In this paper, we have investigated the joint use of PMC and BPC in noisy conditions. PMC was used for obtaining estimates of corrupted HMM parameters. BPC was used to account for uncertainty in these estimates. This technique was tested on set A of the Aurora2 database in a case where only clean models were used and no other information than the utterance to be recognized was available. The tests showed that PMC improved performance significantly for all cases. Further, BPC gave an additional improvement of performance at low SNRs for three of the four noise types in the test set, as well as obtaining comparable performance for high SNRs.

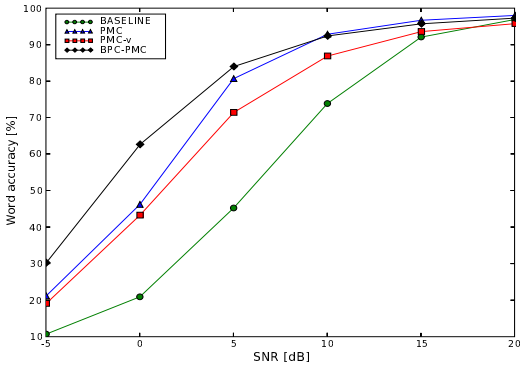

Figure 1: Results, subway noise, log-add approx. for PMC

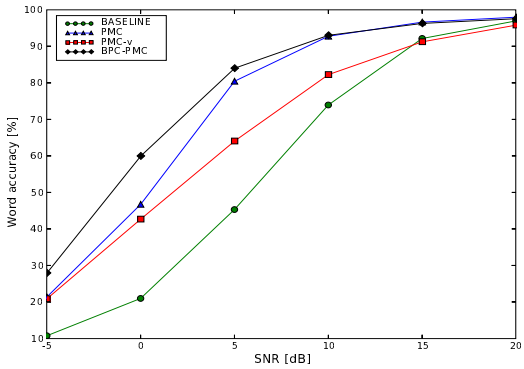

Figure 2: Results, subway noise, log-norm. approx. for PMC

6. Acknowledgements

The work is done as a part of the BRAGE-project, which is organized under the language technology programme KUNSTI and funded by the Norwegian Research Council.

7. References

[1] Hilger, F. and Ney, H., "Quantile Based Histogram Equalization for Noise Robust Speech Recognition", In *Proceedings Eurospeech*, pages 1135-1138, 2001.

[2] Gales, M. F. J. and Young, S. J., "Parallel Model Combination for Speech Recognition in Noise", *Technical report CUED/F-INFENG/TR 135*, June 1993.

[3] Gales, M. F. J., "Predictive Model-Based Compensation Schemes for Robust Speech Recognition", *Speech Communication*, vol. 25, 1998.

[4] Gales, M. F. J. and Young, S. J., "A Fast and Flexible Implementation of Parallel Model Combination", In *Proceedings ICASSP*, pages 133-136, 1995.

[5] Jiang, H., Hirose, K. and Huo, Q., "Robust Speech Recognition Based on a Bayesian Prediction Approach", *IEEE Trans. Speech and Audio Processing*, vol. 7, July 1999.

[6] Myrvoll, T. A. and Nakamura, S., "Online Minimum Mean Square Error Filtering of Noisy Cepstral Coefficients Using a Sequential EM algorithm.", In *Proceedings ICSLP*, pages 117-120, 2004.

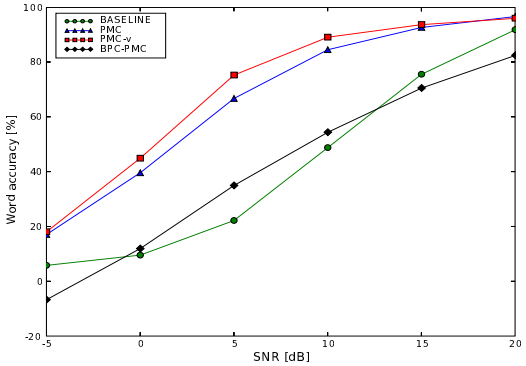

Figure 3: Results, babble noise, log-add approx. for PMC

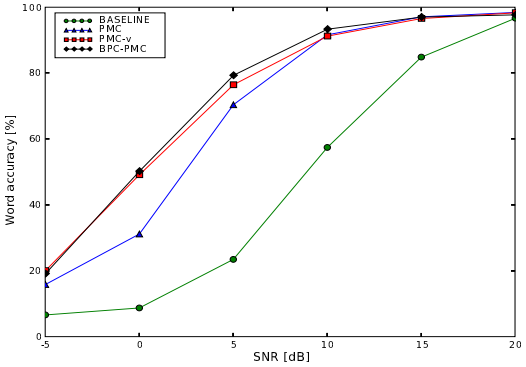

Figure 6: Results, car noise, log-norm. approx. for PMC

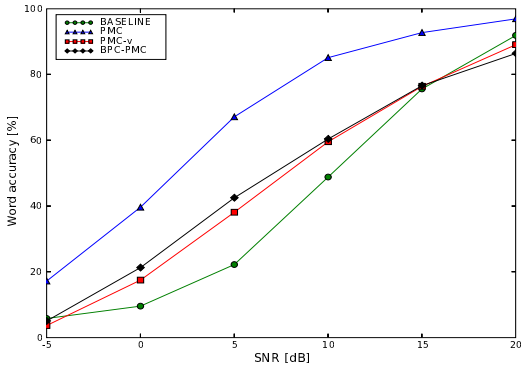

Figure 4: Results, babble noise, log-norm. approx. for PMC

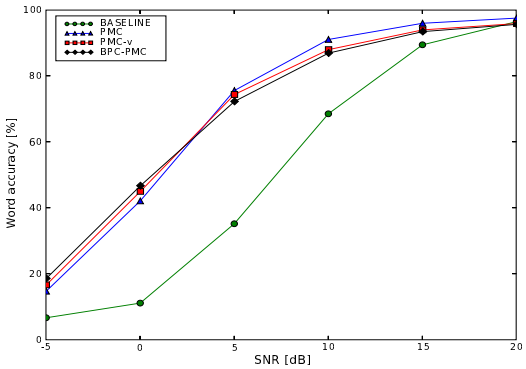

Figure 7: Results, exhibition noise, log-add approx. for PMC

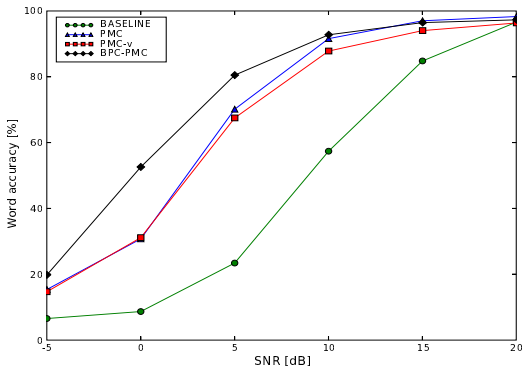

Figure 5: Results, car noise, log-add approx. for PMC

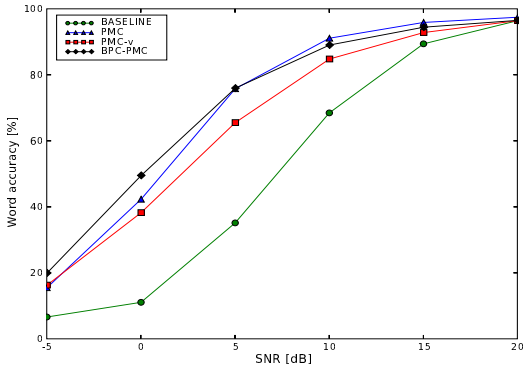

Figure 8: Results, exhibition noise, log-norm. approx. for PMC