IDIAP RESEARCH REPORT

# Direct optimisation of a multilayer perceptron for the estimation of cepstral mean and variance statistics

John Dines [a] Jithendra Vepa [a]

IDIAP–RR 07-13

March 2007

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Martigny, Switzerland

**IDIAP Research Institute** **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-13

# Direct optimisation of a multilayer perceptron for the estimation of cepstral mean and variance statistics

John Dines Jithendra Vepa

March 2007

**Abstract.** We propose an alternative means of training a multilayer perceptron for the task of speech activity detection based on a criterion to minimise the error in the estimation of mean and variance statistics for speech cepstrum based features using the Kullback-Leibler divergence. We present our baseline and proposed speech activity detection approaches for multi-channel meeting room recordings and demonstrate the effectiveness of the new criterion by comparing the two approaches when used to carry out cepstrum mean and variance normalisation of features used in our meeting ASR system.

# 1 Introduction

In automatic speech recognition, it is common practice to carry out the normalisation of features in order to remove unwanted sources of variance such as noise, channel distortion and speaker variabilities. A simple, but effective approach is *cepstral mean and variance normalisation* (CMVN) in which the statistics of cepstrum based speech features are normalised to have zero mean and unit variance. Such normalisation has been shown to be effective in removing convolutional channel distortion, which becomes additive in the cepstral domain [1, 2]. In order to ensure a reliable estimate of mean and variance statistics, silence frames are ignored in the computation, hence in practice, the success of this approach is intrinsically linked with effective speech activity detection (SAD).

An application environment in which robust speech activity detection continues to be a challenging task is in the domain of individual headset microphone recordings (IHM) of meetings to which a significant amount of research effort has been devoted [3, 4, 5]. Drawing on a variety of approaches, this work has demonstrated that the segmentation of IHM meeting room recordings using traditional SAD approaches is generally insufficient and extra care is particularly needed to deal with the presence of cross-talk. While demonstrating improvements over simple SAD schemes, this work has still exhibited substantial performance gaps of 8% to 10% relative WER in comparison to ASR carried out on the reference segmentation, motivating further investigation of this problem.

In our work on the segmentation of IHM recordings we have demonstrated that a direct relationship exists between frame level accuracy of the statistical classifier used in speech activity detection (in this case a multilayer perceptron) and the word accuracy of a first-pass automatic speech recognition (ASR) system [6]. We hypothesised that a possible explanation for this relationship was the improved accuracy of estimation of the CMVN statistics as the frame level accuracy of the classifier increases, thus providing a better match between the acoustic models and normalised features. In order to more directly exploit this relationship, we propose in this paper a new criterion to train the statistical classifier for SAD based on the Kullback-Leibler divergence between normal distributions, with the aim of directly minimising errors in the estimation of the CMVN statistics, and demonstrate the effectiveness of our approach for ASR experiments on IHM meeting room recordings.

The paper is organised as follows: in section 2 we briefly cover background material, including cepstral feature mean and variance normalisation and speech activity detection for IHM meeting room recordings. Section 3 describes in more detail the baseline and proposed approaches for the optimisation of the SAD classifier, followed in section 4 by the presentation of results and discussion from our preliminary experiments. Section 5 rounds up the paper with concluding remarks and ongoing work.

# 2 Background

## 2.1 Cepstral mean and variance normalisation

Cepstrum based features can be made invariant to linear, stationary channel distortions by the subtraction of the mean [1]. This is possible since such distortions are convolutional in the time domain, which become additive in the cepstrum feature space as this is simply a linear transformation of the log-spectrum. Moreover, additional improvement may be obtained by also normalising the variance of the features and only calculating the mean and variance normalisation statistics during periods of speech activity [2]. Such feature normalisation is commonly referred to as cepstral mean and variance normalisation (CMVN). Given the speech/silence segmentation of a recording of $T$ observation vectors $\mathbf{X}_1^T = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)\}$, the estimation and application of cepstral mean and variance normalisation is carried out according to equations 1 – 3.

$$\mu^k = \frac{\sum_{t=1}^{T} I_{\mathbf{x}(t)=\text{speech}} \cdot x^k(t)}{\sum_{t=1}^{T} I_{\mathbf{x}(t)=\text{speech}}} \quad (1)$$

$$\sigma^k = \frac{\sum_{t=1}^{T} I_{\mathbf{x}(t)=\text{speech}} \cdot (\mu^k - x^k(t))^2}{\sum_{t=1}^{T} I_{\mathbf{x}(t)=\text{speech}}} \quad (2)$$

$$\hat{x}^k(t) = \frac{x^k(t) - \mu^k}{(\sigma^k)^{0.5}} \quad (3)$$

where $I_{\mathbf{x}(t)} = \text{speech}$ is an indicator function that is equal to 1 when $\mathbf{x}(t)$ is speech and 0 otherwise. $\mu^k$ and $\sigma^k$ are the estimated observation vector mean and variance statistics respectively, where the superscript indicates the $k$th component of the observation vector $\mathbf{x}(t)$, and $\hat{x}^k(s)$ is the normalised feature vector component.

In training the acoustic models of an ASR system, the cepstral mean and variance statistics are typically collected per speaker using a given speech/silence segmentation (either manually transcribed or, better yet, derived from a forced alignment of the reference transcriptions). However, during recognition the segmentation is unknown and instead must be calculated automatically using SAD. Clearly, a good estimation of the CMVN statistics requires effective speech activity detection, otherwise a mismatch between the acoustic models and features will be introduced.

### 2.2 Speech activity detection in meetings

In our previous work, we have developed a speech activity detection approach for IHM meeting room recordings using a multilayer perceptron (MLP) trained to estimate the posterior probability of speech/silence for a given feature vector input [6]. A relatively sophisticated approach is called for as the speech activity detection task on this data is non-trivial due to there often being a high-level of 'noise' on each participant's microphone, due mostly to either cross-talk from adjacent speakers or non-speech sounds (such as breath and laughter). In the absence of effective speech activity detection, a much higher word error rate (WER) is observed on subsequent ASR.

Training of the MLP is carried out using speech/silence targets based off a forced alignment of the reference transcripts from several meeting data corpora (see [6] for details). MLP training proceeds using error back propagation of the relative-entropy between target and estimated class posterior probabilities (see Section 3.1 for further details). During inference, the recordings are segmented into speech/non-speech portions using a Viterbi search in which scaled likelihoods are generated from the MLP classifier class posterior estimates and class prior probabilities:

$$p(\mathbf{x}(t)|\mathcal{C}_j) = \frac{P(\mathcal{C}_j|\mathbf{x}(t))}{P(\mathcal{C}_j)} \quad (4)$$

where $\mathbf{x}(t)$ is the feature vector input to the MLP[1] and $\mathcal{C}_j, j \in \{sp, sil\}$ represents the speech and silence classes, respectively. Segment minimum duration is imposed via the number of states in the left-right speech/silence model HMM topology. Segment minimum duration, log-insertion penalty, and class priors can all be tuned on development data to optimise the system performance with respect to frame-level speech/silence classification error rate or ASR word error rate.

## 3 Optimisation of the SAD classifier

Error back-propagation training of a multilayer perceptron is carried out through the updating of model parameters via gradient descent by finding expressions for the derivative of the error criterion

[1]Strictly speaking, the input feature vector is $\mathbf{X}_{t-W}^{t+W} = \{\mathbf{x}(t-W), \ldots, \mathbf{x}(t), \ldots, \mathbf{x}(t+W)\}$ since, in practice, we take a window of $2W+1$ concatenated feature vectors centred on time $t$. We use the notation in equation 4 throughout this article for simplicity.

with respect to the connection weights [7]. By using the chain rule, this may be obtained from the product of the partial derivative of the criterion function with respect to the unit activations with the partial derivative of the unit activations with respect to the weights. In this section we briefly describe the criterion used in the baseline system and then describe our proposed criterion and it's application to error back-propagation training of an MLP.

### 3.1 Baseline criterion

In the baseline SAD system used in our previous experiments, the MLP was trained to perform speech/silence classification of the inputs. The criterion used in such a training scenario is commonly the Kullback-Leibler divergence (or relative-entropy) between the target posterior distribution, $r_i(t) = P(\mathcal{C}_i|\mathbf{x}(t))^2$, and that estimated by the MLP, $\phi_i(t) = \hat{P}(\mathcal{C}_i|\mathbf{x}(t))$.

$$Q^b = \sum_{t=1}^{T}\sum_{i=1}^{M} r_i(t) \log \frac{r_i(t)}{\phi_i(t)} \tag{5}$$

$$\frac{\partial Q^b}{\partial \phi_i(t)} = -\frac{r_i(t)}{\phi_i(t)} \tag{6}$$

where $M$ is the number of classes and $\mathcal{C}_i, i \in [1, M]$ are the classes associated with the MLP outputs.

### 3.2 Proposed criterion

We have observed in our previous work that the WER obtained from an automatically derived segmentation of meeting room data was very strongly correlated with the frame level error rate of the MLP based speech activity detection classifier [6]. We ascribed this behaviour as being partially due to the error in estimation of CMVN statistics (with respect to a 'ground truth' estimate made using a forced alignment of the reference transcription of test data). To test this hypothesis we first propose a measure of the accuracy of estimation of the CMVN statistics based on the Kullback-Leibler divergence between two normal distributions with diagonal covariance as shown in equation 7, where we have the reference $\mathcal{N}(\cdot; \mu_r, \sigma_r)$ and estimated CMVN statistics $\mathcal{N}(\cdot; \mu_\phi, \sigma_\phi)$, respectively:

$$Q^c = -\frac{N}{2} + \frac{1}{2}\sum_{k=1}^{N}\Bigg[\log \sigma_\phi^k - \log \sigma_r^k + \frac{\sigma_r^k}{\sigma_\phi^k} + \frac{(\mu_\phi^k - \mu_r^k)^2}{\sigma_\phi^k}\Bigg] \tag{7}$$

where $N$ is the dimension of the multivariate normal distributions.

Using results from our initial work on the NIST rich transcription 2006 evaluation (further details of the AMI RT06s system can be found in [8]), we first used the criterion in equation 7 to measure the distance between the CMVN statistics estimated from the forced alignment reference, $\text{REF}_{fa}$, and those estimated using automatic systems at various system operating points. The results are shown in Figure 1, with a clear trend between the proposed distance metric for the CMVN estimate and the ASR performance being evident. We now describe our algorithm for the direct minimisation of the proposed CMVN criterion in equation 7 by error back-propagation training of an MLP.

Using equations 1 and 2 as a basis and ignoring, for the moment, the minimum duration imposed by the HMM topology (as described in section 2.2), we can make a maximum likelihood estimate of the CMVN statistics using the output of the MLP, where there are two MLP outputs $\phi(t) = P(\mathcal{C}_{sp}|\mathbf{x}(t))$

[2] A *one hot target* is used in which the true class is given a probability of one and the remaining classes are set to zero.

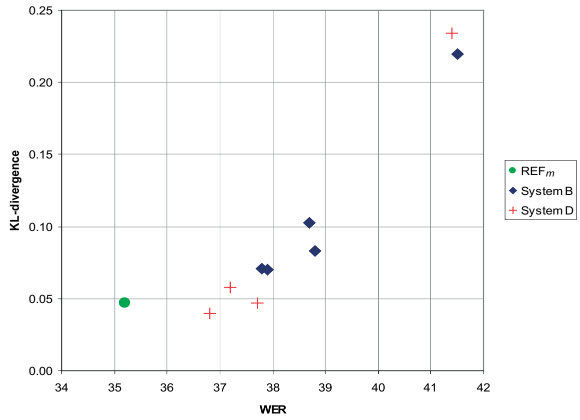


Figure 1: Frame error rate (FER) versus average KL-divergence between CMVN statistics, where test segments are compared against the segments generated from forced alignment of the recording transcripts, $REF_{fa}$. Results are shown for two systems detailed in [6] as well as the manually derived segmentation, $REF_m$

and $\bar{\phi}(t) = P(\mathcal{C}_{sil}|\mathbf{x}(t)) = 1 - \phi(t)$:

$$\mu_\phi^k = \frac{\sum_{s=1}^T \phi(s) x^k(s)}{\sum_{s=1}^T \phi(s)} \tag{8}$$

$$\sigma_\phi^k = \frac{\sum_{s=1}^T \phi(s)(x^k(s))^2}{\sum_{s=1}^T \phi(s)} - \mu_\phi^2 \tag{9}$$

Hence, we can relate the MLP outputs to the new training criterion by the substitution of equations 8 and 9 into equation 7. To minimise $Q^c$ by gradient descent, we take the partial derivative of equation 7 with respect to the MLP output $\phi(t)$ to give:

$$\begin{aligned}\frac{\partial Q^c}{\partial \phi(t)} =& \frac{1}{2}\sum_{k=1}^N \Bigg[ \frac{\partial \sigma_\phi^k}{\partial \phi(t)} \Big[ (\sigma_\phi^k)^{-1} \\ & - (\sigma_\phi^k)^{-2} \big( \sigma_r^k + (\mu_\phi^k - \mu_r^k)^2 \big) \Big] \\ & + 2 \frac{\partial \mu_\phi^k}{\partial \phi(t)} (\mu_\phi^k - \mu_r^k)(\sigma_\phi^k)^{-1} \Bigg]\end{aligned} \tag{10}$$

Then it follows that we need to find $\frac{\partial \mu_\phi^k}{\partial \phi(t)}$ and $\frac{\partial \sigma_\phi^k}{\partial \phi(t)}$. From equation 8 we have:

$$\frac{\partial \mu_\phi^k}{\partial \phi(t)} = \frac{x^k(t) - \mu_\phi^k}{\sum_{s=1}^T \phi(s)} \tag{11}$$

Similarly from equation 9 we have:

$$\frac{\partial \sigma_\phi^k}{\partial \phi(t)} = \frac{\sum_{s=1}^T \phi(s)((x^k(t))^2 - (x^k(s))^2)}{\left(\sum_{s=1}^T \phi(s)\right)^2} - 2\frac{\partial \mu_\phi^k}{\partial \phi(t)} \mu_\phi \tag{12}$$

Using the chain-rule, it is trivial to find:

$$\frac{\partial Q^c}{\partial \bar{\phi}(t)} = -\frac{\partial Q^c}{\partial \phi(t)} \tag{13}$$

Substituting equations 11 and 12 into equations 10 and 13 gives the gradient of the new error criterion $Q^c$ with respect to the MLP outputs $\phi(t)$ and $\bar{\phi}(t)$, enabling us to perform gradient descent training of the MLP parameters directly from the target CMVN statistics.

### 3.3 Practical considerations

In theory, we now have all that we need to update the MLP parameters according to the new training criterion, but there still remain a number of practical issues. First of all, it is immediately apparent from equations 11 and 12 that the criterion depends upon the entire input sequence. Thus, in order to perform back-propagation for a single observation $\mathbf{x}(t)$ we must compute MLP outputs $\phi(t)$ and $\bar{\phi}(t)$ for all $t = [1, T^i]$ for the given sequence $i$ (in our case a single channel from an IHM meeting recording), making a stochastic training regime prohibitively expensive. Unfortunately, a batch training scheme is also unattractive for large training sets, since the parameter update step size will need to be prohibitively small [9]. In this initial investigation we train the baseline MLP using a conventional stochastic approach, while we are obliged to train for our proposed criterion using a mini-batch approach (weights are updated after each individual meeting recording). We hope to address this inherent disadvantage in future work.

## 4 Experiments

### 4.1 System setup and training

The MLPs are each trained from eight minutes of acoustic data from a set of 150 meetings as described in [6]. Each MLP has an input context of nine frames with twenty five hidden units. We use fewer parameters than in our previously reported work in order to reduce training time. This resulted in only a small decrease in classification accuracy of the MLP. The training targets for the baseline MLP are derived from a forced alignment of the reference transcripts, while the training targets for the proposed criterion are the CMVN statistics derived from the aforementioned forced alignments. The ASR system with which we conduct our experiments is described in [8] where the acoustic models have been trained using the same CMVN feature normalisation as used in the proposed MLP training. Our experiments are conducted on the NIST rich transcription 2006 Spring evaluation meeting room data[3].

### 4.2 Results and discussion

During training we keep track of the frame accuracy and KL-divergence between CMVN statistics ($Q_c$) estimated by the MLPs and the respective references. This enables us to compare the two criteria in terms of their convergence properties as shown in Figure 2. We note a number of properties of this graph. First of all the convergence of the proposed system is much slower. As already mentioned, this is due to the need to select a more conservative step size in order to allow for the mini-batch training to converge to a reasonable solution. WE have found that larger step sizes resulted in convergence to poor local minima. Secondly, we note that the systems converge to quite different solutions, with the baseline system converging to a lower FER rate than the proposed approach and conversely the proposed approach converging to a lower $Q^c$. This is as we'd hoped since, it was our intention that the proposed approach would lead to a better, more robust estimator of CMVN statistics by no longer being constrained to match the hard definition of speech/silence classes as defined by the reference

[3] see http://www.nist.gov/speech/tests/rt/rt2006/spring/

segmentation. We also take note that the best $Q^c$ achieved by the baseline system does not coincide with the best FER, in fact, it occurs early in the training after only seven iterations. As a final remark, we also point out that the final $Q^c$ achieved in Figure 2 is still much greater than those presented in Figure 1 due to there currently being no minimum duration constrain in our formulation of the maximum-likelihood estimate of the CMVN statistics.

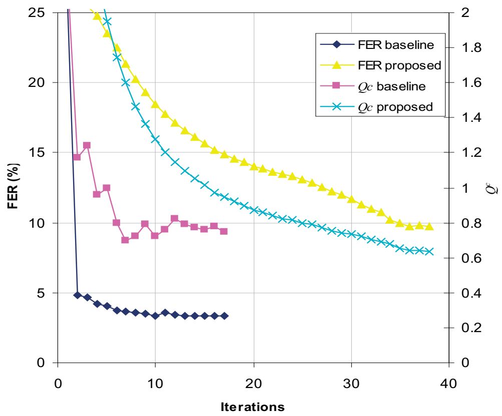


Figure 2: Frame error rate (FER) and CMVN distance ($Q^c$) between automatic systems and reference ($REF_{fa}$)). The first Y-axis shows FER and the second Y-axis shows $Q^c$

The two MLPs were also evaluated using the first pass of the AMI RT06seval ASR system using CMVN statistics calculated from the baseline and proposed MLPs according to equations 8 and 9. A manual segmentation of the test data was used in order to isolate the effect of the CMVN estimation from the speech segmentation problem. The results of these experiments are shown in Table 1 along with those for the 'reference' CMVN statistics. We see from this table that a small, but appreciable improvement is obtained using the proposed approach.

| System | FER | $Q^c$ | Sub | Del | Ins | WER |
|---|---|---|---|---|---|---|
| REF$_{fa}$ | 0 | 0 | 26.1 | 9.0 | 4.3 | 39.4 |
| baseline | 3.3 | 0.75 | 26.6 | 8.6 | 5.0 | 40.2 |
| proposed | 9.8 | 0.63 | 26.0 | 9.5 | 4.3 | 39.9 |

Table 1: ASR and frame-level performance using CMVN statistics calculated using MLPs trained using baseline and proposed criteria and the reference (REF$_{fa}$) CMVN statistics.

# 5 Conclusions and future work

In this paper we have described a novel optimisation criterion for training an MLP based on the Kullback-Leibler divergence between a set of target and estimated cepstral feature mean and variance statistics. We provided details of gradient descent training of the MLP according to the new criterion via back-propagation. Although this work is quite preliminary, we have demonstrated that the new training scheme does indeed converge to a better solution than our baseline system in terms of the proposed criterion and we have demonstrated the implications of this in large vocabulary speech recognition experiments on meeting room data.

In future work we wish to address some short comings of the current approach. Firstly, the mini-batch training scheme should be replaced by a computationally efficient stochastic scheme as is used in the baseline system, which should lead to faster convergence to a better solution. We would also like to incorporate duration constraints into the optimisation process. There are several ways of addressing this that are currently under investigation. We also plan to apply this approach to meeting room data with significantly different recording conditions to the training data, where we hope our proposed approach will exhibit better generalisation than the baseline.

# 6 Acknowledgements

The authors would like to thank Thomas Hain and Sheffield University for access to computing resources for running the ASR experiments. This work was supported by the European Union 6th FWP IST Integrated Project on Augmented Multi-party Interaction (AMI, FP6-506811) and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2).

# References

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.

[2] O. Vikki and K. Laurila, "Ceptral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[3] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proc. ICSLP*, Jeju Island, Korea, 2004.

[4] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, January 2005.

[5] K. Boakye and A. Stolcke, "Improved speech activity detection using cross-channel features for recognition of multiparty meetings," in *Proc. Interspeech (ICSLP)*, 2006.

[6] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Proc. Interspeech (ICSLP)*, 2006.

[7] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1996.

[8] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system: Progress and performance," in *Proc. NIST RT06 Spring workshop*, 2006.

[9] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, December 2003.