Evaluation of Model-Based Feature Enhancement on the AURORA-4 Task

Veronique Stouten[‡]*, Hugo Van hamme, Jacques Duchateau, Patrick Wambacq*

Center for Processing Speech and Images (PSI)
Dept. of Electrical Engineering - ESAT
Katholieke Universiteit Leuven, Belgium
{Veronique.Stouten,Hugo.Vanhamme}@esat.kuleuven.ac.be

Abstract

In this paper we focus on the challenging task of noise robustness for *large vocabulary* Continuous Speech Recognition (LVCSR) systems in *non-stationary* noise environments. We have extended our Model-Based Feature Enhancement (MBFE) algorithm – that we earlier successfully applied to small vocabulary CSR in the AURORA-2 framework – to cope with the new demands that are imposed by the large vocabulary size in the AURORA-4 task. To incorporate a priori knowledge of the background noise, we combine scalable Hidden Markov Models (HMMs) of the cepstral feature vectors of both clean speech and noise, using a Vector Taylor Series approximation in the power spectral domain. Then, a global MMSE-estimate of the clean speech is calculated based on this combined HMM. This technique is easily embeddable in the feature extraction module of a recogniser and is intrinsically suited for the removal of non-stationary additive noise. Our approach is validated on the AURORA-4 task, revealing a significant gain in noise robustness over the baseline.

1. Introduction

It is well-known that Automatic Speech Recognition (ASR) systems are significantly more sensitive to interfering noise than humans are. As a result, ASR-applications often fail to meet user expectations in noisy environments. Since developers are forced to restrict the complexity of ASR-systems, this sometimes leads to applications that are not at all user-friendly, which is illustrated by the use of ASR in a car environment. In this case the vocabulary size of the system has to be kept unreasonably small due to the technical limitations of ASR-technology. It is hence clear that an improved robustness in realistic conditions will open new perspectives for the employment of this technology in a broad range of everyday situations.

The sensitivity of ASR-systems to noise became a major research issue when this technology moved from the laboratory environment to a realistic - and hence noisy - environment. Over the last decades several approaches to mitigate the adverse effects of a mismatch between training and testing conditions were proposed, one of which modifies the noisy features to better resemble those observed under clean conditions. Compensating for the effects of noise in the front-end has as a major advantage that the computational load is independent of the complexity of the back-end. Taking the capabilities of current CPUs into consideration, this is particularly important for large vocabulary speech recognition. Moreover, any refinement of the acoustic model such as adaptation or discriminative training is still valid. On the other hand, the feature enhancement could be suboptimal compared to approaches such as PMC [1] or MFT [2] since the acoustic model of the back-end remains unchanged and hence the modified statistics of the enhanced speech can not be expressed. Also, for small vocabulary tasks, the computational load of adapting the acoustic model to the non-stationarities of the noise may be feasible after all.

We have implemented a Model-Based Feature Enhancement technique [3], which uses one Hidden Markov Model (HMM) with Gaussian observation probabilities for the clean speech cepstral feature vectors, and another Gaussian HMM for the perturbing noise cepstral feature sequence. Based on these statistical models, the parameters of a combined HMM of the noisy speech are estimated by a first order Vector Taylor Series approximation. Subsequently this product HMM is used to calculate the *a posteriori* probabilities of each combined (speech, noise) state corresponding to a sequence of observation vectors. For each combined state pair also an estimate of the corresponding clean speech can be calculated. Finally, the global MMSE-estimate of the clean speech, given the noisy speech, is obtained as a linear combination of these state-conditional estimates weighted by the *a posteriori* probabilities. Because the noise sequence is modeled by an HMM, this technique is intrinsically suited for the removal of time-varying noise, where the application of classic approaches such as spectral subtraction [4, 5] is less effective. In [6] we have evaluated this feature enhancement technique on the AURORA-2 connected digit recognition task and compared several front-end complexities, the best of which is used in this paper to validate its application on the large vocabulary AURORA-4 task.

In section 2 the mathematical details of the MBFE-algorithm are described. Section 3 presents a description of our speech recognition engine, and in section 4 the experimental evaluation of the proposed preprocessing technique on the AURORA-4 continuous speech dictation task can be found. Finally, conclusions and directions for future work are discussed in section 5.

2. Model-Based Feature Enhancement

Since the noisy speech feature vectors are enhanced by the MBFE-algorithm in the cepstral domain, the parametric model for the acoustic environment, used throughout this work, is given by [7]:

$$x_t = f(s_t, n_t) \quad (1)$$

$$= 0.5\, C\, \log\left(\exp\left(2\, C^{+} s_t\right) + \exp\left(2\, C^{+} n_t\right)\right) \quad (2)$$

where x_t and s_t are the distorted and clean speech cepstral vectors of frame t respectively. The noise cepstral vector n_t is assumed to be uncorrelated and additive to the clean speech in the Mel power spectral domain. C is the DCT-matrix which can be non-square and C^{+} denotes its Moore-Penrose generalised inverse.

[‡] Research Assistant of the Fund for Scientific Research - Flanders (Belgium) (F.W.O. - Vlaanderen).

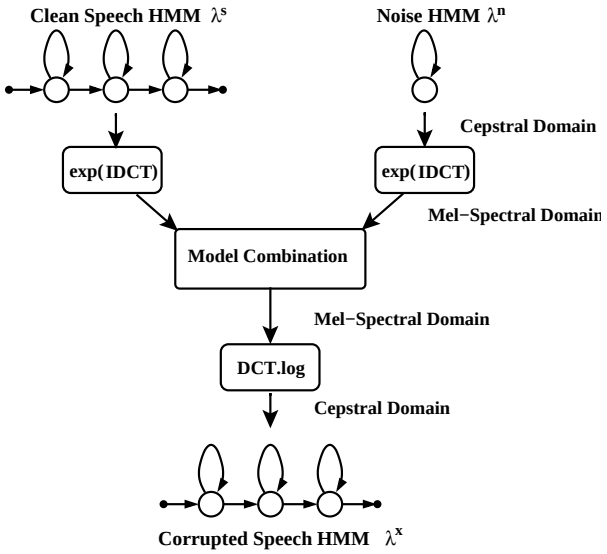

Figure 1: *HMM combination principle.*

Let us further assume that both s_t and n_t can be modeled by Gaussian mixture HMMs, as illustrated in figure 1. Since a mixture of Gaussians can be decomposed into an ergodic single Gaussian HMM, we will use single Gaussian state-conditional probability density functions (pdfs) for simplicity of notation. Let $q_t^s \in \{1, \ldots, M^s\}$ be the speech state at time t, $q_t^n \in \{1, \ldots, M^n\}$ the noise state at time t, λ^s the clean speech HMM and λ^n the noise HMM, then the state-conditional pdfs of clean speech and noise are:

$$p\left[s_t|q_t^s = i\right] = N(s_t; \mu_i^s, \Sigma_i^s) \quad (3)$$

$$p\left[n_t|q_t^n = j\right] = N(n_t; \mu_j^n, \Sigma_j^n) \quad (4)$$

in which Σ_i^s and Σ_j^n are diagonal. Similar to the Parallel Model Combination (PMC) concept [8], we also assume that the corrupted speech x_t can be modeled by a Gaussian mixture HMM λ^x of which the mean $\mu_{(i,j)}^x = E[x|q_t^s = i, q_t^n = j]$ and covariance matrix $\Sigma_{(i,j)}^x = E[(x - \mu_{(i,j)}^x)(x - \mu_{(i,j)}^x)']$ for each combined state (i, j) can be obtained from the combination function (2). For this product HMM the number of states M^x equals $M^s.M^n$, but since MBFE is a preprocessing step the models can be less complex than the models used in the recogniser, which keeps the computational load feasible. In this case the state-conditional pdf becomes:

$$p\left[x_t|q_t^s = i, q_t^n = j\right] = N(x_t; \mu_{(i,j)}^x, \Sigma_{(i,j)}^x) \quad (5)$$

However, since (2) is a non-linear relationship we resort to a first order Vector Taylor Series (VTS) approximation around the means μ_i^s and μ_j^n to linearise it [9]. This yields:

$$x_t = f\left(\mu_i^s, \mu_j^n\right) + F_{(i,j)}\left(s_t - \mu_i^s\right) + G_{(i,j)}\left(n_t - \mu_j^n\right) \quad (6)$$

and hence:

$$\mu_{(i,j)}^x \approx 0.5\,C\,\log\left(\exp\left(2\,C^+\,\mu_i^s\right) + \exp\left(2\,C^+\,\mu_j^n\right)\right) \quad (7)$$

$$\Sigma_{(i,j)}^x \approx F_{(i,j)}\,\Sigma_i^s\,F_{(i,j)}' + G_{(i,j)}\,\Sigma_j^n\,G_{(i,j)}' \quad (8)$$

in which $'$ indicates transpose, and the gradients of the combination function $f(s_t, n_t)$ have the closed form:

$$F_{(i,j)} = C\,\mathrm{diag}\left(\frac{1}{1 + exp\left[2\,C^+\,(\mu_j^n - \mu_i^s)\right]}\right)C^+ \quad (9)$$

$$G_{(i,j)} = I - F_{(i,j)} \quad (10)$$

and I denotes the unity matrix. After model combination, the covariance matrix $\Sigma_{(i,j)}^x$ will no longer be diagonal due to the non-linearity of the log-compression.

For each noisy speech state (i, j) of this combined HMM λ^x a state-conditional MMSE-estimate $\hat{s}_t^{(i,j)}$ of the clean speech, given the corrupted speech at time t, is derived as follows:

$$\hat{s}_t^{(i,j)} = E\left[s_t|x_t, q_t^s = i, q_t^n = j\right] \quad (11)$$

$$= \mu_i^s + \Sigma_i^s\,F_{(i,j)}'\left(\Sigma_{(i,j)}^x\right)^{-1}\left(x_t - \mu_{(i,j)}^x\right) \quad (12)$$

in which the covariance matrix between s_t and x_t is approximated by $\Sigma_i^s\,F_{(i,j)}'$, using the same VTS expansion as in (6).

Additionally, this Cartesian product HMM allows to calculate the *a posteriori* probabilities $\gamma_t^{(i,j)}$ of each combined (speech, noise) state, given the sequence of noisy observation vectors $x_1^T = (x_1, x_2, \ldots, x_T)$:

$$\gamma_t^{(i,j)} = P\left[q_t^s = i, q_t^n = j|x_1^T\right] \quad (13)$$

$$= \frac{\alpha_t(i,j)\,\beta_t(i,j)}{\sum_i \sum_j \alpha_t(i,j)\,\beta_t(i,j)} \quad (14)$$

where $\alpha_t(i, j)$ and $\beta_t(i, j)$ are the forward and backward probability of state (i, j) at time t, respectively. Notice that, whereas the state-conditional estimates $\hat{s}_t^{(i,j)}$ depend only on the noisy speech vector at time t, the *a posteriori* probabilities $\gamma_t^{(i,j)}$ incorporate the temporal structure of the sequence x_1^T.

Finally, a global MMSE-estimate of the clean speech is obtained as a linear combination of the state-conditional estimates [10]:

$$\hat{s_t} = E\left[s_t|x_1^T\right] = \sum_i \sum_j \gamma_t^{(i,j)}\,\hat{s}_t^{(i,j)} \quad (15)$$

in which the *a posteriori* probabilities $\gamma_t^{(i,j)}$ act as a 'soft' selector. These enhanced feature vectors can be fed into the detailed acoustic models of the recogniser.

3. Recognition system

Experiments were conducted with the speaker-independent LVCSR-system that has been developed by the ESAT-PSI speech group of the K.U.Leuven. As opposed to the AURORA-4 reference back-end, this system has a faster experiment turn-around time and a better baseline accuracy.

3.1. Signal processing

To extract the acoustic features from the speech signal, first a power spectrum is calculated every 10 ms on a 32 ms window of the pre-emphasized 16 kHz data (the 8 kHz data was not used in this paper). Then a Hamming window and a mel-scaled triangular filterbank are applied [11]. Finally, the resulting mel spectrum with 24 coefficients is transformed into the log-domain and the proposed noise reduction algorithm, described in section 2, is applied.

3.1.1. Cepstrum-based reference features

To increase the robustness to low frequency noises, the first spectral coefficient is set to 0 in the cepstrum-based reference feature extraction. Since for the AURORA-4 benchmark at 16 kHz all frequencies above 7 kHz were filtered, also the last band is nulled. Then the log-spectrum is transformed into 13 cepstral coefficients ($c0,\ldots,c12$). Finally, after mean normalisation, the first and second order time derivatives are added, which results in a total of 39 features.

3.1.2. MIDA-based features

To improve this cepstrum-based reference, a discriminant feature analysis was included in the feature extraction. After again removing the first and last mel spectra, and mean normalising the remaining 22 spectral coefficients, the first and second order time derivatives are added to result in 66 features. The MIDA-algorithm (an improved LDA-algorithm [12]) then reduces this feature vector to 39 dimensions. Finally, these features are decorrelated (see [13]) because the acoustic modeling uses diagonal covariance Gaussian distributions. Both the feature selection and the decorrelation are made once on the clean data without MBFE-preprocessing.

3.2. Modeling and search

The acoustic modeling is gender independent and based on a phone set with 45 phones, without specific function word modeling. A phonetic decision tree (see [14]) defines the 4961 tied states in the cross-word context-dependent (but position-independent) models.

Based on the available 15 hours of clean speech (corresponding to 7138 sentences), models for the different investigated signal processing algorithms are estimated in 3 training steps. First, acoustic models without tying of the Gaussians are initialized, resulting in a total of 21k Gaussians. Then full tying over all states is allowed, the 2k most promising Gaussians per state are selected based on the distances between Gaussians (to avoid prohibitively large models) and the models are re-estimated in the second training step. Finally, the number of Gaussians is reduced further to an average of 200 per state, using the *occupancy* criterion [15], and the third training step is applied. Note that a further reduction to an average of less than 100 Gaussians per state is possible without loss in recognition performance, but this requires additional training. On the whole, these 3 training steps take about 15 hours on a 2.0 GHz Pentium 4 processor.

The bigram language model for a 5k-word closed vocabulary, as provided by Lincoln Laboratory, was used as a language model.

To decode, a single pass time synchronous beam search algorithm is used [16]. For the clean data, the recogniser runs in real time on a 2.0 GHz Pentium 4 processor, whereas in the worst case (mismatch due to noise and microphone) it is only five times slower. Adjusting the necessary parameters would allow a twofold speed up of the recognition, without loss in recognition performance. However, this parameter tuning would require a development test set.

4. Experiments

4.1. The AURORA 4 database

As the AURORA front-end standardization for distributed speech recognition is moving to large vocabulary speech recognition, the ETSI STQ group has defined the AURORA-4 database derived from the WSJ0 Wall Street Journal 5k-word dictation task.

Test sets are constructed by adding recorded noise to the clean samples at 8 kHz and 16 kHz at an SNR-level that ranges from 5 dB to 15 dB. Seven different types of noise are added to the close talking microphone signal, which each yield a test set of 166 utterances: no noise (set 01), car (set 02), babble (set 03), restaurant (set 04), street (set 05), airport (set 06) and train (set 07). Test sets 08 through 14 are obtained by adding these same noise types to recordings made with 18 different microphones. Finally, the training set contains 7318 sentences recorded with only the close talking microphone and without noise addition. Apart from that, the AURORA-4 task also defines a multi-condition training set, but we will limit our evaluations to the clean-condition training and wideband data.

4.2. MBFE-models

The MBFE-algorithm requires the design of a Hidden Markov Model for the clean speech, as well as for the noise. In parallel experiments on the AURORA-2 database (reported upon in [6]), we found that the recognition accuracy when using a phoneme loop for the speech model and a single Gaussian noise model, approaches the results obtained with a word loop and an 8-Gaussian ergodic noise HMM. Although these results were established on a small-vocabulary digit recognition task and for the noise types of AURORA-2, the sheer complexity associated with these more sophisticated models does not allow us to verify the validity of these conclusions on the present large vocabulary task. Hence, the reported results are obtained with a phoneme loop and a single Gaussian noise model. However, it is quite likely that the MBFE front-end can be improved significantly when changing the model configuration, such that the noise non-stationarities for instance are modeled better.

Whereas using a non-square DCT-matrix is computationally more tractable, superior results are obtained with a square DCT-matrix. Hence the feature vector of the MBFE-models consists of 24 static MFCCs, together with the log-energy.

Based on the clean training data, we estimate the parameters of 45 phoneme models using a fixed segmentation into context-independent HMM states. Each phoneme has 3 emitting states, with a mixture of 4 Gaussians per state. To this end, we start off with a single Gaussian model and split the Gaussian with the largest mixture weight along the main axis of the covariance matrix. We then perform 10 EM-iterations for the weights, the means and the variances and repeat this process until 4 Gaussians are trained.

To estimate the mean of the noise model, we always use the first 30 and the last 30 frames of every sentence. From the same frames, also the variance is estimated for each utterance during testing, but the latter is pooled over all 166 utterances of the noise condition. While this approach simulates a scenario in which some of the noise model parameters can be estimated off-line, we did not investigate the usefulness of this prior knowledge, i.e. whether the performance would suffer from noise variance estimation on the 60 presumed noise frames.

4.3. Recognition results

First the clean training data (not processed by MBFE) was used to build the baseline models. The MBFE-algorithm was then applied to the mel log-spectra of the clean and the noisy test sets to yield an enhanced feature stream, on which the delta and delta-delta spectra were computed. No feature compression or end pointing was performed. Table 1 lists the word error rate (substitutions, deletions and insertions) of each of the test sets using the clean speech models.

The first line presents the baseline accuracy that is obtained

Performance summary: 16 kHz sampling, no compression, no end pointing															
	Close Talk							Far Talk							
Description	01	02	03	04	05	06	07	08	09	10	11	12	13	14	Avg.
cepstra	7.44	25.05	52.15	55.73	49.36	49.06	55.84	26.11	48.25	68.07	70.72	72.63	68.03	73.19	51.55
spec+MIDA	5.38	20.74	39.23	46.15	40.85	35.73	44.05	22.21	42.95	57.61	61.62	64.05	54.33	60.52	42.53
mbfe+MIDA	5.86	11.42	31.12	41.29	31.75	33.70	31.42	20.74	32.78	48.58	56.46	53.37	51.90	49.43	35.70

Table 1: *Word error rates with cepstral features, MIDA-transformed spectra without and with MBFE-enhancement, all using acoustic models trained on the clean speech.*

by using cepstral features. As the second line indicates, a significant improvement in clean accuracy and noise robustness is obtained by using the MIDA feature transformation. Finally, the application of the MBFE method (third line) also improves the robustness. However, in the latter case the clean model accuracy suffers slightly, which can be explained by the mismatch between training (which occurred without MBFE-processing) and testing (with MBFE-processing). Also the projection onto a 13-dimensional cepstral space (applied in the MBFE front-end) could cause this slight deterioration. Since MBFE does not correct for reverberation effects or the differences in microphone, test sets 08 through 14 still show a high error rate. However, significant accuracy gains are obtained here as well, despite the fact that the MBFE speech models were trained on the close talking microphone and hence show a mismatch for these test sets.

5. Conclusions

We have shown that Model-Based Feature Enhancement can be applied to increase the robustness of large vocabulary speech recognisers to additive non-stationary noise. In this paper, a baseline MBFE-configuration was presented that proves to be effective on the AURORA-4 task. However, as mentioned in section 4, the model configuration of the MBFE front-end can still be improved. In future work, we will therefor simplify the speech model further, while the noise model can be made more complex to better model the non-stationarity of the noise. Pruning techniques in the calculation of $\alpha_t(i,\,j)$, $\beta_t(i,\,j)$, $\gamma_t^{(i,j)}$ and $\hat{s_t}$ will also allow a reduction of the computational load of the MBFE-algorithm. Finally, effort will be spent on the development of a training method for the multi-condition training set with MBFE.

6. References

[1] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using Parallel Model Combination," *IEEE Trans. on SAP*, vol. 4, no. 5, pp. 352–359, 1996.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[3] C. Couvreur and H. Van hamme, "Model-based feature enhancement for noisy speech recognition," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, vol. III, pp. 1719–1722.

[4] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Comm.*, vol. 11, no. 2-3, pp. 215–228, 1992.

[5] F. Xie and D. Van Compernolle, "Speech enhancement by nonlinear spectral estimation - a unifying approach," in *Proceedings EUROSPEECH*, Berlin, Germany, Sept. 1993, pp. 617–620.

[6] V. Stouten, H. Van hamme, K. Demuynck, and P. Wambacq, "Robust speech recognition using model-based feature enhancement," in *submitted to Eurospeech 2003*.

[7] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

[8] M.F.J. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, Sept. 1995.

[9] P.J. Moreno, B. Raj, and R.M.J. Stern, "A Vector Taylor Series approach for environment-independent speech recognition," in *Proc. ICASSP*, Atlanta, U.S.A., May 1996, vol. II, pp. 733–736.

[10] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "HMM-based strategies for enhancement of speech signals embedded in non-stationary noise," *IEEE Trans. on SAP*, vol. 6, no. 5, pp. 445–455, 1998.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognitions in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[12] J. Duchateau, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Class definition in discriminant feature analysis," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sept. 2001, vol. III, pp. 1621–1624.

[13] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "Improved feature decorrelation for HMM-based speech recognition," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, vol. VII, pp. 2907–2910.

[14] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Comm.*, vol. 24, no. 1, pp. 5–17, Apr. 1998.

[15] J. Duchateau, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Improved parameter tying for efficient acoustic model evaluation in large vocabulary continuous speech recognition," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, vol. V, pp. 2215–2218.

[16] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Comm.*, vol. 30, no. 1, pp. 37–53, Jan. 2000.