Robust Speech Recognition using Model-Based Feature Enhancement

Veronique Stouten[‡], *Hugo Van hamme, Kris Demuynck, Patrick Wambacq*

Center for Processing Speech and Images (PSI)
Dept. of Electrical Engineering - ESAT
Katholieke Universiteit Leuven, Belgium
{vstouten,hvanhamm,krisdm,wambacq}@esat.kuleuven.ac.be

Abstract

Maintaining a high level of robustness for Automatic Speech Recognition (ASR) systems is especially challenging when the background noise has a time-varying nature. We have implemented a Model-Based Feature Enhancement (MBFE) technique that not only can easily be embedded in the feature extraction module of a recogniser, but also is intrinsically suited for the removal of non-stationary additive noise. To this end we combine statistical models of the cepstral feature vectors of both clean speech and noise, using a Vector Taylor Series approximation in the power spectral domain. Based on this combined HMM, a global MMSE-estimate of the clean speech is then calculated. Because of the scalability of the applied models, MBFE is flexible and computationally feasible. Recognition experiments with this feature enhancement technique on the Aurora2 connected digit recognition task showed significant improvements on the noise robustness of the HTK recogniser.

1. Introduction

The application of Automatic Speech Recognition systems in poorly conditioned environments such as in a car, over the telephone or in industrial surroundings, has promoted noise robustness of these systems to a major issue in current research. In realistic conditions the speech input can no longer be assumed to come from a known microphone through a channel with a high signal to noise ratio. Because it is well known that recognition rates of ASR systems drop considerably when there is a mismatch between the training and the testing conditions, modifications to the system are necessary to compensate for the effects of interfering signals.

To mitigate the effect of environmental noise several approaches have been proposed, which can be broadly divided into 3 categories [1]. Firstly, an increase of the noise robustness can be achieved by extracting speech features that are inherently less distorted by noise. In spite of the limitations of Mel-Frequency Cepstral Coefficients (MFCC), they are often used because of their low correlation and their ability to arrive at a compact and computationally efficient representation of the speech signal. A second approach to reduce the mismatch is to adapt the acoustic models in the recogniser to the changing noise conditions [2, 3]. Although these techniques are very flexible, the major disadvantage is that they lack scalability and also the computational load with large and/or dynamic vocabulary speech recognition systems can become prohibitively large. These problems are avoided by the third approach, in which the features are enhanced before they are fed into the recogniser. This can be achieved either prior to the feature extraction (like speech enhancement techniques, speech dereverberation, ...), or by incorporating extra 'cleaning' steps into the feature extraction module [4, 5]. Such a feature enhancement step is largely independent of the vocabulary size of the recogniser and also does not require an adaptation of the recognition software. In our research we focus on this last approach to increase the noise robustness of ASR-systems in non-stationary noise conditions.

We have implemented a Model-Based Feature Enhancement technique [6], 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 Minimum Mean Square Error (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 [7] is less effective.

A detailed description of the MBFE-algorithm is presented in section 2. We have evaluated the performance of the proposed preprocessing technique on the Aurora2 connected digit recognition task and investigated the effect of changing the front-end complexity. These implementation issues, together with the obtained recognition accuracy, can be found in section 3. Finally, conclusions and directions for future work are discussed in section 4.

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 [8] :

$$\begin{aligned} x_t &= f(s_t, n_t) & (1)\\ &= 0.5\, C \, \log\left(\exp\left(2\, C^+ s_t\right) + \exp\left(2\, C^+ n_t\right)\right) & (2) \end{aligned}$$

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^+ denotes the Moore-Penrose generalised inverse of the non-square DCT-matrix C.

Let us further assume that both s_t and n_t can be modeled by Gaussian mixture HMMs, as illustrated in figure 1.

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

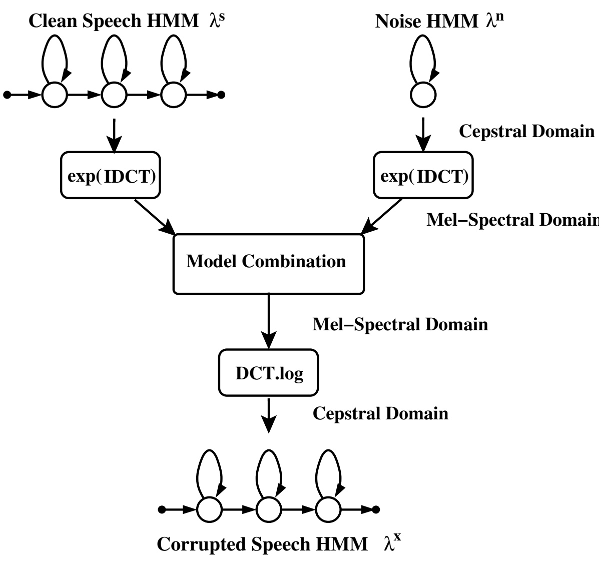

Figure 1: *HMM combination principle.*

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 $q^s_t \in \{1, \ldots, M^s\}$ be the speech state at time t, $q^n_t \in \{1, \ldots, M^n\}$ 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 :

$$p\left[s_t|q^s_t = i\right] = N(s_t; \mu^s_i, \Sigma^s_i) \qquad (3)$$

$$p\left[n_t|q^n_t = j\right] = N(n_t; \mu^n_j, \Sigma^n_j) \qquad (4)$$

in which Σ^s_i and Σ^n_j are diagonal. Similar to the Parallel Model Combination (PMC) concept [9], we also assume that the corrupted speech x_t can be modeled by a Gaussian mixture HMM λ^x of which the mean $\mu^x_{(i,j)} = E[x|q^s_t = i, q^n_t = j]$ and covariance matrix $\Sigma^x_{(i,j)} = E[(x - \mu^x_{(i,j)})\,(x - \mu^x_{(i,j)})^{'}]$ 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 :

$$p\left[x_t|q^s_t = i, q^n_t = j\right] = N(x_t; \mu^x_{(i,j)}, \Sigma^x_{(i,j)}) \qquad (5)$$

However, since (2) is a non-linear relationship we resort to a first order Vector Taylor Series (VTS) approximation around the means μ^s_i and μ^n_j to linearise it [10]. This yields :

$$x_t = f\left(\mu^s_i, \mu^n_j\right) + F_{(i,j)}\,\left(s_t - \mu^s_i\right) + G_{(i,j)}\,\left(n_t - \mu^n_j\right) \qquad (6)$$

and hence :

$$\mu^x_{(i,j)} \approx 0.5\,C\,\log\left(\exp\left(2\,C^+\,\mu^s_i\right) + \exp\left(2\,C^+\,\mu^n_j\right)\right) \qquad (7)$$

$$\Sigma^x_{(i,j)} \approx F_{(i,j)}\,\Sigma^s_i\,F^{'}_{(i,j)} + G_{(i,j)}\,\Sigma^n_j\,G^{'}_{(i,j)} \qquad (8)$$

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

$$F_{(i,j)} = C\,\mathrm{diag}\left(\frac{1}{1 + exp\left[2\,C^+\,(\mu^n_j - \mu^s_i)\right]}\right)C^+ \qquad (9)$$

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

and I denotes the unity matrix. After model combination, the covariance matrix $\Sigma^x_{(i,j)}$ 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 $\hat{s}^{(i,j)}_t$ of the clean speech, given the corrupted speech at time t, is derived as follows :

$$\hat{s}^{(i,j)}_t = E\left[s_t|x_t, q^s_t = i, q^n_t = j\right] \qquad (11)$$

$$= \mu^s_i + \Sigma^s_i\,F^{'}_{(i,j)}\,\left(\Sigma^x_{(i,j)}\right)^{-1}\left(x_t - \mu^x_{(i,j)}\right) \qquad (12)$$

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

Additionally, this Cartesian product HMM allows to calculate the a posteriori probabilities $\gamma^{(i,j)}_t$ of each combined (speech, noise) state, given the sequence of noisy observation vectors $x^T_1 = (x_1, x_2, \ldots, x_T)$:

$$\gamma^{(i,j)}_t = P\left[q^s_t = i, q^n_t = j|x^T_1\right] \qquad (13)$$

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

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 $\hat{s}^{(i,j)}_t$ depend only on the noisy speech vector at time t, the a posteriori probabilities $\gamma^{(i,j)}_t$ incorporate the temporal structure of the sequence x^T_1.

Finally, a global MMSE-estimate of the clean speech is obtained as a linear combination of the state-conditional estimates [11] :

$$\hat{s}_t = E\left[s_t|x^T_1\right] = \sum_i \sum_j \gamma^{(i,j)}_t\,\hat{s}^{(i,j)}_t \qquad (15)$$

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

3. Experiments

The performance of the proposed preprocessing technique is evaluated on the Aurora2 speaker independent connected digit recognition task. The speech recognition results reported here are produced by the complex back-end configuration as defined by the ETSI Aurora group. Whole word digit models were trained on the clean speech training database provided by Aurora2 using the HTK scripts with default settings. The digit models have 16 emitting states with 20 Gaussians per state, while the silence model has 3 states with 36 Gaussians per state. Also, a one-state short pause model, tied with the middle state of the silence model, is used.

Features were extracted by the mel-cepstrum front-end version 2.0. The parameter set used in the MBFE-preprocessing step, consists of the first 13 cepstral coefficients ($c0$, $c1$,..., $c12$) together with the log-energy. The inclusion of $c0$ allows the HMM-parameters to be transformed to the power spectral domain, as required by formulas (7) and (8). Because of the non-square DCT-matrix, the 13-dimensional means and variances are zero padded to the required number. After MBFE-enhancement a cepstral mean subtraction (CMS) reduces the convolutional distortion caused by the use of different microphones. Finally $c0$ and the log-energy are combined in an energy coefficient, which together with the dynamic coefficients yielded 39-dimensional feature vectors for recognition, as explained in [12].

Aurora2, clean training, multicondition testing.					
	Subway	Babble	Car	Exhibit.	Avg.
Clean	99.72	99.70	99.64	99.88	99.74
20 dB	98.50	98.16	98.81	98.58	98.51
15 dB	96.81	96.22	97.17	97.01	96.80
10 dB	93.25	91.60	93.20	92.53	92.65
5 dB	86.74	80.17	83.84	83.83	83.65
0 dB	71.23	54.26	63.91	66.80	64.05
-5 dB	48.79	23.91	36.77	46.59	39.02
Avg.	89.31	84.08	87.39	87.75	87.13

Table 1: *Recognition accuracy after MBFE-enhancement of the static coefficients with 8 Gaussian mixture ergodic noise model and word speech model.*

Aurora2, clean training, multicondition testing.					
	Subway	Babble	Car	Exhibit.	Avg.
Clean	99.72	99.70	99.64	99.88	99.74
20 dB	98.07	98.16	98.66	97.38	98.07
15 dB	95.27	95.80	96.66	95.25	95.75
10 dB	90.24	88.60	92.04	90.56	90.36
5 dB	78.85	71.89	78.53	77.41	76.67
0 dB	59.53	37.06	51.24	56.31	51.04
-5dB	34.02	0.79	22.40	34.90	23.03
Avg.	84.39	78.30	83.43	83.38	82.38

Table 2: *Recognition accuracy after MBFE-enhancement of the static coefficients with single Gaussian ergodic noise model and phoneme speech model.*

Since the computational load of the MBFE-algorithm is proportional to M^x and hence largely dependent on the complexity of the statistical models of both the background noise and the clean speech, the sensitivity of our algorithm to these implementation issues is discussed in the next sections.

3.1. MBFE noise model

In our experiments a one-state Gaussian mixture noise model was trained, based on the noise present in the Aurora dataset. As opposed to [13], this noise model is fixed for one noise condition. We compared both a single Gaussian and a 8 Gaussian mixture, obtained by the EM-algorithm. Whereas the single Gaussian noise model is very efficient in terms of computation, we would expect that a more complicated noise model, such as the one-state 8 Gaussian mixture model, can describe the environmental characteristics in more detail and with more flexibility. However, the difference in recognition accuracy was mainly visible at low SNR, as our results indicate (table 4 and 1), such that in most cases the eight-fold increase in computational load is not justified by the increase in accuracy.

3.2. MBFE speech model

For the speech model both a word model and a phoneme model are considered to gauge the sensitivity of our algorithm to the speech model complexity. An SNR-dependent factor that decreases with the SNR-level, was introduced to increase the transition probability to the silence model. This way we could avoid a significant amount of insertion errors at low SNR-levels.

Firstly, we trained the parameters of the 11 digits (one,…, nine, oh, zero) and the silence and short pause word models. The digit models have 16 emitting states with 3 Gaussians per state and are connected in a loop, while the silence model has 3 states with 6 Gaussians per state and the one-state short pause model is again tied with the middle state of the silence model.

Aurora2, clean training, multicondition testing.					
	Subway	Babble	Car	Exhibit.	Avg.
Clean	99.72	99.70	99.64	99.88	99.74
20 dB	98.45	98.31	98.66	97.84	98.32
15 dB	96.10	96.25	96.57	95.68	96.15
10 dB	91.43	91.23	92.37	91.42	91.61
5 dB	82.44	74.61	79.60	79.45	79.03
0 dB	65.15	42.11	53.68	61.12	55.52
-5dB	37.21	7.50	24.58	37.77	26.77
Avg.	86.71	80.50	84.18	85.10	84.12

Table 3: *Recognition accuracy after MBFE-enhancement of the static coefficients with 8 Gaussian mixture ergodic noise model and phoneme speech model.*

Secondly, we prepared a speech model by connecting the phonemes that occur in the 11 digits in a loop, which reduced the computational requirements considerably. In general, these phonemes were modeled by a 3-state left-to-right HMM with 3 Gaussians per state, except for some of the vowels and diphthongs for which 5-state to 10-state HMMs were used, such that a total of 234 Gaussians is used. The experiments (table 2 and 3) indicate that the accuracy for the connected digit recognition task is somewhat lower when a phoneme speech model is used, which indicates that errors made in the front-end propagate to some extent to the back-end recogniser.

3.3. Dynamic coefficients

When using a word speech model, it turned out that calculating the dynamic parameters on the MBFE-enhanced feature sequence is worse than not changing them at all. Hence for the results in table 4 and 1 only the static parameters are preprocessed, while the deltas and delta-deltas are obtained from the noisy feature vectors. On the other hand, in the case of a phoneme speech model (table 2 and 3) superior recognition accuracy was obtained with the dynamic coefficients calculated on the enhanced vectors. However, we found that in this case the results could be improved even more when the deltas are obtained from the noisy feature vectors, but the delta-deltas are calculated on a smoothed version of the MBFE-enhanced feature sequence. To this end we use a low-pass filter:

$$H(z) = \frac{1}{\left(2 - z^{-1}\right)^2} \tag{16}$$

to smooth the difference between the state-conditional estimates $\hat{s}_t^{(i,j)}$ and the observed noisy speech before combination in (15). As shown in table 5, the recognition accuracy at high SNR-levels becomes only slightly inferior to the accuracy obtained with a digit speech model (table 4).

4. Conclusions

Preprocessing the noisy speech feature vectors with the MBFE-enhancement algorithm before they are decoded by the recogniser, reduces the word error rate of the ASR-system considerably. Moreover, this technique proves to be successfully applicable to non-stationary noise conditions, while the computational load remains feasible. With regard to speeding up the algorithm, pruning could still be considered in the calculation of $\alpha_t(i,j)$, $\beta_t(i,j)$, $\gamma_t^{(i,j)}$ and $\hat{s}_t$. We have shown that in the MBFE-algorithm a trade-off exists between the computational

Aurora2, clean training, multicondition testing.														
	A					B					C			
	Subw.	Babble	Car	Exhibit.	Avg.	Rest.	Street	Airprt	Station	Avg.	Sub.M	Str.M	Avg.	Avg.
Clean	99.72	99.70	99.64	99.88	99.74	99.72	99.70	99.64	99.88	99.74	99.66	99.73	99.70	99.73
20 dB	98.43	98.22	98.93	98.24	98.46	98.74	99.10	98.81	98.83	98.87	97.97	97.58	97.78	98.49
15 dB	96.78	96.25	97.05	96.73	96.70	96.65	96.18	96.69	95.96	96.37	95.73	94.01	94.87	96.20
10 dB	92.72	91.02	93.08	91.51	92.08	90.27	88.54	92.09	89.20	90.03	92.05	86.31	89.18	90.68
5 dB	85.05	78.17	82.91	81.30	81.86	78.78	75.12	81.06	77.63	78.15	82.87	68.77	75.82	79.17
0 dB	69.11	51.66	62.39	62.67	61.46	55.14	52.06	59.02	53.44	54.92	67.45	43.80	55.63	57.67
-5 dB	46.58	20.92	35.82	44.92	37.06	26.83	29.35	30.90	27.28	28.59	42.83	24.67	33.75	33.01
Avg.	88.42	83.06	86.87	86.09	86.11	83.92	82.20	85.53	83.01	83.67	87.21	78.09	82.65	84.44

Table 4: *Recognition accuracy after MBFE-enhancement of the static coefficients with single Gaussian ergodic noise model and word speech model.*

Aurora2, clean training, multicondition testing.														
	A					B					C			
	Subw.	Babble	Car	Exhibit.	Avg.	Rest.	Street	Airprt	Station	Avg.	Sub.M	Str.M	Avg.	Avg.
Clean	99.72	99.70	99.64	99.88	99.74	99.72	99.70	99.64	99.88	99.74	99.66	99.73	99.70	99.73
20 dB	98.46	98.49	99.02	98.06	98.51	98.80	97.58	98.63	98.73	98.44	97.76	97.04	97.40	98.26
15 dB	96.32	95.65	97.38	96.08	96.36	97.33	95.41	96.87	96.48	96.52	95.06	93.44	94.25	96.00
10 dB	91.34	89.66	92.34	90.74	91.02	90.27	86.88	91.98	90.19	89.83	90.18	83.83	87.01	89.74
5 dB	80.44	73.58	78.41	78.96	77.85	76.54	68.65	78.35	75.25	74.70	77.25	61.22	69.24	74.87
0 dB	60.73	40.75	50.52	56.99	52.25	47.71	41.48	50.73	45.51	46.36	57.02	34.67	45.85	48.61
-5 dB	37.06	8.10	23.98	34.90	26.01	14.40	19.53	19.68	17.99	17.90	34.36	17.26	25.81	22.73
Avg.	85.46	79.63	83.53	84.17	83.20	82.13	78.00	83.31	81.23	81.17	83.45	74.04	78.75	81.50

Table 5: *Recognition accuracy after MBFE-enhancement with single Gaussian ergodic noise model and phoneme speech model and smoothing of the delta-deltas.*

load and the accuracy and that recognition accuracy is still high when using a single Gaussian noise model in combination with a phoneme speech model.

As explained in section 3, the dynamic coefficients were not yet enhanced when using a digit model. However, since the deltas and delta-deltas are inaccurate as soon as one of the static parameters within its window is distorted, we are convinced that this enhancement could still improve the performance of our algorithm. Moreover, when different microphones are used the MBFE-speech model will no longer yield an accurate fit. Hence we expect that the incorporation of the effect of convolutional distortions into our speech model will further improve the accuracy.

5. References

[1] Y. Gong, "Speech recognition in noisy environments : A survey," *Speech Comm.*, vol. 16, no. 3, pp. 261–291, 1995.

[2] 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.

[3] A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, Albuquerque, U.S.A., Apr. 1990, pp. 845–848.

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sept. 2001, pp. 217–220.

[5] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, 2001.

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

[7] 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.

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

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

[10] 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, pp. 733–736.

[11] 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.

[12] D. Macho, L. Mauuary, B. Noé, Y.M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proc. ICSLP*, Denver, Colorado, U.S.A., Sept. 2002, pp. 17–20.

[13] M. Ida and S. Nakamura, "HMM composition-based rapid model adaptation using a priori noise GMM adaptation evaluation on Aurora2 corpus," in *Proc. ICSLP*, Denver, Colorado, U.S.A., Sept. 2002, pp. 437–440.