Multiple Stream Model-Based Feature Enhancement for Noise Robust Speech Recognition

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

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{vstouten,hvanhamm,wambacq}@esat.kuleuven.ac.be

Abstract

In this paper, we motivate the introduction of multiple feature streams to cover the gap between the noise-free and the estimated features in the context of Model-Based Feature Enhancement (MBFE) for noise robust speech recognition. Especially at low local SNR-levels the global MMSE-estimate might not be optimal and its uncertainty is large. Therefore, it is first shown how a constrained quadratic optimisation problem can improve the linear combination weights in the MMSE-formula. Alternatively, these weights are then approximated by K Kronecker deltas. Both approaches are compared by recognition experiments on the Aurora2 task. Also, Multiple Stream MBFE is validated on the large vocabulary Aurora4 benchmark task. On the latter, a decrease in average Word Error Rate could be obtained from 37.73% (no enhancement) to 26.13% (single stream MBFE) and finally, to 24.89% (multiple stream MBFE).

1. Introduction

Model-Based Feature Enhancement (MBFE) has proven to be a scalable and efficient technique to jointly reduce the interfering additive and channel noise from a noisy speech utterance before recognition by an Automatic Speech Recognition (ASR) system [1]. In this technique, a model combination with a Vector Taylor Series (VTS) approximation is applied in a front-end preprocessing step. This considerably reduces the computational load compared to e.g. JAC [2] or PMC [3], due to the drop in the required complexity of the models that are adapted. Because the generated MMSE-estimate of clean speech exhibits far less mismatch with the acoustic models (that are trained on clean speech) than the observed noisy speech, a considerable increase in recognition accuracy is obtained. However, until now the back-end recogniser considered the MMSE-estimate as if it was a true clean utterance, while inevitably, in every feature enhancement algorithm some residual uncertainty is left. A few attempts to incorporate information about the uncertainty can be found in [4, 5, 6]. In this paper, we motivate the introduction of multiple feature streams to cover the gap between the true and the estimated features.

The baseline MBFE-algorithm is briefly reviewed in section 2. In section 3, we first explain how the linear combination weights in the MMSE-formula can be improved by solving a constrained quadratic optimisation problem. Alternatively, these weights are then approximated by Kronecker deltas. Experimental evidence of the increased recognition accuracy of the resulting system will be given in sections 4 and 5, where recognition results on both the Aurora2 task and the complete Aurora4 task are presented. Finally, conclusions can be found in section 6.

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

2. Baseline MBFE

The main principles of the MBFE-technique are now briefly reviewed. First, a shifted HMM-model λ^s of the clean speech and an HMM λ^n of the noise are combined in the MBFE front-end, by which an estimate of the noisy speech HMM λ^x is obtained. The state-conditional pdfs of clean speech s_t and noise n_t are assumed to be Gaussian mixtures with means μ_i^s, μ_j^n and diagonal covariance matrices Σ_i^s, Σ_j^n in the cepstral domain, respectively. The non-linearity of the relationship between s_t, n_t, the channel h and the noisy speech x_t is approximated by a first order Vector Taylor Series:

$$\begin{aligned} x_t &= f(s_t, n_t, h) \\ &\approx C \log\left(\exp\left(C^{-1}\left(s_t + h\right)\right) + \exp\left(C^{-1} n_t\right)\right) \quad (1) \\ &\approx f\left(\mu_i^s, \mu_j^n, \overline{h}\right) + F_{(i,j)}\left(s_t - \mu_i^s\right) + G_{(i,j)}\left(n_t - \mu_j^n\right) \quad (2) \end{aligned}$$

in which C denotes the DCT-matrix, and the gradients of the combination function $f(s_t, n_t, h)$ have the closed form:

$$F_{(i,j)} = C\,\mathrm{diag}\left(\frac{1}{1 + exp\left[C^{-1}\left(\mu_j^n - \mu_i^s - \overline{h}\right)\right]}\right) C^{-1} \quad (3)$$

$$G_{(i,j)} = I - F_{(i,j)} \quad (4)$$

and I is the identity matrix. The Gaussian pdf of x_t then has a mean and a covariance matrix:

$$\begin{aligned} \mu_{(i,j)}^x &\approx C \log\left(\exp\left(C^{-1}\left(\mu_i^s + \overline{h}\right)\right) + \exp\left(C^{-1}\,\mu_j^n\right)\right) \\ &\quad + F_{(i,j)}\,\delta h \quad (5) \\ \Sigma_{(i,j)}^x &\approx F_{(i,j)}\,\Sigma_i^s\,F_{(i,j)}^{'} + G_{(i,j)}\,\Sigma_j^n\,G_{(i,j)}^{'} \quad (6) \end{aligned}$$

The shift $(\overline{h} + \delta h)$ of the clean speech HMM is obtained by an iterative EM-algorithm to jointly remove additive and channel noise [7], since any linear filtering operation results in a shift in the cepstral domain. The corresponding update formula is given by:

$$\begin{aligned} \delta h = &\left[\sum_t \sum_{(i,j)} \gamma_t^{(i,j)}\,F_{(i,j)}^{'}\left(\Sigma_{(i,j)}^x\right)^{-1} F_{(i,j)}\right]^{-1} \\ &\cdot \left[\sum_t \sum_{(i,j)} \gamma_t^{(i,j)}\,F_{(i,j)}^{'}\left(\Sigma_{(i,j)}^x\right)^{-1}\left(x_t - \overline{\mu_{(i,j)}^x}\right)\right] \quad (7) \end{aligned}$$

Finally, an MMSE-estimate of the clean speech, given the noisy observation vectors $x_1^T = (x_1, x_2, \ldots, x_T)$, is calculated:

$$\begin{aligned} \hat{s}_t^{MMSE} &= E\left[s_t|x_1^T\right] = \sum_{(i,j)} P[i,j|x_1^T]\, E\left[s_t|x_1^T, i, j\right] \\ &= \sum_{(i,j)} \gamma_t^{(i,j)}\, \hat{s}_t^{(i,j)} \end{aligned} \quad (8)$$

in which (i, j) denotes the combined (speech, noise) state. The state-conditional estimates are given by:

$$\hat{s}_t^{(i,j)} = \mu_i^s + \Sigma_i^s\, F'_{(i,j)} \left(\Sigma^x_{(i,j)}\right)^{-1} \left(x_t - \mu^x_{(i,j)}\right) \quad (9)$$

The posterior probabilities $\gamma_t^{(i,j)}$ are calculated by the forward-backward algorithm. However, in our case the latter becomes trivial due to the use of uniformly weighted ergodic HMMs. Unlike for the state-conditional estimates, the velocity and acceleration features are used to determine the posterior probabilities. Assuming that the gradients F and G in the VTS approximation (2) remain constant across the time-interval on which the deltas are calculated, the values of the first derivative parameters of the combined HMM λ^x, are given by:

$$\mu^{\Delta x}_{(i,j)} \approx F_{(i,j)}\, \mu_i^{\Delta s} + G_{(i,j)}\, \mu_j^{\Delta n} \quad (10)$$

$$\Sigma^{\Delta x}_{(i,j)} \approx F_{(i,j)}\, \Sigma_i^{\Delta s}\, F'_{(i,j)} + G_{(i,j)}\, \Sigma_j^{\Delta n}\, G'_{(i,j)} \quad (11)$$

and similar for the second derivatives.

3. Improving the combination weights

Usually, the Gaussians in the acoustic space are sampled at only one point, namely the one that represents the most likely clean speech estimate ($\hat{s}_t^{MMSE}$) at that time. Since this estimate minimises the mean of the squared error, it can be expected to yield good results. However, especially at low local SNR-levels its uncertainty is large and this MMSE-estimate might not be optimal [8]. After all, due to the fully connected front-end HMMs, almost no time constraints are incorporated in calculating the weights in (8). The only temporal information arises from the dynamic features used in the forward-backward algorithm to calculate the posterior probabilities. Therefore, it can be expected that the combination of the state-conditional estimates $\hat{s}_t^{(i,j)}$ with these $\gamma_t^{(i,j)}$ could still be improved if more information was available. On the other hand, the more detailed back-end acoustic model can make a more profound choice between these estimates, based on the larger context in which each frame occurs. This motivates the next optimisation problem.

3.1. Optimisation problem

For each time instant t, let $\beta_t(m), m = 1 \ldots (M^s.M^n)$ be the sorted states (i, j), obtained by sorting the corresponding posterior probabilities $\gamma_t^{(i,j)}$ in descending order. Here M^s and M^n denote the number of speech and noise states in λ^s and λ^n, respectively. Since the $\gamma_t^{(i,j)}$ can be considered as rough estimates of the optimal combination weights, they give an indication of the dominant terms in (8). In the remaining part, only M of these terms are retained. Hence, for $m = 1 \ldots M$:

$$\hat{s}_{t,m} = \hat{s}_t^{(i,j)} \quad \text{if} \quad (i,j) = \beta_t(m) \quad (12)$$

Let $\alpha = [\alpha_1 \ldots \alpha_M]'$ be the vector with the (unknown) optimal posterior probabilities in (8) and let $S = [\hat{s}_{t,1} \ldots \hat{s}_{t,M}]$ be the matrix with the corresponding sorted state-conditional estimates. For each back-end Gaussian q, α can be calculated by maximising its log-likelihood at that time instant:

$$\max_{\alpha} \{-\frac{1}{2}(S\alpha - \mu_q)'\ (\Sigma_q)^{-1}\ (S\alpha - \mu_q)\} \quad (13)$$

subject to the linear constraints:

$$\alpha_m \geq 0 \quad \text{for} \quad m = 1 \ldots M \quad (14)$$

$$\sum_m \alpha_m = 1 \quad (15)$$

This optimisation is solved with a gradient descent algorithm, in which $\alpha^{(0)}$ is initialised to $[1/M \ldots 1/M]'$. In iteration l we then have:

$$\alpha^{(l)} = \alpha^{(l-1)} - \delta \left(S'\ (\Sigma_q)^{-1} S\alpha^{(l-1)} - S'\ (\Sigma_q)^{-1}\ \mu_q\right) \quad (16)$$

with the step δ chosen to maximise (13) after substituting (16), subject to the constraints (15). This requires $O(DM^2)$ multiplications, with D the dimension of the feature vectors. Since the optimised clean speech estimate $S\alpha$ is calculated on a truncated version of S, both $S\alpha$ and the original MMSE-estimate $\hat{s}_t^{MMSE}$, are evaluated in the acoustic model of the recogniser. At each time instant, only the score of the best matching one is kept for recognition. Alternatively, $\hat{s}_t^{MMSE}$ can be included in S, which gave almost no difference in performance.

3.2. Multiple streams of estimates

Instead of using a computationally rather expensive optimisation, each (or some) of the corners of the polygon can be considered as a candidate solution. In this case, the unknown optimal combination weights are approximated by K Kronecker deltas. Hence, the K feature streams are obtained by selecting only the dominant terms in (8). For $k = 1 \ldots K$:

$$\hat{s}_{t,k} = \hat{s}_t^{(i,j)} \quad \text{if} \quad (i,j) = \beta_t(k) \quad (17)$$

Then, each of these K feature streams, together with the MMSE-estimate $\hat{s}_t^{MMSE}$, give rise to $(K + 1)$ streams that are evaluated in the acoustic model of the recogniser. At each time instant, the score of the best matching one is kept for recognition. In this case, only $(K + 1)$-times more Gaussian evaluations are required in the back-end, which is still feasible since the number of streams can be kept reasonably small, as will be shown in section 5.1.

4. Experiments on Aurora2

To illustrate the superior recognition accuracy that is obtained by implementing these techniques, experiments are conducted on the Aurora2 speaker independent digit recognition task for two noise types (subway and car noise) and SNR-levels between 0dB and 20dB.

Features are extracted by the MFCC front-end, complying to the ETSI ES 201 108 standard without compression. All results are obtained by enhancing the noisy speech by the MBFE-algorithm, using front-end models with 128 fully connected Gaussians for the clean speech model and 1 Gaussian for the noise model. The parameters of both models are obtained offline for each SNR-level. A channel estimate is calculated online by the recursive EM-algorithm described in [7]. Front-end estimates are evaluated by the complex back-end recognition system, with whole word digit models trained on the clean

speech training database of 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.

4.1. Optimal combination weights

In the optimisation problem (3.1), s is truncated to M state-conditional estimates. Because we neither want to exclude some of the correct estimates, nor want to incorporate very unlikely estimates, a trade-off has to be made. We choose $M = 4$, because the $\gamma_t^{(i,j)}$ for states corresponding to $m > 4$ are already quite small. To limit the computational load, only 1 iteration is done in (16). To allow a fair comparison, K is also set to 4 in the Multiple Stream experiments.

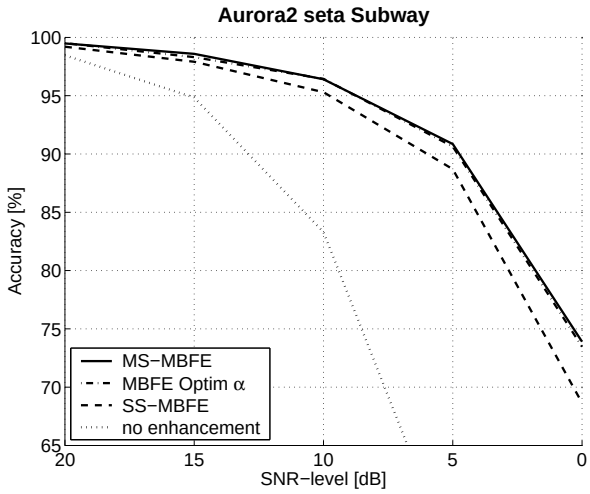

Figure 1: *Accuracy for reference, SS-MBFE* ($K = 0$)*, MS-MBFE* ($K = 4$) *and MBFE with optimisation of* α ($M = 4$).

4.2. Results

Recognition results are shown in figure 1 for the subway noise, and in figure 2 for the car noise condition. Single Stream MBFE ($K = 0$) is compared with Multiple Stream MBFE ($K = 4$) and MBFE for which the optimisation problem is solved iteratively. No significant difference in accuracy is observed between applying 1 iteration of a gradient descent algorithm on the one hand and evaluating each of the corners of the polygon (the MS-MBFE case) on the other hand. Hence, from a computational point of view MS-MBFE is preferable. Note that the corners of the polygon are valid solutions to the optimisation problem. Hence, when the gradient descent algorithm has converged, it will never be inferior to MS-MBFE in the sense of (13). However, the difference in performance between both approaches is in any case small.

5. Experiments on Aurora4

Experiments are also conducted on the Aurora4 large vocabulary database, derived from the WSJ0 Wall Street Journal 5k-word dictation task. In this database, seven different types of noise are added to the close talking microphone signal: 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

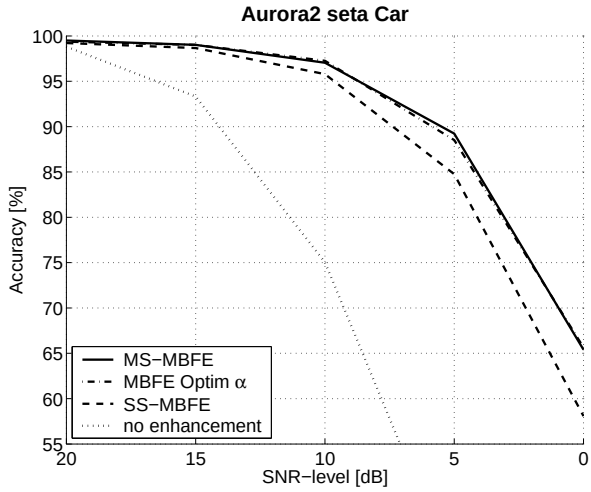

Figure 2: *Accuracy for reference, SS-MBFE* ($K = 0$)*, MS-MBFE* ($K = 4$) *and MBFE with optimisation of* α ($M = 4$).

14 are obtained by adding these same noise types to recordings made with 18 different microphones. For each of the 2x7 test sets, all 330 utterances (with an SNR-level that ranges from 5 dB to 15 dB) are evaluated.

First, the mel-cepstral features are extracted from the speech signal as explained in [7]. Then, the K feature streams, together with the global MMSE-estimate of the clean speech are calculated by the Multiple Stream MBFE-algorithm of section 3.2. Finally, the first and second order time derivatives are added and the MIDA-algorithm is applied to reduce the features to 39 dimensions. The front-end models λ^s and λ^n for speech and noise are trained on the mel-cepstral features. The pdfs consist of 256 and 1 single-Gaussian states, respectively, with diagonal covariance matrices. The noise statistics are obtained from the first 30 and the last 30 frames of each sentence.

Because of its fast experiment turn-around time and good baseline accuracy, the speaker-independent LVCSR-system of the K.U.Leuven–ESAT speech group, is used as a back-end recogniser (see [7] for details).

5.1. Number of feature streams

We now explore how the number of feature streams affects the Word Error Rate (WER). As can be seen from figure 3, the introduction of multiple feature streams ($K > 1$) gives rise to a decrease of the WER. This verifies our previous small vocabulary recognition results. However, when K becomes too large, also many unlikely state-conditional estimates are passed to the recogniser, such that the WER increases again. We observe that the optimal value for the number of streams is $K = 6$.

5.2. Results

The first reference results (labelled MIDA in table 1) are obtained when no explicit noise reduction algorithm is applied. Secondly, features are preprocessed by the standard AFE without compression [9]. Thirdly, recognition is performed with SS-MBFE (using only $\hat{s}_t^{MMSE}$). Finally, we show the recognition results when MS-MBFE is applied (using $\hat{s}_t^{MMSE}$ together with $\hat{s}_{t,1} \ldots \hat{s}_{t,6}$). In table 1, mic1 and mic2 denote the average of the first 7 and the last 7 noise conditions, respectively. By comparing AFE with SS-MBFE, we conclude that SS-MBFE

Aurora4, 16 kHz sampling, no compression, no end pointing; Clean condition training.																	
	Close Talk								Far Talk								
TEST	01	02	03	04	05	06	07	mic 1	08	09	10	11	12	13	14	mic 2	Avg.
MIDA	4.95	17.97	32.84	39.88	36.67	28.21	38.24	28.39	23.59	39.44	50.68	55.35	56.81	47.30	56.42	47.08	37.74
AFE	5.44	17.88	23.07	27.93	26.86	22.90	24.72	21.26	25.31	35.40	42.26	43.62	46.12	42.14	42.87	39.67	30.47
SS-MBFE	5.10	8.11	18.81	27.05	21.69	20.44	22.64	17.69	19.24	26.13	37.21	41.15	41.38	37.29	39.62	34.57	26.13
MS-MBFE	4.91	7.53	18.16	25.56	20.74	17.47	21.76	16.59	18.14	25.28	35.36	40.13	40.05	34.82	38.56	33.19	24.89

Table 1: *Word Error Rates without enhancement, with Advanced Front-End preprocessing, with Single Stream MBFE* $(K = 0)$ *and with Multiple Stream MBFE-enhancement* $(K = 6)$*; Clean condition training.*

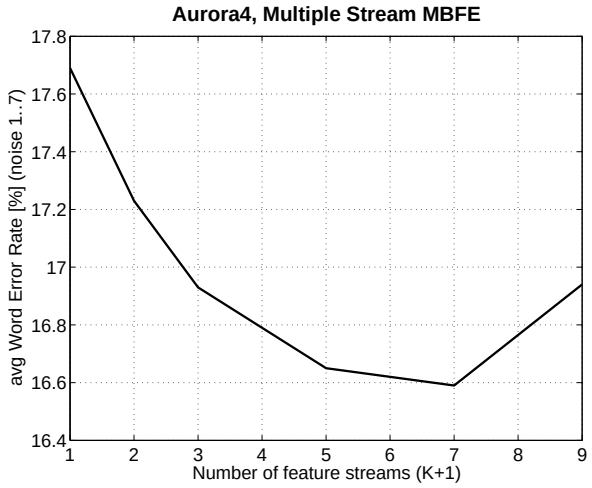

Figure 3: *Effect of the number of feature streams on the average Word Error Rate (Aurora4, test 01 . . . 07).*

obtains a strong baseline performance with a lower WER for each of the 14 noise types. On average, SS-MBFE gives a relative WER-reduction of 14.2% compared to AFE. It is also observed that supplying multiple front-end estimates to the back-end recogniser, proves to be very successful. MS-MBFE can decrease the average WER by another 4.7% relative to SS-MBFE.

6. Conclusions

In this paper we have motivated the need for an improvement of the linear combination weights in the MMSE-formula. To this end, we formulated a constrained quadratic optimisation problem for each back-end Gaussian at each time instant, that was solved by a gradient descent algorithm. In this way, the decision over the best clean speech estimate is postponed until more detailed information is available from the back-end. Because of the rather large computational requirements, we then proposed to approximate the solution K times by a Kronecker delta. In this case, only $(K + 1)$-times more Gaussian evaluations are required, which is still feasible since K can be kept reasonably small. Experimental evidence was given for the similar recognition performance of both approaches on the Aurora2 digit recognition task. From a computational point of view, Multiple Stream MBFE is clearly preferable.

Also, the MS-MBFE-algorithm was tested on the large vocabulary Aurora4 dictation task. The baseline SS-MBFE was compared with the Advanced Front-End (AFE) standard and with the MS-MBFE. The optimal value for K was found to be 6. Experiments confirmed the superior performance that was obtained by generating $(K + 1)$ feature streams.

Future work includes the improvement of the dynamic features, which are the only timing information available to the front-end. This would give rise to more accurate posterior probabilities, and hence could further improve our recognition results. Other approximations to compute the velocity and acceleration parameters of the combined model will be investigated.

7. Acknowledgement

This work has been partially supported by the IST-2001-38299 MUSA project.

8. References

[1] V. Stouten, H. Van hamme, K. Demuynck, and P. Wambacq, "Robust speech recognition using model-based feature enhancement," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 17–20.

[2] A. Bernard, Y. Gong, and X. Cui, "Can back-ends be more robust than font-ends? Investigation over the Aurora-2 database," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 1025–1028.

[3] 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.

[4] T. Kristjansson and B. Frey, "Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition," in *Proc. ICASSP*, Orlando, Florida, May 2002, pp. 61–64.

[5] J. Arrowood and M. Clements, "Using observation uncertainty in HMM decoding," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 1561–1564.

[6] L. Deng, J. Droppo, and A. Acero, "Exploiting variances in robust feature extraction based on a parametric model of speech distortion," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 2449–2452.

[7] V. Stouten, H. Van hamme, and P. Wambacq, "Joint removal of additive and convolutional noise with model-based feature enhancement," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 949–952.

[8] V. Stouten, H. Van hamme, and P. Wambacq, "Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement," in *Proc. ICSLP*, Jeju Island, Korea, Oct. 2004.

[9] ETSI standard doc., "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," *ETSI ES 202 050 v1.1.1 (2002-10)*.