Accounting for the Uncertainty of Speech Estimates in the Context of Model-Based Feature Enhancement

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 present two techniques to cover the gap between the true and the estimated clean speech features in the context of Model-Based Feature Enhancement (MBFE) for noise robust speech recognition. While in the output of every feature enhancement algorithm some residual uncertainty remains, currently this information is mostly discarded. Firstly, we explain how the generation of not only a global MMSE-estimate of clean speech, but also several alternative (state-conditional) estimates are supplied to the back-end for recognition. Secondly, we explore the benefits of calculating the variance of the front-end estimate and incorporating this in the acoustic models of the recogniser. Experiments on the Aurora2 task confirmed the superior performance of the resulting system: an average increase in recognition accuracy from 85.65% to 88.50% was obtained for the clean training condition.

1. Introduction

Previously, we have investigated the capabilities of Model-Based Feature Enhancement (MBFE) in reducing the interfering additive and channel noise from a noisy speech utterance before recognition by an Automatic Speech Recognition (ASR) system [1]. We showed how the principles of Parallel Model Combination (PMC) [2] can be applied in a front-end preprocessing step, which considerably reduces the computational load, since the complexity of the models can drop with a factor of 28 (128 front-end Gaussians compared to 3628 back-end Gaussians in the Aurora2 experiments). 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. In this paper, we explore two techniques to incorporate the uncertainty of the front-end estimate of clean speech, as will be explained in section 3 and section 4, to cover the gap between the true and the estimated features. Experimental evidence of the increased recognition accuracy of the resulting system will be given in section 5, where recognition results on the complete Aurora2 task are presented. Finally, conclusions can be found in section 6.

2. Baseline MBFE

The main principles of the MBFE-technique are now briefly reviewed. First, a shifted HMM-model of the clean speech and an HMM of the noise are combined in the MBFE front-end, by which an estimate of the noisy speech HMM is obtained. The state-conditional pdfs of clean speech s_t and noise n_t are assumed 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 :

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

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

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

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

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

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

$$\Sigma_{(i,j)}^x \approx F_{(i,j)} \Sigma_i^s F_{(i,j)}' + G_{(i,j)} \Sigma_j^n G_{(i,j)}' \quad (6)$$

The shift $(\bar{h} + \delta h)$ of the clean speech HMM is obtained by a novel iterative EM-algorithm that extended the baseline MBFE to jointly remove additive and channel noise (see [3] for details), since any linear filtering operation is reflected by a shift in the cepstral domain. Finally, an MMSE-estimate of the clean speech, given the noisy observation vectors $x_1^T = (x_1, x_2, \dots, x_T)$, is calculated :

$$\hat{s}_t^{MMSE} = E \left[s_t | x_1^T \right] = \sum_{(i,j)} \gamma_t^{(i,j)} \hat{s}_t^{(i,j)} \quad (7)$$

in which $\gamma_t^{(i,j)} = P[i, j | x_1^T]$ are the posterior probabilities for each combined (speech, noise) state (i, j) and the state-conditional estimates are given by:

$$\hat{s}_t^{(i,j)} = \mu_i^s + \Sigma_i^s F_{(i,j)}' \left(\Sigma_{(i,j)}^x \right)^{-1} (x_t - \mu_{(i,j)}^x) \quad (8)$$

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

3. Multiple streams of estimates

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. However, especially at low local SNR-levels its uncertainty is large and this MMSE-estimate might not be optimal. After all, due to the fully connected front-end HMMs, no temporal information is used in calculating the weights in (7). 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 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. To avoid the need for a second iteration and instead of using a computationally expensive optimisation, we use K approximations of the unknown optimal combination weights. Namely, by K Kronecker deltas each of the K terms with a dominant weight in (7) is selected. Finally, each of these state-conditional estimates, together with the MMSE-estimate $\hat{s}_t^{MMSE}$, give rise to $(K + 1)$ feature streams that 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. In this way, the decision over the best clean speech estimate is postponed until more detailed information is available. Note that the calculation of these supplementary estimates does not involve an extra computational load of the MBFE-algorithm, since they are needed anyhow to compute $\hat{s}_t^{MMSE}$. On the other hand, $(K + 1)$ -times more Gaussian evaluations will be required in the back-end, but this is still feasible since the number of streams can be kept reasonably small, as can be observed from figure 1 (where the posterior probability of the 4th state-conditional estimate is already quite small).

Figure 1: Posterior probabilities of the 4 most likely state-conditional estimates.

4. Variance of estimates

Another way to incorporate the uncertainty of the estimate in the recognition process, is by taking the variance of the MMSE-estimate into account. Indeed, while the MMSE-estimate represents an expected value, also all necessary information is available to calculate the variance of the clean speech estimate. The latter can then be used to increase the variance of the acoustic

models in the back-end recogniser, thereby expressing the larger uncertainty at specific time instants. This idea exhibits some similarities with work from other researchers, such as [4, 5, 6], in which also Gaussian pdf descriptions of features are used instead of points in feature space. In the MBFE-context, the covariance matrix of $\hat{s}_t^{MMSE}$ at time t , is given by :

$$\Sigma_{MMSE}^{\hat{s}_t} = E \left[(s_t - \hat{s}_t^{MMSE}) (s_t - \hat{s}_t^{MMSE})' | x_1^T \right] \quad (9)$$

$$= \sum_{(i,j)} \gamma_t^{(i,j)} \Sigma_{(i,j)}^{\hat{s}_t} + \sum_{(i,j)} \gamma_t^{(i,j)} \cdot (\hat{s}_t^{MMSE} - \hat{s}_t^{(i,j)}) (\hat{s}_t^{MMSE} - \hat{s}_t^{(i,j)})' \quad (10)$$

in which the covariance matrices of the state-conditional estimates are given by :

$$\Sigma_{(i,j)}^{\hat{s}_t} = \Sigma_i^s - \Sigma_i^s F_{(i,j)}' (\Sigma_{(i,j)}^x)^{-1} F_{(i,j)} \Sigma_i^s \quad (11)$$

After diagonalisation of (10), a fraction of this variance is added to the variances of the back-end acoustic model. Note that this implies that the resulting variance will be different at each time instant. Figure 2 gives an example taken from the Aurora2 database (setB, street noise) at an SNR-level of 10dB. This clearly illustrates that the variance is high when the local SNR-level is low, or equivalently, when the uncertainty is high. On the other hand, when the energy in the signal is high, the variance is low, indicating a higher confidence in the generated estimate. When we compare figure 2 and 1, we observe that whenever the variance of the MMSE-estimate is low, one of the state-conditional probabilities dominates the others.

Figure 2: c_0 component of the noisy speech and of the MMSE-estimate of clean speech, together with the variance of the latter (scaled for representation only).

5. Experiments

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 the 10 noise types and the 6 SNR-levels that are provided. Features are extracted by the mel-cepstrum front-end version 2.0. 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. A channel estimate is calculated online by the recursive EM-algorithm described in [3]. 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.

5.1. Multiple streams of estimates

Experiments were conducted with a different number of feature streams K . However, the results indicated that using 4 state-conditional estimates is optimal in this case. When less feature streams are considered, performance is closer to the baseline performance (in which only the MMSE-estimate is used), while increasing the value of K forces the MBFE-algorithm to generate very unlikely estimates, such that the performance drops again. Only the static components of the noisy feature vector stream are enhanced by the MBFE-algorithm. The velocity and acceleration components are generated from these enhanced streams. In our experiments, the posterior probabilities are not used explicitly by the back-end. Since K is kept small enough not to generate very unlikely estimates, incorporating these probabilities did not improve the results.

5.2. Variance of estimates

In our experiments, both the static and the velocity components of the variances of the back-end acoustic model are increased, while the acceleration components are left unchanged. Under the assumption of uncorrelated frames, the velocity components of the variance of the speech estimates can be calculated from the corresponding static components in (10) as follows :

$$\Sigma_{MMSE}^{\Delta \hat{s}_t} = \sum_{\tau=-L}^{+L} w_{\tau}^2 \Sigma_{MMSE}^{\hat{s}_{t+\tau}} \quad (12)$$

in which the w_{τ} denote the filter coefficients from the delta-calculation with a window length of $2L + 1$. Although the acceleration components of the variance of the speech estimates can be obtained in a similar way, preliminary results indicated that in this case the approximations become too large and prevent a further increase in performance. A comparable conclusion was drawn in [4]. Finally, the obtained front-end variances are multiplied by a constant factor (in our case this factor was experimentally tuned to 0.1) before adding them to the back-end variances on a frame-by-frame basis.

5.3. Results

Table 1 presents the reference recognition accuracy of the baseline system, when only the MMSE-estimate is evaluated in the back-end acoustic model. Table 2 shows how the performance of this system improves when not only the MMSE-estimate, but also the 4 most likely state-conditional estimates are evaluated, as explained in section 3. The average recognition accuracy increases from 85.65% to 88.17%. Finally, table 3 presents the recognition accuracy after MBFE-enhancement using the MMSE-estimate, together with the 4 most likely state-conditional estimates during recognition, while the back-end

variance is increased with the variance of the MMSE-estimate, as explained in section 4. The average accuracy increases to 88.50%, which on the whole represents a relative digit error rate reduction of 19.8%.

6. Conclusions

In this paper we took advantage of the very flexible way in which the MBFE-algorithm can generate clean speech estimates. We described two techniques to incorporate the uncertainty of the input vectors to the back-end recogniser, both of which increased the recognition performance of the ASR-system considerably. The innovative idea to supply multiple cleaned front-end estimates to the back-end recogniser, proved very successful at a feasible computational load. Indeed, the use of 5 feature streams increases the recognition accuracy from 85.65% to 88.17% for the clean training condition, whereas increasing the variance of the back-end acoustic models has a smaller (additional) effect on the obtained performance. On the whole, we conclude that incorporating the uncertainty of the front-end estimates is definitely beneficial to the accuracy of the system.

Currently, we are evaluating our system on the more complex large vocabulary recognition task of Aurora4, derived from the WSJ0 Wall Street Journal 5k-word dictation task. As an extension, we will investigate to which extent the generated estimates can be further optimised when the front-end directly is supplied with feedback from the back-end recognition process.

7. 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] 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] 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, Accepted for publication.
- [4] 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.
- [5] J. Arrowood and M. Clements, "Using observation uncertainty in HMM decoding," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 1561–1564.
- [6] 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.

Aurora2, clean training, multicondition testing.														
	A					B					C			Avg.
	Subw.	Babble	Car	Exhibit.	Avg.	Rest.	Street	Airprt	Station	Avg.	Sub.M	Str.M	Avg.	
Clean	99.72	99.70	99.67	99.88	99.74	99.72	99.70	99.67	99.88	99.74	99.66	99.73	99.69	99.73
20 dB	99.20	99.18	99.22	99.01	99.15	99.23	98.85	99.22	99.17	99.12	98.93	98.28	98.61	99.03
15 dB	97.91	98.10	98.66	97.41	98.02	98.31	97.43	98.54	97.84	98.03	96.99	96.07	96.53	97.73
10 dB	95.30	94.35	95.79	93.61	94.76	94.20	91.99	95.05	93.89	93.78	93.61	88.36	90.98	93.62
5 dB	88.70	81.41	84.76	84.67	84.89	82.93	77.24	84.43	82.54	81.79	84.92	70.59	77.75	82.22
0 dB	68.65	48.70	58.07	61.56	59.25	54.65	47.88	58.48	51.87	53.22	63.99	42.96	53.48	55.68
-5 dB	36.14	17.50	20.64	31.22	26.38	21.92	17.20	21.98	15.74	19.21	31.29	17.38	24.34	23.10
Avg.	89.95	84.35	87.30	87.25	87.21	85.86	82.68	87.14	85.06	85.19	87.69	79.25	83.47	85.65

Table 1: Recognition accuracy after MBFE-enhancement using only the MMSE-estimate during recognition.

Aurora2, clean training, multicondition testing.														
	A					B					C			Avg.
	Subw.	Babble	Car	Exhibit.	Avg.	Rest.	Street	Airprt	Station	Avg.	Sub.M	Str.M	Avg.	
Clean	99.72	99.70	99.67	99.88	99.74	99.72	99.70	99.67	99.88	99.74	99.66	99.73	99.69	99.73
20 dB	99.48	99.18	99.49	98.98	99.28	99.42	98.97	99.31	99.26	99.24	99.23	98.79	99.01	99.21
15 dB	98.59	98.46	99.02	98.12	98.55	98.13	97.94	98.39	97.93	98.10	98.31	97.55	97.93	98.24
10 dB	96.41	95.44	97.05	95.43	96.08	94.44	93.29	95.62	94.63	94.50	95.86	92.35	94.10	95.05
5 dB	90.85	85.25	89.23	86.92	88.06	85.35	82.41	87.29	85.34	85.10	89.41	77.81	83.61	85.99
0 dB	73.90	56.95	65.40	66.55	65.70	60.73	54.69	65.85	58.75	60.00	70.74	50.18	60.46	62.37
-5 dB	39.70	23.52	25.35	36.44	31.25	26.31	20.95	28.51	19.38	23.79	36.20	20.98	28.59	27.73
Avg.	91.85	87.06	90.04	89.20	89.53	87.61	85.46	89.29	87.18	87.39	90.71	83.34	87.02	88.17

Table 2: Recognition accuracy after MBFE-enhancement using the MMSE-estimate, together with the 4 most likely state-conditional estimates during recognition.

Aurora2, clean training, multicondition testing.														
	A					B					C			Avg.
	Subw.	Babble	Car	Exhibit.	Avg.	Rest.	Street	Airprt	Station	Avg.	Sub.M	Str.M	Avg.	
Clean	99.72	99.70	99.67	99.88	99.74	99.72	99.70	99.67	99.88	99.74	99.66	99.73	99.69	99.73
20 dB	99.48	99.21	99.52	99.07	99.32	99.45	98.97	99.31	99.29	99.26	99.26	98.85	99.06	99.24
15 dB	98.62	98.49	99.02	98.24	98.59	98.10	98.10	98.36	98.06	98.16	98.34	97.64	97.99	98.30
10 dB	96.62	95.41	97.17	95.53	96.18	94.72	93.38	95.68	94.54	94.58	95.89	92.59	94.24	95.15
5 dB	91.31	85.52	89.59	87.32	88.43	85.91	82.92	87.29	86.02	85.53	89.81	78.02	83.91	86.37
0 dB	74.36	58.43	66.78	67.29	66.72	61.68	55.96	67.28	59.64	61.14	71.57	51.24	61.41	63.42
-5 dB	41.76	24.24	25.89	37.06	32.24	27.42	21.95	29.26	19.65	24.57	37.00	21.10	29.05	28.53
Avg.	92.08	87.41	90.42	89.49	89.85	87.97	85.87	89.58	87.51	87.73	90.97	83.67	87.32	88.50

Table 3: Recognition accuracy after MBFE-enhancement using the MMSE-estimate, together with the 4 most likely state-conditional estimates during recognition, while the back-end variance is increased with the variance of the MMSE-estimate.