EFFECT OF PHASE-SENSITIVE ENVIRONMENT MODEL AND HIGHER ORDER VTS ON NOISY SPEECH 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

Model-based techniques for robust speech recognition often require the statistics of noisy speech. In this paper, we propose two modifications to obtain more accurate versions of the statistics of the combined HMM (starting from a clean speech and a noise model). Usually, the phase difference between speech and noise is neglected in the acoustic environment model. However, we show how a phase-sensitive environment model can be efficiently integrated in the context of Multi-Stream Model-Based Feature Enhancement and gives rise to more accurate covariance matrices for the noisy speech. Also, by expanding the Vector Taylor Series up to the second order term, an improved noisy speech mean can be obtained. Finally, we explain how the front-end clean speech model itself can be improved by a preprocessing of the training data. Recognition results on the Aurora4 database illustrate the effect on the noise robustness for each of these modifications.

1. INTRODUCTION

It is well known that distortions introduced by both the acoustic environment and the communication channel can significantly degrade Automatic Speech Recognition (ASR) performance. One class of techniques that addresses this problem consists of model-based techniques that either modify the back-end statistical models [1, 2] or compensate the observed acoustic feature vectors using estimates of the clean speech and/or the background (noise) model parameters [3, 4, 5, 6]. In this paper, we will focus on the latter approach.

Model-Based Feature Enhancement (MBFE) is a scalable and efficient technique to jointly reduce the interfering additive and convolutional noise from a noisy speech utterance before recognition by an ASR system [7, 9]. In this technique, a Hidden Markov Model combination with a first order Vector Taylor Series (VTS) approximation of the non-linear model of the acoustic environment is applied in a front-end preprocessing step. This considerably reduces the computational load compared to e.g. JAC [1] or PMC [2], due to the drop in the required complexity of the models that are adapted. Because the generated 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, the commonly used approximation of the non-linear relationship between the noisy, the clean speech and the noise cepstral feature vectors can be improved. We will show that a phase-sensitive environment model gives rise to a correction term for the combined HMM covariance matrices. Although this phase-sensitive model gave rise to a computationally intractable conditional observation probability density function (pdf) in the SNR-dependent Variance Model of [8], it can easily and efficiently be integrated in the context of Model-Based Feature Enhancement. For the combined HMM means, we will introduce a second order term in the VTS to better simulate the effect of the environment on clean speech feature vectors. Both modifications of our baseline system contribute to a decrease of the Word Error Rate.

The outline of this paper is as follows. First, the main principles of the Multi-Stream MBFE-technique are briefly reviewed. In section 3, the phase-sensitive environment model and the corresponding update of the covariance matrices will be introduced. Then, in section 4 the Vector Taylor Series is expanded up to the second order term. Experiments are conducted on the Aurora4 large vocabulary database, which is described in section 5. Also, a more accurate front-end clean speech model is obtained by applying a channel correction on the training data. This is explained in section 5.2. Finally, results and conclusions can be found in sections 5.4 and 6, respectively.

2. BASELINE MS-MBFE

The main principles of the MBFE-technique are now briefly reviewed. In MBFE, prior knowledge is integrated in the feature enhancement step by using two HMM models, namely λ^s for the clean speech cepstral feature vectors and λ^n for the noise cepstral feature vectors. 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, respectively. The linear filtering operation from the channel h, results in a shift in the cepstral domain of the clean speech model means μ_i^s. Therefore, the first step is to combine a shifted version of λ^s with λ^n in the MBFE front-end, by which an estimate of the noisy speech HMM λ^x is obtained. The often used relationship between s_t, n_t, h and the noisy speech x_t is given by:

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

in which C denotes the DCT-matrix. Its non-linearity is approximated by a first order Vector Taylor Series, with a state-dependent expansion point given by $(\mu_i^s, \mu_j^n, \overline{h})$:

$$x_t \approx f\left(\mu_i^s, \mu_j^n, \overline{h}\right) + F_{(i,j)}\left(s_t - \mu_i^s\right) + G_{(i,j)}\left(n_t - \mu_j^n\right) \quad (2)$$

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

and the gradients of the combination function $f(s_t, n_t, h)$ have the closed form :

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

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

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

$$\mu^x_{(i,j)} \approx C \,\log\left(\exp\left(C^{-1}\left(\mu_i^s + \overline{h}\right)\right) + \exp\left(C^{-1}\,\mu_j^n\right)\right) \quad (5)$$

$$\Sigma^x_{(i,j)} \approx F_{(i,j)}\,\Sigma_i^s\,F'_{(i,j)} + G_{(i,j)}\,\Sigma_j^n\,G'_{(i,j)} \quad (6)$$

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. The corresponding update formula is given by [9] :

$$\delta h = \left[\sum_t \sum_{(i,j)} \gamma_t^{(i,j)}\,F'_{(i,j)}\left(\Sigma^x_{(i,j)}\right)^{-1} F_{(i,j)}\right]^{-1} \cdot \left[\sum_t \sum_{(i,j)} \gamma_t^{(i,j)}\,F'_{(i,j)}\left(\Sigma^x_{(i,j)}\right)^{-1}\left(x_t - \mu^x_{(i,j)}\right)\right] \quad (7)$$

Finally, the undistorted clean speech is estimated from the observed feature stream. However, we do not want to make a hard decision about the correct clean speech in the front-end. Instead, the back-end acoustic model, which is more detailed than the front-end, can use a larger context in the decision process [6]. Therefore, multiple streams of front-end clean speech feature estimates are generated and sent to the (unchanged) back-end recogniser (Multi-Stream MBFE). These $(K+1)$ streams consist of K state-conditional estimates, together with the global MMSE-estimate of clean speech, given the noisy observation vectors $x_1^T = (x_1, x_2, \ldots, x_T)$. The global MMSE-estimate of clean speech is given by :

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

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

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

As can be seen from formula (9), in each state (i,j) we combine the prior speech knowledge (μ_i^s) with the standardised observation $((\Sigma^x_{(i,j)})^{-1}(x_t - \mu^x_{(i,j)}))$ according to the cross-covariance matrix $(\Sigma_i^s\,F'_{(i,j)})$ between clean speech and noisy speech. These estimates also minimise the mean squared estimation error.

3. PHASE-SENSITIVE ENVIRONMENT MODEL

In equation (1), the phase difference between speech and noise is neglected. Because it can be expected that a phase-sensitive environment model is more accurate, its formula is now derived. Assuming additive noise and linear channel distortion, the mel-frequency domain relationship between the noisy speech $X(f)$, the clean speech $S(f)$, the channel $H(f)$ and the noise $N(f)$, is given by :

$$X(f) = S(f)\,H(f) + N(f) \quad (10)$$

For the mel-power spectra, we can write :

$$|X(f)|^2 = |S(f)|^2\,|H(f)|^2 + |N(f)|^2 + 2\cos(\phi)\,|S(f)|\,|H(f)|\,|N(f)| \quad (11)$$

in which ϕ is the phase difference between speech and noise. While the expected value of $\cos(\phi)$ is zero, its variance can not be neglected. In the remaining of this paper, $\cos(\phi)$ will be denoted by α. The distribution of α can be generalised over the noise types and is in this paper estimated on training data using the cepstral domain relationship :

$$\alpha = \frac{\exp\left(C^{-1}\,x_t\right) - \exp\left(C^{-1}\left(s_t + h\right)\right) - \exp\left(C^{-1}\,n_t\right)}{2\,\exp\left(C^{-1}\left(\frac{s_t+h+n_t}{2}\right)\right)} \quad (12)$$

It is well modelled by a zero mean gaussian with frequency-dependent variance $\sigma^2_{\alpha_i}$, as is confirmed in [8] :

$$p[\alpha] = N(\alpha; 0, \Sigma^\alpha) \quad (13)$$

$$\Sigma^\alpha = \mathrm{diag}\left([\sigma^2_{\alpha_1} \ldots \sigma^2_{\alpha_D}]\right) \quad (14)$$

with D the number of mel-frequency bins. In the context of MBFE, we can then write the phase-sensitive cepstral domain environment model $\tilde{f}(s_t, n_t, h, \alpha)$ as :

$$x_t = C\,\log\left(\exp\left(C^{-1}\left(s_t + h\right)\right) + \exp\left(C^{-1}\,n_t\right) + 2\,\alpha\,\exp\left(C^{-1}\left(\frac{s_t + h + n_t}{2}\right)\right)\right) \quad (15)$$

in which C denotes the DCT matrix as before. The linearisation of $\tilde{f}(s_t, n_t, h, \alpha)$ instead of $f(s_t, n_t, h)$ gives rise to a correction term for the combined HMM covariance matrices. In this case :

$$\tilde{\Sigma}^x_{(i,j)} \approx F_{(i,j)}\,\Sigma_i^s\,F'_{(i,j)} + G_{(i,j)}\,\Sigma_j^n\,G'_{(i,j)} + A_{(i,j)}\,\Sigma^\alpha\,A'_{(i,j)} \quad (16)$$

with $A_{(i,j)}$ the first derivative of $\tilde{f}(s_t, n_t, h, \alpha)$ wrt. α, evaluated in the corresponding VTS expansion point $(\mu_i^s, \mu_j^n, \overline{h}, \mu^\alpha)$:

$$A_{(i,j)} = C\,\mathrm{diag}\left(\frac{2\,exp\left(C^{-1}\left(\frac{\mu_i^s + \overline{h} + \mu_j^n}{2}\right)\right)}{exp\left(C^{-1}\left(\mu_i^s + \overline{h}\right)\right) + exp\left(C^{-1}\,\mu_j^n\right)}\right) \quad (17)$$

The combined HMM means are not affected.

4. MEAN CORRECTION

Also, a correction term for the combined HMM means is introduced. Up to now, the non-linear environment model was linearised by a first order VTS. However, this approximation can be improved if higher order terms are incorporated. Here, the second order derivatives of $\tilde{f}(s_t, n_t, h, \alpha)$ wrt. s and wrt. n, respectively, are used to better approximate the combined HMM means. The state-conditional probability density functions of the noisy speech

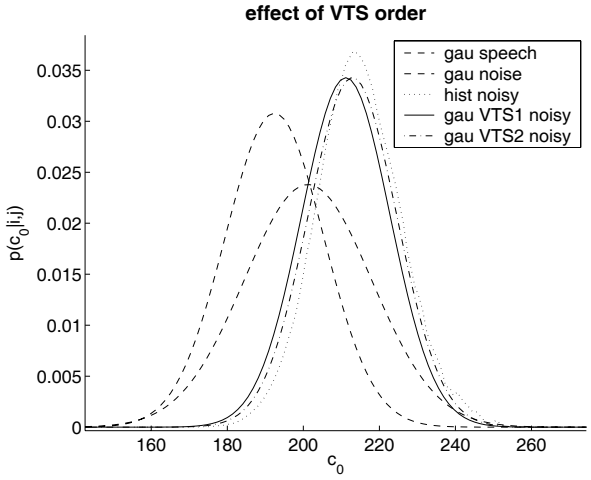

Fig. 1. *Effect of the VTS order on the combined HMM mean (for cepstral coefficient c_0). Gaussian of λ^s, gaussian of λ^n, corresponding histogram of noisy speech samples, gaussian of λ^x with first order VTS, gaussian of λ^x with second order VTS.*

are still assumed gaussian. In this case, the improved means are given by :

$$\tilde{\mu}^x_{(i,j)} \approx \tilde{f}\left(\mu^s_i, \mu^n_j, \overline{h}, \mu^\alpha\right) + \frac{1}{2} H_{(i,j)} \left(\Sigma^s_i + \Sigma^n_j\right) \tag{18}$$

with

$$H_{(i,j)} = \left.\frac{\partial^2 \tilde{f}}{\partial s_t^2}\right|_{\left(\mu^s_i, \mu^n_j, \overline{h}, \mu^\alpha\right)} = \left.\frac{\partial^2 \tilde{f}}{\partial n_t^2}\right|_{\left(\mu^s_i, \mu^n_j, \overline{h}, \mu^\alpha\right)} \tag{19}$$

$$= C \left(\frac{exp\left(C^{-1}\left(\mu^n_j - \mu^s_i - \overline{h}\right)\right)}{\left(1 + exp\left(C^{-1}\left(\mu^n_j - \mu^s_i - \overline{h}\right)\right)\right)^2} \right) C^{-2} \tag{20}$$

Figure 1 illustrates the effect of this mean correction for cepstral coefficient c_0. The histogram of noisy speech samples is obtained by combining samples taken from the gaussian of the clean speech and the noise model, respectively, using equation 15 (with $\alpha = 0$). Clearly, the mean which is calculated with a second order VTS resembles more closely the mean of the histogram than the mean of the first order VTS does. Hence, it can be expected that the higher order VTS yields a better performance. Also, the increase in computational load is limited, since several of the quantities in (20) are already known. In section 5, the effect of these new formulae (16) and (18) will be illustrated, based on speech recognition experiments with the modified MS-MBFE front-end.

5. EXPERIMENTS

Experiments are 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 14 are obtained by adding these same noise types to recordings made with 18 different microphones. Because the channel distortion is not our main interest, the latter test sets are not included in our experiments. For each of the first 7 test sets, all 330 utterances (with an SNR-level that ranges from 5 dB to 15 dB) are evaluated. No compression or end pointing is performed.

5.1. Signal processing

First, the mel-cepstral features are extracted from the speech signal as explained in [9]. Then, K state-conditional estimates, together with the global MMSE-estimate of the clean speech are calculated by the baseline or the improved MS-MBFE front-end. The number of feature streams ($K+1$) is optimised separately for the reference system and for each of the improved systems. Finally, the first and second order time derivatives are added to each of the streams and the MIDA-algorithm is applied to reduce the features to 39 dimensions.

5.2. Front-end models

Experimental evidence was found for the benefit of applying a channel correction on the clean speech training data, before clustering them in a clean speech model. To this end, an MBFE preprocessing with an initial (uncorrected) speech model, is applied. Because the data does hardly contain additive noise, the convolutional noise removal will dominate in this preprocessing step to compensate for differences in loudness between the speakers, etc.

The speech HMM λ^s consists of 256 states with single-Gaussian pdfs and diagonal covariance matrices in the mel-cepstral domain. The noise HMM λ^n consists of 1 single-Gaussian state, whose noise statistics are obtained from the first 30 and the last 30 frames of each noisy speech sentence. In our experiments, this noise model is kept fixed for the entire utterance, although it could be adapted with the available (noise only) frames.

5.3. Back-end recogniser

The speaker-independent LVCSR-system that has been developed by the ESAT speech group of the K.U.Leuven, is used as a back-end recogniser (details can be found in [9]). Note that the multiple front-end feature streams do not require a change of the back-end acoustic model. Instead, each of the streams is evaluated and for each time instant the best matching one (in terms of maximum observation probability) is kept on state level.

5.4. Results

In this section, experimental evidence is given for each of the proposed improvements of our baseline MBFE system separately, namely the correction term for the combined HMM covariance matrices (section 3), the correction term for the combined HMM means (section 4), and the channel corrected clean speech front-end model (section 5.2). Also, the recognition results are shown for the combined system that includes all three modifications.

In table 1, the Word Error Rates are shown for the Advanced Front-End (AFE) ETSI standard without compression [10], together with the Single-Stream (using only the global MMSE clean speech estimate, $K = 0$) and the Multi-Stream MBFE front-end. As can be seen from the results, the performance improvement of SS-MBFE is small for each of the modifications. Introducing the phase-sensitive environment model yields the largest improvement, with an absolute average WER decrease of 0.2% for

Aurora4, 16 kHz; Clean condition training.										
			Test Set							
		K	01	02	03	04	05	06	07	Avg
AFE		0	5.44	17.88	23.07	27.93	26.86	22.90	24.72	21.26
SS-MBFE	Baseline	0	5.10	8.11	18.81	27.05	21.69	20.44	22.64	**17.69**
	Phase-sensitive	0	5.08	7.86	18.98	26.90	21.97	19.82	22.25	17.55
	2^{nd} order VTS	0	5.14	8.22	19.09	26.49	22.29	19.95	22.72	17.70
	h-corrected λ^s	0	5.14	8.28	18.79	27.07	21.99	19.99	22.32	17.65
	Combined	0	5.19	8.26	18.77	26.42	21.93	19.76	22.08	**17.49**
MS-MBFE	Baseline	6	4.91	7.53	18.16	25.56	20.74	17.47	21.76	**16.59**
	Phase-sensitive	10	4.86	7.58	18.16	24.70	19.84	17.00	21.73	16.27
	2^{nd} order VTS	10	4.93	7.88	17.90	24.62	20.27	17.32	21.48	16.34
	h-corrected λ^s	8	5.04	7.64	18.05	23.97	19.97	17.28	21.52	16.21
	Combined	14	4.95	7.77	17.58	23.78	20.14	16.53	21.24	**16.00**

Table 1. *Word Error Rates with the Advanced Front-End, with Single-Stream MBFE and with Multi-Stream MBFE enhancement: baseline, correction of covariance matrices with phase-sensitive model, mean correction with second order VTS, channel-corrected λ^s and the combined system (with all 3 modifications).*

the combined system. However, it has been shown [6] that making a soft decision on the clean speech estimate in the front-end is beneficial, since more detailed information is available in the back-end. Indeed, a larger gain is obtained when these modifications are applied in the MS-MBFE algorithm. The value of K is a trade-off between excluding correct estimates and incorporating unlikely estimates, which can be different for each of the systems. The optimal value of K is shown in the first column of table 1. Its increase confirms that better estimates are being generated.

As can be seen from this table, each of the proposed modifications contributes to a decrease of the Word Error Rate. Moreover, the combined system achieves a higher robustness (a lower WER) than each of them separately. On the whole, an absolute average WER decrease of 0.59% is obtained, which is equivalent to a relative improvement of 3.5%.

6. CONCLUSIONS

In this paper, we have improved some of the approximations made in MBFE to obtain more accurate versions of the noisy speech statistics of the combined HMM. First, we showed how the phase difference between speech and noise (that is often neglected in the acoustic environment model) gives rise to an additional term in the calculation of the covariance matrices for the noisy speech. Then, we introduced an improvement of the noisy speech means by expanding the Vector Taylor Series up to the second order term. Also, a more accurate front-end speech model was obtained by a preprocessing of the training data. Experimental evidence was given for the superior noise robustness of the modified Multi-Stream MBFE system. On the whole, an average relative decrease of 3.5% on the Aurora4 dictation task (test set 01 to 07) was obtained.

Future work includes the investigation of an online adaptation of the noise model, instead of using a fixed model which is trained offline (or from the first 'silence' frames of the utterance).

7. ACKNOWLEDGEMENT

This work was partly supported by 'Research Fund (Onderzoeksfonds) K.U.Leuven', project no. OT/03/32/TBA.

8. REFERENCES

[1] 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.

[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] 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.

[4] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 965–968.

[5] L. Deng, J. Droppo, and A. Acero, "Log-domain speech feature enhancement using sequential MAP noise estimation and a phase-sensitive model of the acoustic environment," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 1813–1816.

[6] 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, pp. 105–108.

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

[8] J. Droppo, A. Acero, and L. Deng, "A non-linear observation model for removing noise from corrupted speech log mel-spectral energies," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 182–185.

[9] 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, pp. 949–952.

[10] 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)*.