JOINT REMOVAL OF ADDITIVE AND CONVOLUTIONAL NOISE WITH MODEL-BASED FEATURE ENHANCEMENT

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

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

ABSTRACT

In this paper we describe how we successfully extended the Model-Based Feature Enhancement (MBFE)-algorithm to jointly remove additive and convolutional noise from corrupted speech. Although a model of the clean speech can incorporate prior knowledge into the feature enhancement process, this model no longer yields an accurate fit if a different microphone is used. To cure the resulting performance degradation, we merge a new iterative EM-algorithm to estimate the channel, and the MBFE-algorithm to remove non-stationary additive noise. In the latter, the parameters of a *shifted* clean speech HMM and a noise HMM are first combined by a Vector Taylor Series approximation and then the state-conditional MMSE-estimates of the clean speech are calculated. Recognition experiments confirmed the superior performance on the Aurora4 recognition task. An average relative reduction in WER of 12% and 2.8% on the clean and multi condition training respectively, was obtained compared to the Advanced Front-End standard.

1. INTRODUCTION

To cure the performance degradation of automatic speech recognition systems in the presence of both additive noise and channel variations, several compensation techniques are often combined (e.g. Spectral Subtraction and CMS, Wiener filtering and blind equalisation [1]). However, results indicate that a joint estimation of both types of noise is feasible [2, 3]. In this paper we focus on a technique that simultaneously removes additive and convolutional noise from the acoustic feature sequence prior to recognition.

Previously we have implemented an MBFE-algorithm for noise robust speech recognition [4], the ideas of which were first introduced by Ephraim [5] in the context of speech enhancement. In this technique we use 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. To this end, the non-linear model of the acoustic environment in the cepstral domain is approximated by a first order Vector Taylor Series. Subsequently, the resulting 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.

In this work a valuable extension to the MBFE-algorithm is proposed, that enables us to simultaneously remove additive background noise and convolutional channel distortions. Especially for MBFE, such a joint noise removal is superior to successively removing these 2 types of mismatch effect between training and testing conditions. The reason is that the MBFE-speech model, trained with one microphone, no longer yields an accurate fit if a different microphone is used. Moreover, this model mismatch also affects the additive noise parameter estimation. Hence, we are convinced that incorporating the effect of convolutional distortions into our speech model will further improve the accuracy. To this end, we propose an iterative EM-algorithm that updates an initial channel estimate to maximise the likelihood of the observed data. Once initialised, our algorithm proves to generate a stable channel estimate, even when only silence frames are observed (and hence no speech or channel information are present in the observed data).

Section 2 presents a detailed description of the extended MBFE-algorithm to simultaneously remove additive and convolutional noise. An evaluation of the performance of the resulting preprocessing technique on the Aurora4 large vocabulary dictation task and the obtained recognition accuracy, can be found in section 3. Finally, conclusions and directions for future work are discussed in section 4.

2. CONVOLUTIONAL MBFE

2.1. Effect of convolutional noise

The parametric model of the acoustic environment, used in this work, is very similar to [6], and is shown in figure 1. Since the enhancement takes place in the cepstral domain, the approximate relationship between the distorted speech vector x_t, the additive noise n_t, the channel h and the clean speech s_t of frame t, is given

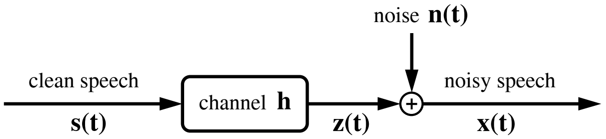

Fig. 1. *Model of the acoustic environment.*

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

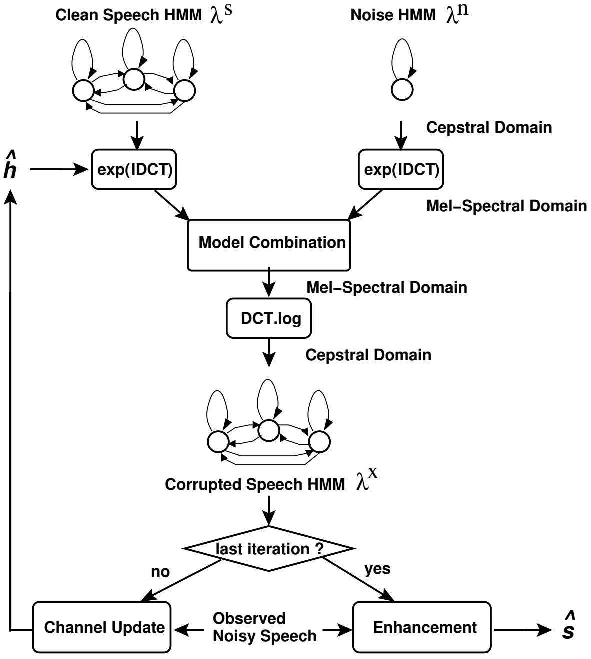

Fig. 2. *HMM combination principle with iterative channel estimation.*

by:

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

in which C^{-1} denotes the inverse of the DCT-matrix C. In MBFE both s_t, n_t as well as x_t are modeled by HMMs in the cepstral domain with Gaussian observation pdfs for each state q, as explained in [4]:

$$
\begin{aligned}
p\left[s_t | q_t^s = i\right] &= N(s_t; \mu_i^s, \Sigma_i^s) && (2)\\
p\left[n_t | q_t^n = j\right] &= N(n_t; \mu_j^n, \Sigma_j^n) && (3)\\
p\left[x_t | q_t^s = i, q_t^n = j\right] &= N(x_t; \mu_{(i,j)}^x, \Sigma_{(i,j)}^x) && (4)
\end{aligned}
$$

Here, the parameters of the corrupted speech HMM λ^x are obtained by using a first order Vector Taylor Series (VTS) approximation around the means μ_i^s and μ_j^n to linearise eq. (1) for each state. From eq. (1) it is clear that the channel h causes a shift of the clean speech model means μ_i^s, such that this MBFE-model will no longer yield an accurate fit if a different microphone is used. However, once this channel is known, we can simply shift our speech model and apply MBFE as before to remove the non-stationary additive noise. A global scheme of this procedure is depicted in figure 2. Briefly, the global MMSE-estimate of the clean speech is obtained as a linear combination of the state-conditional MMSE-estimates, where the weights are given by the a posteriori probabilities. To account for correlation effects, we take the square root (experimentally tuned) of the observation probabilities in the forward-backward algorithm, in which also an approximation of the dynamic parameters of x_t is used. To obtain the latter, the gradients F and G of eq. (1) to s_t and n_t, are assumed to remain constant across the time-interval on which the deltas are calculated, such that the mean and the covariance of the deltas become:

$$
\begin{aligned}
\mu_{(i,j)}^{\Delta x} &\approx F_{(i,j)}\, \mu_i^{\Delta s} + G_{(i,j)}\, \mu_j^{\Delta n} && (5)\\
\Sigma_{(i,j)}^{\Delta x} &\approx F_{(i,j)}\, \Sigma_i^{\Delta s}\, F_{(i,j)}^{'} + G_{(i,j)}\, \Sigma_j^{\Delta n}\, G_{(i,j)}^{'} && (6)
\end{aligned}
$$

and similarly for the delta-deltas. The next section describes the iterative EM-algorithm to estimate the channel online.

2.2. Channel estimation

The auxiliary function that is optimised by the iterative EM-algorithm, is given by [7]:

$$
Q(h'|h) = \sum_{(i,j)} p(i, j | x, h, \lambda^x)\, log\left(p(x, i, j | h, \lambda^x)\right) \qquad (7)
$$

or alternatively, the following function needs to be maximised:

$$
\sum_{(i,j)} \gamma_t^{(i,j)}\, log\left(p(x | i, j, h, \lambda^x)\right) \qquad (8)
$$

Here i and j are hidden variables that denote the speech and the noise state sequence respectively, x is the observed noisy speech, and $\gamma_t^{(i,j)}$ are the a posteriori probabilities as mentioned before. Because of the Gaussian form of $p(x|i, j, h, \lambda^x)$ (eq. (4)), this problem can be rewritten as a maximisation of

$$
\sum_{(i,j)} \sum_t -\frac{1}{2}\gamma_t^{(i,j)} \left(x_t - \mu_{(i,j)}^x\right)^{'} \left(\Sigma_{(i,j)}^x\right)^{-1} \left(x_t - \mu_{(i,j)}^x\right) \qquad (9)
$$

Then the following approximations are applied. Firstly, since F and G are non-linear functions of h, we neglect the change of $\Sigma_{(i,j)}^x$ with different values of h. Secondly, we linearise the dependency of $\mu_{(i,j)}^x$ around $\overline{h}$, such that we can write

$$
\begin{aligned}
\mu_{(i,j)}^x &\approx C \log\left(\exp\left(C^{-1}(\mu_i^s + \overline{h})\right) + \exp\left(C^{-1}\mu_j^n\right)\right) \\
&\quad + F_{(i,j)}\, \delta h && (10)\\
&\approx \overline{\mu_{(i,j)}^x} + F_{(i,j)}\, \delta h && (11)\\
\Sigma_{(i,j)}^x &\approx F_{(i,j)}\, \Sigma_i^s\, F_{(i,j)}^{'} + G_{(i,j)}\, \Sigma_j^n\, G_{(i,j)}^{'} && (12)
\end{aligned}
$$

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

$$
\begin{aligned}
F_{(i,j)} &= C\, \text{diag}\left(\frac{1}{1 + exp\left[C^{-1}(\mu_j^n - \mu_i^s - \overline{h})\right]}\right) C^{-1} && (13)\\
G_{(i,j)} &= I - F_{(i,j)} && (14)
\end{aligned}
$$

and I denotes the unity matrix. This way, $\frac{\delta Q}{\delta h} = 0$ yields

$$
0 = \sum_{(i,j)} \sum_t \gamma_t^{(i,j)}\, F_{(i,j)}^{'} \left(\Sigma_{(i,j)}^x\right)^{-1} \left(x_t - \overline{\mu_{(i,j)}^x} - F_{(i,j)}\, \delta h\right) \qquad (15)
$$

and hence the channel update is given by:

$$
\begin{aligned}
\delta h = &\left[\sum_{(i,j)} F_{(i,j)}^{'} \left(\Sigma_{(i,j)}^x\right)^{-1} F_{(i,j)} \sum_t \gamma_t^{(i,j)}\right]^{-1} \\
&\cdot \left[\sum_{(i,j)} F_{(i,j)}^{'} \left(\Sigma_{(i,j)}^x\right)^{-1} \sum_t \gamma_t^{(i,j)} \left(x_t - \overline{\mu_{(i,j)}^x}\right)\right] \qquad (16)
\end{aligned}
$$

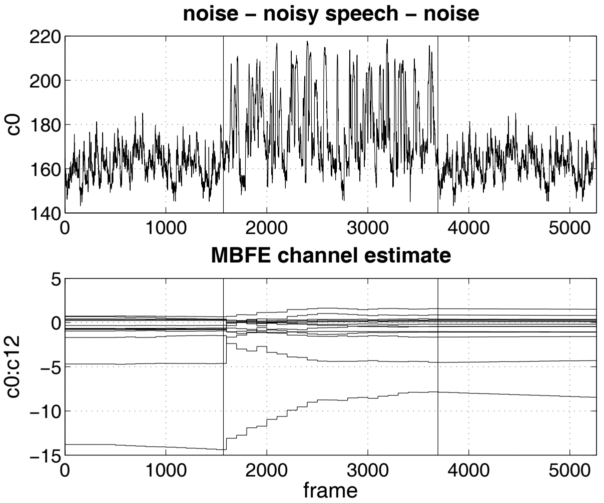

Fig. 3. *Effect of noise frames on MBFE channel estimate. (top): Noisy speech utterances embedded in noise-only frames. (bottom): MBFE channel estimate in cepstral domain.*

To come to a real-time implementation, this channel update formula is implemented with an exponential weighting of numerator and denominator. One of the advantages of eq. (16), compared to CMS for instance, is that it does not need a voice activity detector. Indeed, in case the input data contains only additive noise (no speech and no channel) the gradient $F_{(i,j)}$ approximates zero, such that the update δh automatically becomes very small. Although eq. (16) exhibits a strong similarity with the one in [8] to remove convolutional filtering effects, the latter does not have this advantage. Experiments confirmed that a good recognition accuracy can be obtained with zero as an initial channel estimate, then applying 10 EM-iterations on the first 500 frames, while for every next 100 frames exponential weighting and only 5 EM-iterations are applied. The stability of our channel estimate in the presence of noise, can clearly be seen in figure 3. This plot shows an utterance (between the vertical bars) from the Aurora2 database (setC, N1, SNR15), that is embedded in noise-only frames from the same noise condition. On the left of the vertical bars, we see that $\hat{h}$ is not updated after initialisation, while in the middle part the channel estimate converges to a stable value during the noisy speech frames. Finally, on the right of the bars, this obtained value is again hardly changed during the noise-only input frames. We conclude that the influence of the noise frames on the obtained channel estimate is very small.

3. EXPERIMENTS

3.1. Front-end processing

Experiments were conducted on the Aurora4 large vocabulary database, derived from the WSJ0 Wall Street Journal 5k-word dictation task. For each of the 2x7 test sets (no noise, car, babble, restaurant, street, airport, train), all 330 utterances, with an SNR-level that ranges from 5 dB to 15 dB, are evaluated.

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. Then a Hamming window and a mel-scaled triangular filter-bank are applied, and the resulting mel spectrum with 24 coefficients is transformed into the cepstral domain. From this static parameter set, the global MMSE-estimate of the clean speech is then calculated by the extended MBFE-algorithm. Afterwards, each sentence is smoothed by the low-pass filter $H(z) = 1/\left(2 - z^{-1}\right)^2$. As in [4], the first and last mel spectra are then removed and the remaining 22 spectral coefficients are mean normalised. Finally, the 66 features that result from adding the first and second order time derivatives, is reduced by the MIDA-algorithm to 39 dimensions, which are then decorrelated.

3.2. 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 because of its fast experiment turn-around time and good baseline accuracy.

The gender independent acoustic modeling is based on a set of 45 phones, without specific function words. A phonetic decision tree, developed for the clean and multi condition training data respectively, defines the 4961 tied states in the cross-word context-dependent (but position-independent) models. In the first training step, 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 further reduced to an average of 200 per state, using the occupancy criterion, and the third training step is applied.

A bigram language model for a 5k-word closed vocabulary is provided by Lincoln Laboratory, while decoding is done with a time-synchronous beam search algorithm.

3.3. MBFE front-end models

The design of the MBFE front-end noise model and clean speech model is now described. In our experiments the noise model consists of a one-state single-Gaussian HMM. This HMM is obtained by estimating the mean on the first 30 and the last 30 frames of each sentence. As before [4], the variance of this Gaussian is estimated from the same 60 frames, but is pooled over all 330 sentences of the noise type, thereby simulating the scenario in which some of the noise model parameters can be estimated offline. The choice of this noise HMM topology is motivated by previous results, which indicated little performance loss as compared to more complex noise models, while offering a tractable computational load.

This noise model is combined with a very simple speech model, namely instead of a phoneme HMM we use an 128 Gaussian ergodic HMM. The latter is obtained by EM-clustering the clean training dataset provided in the Aurora4 database. Although an ergodic HMM incorporates less prior knowledge on the allowed state sequence during decoding, incorrect decisions can easier be corrected by the more detailed acoustic models in the back-end recogniser. Experimental results showed that the recognition accuracy is hardly affected with this simpler speech model, even though in this model less Gaussians are used. It was also verified that clustering the clean training dataset to more Gaussians (256 or 512) yielded no significant accuracy gain. This implies that not only

Aurora4, 16 kHz sampling, no compression, no end pointing.																	
	Close Talk								Far Talk								
TEST	1	2	3	4	5	6	7	mic 1	8	9	10	11	12	13	14	mic 2	Avg.
clean																	
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
MBFE	4.88	8.67	20.85	28.99	23.95	22.10	23.15	18.94	21.91	30.30	39.42	44.55	42.93	39.62	42.24	37.28	28.11
conv MBFE	4.93	8.43	20.51	28.81	22.77	20.68	23.30	18.49	19.24	26.98	37.40	42.91	40.87	38.09	40.11	35.09	26.79
multi																	
MIDA	7.77	9.04	15.86	19.54	17.88	14.65	19.93	14.95	15.43	18.53	30.17	31.44	33.03	28.34	34.17	27.30	21.13
AFE	6.84	10.84	15.28	19.19	18.05	14.46	17.62	14.61	16.27	22.03	29.42	31.31	32.43	28.41	30.97	27.26	20.94
MBFE	7.15	8.35	15.73	21.02	17.84	15.82	17.99	14.84	17.39	21.11	30.75	33.53	33.83	30.64	33.96	28.74	21.79
conv MBFE	7.25	8.03	15.75	20.21	16.91	15.65	17.04	14.41	13.21	17.58	28.99	32.22	31.23	28.96	32.06	26.32	20.36

Table 1. *Word error rates without enhancement, with Advanced Front-End preprocessing, with MBFE-enhancement and with convolutional MBFE-enhancement; clean and multi condition training.*

the computational load, which is proportional to the total number of Gaussians, is decreased, but also the forward-backward algorithm becomes trivial and the training of the speech model is less complex.

3.4. Experimental results

The first reference results (labeled MIDA in table 1) are obtained by leaving out the MBFE-enhancement and the smoothing from the processing steps described in 3.1. For these features, no explicit (additive) noise reduction algorithm is applied. Secondly, features are preprocessed by the standard AFE, without compression [9]. However, to increase the level of comparability, the dynamic coefficients are not calculated by the reference scripts. Also, no frames are dropped by a feature vector selection. After this enhancement, again the first and last mel-spectra are removed, and the MIDA-algorithm and a decorrelation are applied. Finally, we show the recognition results when MBFE is applied, without and with channel estimation.

In table 1, mic1 and mic2 denote the average of the first 7 and the last 7 noise conditions, respectively. The results indicate that the WER decreases significantly when the channel estimation process is integrated in the MBFE-algorithm. This confirms the better match of the MBFE-speech model when the effect of a different microphone is estimated and accounted for in the front-end processing. For the clean training condition, a relative reduction in WER of 29% compared to no enhancement, and 12% compared to the AFE is obtained. Finally, when the back-end acoustic models are trained on noisy, preprocessed data (multi condition training), MBFE outperforms the AFE with a 2.8% relative WER-reduction.

4. CONCLUSIONS

In this paper we have shown how the MBFE-algorithm successfully can be extended to jointly remove additive and convolutional noise. The presented EM-algorithm yields a stable estimate of the channel that is hardly affected by the non-speech frames. Experimental results showed the superior performance of MBFE compared to the AFE for the Aurora4 large vocabulary recognition task, which confirms the better match between the clean speech estimates and the back-end acoustic models.

Nevertheless, further optimisations in the context of MBFE can still be considered, such as finding the optimal method to train the noise model. Currently, the noise HMM is fixed for each noise condition (i.e. SNR-level and noise type). However, with an online adaptation of the noise model, frames that are highly likely to contain only noise could be used to adapt the mean of the noise HMM to the varying environment.

5. REFERENCES

[1] 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.

[2] T. Kristjansson, B. Frey, and L. Deng, “Joint estimation of noise and channel distortion in a generalized EM framework,” in *Proc. ASRU*, Madonna di Campiglio, Italy, Dec. 2001.

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

[4] V. Stouten, H. Van hamme, J. Duchateau, and P. Wambacq, “Evaluation of model-based feature enhancement on the AURORA-4 task,” in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 349–352.

[5] Y. Ephraim, “A minimum mean square error approach for speech enhancement,” in *Proc. ICASSP*, New Mexico, USA, Apr. 1990, pp. 829–832.

[6] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, “HMM adaptation using Vector Taylor Series for noisy speech recognition,” in *Proc. ICSLP*, Beijing, Oct. 2000, pp. 869–872.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM-algorithm,” *J. of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[8] A. Sankar and C.-H. Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” *IEEE Trans. on SAP*, vol. 4, no. 3, pp. 190–202, May 1996.

[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).*