ON SEPARATING ENVIRONMENTAL AND SPEAKER ADAPTATION

*Carlos Lima, Carlos Silva, Adriano Tavares and Jorge Oliveira**

Department of Industrial Electronics of University of Minho, Portugal
{carlos.lima, adriano.tavares, carlos.silva}@dei.uminho.pt

*Department of Electrical Engineering, Polytechnic Institute of Leiria, Portugal
oliveira@estg.iplei.pt

ABSTRACT

This paper presents a maximum likelihood (ML) approach, concerned to the background model estimation, in noisy acoustic non-stationary environments. The external noise source is characterised by a time constant convolutional and a time varying additive components. The HMM composition technique, provides a mechanism for integrating parametric models of acoustic background with the signal model, so that noise compensation is tightly coupled with the background model estimation. However, the existing continuous adaptation algorithms usually do not take advantage of this approach, being essentially based on the MLLR algorithm. Consequently, a model for environmental mismatch is not available and, even under constrained conditions a significant number of model parameters have to be updated. From a theoretical point of view only the noise model parameters need to be updated, being the clean speech ones unchanged by the environment. So, it can be advantageous to have a model for environmental mismatch. Additionally separating the additive and convolutional components means a separation between the environmental mismatch and speaker mismatch when the channel does not change for long periods. This approach was followed in the development of the algorithm proposed in this paper.

One drawback sometimes attributed to the continuous adaptation approach is that recognition failures originate poor background estimates. This paper also proposes a MAP-like method to deal with this situation.

1. INTRODUCTION

As speech recognition of broadcast news has received a great deal of attention, the adaptation requirement of the existing recognition systems to non-stationary noisy conditions increases. In general, the existing recognition systems are not adequate to deal with non-stationary conditions due, for example, to the presence of music in the background [1]. Actual telephone data have shown that the convolutional distortions observed on the telephone line are almost constant for a given call and vary between calls [2]. Therefore, one can conclude that the telephone channel can be roughly characterised by a non-stationary additive noise and a stationary convolutional noise, for the same call.

The mismatch between the training and testing conditions of an automatic speech recogniser can be efficiently reduced by adapting the parameters of the recogniser to the testing conditions. Recently, a family of online or incremental adaptation algorithms for continuous density hidden Markov models (HMM) based speech recognisers have appeared that are based on constrained re-estimation of the distribution parameters [3][4]. These algorithms can be used in unsupervised adaptation mode and can adapt to new conditions automatically, based on the recogniser's hypothesis. However, while these algorithms are designed to operate in non-stationary environmental conditions, they are based on the assumption that the environmental mismatch can be modelled by an affine transform on the means and variances of the clean speech distributions, even when the background is modelled by an HMM [5]. This assumption seems not to make much sense since it is considered that the clean speech is not frequently changed by the environment, instead is the environment that changes continuously. Additionally, non-stationary environments can be more accurately modelled by an HMM. This means that the noisy speech HMMs are an expanded version in the number of both states and mixture components, of the clean speech HMMs. Hence, it would make much more sense to have a model of environmental mismatch, where environmental adaptation could be done by updating the environmental model parameters keeping up the clean speech distribution model parameters unchanged. Consequently fewer parameters have to be updated since only the environmental model needs to be adapted, while a broad range of environmental conditions from stationary to non-stationary can be handled. This approach was followed in the development of the algorithm proposed in this paper.

2. SPEECH AND NOISE JOINT MODELLING

This section introduces the integrated model of signal and background assuming convolutional and additive distortions. The telephone channel can be roughly characterised by a non-stationary additive noise and a

stationary convolutional noise for the same call, whereas in the spectral domain the speech is degraded by a multiplicative time constant vector and by an additive time dependent stochastic process. A stochastic model compatible with the non-stationary property of the noise process is the Hidden Markov Model (HMM). HMM composition approach allows recognising concurrent signals simultaneously, assuming that both are HMM modelled and combined by an *a priori* known function. Pairs of states of clean speech and noise models form each state of the noisy speech model.

Assuming that the observed stochastic process $\mathbf{x}_t$ (clean speech) has independent and identically distributed random variables (Gaussian), the auxiliary function is given by [6]

$$Q(\lambda_X, \lambda'_X) = \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\left\{\log a_{s_{t-1}s_t} + \log c_{s_t,c_t} - \sum_{i=1}^{D}\left[\frac{1}{2}\log\ \sigma^{2}_{n,m,i} + \frac{(x_{t,i} - \mu_{n,m,i})^2}{2\sigma^{2}_{n,m,i}}\right]\right\} \quad (1)$$

where $\gamma_t(n,m)$ is the joint probability of the observation vector $\mathbf{x}_t$, the state n and the mixture component m, D is the observed vector dimensionality, a and c refer respectively to the transition state probability and mixture coefficient. If the clean speech is corrupted by a stationary channel, whose frequency response in the Power Spectrum Density is given by a time constant vector (**w**), and by an additive process then the noisy speech is given by $\mathbf{z}_t = \mathbf{w}\mathbf{x}_t + \mathbf{y}_t$ (2)

where the product of vectors represents an element by element product.

If the noise process $\mathbf{y}_t$ is also modelled by an HMM, assuming the noise is statistically independent of the signal, the probability density function for the noisy speech can be obtained from the composition of the two Markov models.

For non-stationary environments with stationarity comparable to the one of the speech, noise models with more than one state are required. In our experiment the additive noise model has only one state.

Future developments of the algorithm will hold noises with variability similar to speech. For simplicity we have only used Gaussian noise, therefore the additive noise model has only one component in the mixture. The extension for more complex noise models with more states and more mixtures is straightforward. Under these circumstances the distribution parameters of the noise becomes only dependent on the variable *i* (component of the observation vector), which means that the mean and noise variance are state and mixture independent. Therefore, the auxiliary function (1) regarding to the model that integrates the speech, the convolutional and additive distortions becomes

$$Q(\lambda_Z, \lambda'_Z) = -\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\sum_{i=1}^{D}\left\{\frac{1}{2}\log\left(w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)\right) + \frac{\left(z_{t,i} - \left(w_i\mu_x(n,m,i) + \mu_y(i)\right)\right)^2}{2\left(w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)\right)}\right\} \quad (3)$$

3. BACKGROUND MODELS ESTIMATION

Maximising the Q function (equation (3)) in order to the model parameters does not result in a closed form expression for some of these parameters. This happens in the estimation of the channel vector and it could be remedied using the called alternate Q function [7]. This approach has been taken to derive general expressions for the parameters estimates of the original speech model given noisy observations in [7], where both the speech and distortions were modelled by a mixture of Gaussians. The problem addressed in this section is the reverse problem of finding the parameters of the distortions model given the distorted speech.

Channel Estimation: Adapting the alternate Q function to our case of additive and convolutional distortions by assuming that all the components of the vector **w** are not null, a solution for each component of **w** exists and is given by

$$\frac{dQ(\lambda_X, \lambda'_X)}{dw_i} = -\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\left[w_i^2 + \frac{\mu_x(n,m,i)}{\sigma_x^2(n,m,i)}E\{x_t / z_t, n, m, \lambda_X\}w_i - \frac{E\{x_t^2 / z_t, n, m, \lambda_X\}}{\sigma_x^2(n,m,i)}\right] \quad (4)$$

Equation (4) is only valid if the channel distortion is state and mixture independent, with all frequency components not null. Otherwise, the maximisation of the alternate Q function does not result in a closed form expression for the channel estimate. Since the channel frequency response is by nature positive because it is the square of the modulus of the frequency response, then we are only interested in the positive root of equation (4). It can be easily shown [9] that this equation always has a positive and a negative real root.

Additive Noise Model Estimation: Given the constant vector channel distortion and the distorted speech, the additive noise model parameters can be estimated by maximising the Q function in order to these parameters, as usual. However, maximising function (3) in order to $\lambda' = (\mu_y, \sigma_y^2)$ does not result in a closed-form expression for the noise variance. In an experiment where the goal was to compensate for the channel distortions in the cepstral domain given a relatively small amount of adaptation (distorted) speech, Sankar [6] used the alternate Q function and the derivations of Rose [7] to get the re-estimation formulas. However, Sankar noted that if the noise variance is small, then the convergence of the EM algorithm is slow. In the limit, when the noise variance is null the estimate will not change at all. This was found to be the case in the Sankar's experiment, where the variance in the mismatch due to the different transducers and transmission channels was small. Even in our case, where the additive noise has relatively bigger variances, but still smaller than the speech variance, for higher SNR, the convergence of the EM algorithm becomes very slow. However, in our case, due to the on-line parameters estimation, the speed of convergence is more critical than the Sankar's one because it retards the recognition.

Sankar remedied this situation by using equation (5) to estimate the noise mean. Equation (5) is derived by maximising equation (3) in order to the noise mean, once that the noise variance does not have a closed form solution. Equation (5) is derived in [6] and is given by

$$\mu_y(i) = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\dfrac{z_{t,i} - w_i\mu_x(n,m,i)}{w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\dfrac{\gamma_t(n,m)}{w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)}} \quad (5)$$

However, this procedure only solves the slowness of the convergence on the mean, but leaving the convergence of the variance slow. This can be verified examining the equations derived by Sankar, and is confirmed by our experimentation.

Equation (6) although not being an exact solution for the maximisation of equation (3) in order to the noise variances, which is only a reasonable approximation for high SNR, has shown very useful especially relative to the above described limitations of the Sankar's procedure. Equation (6) can be derived similarly to the derivation of equation (5) assuming however, that the speech variance is much larger than the noise variance [9].

$$\sigma_y^2(i) = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\left[\dfrac{\left(z_{t,i} - w_i\mu_x(n,m,i) - \mu_y(i)\right)^2 - w_i^2\sigma_x^2(n,m,i)}{\left(w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)\right)^2}\right]}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\dfrac{\gamma_t(n,m)}{\left(w_i^2\sigma_x^2(n,m,i) + \sigma_y^2(i)\right)^2}} \quad (6)$$

4. WEIGHT UPGRADE OF THE BACKGROUND MODELS

The update of the additive noise model (equations (5) and (6)) and the channel frequency response (equation (4)) can be really effective in supervised mode. However, most interesting applications require unsupervised adaptation, where the algorithm can adapt to new conditions automatically, based on the recogniser's hypothesis. In the continuous speech recognition paradigm, if the hypothesis is incorrect, however, the benefit may come from the fact that only a part of the hypothesis is incorrect. Hence, convergence of the environmental model could be very dependent on the initial word error rate. However, this could not always be the case in continuous speech recognition applications and will never be the case in isolated word recognition applications. Therefore, a procedure for environmental update that takes into consideration the recogniser's certainty, which is proportional to the score differences among the various hypothesis, can constitute a better solution than assuming always correct hypothesis. This MAP-like environmental estimation based on the N-best hypothesis when applied to the isolated word recognition case involves simultaneous calculations in more than one HMM. For the case where one predict, for example, that the correct hypothesis is in the three more probable ones, the function to be maximised is $P(Z/\lambda_1 \vee \lambda_2 \vee \lambda_3)$ instead of $P(Z/\lambda)$ where λ is the set of parameters of the model that represents the recognised class on noisy speech. Applying the Bayes rule to the function to be maximised one obtains

$$P(Z/\lambda_1 \cup \lambda_2 \cup \lambda_3) = \frac{P(\lambda_1)P(Z/\lambda_1) + P(\lambda_2)P(Z/\lambda_2) + P(\lambda_3)P(Z/\lambda_3)}{P(\lambda_1) + P(\lambda_2) + P(\lambda_3)} \quad (7)$$

It can be easily shown that solution of equation (7) in the context of Markov models, in regards to the channel is as for equation (4), the positive root of a second grade equation. However, in this case the calculations are simultaneously made on the three most probable models [9].

Regarding to the additive noise model the solution of equation (7) is also very similar to equations (5) and (6), assuming also the noise variance is much smaller than the speech one. Otherwise a closed form solution to the noise variance does not exist. The solutions are also made by simultaneously performing calculations in the three most probable Markov models [9].

5. EXPERIMENTAL RESULTS

The proposed algorithm was tested in an Isolated Word Recognition system where continuous speech recognition was simulated by recognising continuously isolated words. The used parameterisation is obtained by grouping 16 contiguous spectral components (Power Spectral Density components) forming 16 bands for a 512 points FFT. This kind of features is unusual nowadays since cepstrum based features are really more effective on speech modelling than spectral based features. Future developments of the algorithm will be based on cepstral features but under the same approach, which foresees changes in the structure of the clean speech models in the noisy speech modelling.

The used speech was acquired under controlled environmental conditions band-pass filtered from 100 to 3200 Hz, sampled at a 6.67 kHz and analysed in segments of 45 ms duration at a frame rate of 66.67 windows/sec.

The recognising speech was computationally contaminated with a constant multiplicative distortion in the frequency domain (convolutional noise) and a time varying additive distortion. The multiplicative distortion and a rough approximation used to evaluate the recovery of the algorithm from poor initial channel estimates are shown in figure 1. Additive distortions were generated for an SNR of 10 and 5 dB considering the power of the first recognising digit. The recognising digits were contaminated alternately in such away that 20 contiguous digits have the same noise level, and contiguous groups of 20 digits were alternatively contaminated by the two levels of generated noise. This type of non-stationarity was chosen because it hinders the recogniser task by allowing abrupt variations in the noise level. Additionally noise level variations occurs jointly with the change of the speaker, which constitute more one difficulty for the recogniser given that the channel adaptation means also speaker adaptation.

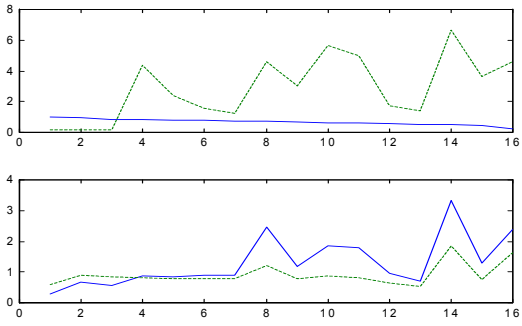

Figure 1. Upper: true (line) and initial channel estimate. Lower: first (line) and second channel estimates. The initial noise estimate was 10^5 times the true value for mean and variance.

Wrong initial noise estimates were 10^5 times greater and 10^5 times smaller than the true values. No differences in performance were obtained for this two initial noise estimates. Table 1 shows the recognition results obtained for some of the combinations between the additive and convolutional noises.

Table 1. Recognition performance adapting on-line the Environmental parameters

Case	Failures in 400
Training on noisy speech (SNR=10dB)	21
True initial channel, (1)	153
True init. channel, on-line noise adaptatic	32
Wrong init. estimates, full adapt. on S. M	21
Wrong init. estimates, full adapt. on U. M	36
Using weight update suggested in sect. 4	25
True init. models and weight update	23

1) Constant noise power, case of an SNR of 10 dB. S. M. and U. M. means respectively adaptation in supervised and unsupervised mode.

The results show that the recogniser can adapts automatically to varying environmental conditions even when the current environmental estimates are poor. In supervised adaptation mode and using very poor environmental estimates the recogniser reaches the theoretical best possible performance, just that obtained by training on stationary noisy speech (fourth entry of table 1). Adapting both the additive noise and the channel is superior to adapt the additive noise and using the true channel distortion. This occurrence can be related with the speaker characteristics, which are also roughly modelled by a multiplicative distortion in the PSD domain.

Finally, the proposed weight update of the environmental model improves the performance of the algorithm by 3.75% for very poor initial estimates, however, as expected degrades the performance (0.5 %) relatively to the supervised adaptation case.

6. DISCUSSION

A continuous adaptation algorithm where a distinct model for the environment exists has been presented. This approach seems to make more sense than its MLLR counterpart in the following aspects:

1) Given the non-stationary nature of the most distortions found in practical applications it makes some sense model this non-stationarity by an HMM. Hence, the structure of the clean speech model has to be changed (increasing the number of states) in order to accommodate non-stationary environmental distortions.

2) Since the algorithm can adapt automatically to new environmental conditions, only the environmental model needs to be updated, which reduces the number of parameters to be estimated and improves the accuracy of the estimates in on-line applications.

3) When isolated noise samples are available, typically collected in the beginning of the speech segment, the recogniser can be instantaneously adapted decreasing the initial word error rate, which is important when using unsupervised adaptation.

4) Distinguishing between speaker mismatch and environment mismatch could be useful to provide a speaker adapted system that was independent of the acoustic environment.

REFERENCES

[1] Raj, B., Parikh, V. N. and Stern, Richard M. (1997). The effects of background music on speech recognition accuracy. In Proc. ICASSP'97 pages 851-854.

[2] Mokbel, C., Monné, J. and Jouvet, D. (1993), On line adaptation of a speech recognizer to variations in telephone line conditions. IN EUROSPEECH, pages 1247–1250.

[3] Leggetter, C. J., and Woodland, P. C. (1995). Flexible speaker adaptation using maximum-likelihood linear regression. In Proc. ARPA Workshop on Spoken Language Technology, Jan. 1995, pp. 110-115.

[4] Woodland, P. C., Gales, M. J. F. and Pye, D. (1996). Improving environmental robustness in large vocabulary speech recognition. In Proc. ICASSP'96 Vol. I pp. 65-68.

[5] Wang, Z.-H. and Kenny, (1998). Speech Recognition In Non-Stationary Adverse Environments. In Proc. ICASSP'98 pages -.

[6] Sankar, A., and Lee, C.-H. (1996). A maximum likelihood approach to stochastic matching for robust speech recognition. IEEE Transactions on Speech and Audio Processing, vol. 4, no. 3.

[7] Rose, R., Hofstetter, E., and Reynolds, D. (1994). Integrated models of speech and background with application to speaker identification in noise. IEEE Trans. Speech Audio Processing, vol. 2, pages 245-257.

[8] Gales, M. J. F. and Young, S. J. (1993). PMC for speech recognition in additive and convolutional noise. Technical Report 154.

[9] Lima, Carlos, Almeida, Luis B. and Monteiro, J. L. (2001). On-line adaptation of Speech Recognisers to Non-stationary environments. Technical Report 001.