Separating speaker and environment variabilities for improved recognition in non-stationary conditions

Luca Rigazio, Patrick Nguyen, David Kryze, Jean-Claude Junqua

Panasonic Speech Technology Laboratory
3888 State Street, Santa Barbara, CA 93105
{rigazio, nguyen, kryze, jcj}@research.panasonic.com

Abstract

In this paper we address the problem of speaker adaptation in noisy environments. We estimate speaker adapted models from noisy data by combining unsupervised speaker adaptation with noise compensation. We aim at using the resulting speaker adapted models in environments that differ from the adaptation environment, without a significant loss in performance. The key idea is to separate speaker and environment variabilities and associate them to independent models. We show that linear models for both speaker and environment are critical for achieving this goal. Experiments for 2000 and 4000 isolated word tasks on real car noise show that unsupervised speaker adaptation combined with noise compensation can provide more than 20% error rate reduction compared with noise compensation only, and more than 50% error rate reduction compared with speaker adaptation only.

1. Background

In recent years important improvements were obtained with the adaptation of acoustic models to new test conditions. General techniques were successfully applied for the adaptation to new speakers [1] and noise compensation was successfully used to deal with background noise [2, 3]. However, few have attempted to model speaker dependent effects in noisy environments [4]. Our approach deals with speaker variability and environment variability separately. The main advantage is that speaker adapted acoustic models may be estimated in one acoustic environment and used in others. General adaptation methods, such as MLLR, may be used to compensate for environment mismatch; however, they ignore information about the environment model. Noise compensation algorithms rely on that model to provide very fast and accurate adaptation. Our method is based on a first order approximation to compensate for the environment [5] and on a global adaptation scheme to adapt to the speaker [1, 4]. We show that a linear model of the environment is crucial to achieve the separation of speaker and environment variabilities.

2. Speaker adaptation

Speaker adaptation has to deal with the problem of estimating reliable models from small amounts of data. A variety of speaker adaptive algorithms have been proposed. Here we consider only MLLR [1] and MAP|MLLR [4]. In the MAP|MLLR scheme MLLR is applied first:

$$\mu_{MLLR} = \left[\arg\max_{W} p\left(O|W\mu_0\right)\right]\mu_0;$$

then a MAP smoothing is applied to relax the constraints imposed by the linear regression:

$$\mu_{MAP|MLLR} = \arg\max_{\mu} p(O|\mu)p_0\left(\mu|\mu_{MLLR}\right).$$

In the previous equations μ_0 is the speaker independent mean, W is the regression matrix, $p(O|\mu)$ is the likelihood and $p_0\left(\mu|\mu_{MLLR}\right)$ is the likelihood conjugate prior centered around μ. For the MLLR step we used a single regression class. Throughout the experiments, MLLR adaptation provided results close to MAP|MLLR, but consistently worse. For this reason we report only MAP|MLLR results. Notice both MLLR and MAP|MLLR adaptations are linear operators $\hat{\mu} = A\{O, \mu\}$.

3. Noise compensation

Let X be a spectral vector, and let $C(X) = F\log(X)$ be the cepstral operator, where F is the DCT matrix and $\log(X)$ is intended to be component-wise. Noise compensation aims at modifying the acoustic model parameters to match to a certain acoustic environment. Compensation of additive noise in the cepstral domain is a non linear operation. As suggested in [2], noise compensation of model first order statistics can be carried out according to $C(S+N) = C(C^{-1}(C(S)) + N)$, where $C(S)$ is the clean speech cepstrum (or equivalently the gaussian means), $C(S+N)$ is the estimate of the speech cepstrum subject to the estimated noise N. This compensation scheme in its general form is called PMC, and under certain assumptions can be proven to be optimal in a maximum likelihood sense. However, PMC can be computationally expensive and complex to integrate with speaker adaptation. The PMC operator can be approximated with a first order Taylor expansion as:

$$C(S+N_1) \approx C(S+N_0) + J(S,N_0)\Delta C(N), \quad (1)$$

$$J(S,N_0) = \left.\frac{\partial C(S+N)}{\partial C(N)}\right|_{N=N_0} = F\frac{N_0}{S+N_0}F^T, \quad (2)$$

where N_0, N_1 are the training and test background noises, $\Delta C(N) = C(N_1) - C(N_0)$, and $\frac{N_0}{S+N_0}$ is intended to be a diagonal matrix [3]. Notice that the noise at training time has to be non zero to guarantee the Jacobian matrix $J(S,N_0)$ to be full rank. If $J(S,N_0)$ is singular the compensation may fail since some component (in the log spectral domain) may not be affected by the adaptation, regardless the amount of test noise. When acoustic models are trained in clean environments, the norm of N_0 can become very small, making $J(S,N_0)$ almost singular. For this reason a noise overestimation factor should be used to compute $J(S,N_0)$ for clean acoustic models [5]. From now on we will not distinguish cepstrum and dynamic cepstrum since linear approximations exists for the compensation of dynamic cepstral coefficients [3].

4. Joint speaker adaptation and noise compensation

Our target is to estimate speaker adapted models from noisy data. We are concerned only with first order statistics. We want to separate speaker and environment effects, to improve their estimation and to allow the use of speaker adapted models in environments that differ from the adaptation environment. Let S_I, S_D be the speaker independent and the speaker dependent speech spectra, and let $\{O\}$ be the observed sequence cepstral vectors uttered by one speaker (S_D) in one test environment (N_1). By taking the expectation of the observations, and by using equation (1) we have:

$$E\{O\} = C(S_D + N_1),$$

$$C(S_D+N_1) \approx C(S_D+N_0) + J(S_D,N_0)\Delta C(N).$$

From the linearity of the expectation operator $E\{\cdot\}$ we have:

$$C(S_D+N_0) \approx E\left\{O - J(S_D,N_0)\Delta C(N)\right\}. \quad (3)$$

This means we can compute speaker dependent models for the training environment N_0 by taking the expectation of the modified observations $O' = O - J(S_D,N_0)\Delta C(N)$. Notice that the result holds for any linear generalized expectation operator, including the adaptation operator $A\{O,\mu\}$. Unfortunately equation (3) does not directly solve (in general) since S_D is needed to compute $J(S_D,N_0)$ and vice versa. Consider a series of estimates of the speaker dependent spectra S_D^t for which we assume that the associated Jacobian matrices are slowly changing, i.e. $J(S_D^{t+1},N_0) \approx J(S_D^t,N_0)$. From equation (3) we have:

$$\begin{aligned} C(S_D^{t+1}+N_0) &\approx E\left\{O - J(S_D^{t+1},N_0)\Delta C(N)\right\} \\ &\approx E\left\{O - J(S_D^{t},N_0)\Delta C(N)\right\} \quad (4) \end{aligned}$$

From equations (2) and (4) we can now provide an iterative solution to equation (3):

$$\begin{aligned} \hat{\mu}^{t+1} &= E\left\{O - \hat{J}^t\Delta C(N)\right\} \\ \hat{J}^t &= F\frac{N_0}{C^{-1}(\hat{\mu^t})}F^T \end{aligned}$$

The initialization should be based on the best available estimate of $J(S_D,N_0)$. In practice that means the matrix associated to the models obtained from the last adaptation increment, or the speaker independent matrix for the first adaptation increment ($S_D^0 = S_I$). The algorithm should converge in few iterations, as long the assumptions on $J(S_D^{t+1},N_0)$ are respected. Also, if we make the stronger assumption that $J(S_D^t,N_0)$ is constant in t, it follows that $J(S_D^t,N_0) = J(S_I,N_0)$, that the Jacobian matrices do not have to be recomputed and that equation (3) can be solved directly. This assumption would drastically reduce the computational complexity of the algorithm since recomputing Jacobian matrices is very expensive (it requires exponentiations, divisions and matrix multiplications). In section 5 we will assess the practicality of these assumptions in term of recognition results. Since we are interested in adapting gaussian means of Hidden Markov Models, the expectation operator has to be computed from incomplete data by integrating over the hidden states q via Expectation Maximization [6]. In practice we make the approximation of considering only the best path (Viterbi), and of using for each frame the Jacobian matrix associated to the winning gaussian to compute O'. The alignment is computed using the last acoustic models ($\hat{\mu^t}$) compensated for the test noise with the last Jacobian matrices ($\hat{J}^t$).

5. Experimental results

To test the proposed method we used a database of isolated words recorded at PSTL. The training database is recorded at 11kHz and consist of 156 American speakers recorded in a clean environment, each uttering 150 isolated words. 130 speakers are used as training set, and 26 speakers as development set, for a total of about 10 hours of speech. The test database consists of 14 speakers recorded in a car driving at 30MPH and 60MPH, each uttering 150 words per session, for a total of about 2 hours of speech. The average signal to noise ratios are about 12dB for the 30MPH session and 7dB for 60MPH session. The vocabulary size is 2100 words per session, i.e.

speakers utter different words in each session. The feature extraction is based on the Subband front-end analysis [7]. Twelve cepstral coefficients are computed from an unbalanced decomposition tree with an average frame rate of 86Hz, and an average window size of 17 milliseconds. First order derivatives are then computed on a regression window of five frames. Tree clustered context dependent HMMs were trained on this database, with a total of 924 states and 5872 gaussians. Noise compensation was performed based on the noise estimated during the first 25 frames of the sentence and using modified Jacobian matrices with a noise overestimation factor $\alpha = 2.5$. The compensation was carried out for both static and dynamic coefficients. Cepstral mean adaptation was optionally applied to compensate for channel mismatch. The baseline recognition results for the clean development set (DEV), the 30MPH and 60MPH noisy test sets are reported in table 1. Results without noise compensation are very low because of the large mismatch between training and test environment, and results for Jacobian are close to CMA because the channel is mostly stationary (the database is recorded using the same microphone).

	NONE	JAC	CMA
DEV	6.2	6.1	5.9
30MPH	87.5	12.8	12.0
60MPH	95.8	18.0	17.5

Table 1: Baseline word error rates without noise compensation (NONE), with Jacobian (JAC) and with Jacobian plus cepstral mean adaptation (CMA).

5.1. Results for stationary environments

In this section we test the proposed method for joint speaker adaptation and noise compensation, and we compare it with speaker adaptation on clean and noisy data. With stationary environments we refer to data collected at a fixed car speed: the car noise itself is quite stationary and the amount of noise is also stationary within a recognition session. This setting may facilitate speaker position estimation, especially for the speaker adaptation algorithm, because the perturbations of the noise are stationary and may be averaged out over long periods of time. Table 2 shows recognition results for the proposed method, MAP$|$MLLR$|$JAC, and for the MAP$|$MLLR speaker adaptation. We used unsupervised incremental speaker adaptation, with increment steps of 10 sentences, and a single iteration for the estimation of the speaker adapted Jacobian matrices $\hat{J}^t$ (more iterations did not provide significant improvements). Results show that MAP$|$MLLR$|$JAC improves significantly compared with MAP$|$MLLR in noisy conditions (an average of 55% relative error rate reduction for the 30MPH and 60MPH), and degrades only marginally on the clean development set. Notice that no customizations of speaker adaptation or of noise compensation algorithms were done. Comparing the results of Table 1 with those of Table 2 we notice that MAP$|$MLLR$|$JAC provided an average error rate reduction of 21% compared with CMA and 24% compared with JAC, confirming that the proposed method learns better than the two individual building blocks (speaker adaptation and noise compensation).

	MAP$\|$MLLR	MAP$\|$MLLR$\|$JAC
DEV	4.8	5.1
30MPH	22.6	9.4
60MPH	29.2	13.9

Table 2: Word error rates for speaker adaptation and for joint speaker adaptation and noise compensation in stationary environments.

The iterative algorithm relies on the assumption that the Jacobian matrices do not change significantly during incremental adaptation. The good results obtained here seem to confirm that this assumption is valid. It is of interest to verify the viability of the stronger form of this assumption, i.e. that Jacobian matrices can be considered constant and can be approximated with the initial matrices of speaker independent models, $J(S_I, N_0)$. Table 3 shows results obtained with the MAP$|$MLLR$|$JAC algorithm based on the stronger assumption that Jacobian matrices are unaffected by the speaker adaptation. Although we can notice a performance degradation, the algorithm can still deliver a large improvement over MAP$|$MLLR and JAC or CMA. This discovery may seem counter intuitive, however it is very important for reducing the complexity of the algorithm. Further investigations are needed to understand why this approximation holds so well in real conditions.

	$\hat{J}^t = J(S_I, N_0)$
DEV	4.9
30MPH	9.8
60MPH	14.5

Table 3: Word error rates for joint speaker adaptation and noise compensation for stationary environments, without the update of the Jacobian matrices.

5.2. Results for non stationary environments

As underlined before, recognition experiments on homogeneous sessions are somewhat of a simplification of realistic environments. In real applications, the amount of noise may vary largely from sentence to sentence. By constraining the data to belong to one session, we help the algorithm to learn the combined speaker and environment effects. Intuitively this may introduce a database bias in favor of MAP|MLLR, since for this data separating the effects is not really crucial. To deal with this problem we merged the 30MPH and the 60MPH data, by interleaving sentences. The interleaving lengths were chosen to be a powers of two, $I = 2^k$ with $k = 0 \ldots 6$. We also have increased the lexicon size to 4200 words, since words pronounced during the two sessions are different. This makes absolute recognition rates difficult to compare with previous results. Table 4 shows recognition results averaged across interleaving lengths. Notice that MAP|MLLR|JAC delivers 52% relative error rate reduction compared with MAP|MLLR, 19% compared with CMA, and 21% compared with JAC.

JAC	CMA	MAP|MLLR	MAP|MLLR|JAC
20.7	20.2	34.2	16.3

Table 4: Average word error rates for simulated non stationary environments.

Figure 1 shows recognition results given the interleaving length in a logarithmic scale. The interleaving length can be interpreted as a factor of non-stationarity for the simulated environment (a small k induces a less stationary environment). Obviously non-incremental methods like JAC or CMA are not affected by k, however incremental methods in principle may be influenced. Notice that MAP|MLLR is very sensitive to k and that word error rates increase significantly with k. We believe that for large k the speaker adaptation (that is modeling both speaker and environment) overfits to the stronger environment effects and loses speaker adaptive power. This undesirable behavior is not shown by MAP|MLLR|JAC that delivers a performance almost independent from k. This confirms that speaker and environment effects have been correctly separated, and that this separation resulted in a more robust system and in enhanced capability to estimate persistent speaker dependent effects.

6. Conclusion

Speaker and environment variability are related to different sources. Even if the environment variability can be compensated by general adaptation methods, model-based algorithms are faster and more accurate. Speaker adaptive systems may profit from the separation of speaker and environment variabilities if a model of the environment is used. We introduced a joint environment and speaker adaptation algorithm based on first order approximations. For 2000 and 4000 isolated word recognition tasks on real car noise we obtained between 20% and 50% error rate reduction compared to standard speaker adaptation and noise compensation methods.

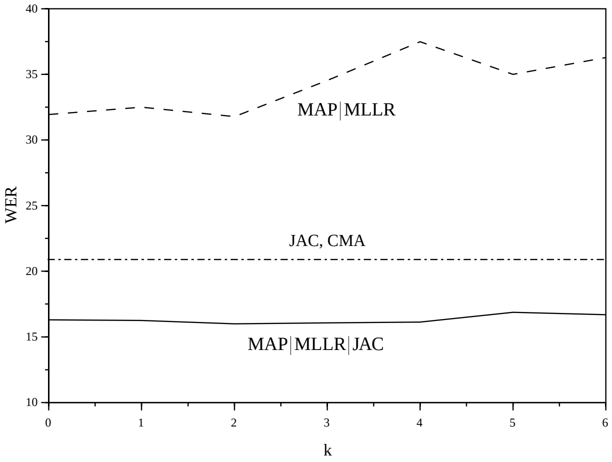

Figure 1: Word error rates for speaker adaptation and for joint speaker adaptation and noise compensation with different interleaving lengths.

7. References

[1] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[2] M. Gales, “Predictive model-based compensation schemes for robust speech recognition,” *Speech Communication*, vol. 25, pp. 49–74, 1998.

[3] S. Sagayama, Y. Yamaguchi, and S. Takahashi, “Jacobian Adaptation of Noisy Speech Models,” in *Proc. of ASRU*, Santa Barbara, CA, Dec. 1997, pp. 396–403.

[4] P. Nguyen and C. Wellekens, “Maximum likelihood Eigenspace and MLLR for speech recognition in noisy environments,” in *Proc. of Eurospeech*, Sep. 1999, vol. 6, pp. 2519–2522.

[5] C. Cerisara, L. Rigazio, R. Boman, and J.-C. Junqua, “Transformation of Jacobian matrices for noisy speech recognition,” in *Proc. of ICSLP*, Beijing, China, Oct. 2000, pp. 1369–1373.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum-Likelihood from Incomplete Data via the EM algorithm,” *Journal of the Royal Statistical Society B*, pp. 1–38, 1977.

[7] D. Kryze, L. Rigazio, T. Applebaum, and J.-C. Junqua, “A New Noise-robust Subband Front-end And Its Comparison To PLP,” in *Proc. of ASRU*, Keystone, CO, Dec. 1999.