Noise-Compensated Hidden Markov Models

Ivandro Sanches, *Member, IEEE*

***Abstract*—The technique of hidden Markov models has been established as one of the most successful methods applied to the problem of speech recognition. However, its performance is considerably degraded when the speech signal is contaminated by noise. This work presents a technique which improves the performance of hidden Markov models when these models are used in different noise conditions during the speech recognition process. The input speech signal enters unchanged to the recognition process, while the models used by the recognition system are compensated according to the affecting noise characteristics, power and spectral shape. Hence, the compensation stage is independent of the recognition stage, allowing the models to be continually adjusted.**

The models used in this work are from a continuous density hidden Markov algorithm, having cepstral coefficients derived from linear predictive analysis as state parameters. It is used only static features in the models in order to show that, when properly compensated for the noise, these static features contribute significantly to improve noisy speech recognition. It is observed from the results that the parameters kept their capability to discriminate among different classes of signals, indicating that, in the context of speech recognition, the use of autoregressive-derived parameters with noisy signals does not represent an impediment. A matrix-way of converting from autoregressive coefficients to normalized autocorrelation coefficients is presented.

The affecting noise is assumed additive and statistically independent of the speech signal. Although the noise dealt with should also be stationary, good performance was achieved for nonstationary noise, such as operations room noise and factory environment noise. The concept of intra-word signal-to-noise ratio is presented and successfully applied. The resulting compensated models revealed to be less dependent on the training data set when compared to the trained hidden Markov models. Due to the computational simplicity, the time required to adjust a model is significantly shorter than the time to train it.

***Index Terms*—Background noise, hidden Markov models, intra-word signal-to-noise ratio, speech recognition.**

I. Introduction

SOLUTIONS to the problem of speech recognition in noise may fall in two categories, both of them aim to reduce the mismatch between training and recognition situations. One category focuses on the restoration of the clean speech signal from the noisy one (speech enhancement), while the other category can cope with the presence of noise in the recognition process, which is the approach of this work.

Manuscript received February 14, 1998; revised November 2, 1999. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazilian Government Founding Council. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Picone.

The author is with the Departamento de Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, CEP 05508-900, São Paulo, S.P., Brazil (e-mail: isanches@usp.br).

Publisher Item Identifier S 1063-6676(00)06981-9.

There is not a clear boundary between these two categories, which can have their methods combined to improve recognition performance. Some successful methods of speech enhancement include: *spectral subtraction* (SS) suggested by Boll [1], [2]; *signal restoration by spectral mapping* introduced by Juang and Rabiner [3]; adaptive filtering techniques as *Kalman filtering* [4] and *all-pole modeling of degraded speech*, proposed by Lim and Oppenheim [5], that conveniently combines Wiener filtering with the linear prediction technique; and *multiple microphone arrays* which aims to improve the signal-to-noise ratio at the input of the recognition system [6].

Along the boundary of the mentioned categories we have *cepstral mean subtraction* (CMS) [7], [8] which is simple and effective to reduce distortions introduced by microphones and transmission channels. Also, robust metrics and features have shown to be very useful and important at improving recognition rates of noisy speech. The *projection-based likelihood measure* [9], [10] uses the fact that additive white noise causes the cepstral vector norm to diminish but leaving its orientation practically intact. Features that exploit the time varying properties of speech spectra (dynamic features) [11] brought significant improvement to the problem as well as features that exploit the human auditory perception model, like RASTA-PLP [12] that improves speaker independency, reduces the influence of background noise, attenuates the influence of the acoustic channel variations and deals with noisy Lombard speech [13].

This work differs from the mentioned methods in the sense that it depends on the specific noise spectral characteristics (intensity and spectral shape) affecting the recognition procedure. In this way, it is similar to other kinds of model compensation strategies like: *noise-adaptive prototypes* [14] which trains a mapping between noisy and clean features; *state-based filtering* [15] where Wiener filters, designed for each state of a hidden Markov model, are used to filter the sequence of noisy speech observation vectors; *model decomposition* [16] which is a generalization of conventional hidden Markov modeling that provides an optimal method of decomposing simultaneous processes (noisy speech into clean speech and noise signals); and *parallel model combination* (PMC) that is similar to our work in the sense that the speech models are modified to be more representative of the speech in the new acoustic environment given an estimate of the additive noise. In the *data-driven parallel model combination* (DPMC) [17] the speech models are used to generate separate samples of speech and noise which are then combined appropriately to obtain the noise corrupted speech samples which are used to estimate the compensated models.

Our proposed approach to the problem is strongly founded on basic principles, as it will be seen, resulting in a very effective method, allowing a straight forward implementation and short execution time. If a label should be given to this kind of

1063–6676/00$10.00 © 2000 IEEE

approach it could be *spectral addition*, as opposed to the basic principles ruling *spectral subtraction*.

The next section presents the proposed method, followed by the section showing the results achieved employing the NOISEX-92 database [18]. The work is finished with some conclusions.

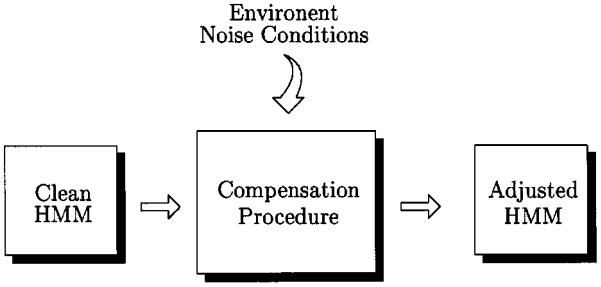

Fig. 1. Creation of an adjusted model.

II. Compensating the Models

In the hidden Markov model (HMM) technique, as in all the other techniques of automatic speech recognition, there are two phases involved in the recognition process, a first phase where the speech recognition system is trained, and a second phase of pattern matching, which is the recognition itself. When training the system, every set of input signals must have good representatives of the word to be modeled, having the least amount of extra noise. Models created (trained) with speech signals having the least amount of noise will be referred in this work as *clean models*. In the recognition phase an input word is compared with each of the trained models and deemed to be the one corresponding to the model which matches most closely. However, in practice the incoming signal in the recognition procedure will contain spurious components (noise), not present in the training phase. As the amount of noise increases, the clean models lose the property of characterizing the input signals accurately. One solution to this problem would be to train all the models of the vocabulary with the corrupted signals. The resulting models of such training procedure will be referred as *noisy models*, that is, models trained by the signals corrupted by the noise affecting the current recognition. Such solution could be applied if the computational cost of the training procedure was not so expensive, since a new training session would be required whenever the noise spectral properties change. A more practical solution would be the creation of a *compensated*, or *adjusted model*, from the clean model, which has the probabilistic information of the noise-free speech conveniently combined with information of the affecting noise, as illustrated in Fig. 1. The great advantage now is the generation of a new model, without the need of an expensive training session. It is expected that an adjusted model will perform better in the degraded situation than a clean model. Thus, the performance of a clean model is a lower limit to the acceptability of an adjusted model, which has as its target the performance of the corresponding noisy model.

The basis of the method is the fact that the autocorrelation function of the signal resulting from the addition of two statistically independent signals is equal to the sum of their individual autocorrelation functions. Therefore, in adjusting a clean model, its state spectral representation is transformed from the autoregressive, or cepstral, domain to the autocorrelation domain. Then, the autocorrelation of the clean model is added to a sample of the autocorrelation of the affecting noise, resulting in the autocorrelation of the noisy signal, which is transformed back to the original spectral representation. At the end of this process, an adjusted model results with better capabilities of handling the noisy signal. Due to the uncorrelation assumption and the nature of the experiments, the Lombard effect[1] can not be addressed.

It was observed that the noise influence affects mainly the spectral characterization of a model, leaving its temporal properties effectively unaltered. In other words, it is the emission probabilities of the states of the model that we are interested in compensate for the noise, while the transition probabilities do not need to be altered.

As we are employing a continuous density HMM, the emission probabilities of the states of a model are characterized by multivariate Normal distributions, consequently, the model compensation method results in the adjust of the mean vector and covariance matrix of each Normal distribution. Next we show the compensation method for multivariate Normal distributions using cepstral coefficients derived from the linear predictive analysis as coefficients.

A. Adjusting the Means

Let r_i, a_i, and c_i, $i = 1, \cdots, p$, be the normalized autocorrelation coefficients ($r_0 = 1$), the linear prediction coefficients (lpc), and the cepstral coefficients, respectively. Let $\boldsymbol{c}_s = [c_1\ c_2\ \cdots\ c_p]^T$ be the mean vector of the multivariate Normal distribution. The corresponding vector of linear predictive coefficients can be obtained from [7]

$$\begin{aligned} a_1 &= -c_1, \\ a_k &= -c_k - \sum_{j=1}^{k-1} \left(1 - \frac{j}{k}\right) a_j c_{k-j}, \qquad 2 \le k \le p. \end{aligned} \tag{1}$$

Once the linear prediction coefficients, $\boldsymbol{a}_s$, are determined, the corresponding normalized autocorrelation coefficients, $\boldsymbol{r}_s$, can be computed by

$$\boldsymbol{A}\boldsymbol{r}_s = -\boldsymbol{a}_s \tag{2}$$

where $\boldsymbol{r}_s = [r_1\ r_2\ \cdots\ r_p]^T$, $\boldsymbol{a}_s = [a_1\ a_2\ \cdots\ a_p]^T$, and we have found that $\boldsymbol{A}$ can be defined as described in Appendix I. Another way of computing $\boldsymbol{r}_s$ from $\boldsymbol{a}_s$ is to use the recursive step-down procedure in conjunction with the Levinson algorithm [19]. The inverse procedure is done by the well known relation (autocorrelation method)

$$\boldsymbol{R}\boldsymbol{a}_s = -\boldsymbol{r}_s \tag{3}$$

[1]The natural speaker response to stress the voice when speaking in a noisy environment.

where $\boldsymbol{R}$ is the symmetric Toeplitz matrix formed by the values $1, r_1, r_2, \cdots, r_{p-1}$.

Assuming zero mean signals, the signal-to-noise ratio can be defined as

$$\text{SNR} = 10 \log \frac{E_s}{E_n} \tag{4}$$

where E_s and E_n are the average energy of the speech signal and noise, respectively. Letting

$$\alpha = \frac{E_n}{E_s} = 10^{-\text{SNR}/10}$$

the normalized autocorrelation coefficients of the noisy signal, $\boldsymbol{r}_{s+n}$, can be approximated by

$$\boldsymbol{r}_{s+n} = \frac{1}{1+\alpha}\left(\boldsymbol{r}_s + \alpha \boldsymbol{r}_n\right)$$

where $\boldsymbol{r}_n$ is the vector with the normalized autocorrelation coefficients of the noise, which can be estimated at some instant preceding a recognition process. Once $\boldsymbol{r}_{s+n}$ is defined, the corresponding $\boldsymbol{a}_{s+n}$ is computed by (3), and $\boldsymbol{c}_{s+n}$ is obtained by the inverse procedure of equation (1), which is explained in [7]. Repeating this procedure to the mean vector of every Normal distribution of all the models we complete the process of compensating the means for the affecting noise. The mean compensation process is illustrated in Fig. 2.

B. Adjusting the Covariances

The idea is to transform the covariance matrix of the model from the cepstral domain to the log-energy domain, where we can combine it to the covariance matrix of the noise in this domain. When this is done, we obtain the noisy covariance matrix in the log-energy domain, which can be transformed back to the cepstral domain as we show now. The transformation from the cepstral domain to the log-energy domain is effected by the application of the discrete cosine transform (DCT) [20]. The DCT of a discrete sequence $x(m)$, $m = 0, 1, \cdots, (M-1)$ is defined by [21]

$$\begin{aligned} G_x(0) &= \frac{\sqrt{2}}{M} \sum_{m=0}^{M-1} x(m), \\ G_x(k) &= \frac{2}{M} \sum_{m=0}^{M-1} x(m) \cos \frac{(2m+1)k\pi}{2M}, \\ & \qquad k = 1, 2, \cdots, (M-1) \end{aligned} \tag{5}$$

where $G_x(k)$ is the kth DCT coefficient. Representing in a matrix form, we have

$$\boldsymbol{G_x} = \boldsymbol{C}\boldsymbol{x}$$

where $\boldsymbol{C}$ is the DCT matrix.

Let $\boldsymbol{c}_s$ and $\boldsymbol{c}_n$ be the means of the first p cepstral coefficients of clean speech (clean model) and noise (estimated), respectively, and $\sigma^2(\boldsymbol{c}_s)$ and $\sigma^2(\boldsymbol{c}_n)$ the corresponding covariance matrices. As mentioned before, the transformation to the log-energy domain is done by

$$\begin{aligned} \boldsymbol{l}_s &= p\boldsymbol{C}\boldsymbol{c}_s \quad \sigma_s^2 = p^2\boldsymbol{C}\sigma^2(\boldsymbol{c}_s)\boldsymbol{C}^T \\ \boldsymbol{l}_n &= p\boldsymbol{C}\boldsymbol{c}_n \quad \sigma_n^2 = p^2\boldsymbol{C}\sigma^2(\boldsymbol{c}_n)\boldsymbol{C}^T \end{aligned}$$

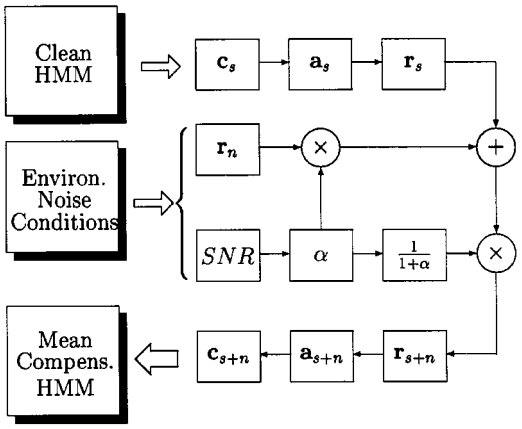

Fig. 2. Diagram of the mean compensation process.

where $\boldsymbol{l}_s$ and $\boldsymbol{l}_n$ are the log-energy components of clean speech and noise, with the respective covariance matrices σ_s^2 and σ_n^2. Now we show how to combine both of these matrices to get σ_{s+n}^2, the covariance matrix of the noisy signal in the log-energy domain. Once σ_{s+n}^2 is determined, the covariance matrix of the noisy signal in the cepstral domain is computed by

$$\sigma^2(\boldsymbol{c}_{s+n}) = p^{-2}\boldsymbol{C}^{-1}\sigma_{s+n}^2\boldsymbol{C}^{-T}. \tag{6}$$

It only remains to show how to determine σ_{s+n}^2. This matrix is defined by

$$\begin{aligned} \sigma_{s+n}^2(i,j) = &\boldsymbol{E}[\log(S_i + N_i)\log(S_j + N_j)] \\ &- \boldsymbol{E}[\log(S_i + N_i)]\boldsymbol{E}[\log(S_j + N_j)]. \end{aligned} \tag{7}$$

Where S_i, N_i, $i = 1, 2, \cdots, p$, represent the components of the energy spectra of clean speech and noise, respectively, satisfying a given SNR. That is,

$$\boldsymbol{l}_s = [\log S_1 \ \log S_2 \ \cdots \ \log S_p]^T,$$

and

$$\boldsymbol{l}_n = [\log N_1 \ \log N_2 \ \cdots \ \log N_p]^T.$$

Based on the observation that the energy in a frequency band is dominated either by signal (clean speech) energy or by noise energy, we model the energy as the larger of the separate energies of signal and noise in the band. Mathematically, at band i, we have that[2]

$$\log(S_i + N_i) \approx \max\{\log S_i, \log N_i\}$$

which is applied to (7) in four distinct spectral conditions:

- $S_i > N_i$, $S_j > N_j$: $\sigma_{s+n}^2(i,j) = \boldsymbol{E}[\log S_i \log S_j] - \boldsymbol{E}[\log S_i]\boldsymbol{E}[\log S_j] = \sigma_s^2(i,j)$;
- $S_i < N_i$, $S_j < N_j$: $\sigma_{s+n}^2(i,j) = \boldsymbol{E}[\log N_i \log N_j] - \boldsymbol{E}[\log N_i]\boldsymbol{E}[\log N_j] = \sigma_n^2(i,j)$;
- $S_i > N_i$, $S_j < N_j$: $\sigma_{s+n}^2(i,j) = \boldsymbol{E}[\log S_i \log N_j] - \boldsymbol{E}[\log S_i]\boldsymbol{E}[\log N_j] = 0$;
- $S_i < N_i$, $S_j > N_j$: $\sigma_{s+n}^2(i,j) = \boldsymbol{E}[\log N_i \log S_j] - \boldsymbol{E}[\log N_i]\boldsymbol{E}[\log S_j] = 0$.

The above can be visually interpreted in Fig. 3.

In Fig. 3, the large square with dotted outline represents the covariance matrix of the noisy signal in the log-energy domain,

[2]This approximation is also used, for instance, in [14] and [16].

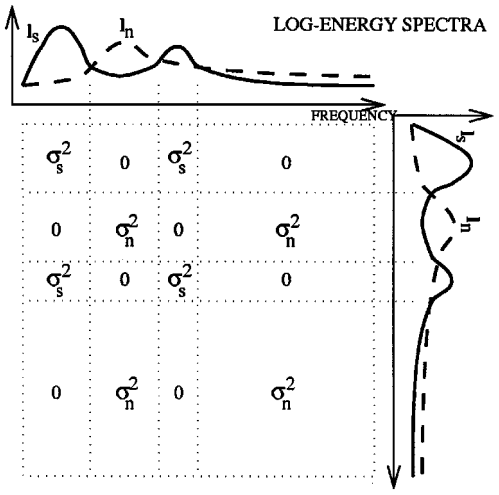

Fig. 3. Composing the noisy covariance matrix σ^2_{s+n}, at a given SNR.

σ^2_{s+n}, and every inner dotted rectangle is the corresponding block taken from σ^2_s or σ^2_n. Rectangles with "0" inside represent null matrices of the corresponding block dimensions. Since σ^2_s and σ^2_n are at least positive semi-definite, the same is also true for the whole matrix.

If we define the Δ-function

$$\Delta_i = \begin{cases} 0, & \text{if } S_i/N_i < 1 \\ 1, & \text{if } S_i/N_i \geq 1 \end{cases} \qquad i = 1, 2, \cdots, p$$

the covariance matrix of the noisy signal in the log-energy domain is computed by

$$\sigma^2_{s+n}(i, j) = \Delta_i \Delta_j \sigma^2_s(i, j) + (1 - \Delta_i)(1 - \Delta_j)\sigma^2_n(i, j). \quad (8)$$

Then, using this last result in (6), we can compute $\sigma^2(\boldsymbol{c}_{s+n})$. If the original cepstral covariance matrices of clean speech, $\sigma^2(\boldsymbol{c}_s)$, and noise, $\sigma^2(\boldsymbol{c}_n)$, were assumed diagonal, the resulting matrix $\sigma^2(\boldsymbol{c}_{s+n})$ has the off-diagonal terms set to zero. Fig. 4 illustrates the process of compensating the covariance matrix.

C. Intra-word SNR

In this section we present the concept of intra-word signal-to-noise ratio, and how it can be used with advantage in the proposed compensation technique.

For a given SNR in a recognition process, we can easily notice that there is a change in SNR along an utterance, for instance, due to the different energy levels of voiced and unvoiced portions along it, the SNR will be higher in the voiced portion than in the unvoiced one. Relying on the capability of the hidden Markov technique to segment the utterance into distinctive portions, and including energy information into the states of a model, an *intra-word* SNR, SNR'_i, can be defined for every state i, $i = 1, 2, \cdots, N$, of a model

$$\text{SNR}'_i = 10 \log \frac{E^i_s}{E_n}$$

where E^i_s, $i = 1, 2, \cdots, N$, is the attached energy value of state i, which is computed in the training procedure of the model. To be practical, this expression should be modified to take account

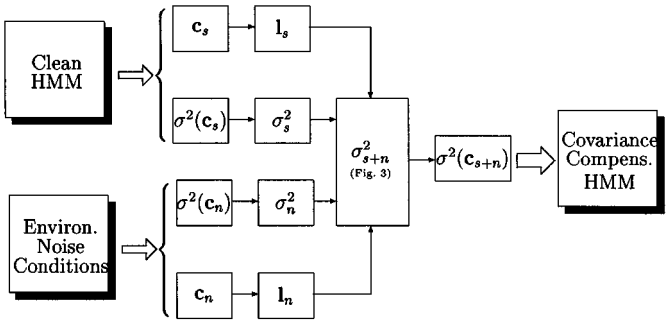

Fig. 4. Diagram of the covariance matrix compensation process.

of the actual speech signal energy in the recognition process. This can be made considering the relationship between E_s and

$$\overline{E_s} = \sum_{i=1}^{N} \frac{E^i_s}{N}$$

which can be regarded as an ensemble average energy of the signal given by the trained model. Supposing fairly stable acoustic conditions, on average, the ratio $E_s/\overline{E_s}$ is one. Thus, the actual *intra-word* SNR, SNR_i, for the state i of a model can be defined as

$$\begin{aligned} \text{SNR}_i &= 10 \log \frac{E_s}{E_n} \frac{E^i_s}{\overline{E_s}} \\ &= 10 \log \frac{E_s}{E_n} + 10 \log \frac{E^i_s}{\overline{E_s}} \\ &= \text{SNR} + \Delta\text{SNR}_i \end{aligned} \quad (9)$$

where the last term of (9), ΔSNR_i, depends only on values defined by the model.

Thus, instead of using (4), the intra-word SNR, SNR_i, given by (9), can be used in the compensating procedures described in Sections II-A and II-B. So, each state of the model will be compensated according to its particular energy level, which is related to the energy level of the corresponding portion of the utterance.

In the next section we present the recognition results when the described procedures are applied.

III. Results

A. Experimental Data

This section shows the results of the application of the technique described. The training of clean and noisy models and the recognition process are done through the use of the speech recognition software HTK [22]. Speech and noise signals are from the NOISEX-92 [18] database, which provides a carefully controlled set of experiments where speech and noise signals have been added together at several values of SNR. It provides a set of control data that can be easily used and for which comparative experimental results become available.

The speech data is partitioned into training and testing data sets. The training data set consists of two sequences of 100 digits, one recorded by a male speaker and one by a female speaker. Every sequence of 100 digits is made up by ten repetitions of each one of the ten digits, from zero to nine. The testing data set is achieved on the same way. In this work we are

employing the data from the male speaker. So, there are 20 repetitions of each digit. Half of the repetitions was used to train the models (training data set) and the other half to test the proposed technique (testing data set).

The noise data was taken from the same database. Every uttered digit corresponds to a defined piece of noise in the noise data set. The speech and noise were added together digitally at five different signal-to-noise ratios: 18, 12, 6, 0, and -6 dB. The 0 dB SNR noisy speech is created simply by adding together the provided clean speech and noise signal. For SNRs above 0 dB, i.e., less noise, 6 dB steps were obtained by successive multiplication of the noise signal samples by 0.5 (hence, 0.5, 0.25, 0.125, for SNR values of respectively 6, 12, and 18 dB); for noisy speech at a SNR of -6 dB the noise signal samples were multiplied by 2.0, before adding them to the clean speech signal. We present the results for operations room and factory noises at these values of SNR.

Both speech and noise were sampled at 16 kHz. The speech data were pre-processed using a window of 25 ms ($L = 400$ samples per window), at every 10 ms, and $p = 20$ lpc-cepstral coefficients computed. The autocorrelation of the noise was computed with the same expression as the autocorrelation of the speech signal in the predictive analysis. The actual normalized autocorrelation coefficients of the noise signal, $\boldsymbol{r}_n$, were achieved from the average of a number, m, of normalized autocorrelation vectors taken randomly from the noise signal. In the experiments, an average of $m = 10$ normalized autocorrelation vectors produced $\boldsymbol{r}_n$.

The models have ten left-to-right states, and every state with a continuous Normal output probability density. The covariance matrices are assumed diagonal. The training process produces, for each digit, the *clean model* (trained with clean speech) and a *noisy model* trained with noisy speech for each SNR under consideration in our experiments in the next section. The clean model is used by the compensation technique to produce a *compensated model* for each SNR, and the noisy model is used to have its performance compared to the performance of the compensated model.

B. Recognition Results

The tables are divided in two parts: *train* and *test*. Under *train*, we have the recognition percentage when the training data set was employed in the recognition process. Under *test*, we employed the testing data set, which is independent from the trained models. As there are a total of 100 digits for training and 100 digits for testing, the number of correctly recognized digits represents the recognition percentage.

Table I shows the recognition percentage results for operations room noise. The terms *noisy*, and *clean* refer to the trained models with noisy and clean speech, respectively. The terms *intra*, and *comp.* refer to the compensated models, where we applied the concept of intra-word SNR to the former. Table II presents the recognition percentage results for factory noise.

In order to examine further the performance of the proposed compensation technique, we observed the behavior of the values of the coefficients of mean vector and covariance matrix under different noise conditions. For instance, Fig. 5 presents the mean values of cepstral coefficient of index 4 for five different vowels,

TABLE I
OPERATIONS ROOM NOISE

	train			
SNR (dB)	noisy	intra	comp.	clean
18	100	100	100	100
12	100	100	100	95
6	100	97	97	77
0	100	89	88	34
-6	100	60	59	12
	test			
SNR (dB)	noisy	intra	comp.	clean
18	100	100	100	99
12	100	100	97	90
6	100	97	96	70
0	94	88	83	22
-6	58	55	52	10

TABLE II
FACTORY NOISE

	train			
SNR (dB)	noisy	intra	comp.	clean
18	100	100	100	100
12	100	100	98	97
6	100	98	95	79
0	100	96	93	51
-6	100	69	68	20
	test			
SNR (dB)	noisy	intra	comp.	clean
18	100	100	100	100
12	99	100	99	92
6	96	97	97	77
0	86	89	86	42
-6	55	64	59	20

namely, /a/ (h**a**t), /e/ (g**e**t), /i/ (sh**e**), /o/ (h**o**t), and /u/ (t**oo**). The straight dash-dotted line represents the mean values for the clean models for each vowel. The dotted line represents the mean values for the noisy models when the SNR varies from 18 to -6 dB in steps of 6 dB, for each of three different kinds of noise: Lynx, F16, and car noises [18]. That is, a total of 75 mean values forms the dotted line for all five vowels. Lastly, the solid line are the mean values for the compensate models. We can notice that the mean values of the compensated models are tracking fairly well the means of the noisy models (assumed as target).

Fig. 6 shows the change in the variance values for the same coefficient index, in the same experimental situation described for Fig. 5. We see that the tracking is not as precise as for the means, but the variance values of the compensated models are slightly bigger than the noisy ones most of the time. This may explain why the compensated models had superior performance to the noisy models in some situations in Table II, when using

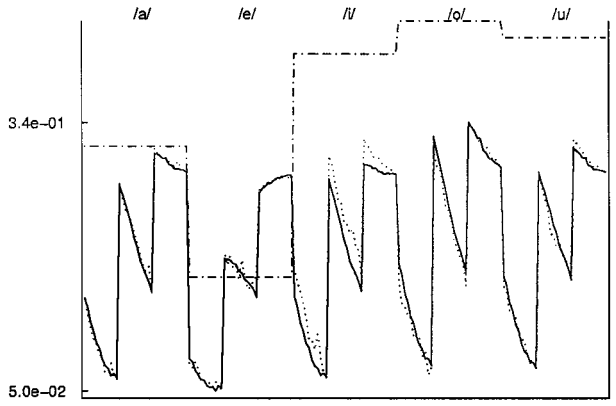

Fig. 5. Mean values changes of cepstral coefficient of index 4.

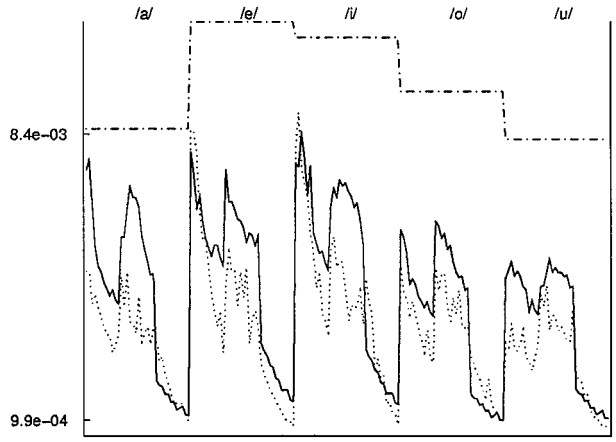

Fig. 6. Variance values changes of cepstral coefficient of index 4.

the testing data set: a higher value in the variance should allow more flexibility when dealing with the (independent) testing data set.

IV. Conclusions

From the results it is clear the advantage in using the compensated models over the clean models. It is clear also the superiority of the noisy models in relation to the compensated models, but such superiority is reduced when we employed the testing data set. We see that there is a loss of performance of the noisy models when we change from the training data set to the testing data set. Such loss of performance (from training to testing data set) is less noticed with the compensated models. This shows that the compensated models are less dependent on the training data set than the noisy models. Since real data are independent from the training data set, we can infer an almost equivalence between noisy and compensated models. The idea of intra-word SNR seems to be reasonable, since we achieved improvements in terms of recognition percentage when we compare the results between **intra** and **comp.** models (Tables I and II). It must be mentioned that the extra energy coefficients did not participate in the training and recognition processes as added parameters in the computation of output state probabilities. They solely entered in the computation of the intra-word SNR.

It is not asserted that autoregressive-derived coefficients produce good models of noisy signals. What has been shown by the results however is that, in the context of speech recognition, representing the signal by such coefficients does not constitute an impediment to achieving good improvements when compensating techniques are used. It is known that dynamic features perform better than uncompensated static features under noisy conditions [23]. Using only static features in this work, and noting the superior performances of the compensated models over the clean ones, we have shown that the static features from the clean models also proved to be useful when appropriately compensated for the noise.

One advantage of the approach is that the input speech signal does not need to be pre-processed, making the model compensation independent of the recognition stage. This allows the compensation to be made at moments when the recognition system is inactive, making the adjusted models compensated for the last affecting background noise characteristics. Using a second microphone, directed to the specific source of noise, and a reliable measure of the current SNR, the compensation can be accomplished even during moments of system activity.

The compensation technique is $O(p^3 + pLm)$ per Normal probability density, where p is the linear prediction order, L is the number of signal samples in one window and m is the number of normalized autocorrelation vectors averaged to produce $\boldsymbol{r}_n$. In $O(p^3 + pLm)$

- the term p^3 is due to the solution of the system of p equations (3) and the multiplication of matrices as in expression (6);
- the term $p\ L\ m$ is from the estimation of $\boldsymbol{r}_n$, the normalized autocorrelation of noise.

All other transformations in the compensation method are $O(p^2)$ or less. The estimation of $\boldsymbol{r}_n$ do not need to be done on the compensation of each probability density, but only once in one compensation session, since the same $\boldsymbol{r}_n$ will be used to compensate every probability density of all models of the recognition system. So, the complexity to compensate k probability density functions is $O(kp^3 + pLm)$.

We believe that the validity of the approach's principle was justified by the experimental results achieved, its computational simplicity, and the sound potential of its applicability.

Appendix
Conversion Matrix

This appendix is related to Section II-A, expression 2.

Expanding the matrix expression (3), putting the r_k's to the left-handed side we have that

$$\begin{array}{ccccccccc} r_1 + a_1 & + & r_1 a_2 & + & \cdots & r_{p-1} a_p & = & 0 \\ r_1 a_1 & + & r_2 + a_2 & + & \cdots & r_{p-2} a_p & = & 0 \\ r_2 a_1 & + & r_1 a_2 & + & \cdots & r_{p-3} a_p & = & 0 \\ \vdots & & \vdots & & \ddots & \vdots & = & \vdots \\ r_{p-1} a_1 & + & r_{p-2} a_2 & + & \cdots & r_p + a_p & = & 0 \end{array}. \tag{10}$$

Row i of this set of equations, defining

$$a_k = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k < 0 \text{ or } k > p \end{cases} \tag{11}$$

can be given by

$$(a_{i-1}+a_{i+1})r_1+(a_{i-2}+a_{i+2})r_2+\cdots(a_0+a_{2i})r_i\cdots +(a_{i-p}+a_{i+p})r_p=-a_i. \quad (12)$$

Representing this set of $i=1,2,\cdots,p$ equations in a matrix form, having r_i's as the independent vector, we obtain

$$\begin{bmatrix} a_0+a_2 & a_{-1}+a_3 & \cdots & a_{-(p-1)}+a_{p+1} \\ a_1+a_3 & a_0+a_4 & \cdots & a_{-(p-2)}+a_{p+2} \\ a_2+a_4 & a_1+a_5 & \cdots & a_{-(p-3)}+a_{p+3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p-1}+a_{p+1} & a_{p-2}+a_{p+2} & \cdots & a_0+a_{2p} \end{bmatrix} \times \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_p \end{bmatrix} = \begin{bmatrix} -a_1 \\ -a_2 \\ -a_3 \\ \vdots \\ -a_p \end{bmatrix}.$$

The above $p\times p$ matrix can be rewritten as

$$\begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-(p-1)} \\ a_1 & a_0 & a_{-1} & \cdots & a_{-(p-2)} \\ a_2 & a_1 & a_0 & \cdots & a_{-(p-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{p-1} & a_{p-2} & a_{p-3} & \cdots & a_0 \end{bmatrix} + \begin{bmatrix} a_2 & a_3 & \cdots & a_p & a_{p+1} \\ a_3 & a_4 & \cdots & a_{p+1} & a_{p+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_p & a_{p+1} & \cdots & a_{2p-2} & a_{2p-1} \\ a_{p+1} & a_{p+2} & \cdots & a_{2p-1} & a_{2p} \end{bmatrix}.$$

Applying (11) to the above, we arrive to the desired matrix $\boldsymbol{A}$

$$\boldsymbol{A}=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & 1 & 0 & \cdots & 0 \\ a_2 & a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{p-1} & a_{p-2} & a_{p-3} & \cdots & 1 \end{bmatrix} + \begin{bmatrix} a_2 & a_3 & \cdots & a_p & 0 \\ a_3 & a_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_p & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

It can be shown that $\boldsymbol{A}$ has an inverse if $a_k,\ k=1,2,\cdots,p$ are such that the filter

$$\frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}}$$

is stable.

ACKNOWLEDGMENT

The author wishes to thank the reviewers for their valuable comments and suggestions.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.
[2] ——, "Speech enhancement in the 1980s: noise suppression with pattern matching," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Decker, 1992.
[3] B. H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1987, pp. 2368–2371.
[4] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1987, pp. 177–180.
[5] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," in *Speech Enhancement*, J. S. Lim, Ed.. Englewood Cliffs, NJ, 1983.
[6] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Norwell, MA, 1997.
[7] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, June 1974.
[8] F.-H. Liu, R. M. Stern, A. Acero, and P. J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, 1994, pp. 61–64.
[9] B. A. Carlson and M. A. Clements, "Speech recognition in noise using a projection-based likelihood measure for mixture density HHM's," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Mar. 1992, pp. 237–240.
[10] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 1659–1671, Nov. 1989.
[11] B. A. Hanson, T. H. Applebaum, and J.-C. Junqua, "Spectral dynamics for speech recognition under adverse conditions," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Norwell, MA: Kluwer, 1997.
[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.
[13] J.-C. Junqua and J.-P. Haton, *Robusteness in Automatic Speech Recognition: Fundamentals and Applications*. Norwell, MA: Kluwer, 1996.
[14] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1495–1503, Oct. 1989.
[15] V. L. Beattie and S. J. Young, "Noisy speech recognition using hidden Markov model state-based filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 1991, pp. 917–920.
[16] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustics Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 845–848.
[17] M. J. F. Gales and S. J. Young, "A fast and flexible implementation of parallel model combination," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1995, pp. 133–136.
[18] A. P. Varga *et al.*, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," DRA Speech Res. Unit, 1992.
[19] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[20] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Mar. 1992, pp. 233–236.
[21] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 2, pp. 90–93, Jan. 1974.
[22] *HTK version 1.2: Reference Manual*, Speech Group, Eng. Dept., Cambridge Univ., Cambridge, U.K., 1990.
[23] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 857–860.

Ivandro Sanches (S'89–M'99) was born in São Paulo, Brazil, in 1964. He received the B.S. and M.S. degrees from Escola Politécnica, University of São Paulo, in 1987 and 1989, respectively, and the Ph.D. degree from Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1994, all in electrical engineering.

He joined the Department of Electrical Engineering, Escola Politécnica, University of São Paulo, in 1989, where he is currently an Associate Professor. He was on leave at Imperial College from 1990 to 1994 as a doctoral student with a scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazilian Government Founding Council. His research interests include robust speech recognition, speaker adaptation and large-vocabulary speech recognition.