ELSEVIER

Speech Communication 24 (1998) 267–285

SPEECH COMMUNICATION

Data-driven environmental compensation for speech recognition: A unified approach

Pedro J. Moreno *, Bhiksha Raj, Richard M. Stern

Department of Electrical and Computer Engineering & School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received 26 July 1996; received in revised form 3 November 1997; accepted 8 May 1998

Abstract

Environmental robustness for automatic speech recognition systems based on parameter modification can be accomplished in two complementary ways. One approach is to modify the incoming features of environmentally-degraded speech to more closely resemble the features of the (normally undegraded) speech used to train the classifier. The other approach is to modifying the internal statistical representations of speech features used by the classifier to more closely resemble the features representing degraded speech in a particular target environment. This paper attempts to unify these two approaches to robust speech recognition by presenting several techniques that share the same basic assumptions and internal structure while differing in whether they modify the features of incoming speech or whether they modify the statistics of the classifier itself. We present the multivaRiate gAussian-based cepsTral normaliZation (RATZ) family of algorithms which modify incoming cepstral features, along with the STAR (STAtistical Reestimation) family of algorithms, which modify the internal statistics of the classifier. Both types of algorithms are data driven, in that they make use of a certain amount of adaptation data for learning compensation parameters. The algorithms were evaluated using the SPHINX-II speech recognition system on subsets of the Wall Street Journal database. While all algorithms demonstrated improved recognition accuracy compared to previous algorithms, the STAR family of algorithms tended to provide lower error rates than the RATZ family of algorithms as the SNR was decreased.

Résumé

Dans les systèmes de reconnaissance de la parole, la robustesse à l'environnement par adaptation des paramètres peut être obtenue de deux façons complémentaires. Une première approche consiste à modifier les paramètres acoustiques de la parole dégradée par l'environnement de façon à ressembler aux paramètres de la parole (habituellement non dégradée) qui a été utilisée lors de l'entraînement. La deuxième solution est de modifier les paramètres statistiques internes au reconnaisseur de façon à mieux représenter les caractèristiques de la parole dégradée dans un environnement cible particulier. Le présent papier tente d'unifier ces deux approches de reconnaissance robuste de la parole en présentant plusieurs techniques qui partagent les mêmes hypothèses de base et la même structure, tout en différant dans le choix de savoir si elles modifient les paramètres d'entrée ou les paramètres statistiques du reconnaisseur. Nous présentons ici la famille d'algorithmes basés sur la normalisation cepstrale gaussienne multi-variable (RATZ) qui modifient les caractéristiques cepstrales d'entrée, ainsi que les algorithmes STAR (re-estimation statistique), qui modifient les

* Corresponding author. Current address: Speech Interaction Technology Group, Cambridge Research Laboratory, Digital Equipment Corporation, One Kendall Square, Bldg. 700, Cambridge, MA 02139-1562, USA. Tel.: +1 617 692 7692; fax: +1 617 692 7650; e-mail: pjm@crl.dec.com.

0167-6393/98/$19.00
PII: S0167-6393(98)00025-9

paramètres internes du reconnaisseur. Les deux types d'algorithmes sont basés sur les données et utilisent une certaine quantité de donnée d'adaptation pour estimer les paramètres de compensation. Les algorithmes ont été évalués en utilisant le système de reconnaissance SPHINX-II sur un sous-ensemble de la base de donnée Wall Street Journal. Bien que tous les algorithmes conduisaient une amélioration des performances en comparaison des algorithmes précédents, la famille d'algorithmes STAR donnait généralement des taux d'erreur plus faibles que la famille d'algorithmes RATZ lorsque le rapport signal/bruit diminuait.

Keywords: Robust speech communication; Environmental compensation; Data-driven algorithms

1. Introduction

As speech recognition technology is transitioning from the laboratory to practical applications the importance of environmental robustness in speech recognition is becoming increasingly appreciated. Many research groups have described algorithms that achieve a greater degree of environmental robustness using a variety of techniques. For example, Acero and Stern (1990), Varga and Moore (1990), Ephraim (1992), Gales and Young (1993), Leggeter and Woodland (1995), and Sankar and Lee (1995) have all described compensation strategies that involve the use of a para-metric model of degradation, combined with optimal estimation of the parameters of the model. Other approaches to robustness include the use of peripheral features based on models of human perception of speech signals (e.g. Ghitza, 1986; Seneff, 1988; Hermansky, 1990; Hermansky and Morgan, 1994), as well as the use of arrays of microphones to separate noise sources from speech signal arriving from different spatial locations (e.g. Flanagan et al., 1985; Sullivan and Stern, 1993). The reviews of Juang (1991) and Stern et al. (1996), and Junqua and Haton (1996) are among a number of recent overviews of the robustness field.

In this paper we consider compensation methods that achieve environmental robustness by directly observing the effects of environmental degradation without the use of structural assumptions about the nature of the degradation. These modifications are generally based on empirical comparisons of parameters derived from high-quality speech used to train the system to parameters derived from degraded speech sampled in the target environment. Some of these techniques have attempted to address the problem by applying compensation factors to the incoming cepstrum vectors (e.g. Acero and Stern, 1990; Liu et al., 1994; Neumeyer and Weintraub, 1994; Moreno et al., 1996) while other techniques have modified parameters characterizing the statistical model of speech sounds such as the means and covariances of the distributions of HMMs (e.g. Gales, 1995; Sankar and Lee, 1995). However, both kind of approaches have usually been presented as separate and distinct techniques.

In this paper we present a series of data-driven empirical compensation methods that use the same unified mathematical scheme to compensate both the incoming data and the internal statistical representations in the classifier. It thus becomes possible to either modify the incoming cepstral vectors or modify the parameters (means and variances) of the distributions of the HMMs using the same basic mathematical formulation of the problem.

The compensation techniques described in this paper are not constrained to operate on any particular feature set and could be applied to any feature representation such as cepstra, LPC-cepstra or even auditory-based representations. However, all the experiments described in this paper were performed on cepstral vectors.

In Section 2 of this paper we make some observations about the effects of the environment on the distributions of log spectra of clean speech by considering simulations using artificially-generated data. We also discuss the reasons for the degradation in recognition accuracy introduced by the environment. In Section 3 we develop the unified mathematical scheme for our data-driven environmental compensation methods that can modify either the incoming feature vectors of degraded speech or the internal distributions characterizing clean speech. In Section 4 we present the

multivaRiate gAussian-based cepsTral normaliZation (RATZ) family of algorithms, which provides compensation of the incoming features. We describe in detail the mathematical structure of the algorithms, and we present experimental results that explore some of the dimensions of the algorithm. In Section 5 we present the STAR (STAtistical Reestimation) family of algorithms, which modify the internal statistics of the classifier. We describe the algorithms and present experimental results. We present comparisons of the STAR and RATZ algorithms and we show that STAR compensation results in greater recognition accuracy. Finally, in Section 6 we summarize our findings.

2. Effects of the environment on distributions of clean speech

"Environment" in the context of speech recognition systems could include such varied effects as low quality microphones, non-linearities on a telephone network, line noise, reverberation, and competing speakers. It is therefore impossible to enumerate each of these conditions and analyze their effects on the distributions of speech individually. For this reason we generically model the effects of the environment on the distribution of clean speech as additive terms to the means and variances of the distributions of clean speech. To justify this approach we present the very simple example of the effect of a linear time-invariant channel and additive uncorrelated noise on the log-spectra of speech. Although our example is presented in terms of log-spectral features, the approaches discussed in this paper are general and work in principle for any types of features. (Cepstral features in particular are related to log-spectral features through a linear transformation, so similar results will directly apply for cepstral features as well.) Finally, using this approach we propose two generic solutions to the problem of robustness.

2.1. The effect of linear filtering and additive noise

A reasonable model for the environment for the case of *linear filtering and additive noise* is given in Fig. 1. This representation was proposed by Acero (1991) and later used by Gales (1995), among many other researchers. The effect of the noise and filtering on clean speech in the power spectral domain can be represented as:

$$P_Y(\omega_k) = |H(\omega_k)|^2 P_X(\omega_k) + P_N(\omega_k), \quad (1)$$

where $P_Y(\omega_k)$ represents the power spectra of the degraded speech, $y[m]$, $P_N(\omega_k)$ represents the power spectra of the noise $n[m]$, $P_X(\omega_k)$ represents the power spectra of the clean speech $x[m]$, $|H(\omega_k)|^2$ represents the squared magnitude of the transfer function of the filter with impulse response $h[m]$ representing channel effects, and ω_k represents a particular mel-spectral band. Our assumptions are that the noise is stationary and uncorrelated with the speech signal, and that the channel is time invariant and independent of the signal level.

Transforming to the log-spectral domain, we obtain:

$$\log[P_Y(\omega_k)] = \log[|H(\omega_k)|^2 P_X(\omega_k) + P_N(\omega_k)]. \quad (2)$$

Defining the log spectra of degraded speech, noise, clean speech, and the log of the squared magnitude of the transfer function of the filter as

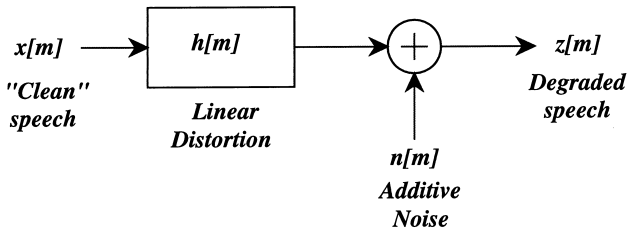

Fig. 1. A model of the environment for additive noise and filtering by a linear channel. $x[m]$ represents the clean speech signal, $n[m]$ represents the additive noise and $y[m]$ represents the resulting degraded speech signal. The function $h[m]$ represents the effects of a linear channel.

$$\begin{aligned}\tilde{Y}[k] &= \log[P_Y(\omega_k)],\\ \tilde{N}[k] &= \log[P_N(\omega_k)],\\ \tilde{X}[k] &= \log[P_X(\omega_k)],\\ \tilde{H}[k] &= \log[|H(\omega_k)|^2],\end{aligned} \tag{3}$$

we obtain the following equations

$$\begin{aligned}\tilde{Y}[k] &= \tilde{X}[k] + \tilde{H}[k]\\ &\quad + \log(1 + \exp(\tilde{N}[k] - \tilde{X}[k] - \tilde{H}[k])).\end{aligned} \tag{4}$$

This expression can be shortened to

$$\tilde{Y}[k] = \tilde{X}[k] + f(\tilde{X}[k], \tilde{H}[k], \tilde{N}[k]) \tag{5}$$

or in vector form

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}), \tag{6}$$

where

$$\begin{aligned}&f(\tilde{X}[k], \tilde{H}[k], \tilde{N}[k])\\ &\quad = \tilde{H}[k] + \log(1 + \exp(\tilde{N}[k] - \tilde{X}[k] - \tilde{H}[k])).\end{aligned} \tag{7}$$

Let us assume that the log-spectral vectors that characterize clean speech follow a Gaussian distribution, $N_x(\mu_x, \Sigma_x)$ and that the noise and channel are perfectly known. Eq. (7) provides a transformation of random variables leading to a new distribution for the log spectra of degraded speech equal to

$$\begin{aligned}p(\boldsymbol{y}|\mu_x, \Sigma_x, \boldsymbol{n}, \boldsymbol{h}) &= \Big\{((2\pi)^D|\Sigma_x|)^{1/2}|(I\\ &\quad -\mathrm{diag}(\mathrm{e}^{\boldsymbol{n}-\boldsymbol{y}}))|\Big\}^{-1} \exp\left\{\frac{1}{2}(\boldsymbol{y} - \boldsymbol{h} - \mu_x + \log(i\right.\\ &\quad \left.-\mathrm{e}^{\boldsymbol{n}-\boldsymbol{y}}))^{\mathrm{T}}\Sigma_x^{-1}(y - h - \mu_x + \log(i - \mathrm{e}^{\boldsymbol{n}-\boldsymbol{y}}))\right\},\end{aligned} \tag{8}$$

where D is the dimensionality of the random variable that characterizes the log spectrum, i is the unitary vector, I is the identity matrix, and $\mathrm{diag}(\mathrm{e}^{\boldsymbol{n}-\boldsymbol{y}})$ is the diagonal matrix whose ith diagonal element is the ith element of $\mathrm{e}^{\boldsymbol{n}-\boldsymbol{y}}$.

The resulting distribution $p(\boldsymbol{y})$is non-Gaussian. However, since a Gaussian distribution assigned to $p(\boldsymbol{y})$ can still capture part of the effect of the channel and noise, we will continue to assume Gaussian distributions. To characterize the Gaussian distributions we need only compute the mean vector and covariance matrices of these new distributions. The new mean vector can be computed as:

$$\mu_y = \mu_x + \int_X f(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}) N_x(\mu_x, \Sigma_x)\, \mathrm{d}\boldsymbol{x} \tag{9}$$

and the covariance matrix can be computed as:

$$\begin{aligned}\Sigma_y &= E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) + E(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})^{\mathrm{T}})\\ &\quad + 2E(\boldsymbol{x}\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})^{\mathrm{T}}) - \mu_y\mu_y^{\mathrm{T}}\\ &= E(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) + \int_X (2\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}))\\ &\quad \times f(x, h, n)^{\mathrm{T}} N_x(\mu_x, \Sigma_x)\, \mathrm{d}\boldsymbol{x} - \mu_y\mu_y^{\mathrm{T}}.\end{aligned} \tag{10}$$

The integrals in Eqs. (9) and (10) do not have closed-form solutions, so numerical methods must be used to estimate the mean vector and covariance matrix of the distribution.

Unfortunately, the previous assumptions about the noise are too simple. In most cases the noise must be estimated and is not known a priori. A more realistic model would be to assign a Gaussian distribution $N_n(\mu_n, \Sigma_n)$ to the noise. To simplify the resulting equations we can also assume that the noise and the speech are statistically independent. Under these assumptions the mean vector and the covariance matrix of the log spectrum of the degraded speech will have the form

$$\mu_y = \mu_x + \int_X N_x(\mu_x, \Sigma_x) \int_N \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}) N_n(\mu_n, \Sigma_n)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{n}, \tag{11}$$

$$\begin{aligned}\Sigma_y &= \Sigma_x + \mu_x\mu_x^{\mathrm{T}}\\ &\quad + \int_X N_x(\mu_x, \Sigma_x) \int_N (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})^{\mathrm{T}})\\ &\quad \times N_n(\mu_n, \Sigma_n)\, \mathrm{d}\boldsymbol{n}\, \mathrm{d}\boldsymbol{x} + \int_X N_x(\mu_x, \Sigma_x) \int_N (x\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})^{\mathrm{T}}\\ &\quad + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})x^{\mathrm{T}}) \times N_n(\mu_n, \Sigma_n)\, \mathrm{d}\boldsymbol{n}\, \mathrm{d}\boldsymbol{x} - \mu_y\mu_y^{\mathrm{T}}.\end{aligned} \tag{12}$$

These equations also have no closed-form solution, and we must estimate the resulting mean vector and covariance matrix through numerical methods. Gales (1995) presents several approximations to solve these equations.

2.2. One-dimensional simulations of noise-corrupted Gaussian data

To help visualize the resulting distributions of degraded data we plot artificially-produced one-dimensional simulations of Gaussian random variables corrupted by additive noise. These plots simulate the simplified case of a one-dimensional log spectral feature vector of speech. Artificial clean data were produced according to the Gaussian distribution $N_x(\mu_x, \sigma_n^2)$ and contaminated with artificially-produced Gaussian noise with the distribution $N_n(\mu_n, \sigma_n^2)$ and a channel h. The artificially-produced clean data, noise, and channel were combined according to Eq. (4) producing the degraded data set $\{y_0, y_1, \ldots, y_{N-1}\}$. From this degraded data set we directly estimated the mean and variance of a maximum likelihood (ML) fit as:

$$\mu_{y,\mathrm{ML}} = \frac{1}{N}\sum_{i=0}^{N-1} y_i,$$

$$\sigma^2_{y,\mathrm{ML}} = \frac{1}{N}\sum_{i=0}^{N-1} (y_i - \mu_{y,\mathrm{ML}})^2. \qquad (13)$$

We also computed a histogram of the degraded data to estimate its real distribution. The contamination was performed at different signal-to-noise ratios (SNRs) defined by $\mu_x - \mu_n$. Fig. 2 shows an example of the original distribution of the original clean parameter x (solid curve), the corrupted value of that parameter y after going through the transformation of Eq. (4) that produces the distribution of Eq. (8) (dashed curve), and the best Gaussian fit to that distribution (dotted curve). The SNR of the noisy signal is 3 dB.

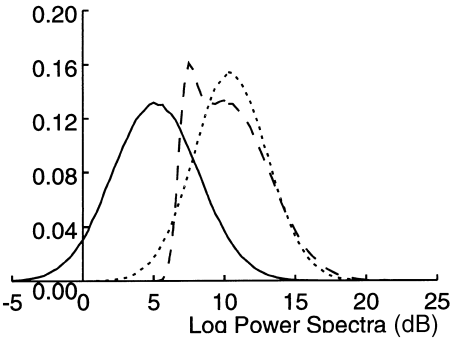

Fig. 2. Estimates of the distribution of degraded Gaussian data via Monte Carlo simulation. The solid line represents the PDF of the clean signal. The dashed line represents the actual PDF of the noise-contaminated signal. The dotted line represents the best Gaussian approximation to the PDF of the degraded signal. The original clean signal had a mean of 5.0 and a variance of 3.0, and the parameter representing the channel effect was set equal to 5.0. The mean of the noise was 7.0 and its variance was 0.5.

We observe that while the resulting probability density function (PDF) is non-Gaussian and sometimes bimodal (as in Fig. 2), a Gaussian fit to the PDF captures some of the effects of the environment on the clean signal. The effect on the parameters of the PDF can be reasonably accurately modeled as a *shift* in the mean of the PDFs and a *decrease* in the variance of the resulting PDF. Intuitively, the shift in the mean occurs mainly because of the effect of the channel on the clean speech signal. The compression in variance occurs because of the logarithmic transformation being applied on the spectrum: the lower region of the distribution of spectral power gets shifted more than the upper region, thereby causing the overall density to shrink in width. Note, however, that this compression of the variance will occur only if the variance of the distribution of the noise is smaller than the variance of the distribution of the clean signal. This change in variance can be represented by an additive factor in the covariance matrix.

2.3. Modeling the effects of the environment as correction factors

In summary, the one-dimensional simulations of the effects of noise on the data imply the following effects

- The PDF of the degraded signal is non-Gaussian, and may exhibit a bimodal shape.
- The PDF of the signal shifts according to the SNR.
- The PDF of the signal is compressed if $\Sigma_x > \Sigma_n$ or expanded if $\Sigma_x < \Sigma_n$. The amount of compression or expansion depends on the SNR.

Since most speech recognition systems model the statistics of speech as mixtures of Gaussian distributions, it is still convenient to continue modeling the resulting PDFs of degraded speech as Gaussian. A simple way to achieve this is by (1) modeling the mean of the distribution of degraded speech as the mean of the clean signal plus a correction vector

$$\mu_y = \mu_x + \boldsymbol{r} \tag{14}$$

and (2) by modeling the covariance matrix of the distribution of degraded speech as the covariance matrix of the clean speech plus a correction matrix

$$\Sigma_y = \Sigma_x + \boldsymbol{R}. \tag{15}$$

The $\boldsymbol{R}$ matrix will be symmetric and will have positive or negative elements according to the value of the covariance matrix of the noise compared with that of the clean signal. It must also be of a form that guarantees that the resulting Σ_y matrix remains a valid covariance matrix. This approach will be used for the algorithms presented throughout this paper.

3. A unified view of data-driven environment compensation

In Section 2 we have suggested that the effect of the environment on the distributions of the log spectra or cepstra of clean speech can be modeled by additive correction factors applied to the mean vectors and covariance matrices. In this section we present techniques that attempt to learn the parameters that characterize these effects directly from sample data, by comparing sets of degraded and clean vectors. This approach does not explicitly assume any model of the environment, but uses empirical observations to infer environmental characteristics.

We first introduce a unified view of environmental compensation and then provide solutions for the compensation factors. We then particularize these solutions for the case of adaptation datasets that were simultaneously recorded using clean speech and degraded speech from the target environment (“stereo” data). Subsequently we particularize the generic solutions for two family of techniques; the Multivariate Gaussian Based Cepstral Normalization (RATZ) techniques (Moreno et al., 1995a, b) and the Statistical Reestimation (STAR) techniques (Moreno et al., 1995b; Moreno, 1996). The performance of these techniques for several databases and experimental conditions is then explored.

3.1. Basic assumptions

We model the distribution of the tth vector $\boldsymbol{x}_t$ of a cepstral vector sequence of length T, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, generically as:

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{k} a_k(t) N_x(\mu_{x,k}, \Sigma_{x,k}). \tag{16}$$

In other words, $p(\boldsymbol{x_t})$is modeled as a summation of K Gaussian components with a priori probabilities that are time dependent. Assuming that each vector $\boldsymbol{x_t}$ is independent and identically distributed (i.i.d.), the overall likelihood for the full observation sequence X becomes

$$l(\boldsymbol{X}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t)$$

$$= \prod_{t=1}^{T} \sum_{k} a_k(t) N_x(\mu_{x,k}, \Sigma_{x,k}). \tag{17}$$

The above likelihood equation offers a double interpretation, depending on whether the compensation is to be performed by modfying the incoming features or by modifying the internal statistical representation used by the recognition system. For convenience, we refer to compensation methods that modify incoming features as “data compensation methods” and compensation methods that modify internal statistical representations as “model compensation methods”. In the sections below we describe two separate but closely-related families of algorithms for environmental compensation: the multivaRiate gAussian-based cepsTral normaliZation (RATZ) family of algorithms, which provides data compensation, and the STAR (STAtistical Reestimation) family of algorithms, which provides model compensation.

For data compensation methods, we set the a priori probabilities $a_k(t)$ of the equations above to be independent of t. This defines a conventional mixture of Gaussian distributions for the entire training set of cepstral vectors.

The interpretation is slightly different for model compensation methods. We assume that the cepstral speech vectors are emitted by an HMM with K states in which each state emission PDF is

composed of a single Gaussian. In this case the $a_k(t)$ terms define the probability of being in state k at time t and follow the state probabilities of a Markov chain defined by the state initial probabilities and the transition matrix of the HMM. Under these assumptions the expression of the likelihood for the full observation sequence is exactly as expressed in Eq. (17).

The assumption of a single Gaussian per state is not limiting at all. Specifically, any state with a mixture of Gaussians for emission probabilities can also be represented by multiple states where the output distributions are single Gaussians and where the incoming transition probabilities of the state are the same as the a priori probabilities of the Gaussians $a_k(t)$ and the exiting transition probability is unity. Fig. 3 illustrates this idea for a specific arbitrary state.

The probabilities $a_k(t)$ depend only on the Markov chain topology and are represented by

$$a(t) = [a_1(t)a_2(t)\dots a_k(t)]^{\mathrm{T}} = A^t\pi, \tag{18}$$

where A^t represents the transition matrix after t transitions and π represents the initial state probability vector of the HMM. The terms $N_x(\mu_{x,k}, \Sigma_{x,k})$ of Eq. (16) refer to the Gaussian densities associated with each of the K states of the HMM.

As we have mentioned before, the effects of noise and linear filtering on the mean vectors and covariance matrices can be expressed as

$$\mu_{y,k} = \boldsymbol{r}_k + \mu_{x,k}, \tag{19}$$

$$\Sigma_{y,k} = \boldsymbol{R}_k + \Sigma_{x,k}, \tag{20}$$

where $\boldsymbol{r}_k$ and $\boldsymbol{R}_k$ represent the corrections applied to the mean vector and covariance matrix respectively of the kth Gaussian. These two correction factors account for the effects of the environment on the distributions of the cepstra of clean speech. The first step of the RATZ and STAR algorithms is to estimate these correction factors.

3.2. Solutions for the correction factors $\mathbf{r}_k$ and $\boldsymbol{R}_k$

Historically, the development of empirical data-driven environmental compensation algorithms has been motivated and accelerated by the collection and dissemination of databases containing speech that is simultaneously recorded in high-quality "clean" environments and in degraded target environments (Liu et al., 1994). In the RATZ and STAR family of algorithms, approaches to solutions for the correction factors $\boldsymbol{r}_k$ and $\boldsymbol{R}_k$ will depend on whether or not such "stereo" data are available in a specific task. We first describe the generic solution for the case in which only samples of degraded speech are available, which we refer to as the "blind" case. We then describe how to particularize these solutions for the case when stereo data are in fact available.

We make extensive use of the EM algorithm (Dempster et al., 1977; Huang et al., 1993) for estimating the correction factors $\boldsymbol{r}_k$ and $\boldsymbol{R}_k$. In this paper we describe how the EM algorithm can be applied to solve for the correction parameters $\boldsymbol{r}_k$ and $\boldsymbol{R}_k$.

3.2.1. Non-stereo-based solutions

We begin with an observed set of T degraded vectors $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T\}$, assuming that these vectors have been produced by a probability density function

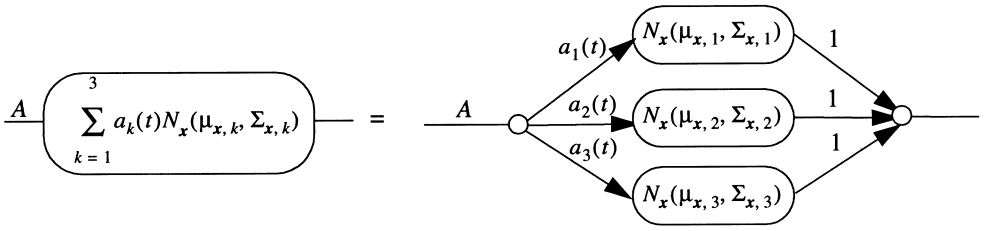

Fig. 3. A state with a mixture of three Gaussians is equivalent to a set of three states, each with a single Gaussian where the transition probabilities are the a priori probabilities of each of the Gaussian mixtures. The constant A represents the transition probability into the state.

$$p(\boldsymbol{y}_t) = \sum_{k=1}^{K} a_k(t) N_y(\mu_{y,k}, \Sigma_{y,k}) \tag{21}$$

which is a summation of K Gaussians where each component relates to the corresponding kth Gaussian of clean speech according to Eq. (19). We define a likelihood function $l(\boldsymbol{Y})$ as

$$l(\boldsymbol{Y}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t) = \prod_{t=1}^{T} \sum_{k} a_k(t) N_y(\mu_{y,k}, \Sigma_{y,k}). \tag{22}$$

We can also express $l(\boldsymbol{Y})$ in terms of the original parameters of clean speech and the correction terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$

$$l(\boldsymbol{Y}) = l(\boldsymbol{Y}|\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_K) = \prod_{t=1}^{T} p(\boldsymbol{y}_t) = \prod_{t=1}^{T} \sum_{k} a_k(t) N_y(\boldsymbol{r}_k + \mu_{x,k}, \boldsymbol{R}_k + \Sigma_{x,k}). \tag{23}$$

For convenience we express the above equation in the logarithm domain defining the log likelihood $L(\boldsymbol{Y})$ as

$$L(\boldsymbol{Y}) = \log(l(\boldsymbol{Y})) = \sum_{t=1}^{T} \log(p(\boldsymbol{y}_t)) = \sum_{t=1}^{T} \log\left(\sum_{k} a_k(t) N_y(\boldsymbol{r}_k + \mu_{x,k}, \boldsymbol{R}_k + \Sigma_{x,k}) \right). \tag{24}$$

Our goal is to find the complete set of K terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$ that maximize the likelihood (or log likelihood). As it turns out there is no direct solution to this problem and some indirect method is necessary. The expectation-maximization (EM) algorithm is one of these methods.

In Appendix A we outline the solution for the correction terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$. The final solutions of the maximum likelihood equation are:

$$\bar{\boldsymbol{r}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{y}_t, \phi]\boldsymbol{y}_t}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{y}_t, \phi]} - \mu_{x,k}, \tag{25}$$

$$\bar{\boldsymbol{R}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{y}_t, \phi]((\boldsymbol{y}_t - \mu_{x,k} - \bar{\boldsymbol{r}}_k)(\boldsymbol{y}_t - \mu_{x,k} - \bar{\boldsymbol{r}}_k)^{\mathrm{T}})}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{y}_t, \phi]} - \Sigma_{x,k}, \tag{26}$$

where ϕ represents the set of parameters $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$ learned in the previous iteration. Eqs. (25) and (26) form the basis for an iterative algorithm. The EM algorithm guarantees that each iteration increases the likelihood of the observed data.

3.2.2. Stereo-based solutions

When simultaneously-recorded clean and degraded speech data ("stereo" data) are available, the information about the environment is encoded in the stereo pairs. By observing how each clean speech vector is transformed into a degraded speech vector we can learn the correction factors more directly.

We can readily assume that the a posteriori probabilities $P[s_t(k)|\boldsymbol{y}_t, \phi]$ can be directly estimated by $P[s_t(k)|\boldsymbol{x}_t]$. This is equivalent to assuming that the probabilities of a vector being produced by each of the underlying classes do not change due to the environment. We call this assumption *a posteriori invariance*. This assumption, although not strictly correct, seems to be a good approximation.

If we expand the $P[s_t(k)|\boldsymbol{y}_t, \phi]$ and $P[s_t(k)|\boldsymbol{x}_t]$ terms we obtain

$$P[s_t(k)|\boldsymbol{y}_t, \phi] = \frac{P[s_t(k)]p(\boldsymbol{y}_t|s_t(k), \phi)}{\sum_{j=1}^{K} p(\boldsymbol{y}_t|s_t(j), \phi)P[s_t(j)]} \tag{27}$$

and

$$P[s_t(k)|\boldsymbol{x}_t] = \frac{P[s_t(k)]p(\boldsymbol{x}_t|s_t(k))}{\sum_{j=1}^{K} p(\boldsymbol{x}_t|s_t(j))P[s_t(j)]}. \tag{28}$$

For the above two expressions to be equal, each of the terms in the summation must be equal. This would imply that each Gaussian is shifted by exactly the same amount and not compressed. However, at high SNRs the shift for each Gaussian is quite similar and the compression in the variances is almost zero. Therefore, at high SNRs the assumption of a posteriori invariance is approximately valid and at lower SNRs it is less valid. In addition, this assumption avoids the need to iterate Eqs. (25) and (26).

This leads to

$$\bar{\boldsymbol{r}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t](\boldsymbol{y}_t - \mu_{x,k})}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]} = \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t](\boldsymbol{y}_t - \boldsymbol{x}_t + \boldsymbol{x}_t - \mu_{x,k})}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]}, \tag{29}$$

$$\bar{\boldsymbol{r}}_k = \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t](\boldsymbol{x}_t - \mu_{x,k})}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]} + \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t](\boldsymbol{y}_t - \boldsymbol{x}_t)}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]}. \tag{30}$$

The expected value of the first term on the left-hand side of Eq. (30) is zero. We therefore approximate this term to be zero. This results in

$$\bar{\boldsymbol{r}}_k \cong \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t](\boldsymbol{y}_t - \boldsymbol{x}_t)}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]}. \tag{31}$$

Similarly, we obtain

$$\bar{\boldsymbol{R}}_k \cong \frac{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]((\boldsymbol{y}_t - \boldsymbol{x}_t - \bar{\boldsymbol{r}}_k)(\boldsymbol{y}_t - \boldsymbol{x}_t - \bar{\boldsymbol{r}}_k)^{\mathrm{T}})}{\sum_{t=1}^{T} P[s_t(k)|\boldsymbol{x}_t]} - \Sigma_{x,k}. \tag{32}$$

3.3. Comparison with previous algorithms

The general solutions above can be particularized for cepstral compensation (data compensation) or for adaptation of HMMs (model compensation). When particularized for data compensation these solutions can be directly compared to fixed codeword-dependent cepstral normalization (FCDCN) (Acero, 1991). When particularized for HMM adaptation they can be compared to dual-channel codebook adaptation (DCCA) (Liu et al., 1994), maximum likelihood linear regression (MLLR) (Leggeter and Woodland, 1995), parallel model combination (PMC) (Gales, 1995), and stochastic matching (Sankar and Lee, 1995).

FCDCN is a data-driven cepstral compensation algorithm introduced by Acero (1991) and further developed by Liu (Liu et al., 1994). In the FCDCN algorithm the data are represented by a codebook computed using vector quantization (VQ). All data belonging to a particular VQ cell of the clean speech codebook are assumed to have been generated by a corresponding cell in the codebook representing degraded speech. Furthermore, FCDCN assumes that the difference between the data in a cell of the degraded speech and the corresponding data for clean speech is the same as the difference between the codewords representing the degraded speech cell and the clean speech cell. This can be viewed as a degenerate case of a modification of the equations in Section 3.2 in which the posterior probabilities of the Gaussians have been rounded off to 1 or 0 (with the most likely Gaussian getting an a posteriori probability of 1, the rest 0). Since the FCDCN method is quantization based, there can be no learning of variances.

DCCA is a data-driven HMM adaptation algorithm similar to FCDCN and was also introduced by Liu (Stern et al., 1994). DCCA adapts the codewords in the HMM codebooks representing the Gaussians using the same principle as FCDCN. It estimates an additive correction term for each of the codewords in the codebook as the difference in the centroids of all observations from clean speech belonging to that centroid and the centroid of their degraded counterparts. This correction is then applied to the codewords before recognition. This can be viewed as a degenerate case of the particularization of the equations to HMM adaptation using Eqs. (31) and (32) where the posterior probabilities of the Gaussians in the HMM codebooks have been quantized to 1 or 0. As in the case of FCDCN, there is no learning of variances.

MLLR (Leggeter and Woodland, 1995) was initially designed for speaker adaptation but it has proven to be quite effective in the field of environmental robustness as well. MLLR models the effect of the environment on the means of the statistics representing clean speech as a linear shift and a rotation matrix. In its latest implementation it also models the effect of the environment on the covariance matrices as a multiplicative matrix. Such a model of the effect of the environment on speech distributions is clearly more flexible that the one proposed in this paper. The STAR algorithm when implemented without stereo data can be thought of as a special case of the MLLR algorithm.

PMC (Gales, 1995) attempts to solve Eqs. (11) and (12) using numerical approximations. The main difference between PMC and the algorithms

proposed here is that PMC attempts to obtain an analytical solution for the correction factors rather than learning them directly from data as does MLLR. This is probably one of the virtues and limitations of all model-based algorithms like PMC. If the structural assumptions for a given model are not valid in a particular environment they can produce erroneous estimates for the correction terms. On the other hand, data-driven algorithms like MLLR or the RATZ and STAR algorithms proposed in this paper bypass this limitation by learning the correction factors directly from the data.

The Stochastic Matching algorithm (Sankar and Lee, 1995) is the closest in spirit to the STAR algorithm when no stereo data are provided. Some of our solutions are identical to the ones provided by the stochastic matching algorithm. The main difference lies in the ability of the algorithms proposed here to take advantage of stereo-recorded data. As we will show in our experiments, the use of stereo data provides for much better solutions for the correction terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$

4. The RATZ family of algorithms

In this section we particularize the general solutions described in Section 3 for the case of the Multivariate-Gaussian-Based Cepstral Normalization (RATZ) family of algorithms which operate on the incoming features. We present an overview of the algorithms and describe in detail the steps followed in RATZ-based compensation. We describe the general stereo-based and blind versions of the algorithms. Finally, we compare experimental results using several databases and environmental conditions.

4.1. Overview of RATZ and blind RATZ

The RATZ algorithms work in three stages: (1) estimation of the statistics of clean speech, (2) estimation of the statistics of degraded speech, and (3) compensation of degraded speech. We describe these steps in detail.

Estimation of the statistics of clean speech. The PDF for the features of clean speech is modeled as a mixture of multivariate Gaussian distributions. Under these assumptions the distribution of the cepstral vectors of clean speech can be written as

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} a_k N_x(\mu_{x,k}, \Sigma_{x,k}), \tag{33}$$

which is equivalent to Eq. (21) for the case of $a_k(t)$ being time independent. The parameters a_k, $\mu_{x,k}$ and $\Sigma_{x,k}$ represent, respectively, the *a priori* probabilities, mean vector, and covariance matrices of each multivariate Gaussian mixture component k. These parameters are learned through traditional maximum likelihood EM methods (Dempster et xal., 1977). The covariance matrix is assumed to be diagonal.

Estimation of the statistics of degraded speech. We assume that the effect of the environment on the statistics of speech can be accurately modeled by applying the proper correction factors to the mean vectors and covariance matrices. If stereo data are not available, we obtain as solutions Eqs. (25) and (26) with $P[s_t(k)|\boldsymbol{y}_t, \phi]$ equal to $P[k|\boldsymbol{y}_t, \phi]$, since in this case the PDF for the features of clean speech is modeled by a mixture of multivariate Gaussian distributions.

The term $P[k|\boldsymbol{y}_t, \phi]$ represents the a posteriori probability of an observed degraded vector $\boldsymbol{y_t}$ being produced by Gaussian k. The solutions are iterative and each iteration guarantees that the likelihood of the data does not decrease.

If stereo data are available we obtain as solutions Eqs. (31) and (32) with $P[s_t(k)|\boldsymbol{x}_t]$ equal to $P[k|\boldsymbol{x}_t]$ since, as before, the PDF for the features of clean speech is modeled by a mixture of multivariate Gaussian distributions. In this case the solutions are non-iterative.

Compensation of degraded speech. The solution for the correction factors $\{r_1, \ldots, \boldsymbol{r}_K, R_1, \ldots, \boldsymbol{R}_K\}$ helps us learn the new distributions of cepstral vectors of degraded speech. With this knowledge we can estimate the best correction to apply to each incoming degraded vector y to obtain an estimated clean vector $\hat{\boldsymbol{x}}$. To do so we use a minimum mean-squared error (MMSE) estimator

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = E(\boldsymbol{x}|\boldsymbol{y}) = \int_{\boldsymbol{X}} xp(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}. \tag{34}$$

Because this equation requires knowledge of the marginal distribution $p(\boldsymbol{x}|\boldsymbol{y})$, and because closed-form solutions are frequently difficult or impossible to obtain, some simplifications are needed. We model the effect of the environment on clean speech as an additive factor $\boldsymbol{r}(\boldsymbol{x})$ to the clean vector $\boldsymbol{x}$ to give the degraded vector $\boldsymbol{y}$. Consequently, the clean vector $\boldsymbol{x}$ can be obtained from the degraded vector $\boldsymbol{y}$ as $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{r}(\boldsymbol{x})$. In this case Eq. (34) simplifies to

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathrm{MMSE}} &= \boldsymbol{y} - \int_{\boldsymbol{X}} \boldsymbol{r}(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{y} - \int_{\boldsymbol{X}} \sum_{k=1}^{K} \boldsymbol{r}(\boldsymbol{x}) p(\boldsymbol{x}, k|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{y} - \sum_{k=1}^{K} P[k|\boldsymbol{y}] \int_{\boldsymbol{X}} \boldsymbol{r}(\boldsymbol{x}) p(\boldsymbol{x}|k, \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \\ &\cong \boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{r}_k P[k|\boldsymbol{y}] \int_{\boldsymbol{X}} p(\boldsymbol{x}|k, \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \\ &\cong \boldsymbol{y} - \sum_{k=1}^{K} \boldsymbol{r}_k P[k|\boldsymbol{y}], \end{aligned} \tag{35}$$

where we have further simplified the $\boldsymbol{r}(\boldsymbol{x})$ expression to $\boldsymbol{r}_k$. This is equivalent to assuming that the $\boldsymbol{r}(\boldsymbol{x})$ term can be well approximated by a constant value within the region in which $p(\boldsymbol{x}|k,\boldsymbol{y})$ has a significant value.

4.2. Experimental results

In this section we describe several experiments designed to evaluate the performance of the RATZ family of algorithms. We explore several of the dimensions of the algorithms including the impact of the number of adaptation sentences on recognition accuracy and the optimal number of Gaussian mixtures.

The experiments described here were performed using the 5000-word Wall Street Journal (WSJ) 1993 database (Paul and Baker, 1992). We use the speaker-independent training corpus, referred to as WSJ0- si_trn, which contains 7240 utterances from 84 speakers reading sentences from the WSJ. These sentences were recorded using a Sennheiser close-talking noise-cancelling headset. The training utterances were collected from 84 speakers. These data are used to build a single set of gender-independent HMMs to train the SPHINX-II system.

The testing set contained 215 sentences with a total number of 4066 words belonging to 10 different native speakers of American English, again reading sentences from the WSJ. These sentences were recorded by NIST using a Sennheiser HMD-410 or HMD-414 close-talking, headset-mounted noise-cancelling microphone. These data were contaminated by us with additive white Gaussian noise at several SNRs. Sentences were corrupted by adding noise scaled on a sentence-by-sentence basis to an average power value computed to produce the required SNR.

For all the experiments using this database, the upper dotted line in the figures represents the performance of the system when fully trained on degraded data, which is a reasonable estimate of the best possible performance to be expected for a particular recognition system at a given SNR. The lower dotted line represents the performance of the system when no compensation other than baseline cepstral mean normalization (CMN) is applied.

The SPHINX-II speech recognition engine was used for our experiments. SPHINX-II is a large-vocabulary, speaker-independent, semicontinuous Hidden Markov Model (SC-HMM)-based continuous speech recognition system and was one of the first systems to demonstrate the feasibility of accurate, speaker-independent, large-vocabulary continuous speech recognition (Huang et al., 1991). The feature set used in all the experiments described in this paper was composed of four streams. The first stream contains 12 mel scaled cepstrum components. The second and third streams contain the cepstrum time derivatives and the second time derivatives respectively. The final stream contains the logarithm of the frame energy, its time derivative and its second time derivative.

4.2.1. Effect of the number of adaptation sentences

Fig. 4 describes recognition accuracy of the RATZ algorithm as a function of the number of

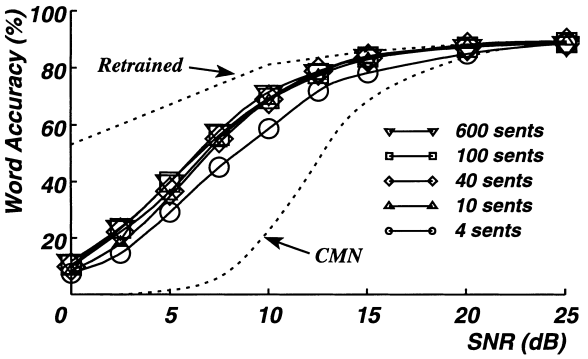

Fig. 4. The effect of the number of adaptation sentences on recognition accuracy obtained using an implementation of the RATZ algorithm with 128 Gaussians.

adaptation sentences. It can be seen that the RATZ algorithm is able to compensate for the effect of the environment even with a very small number of adaptation sentences. In fact, the performance appears to be quite insensitive to the number of adaptation sentences except when the number of sentences becomes less than 10. As in many other learning problems, the number of parameters that can be "adequately" observed depends on the amount of training data available. In this case it appears that 10 sentences are adequate. Perhaps some kind of clustering technique, such as the techniques used in MLLR (Leggeter and Woodland, 1995) where the parameters to be learned are tied could help in enabling RATZ to become more effective with less than 10 sentences.

4.2.2. Effect of the number of Gaussian mixtures

Fig. 5 describes the effect that the number of Gaussian mixtures has on recognition accuracy obtained using RATZ. Several configurations of the algorithm with 256, 64, and 16 Gaussian mixture components were implemented with correction factors learned from 100 stereo adaptation sentences for this study. Recognition accuracy is seen to increase as the number of Gaussian mixtures increases, particularly at lower SNRs.

4.2.3. Stereo based RATZ vs. blind based RATZ

Fig. 6 compares the effect of having stereo data to learn the correction factors on the recognition accuracy. 256 Gaussian mixtures were used in implementing both algorithms, RATZ and blind

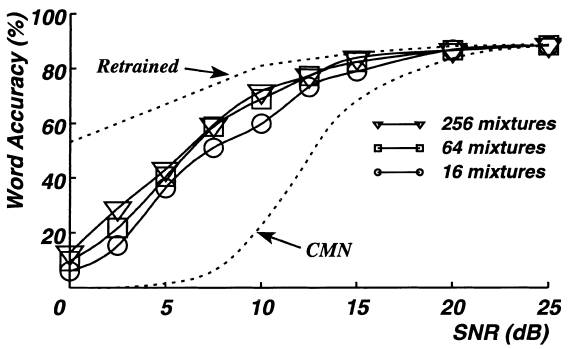

Fig. 5. The effect of the number of Gaussians on recognition accuracy of the RATZ algorithms. Results were obtained using 100 adaptation sentences.

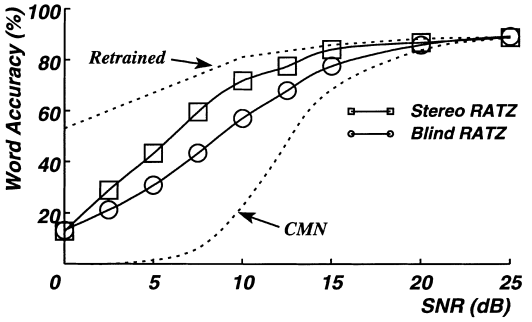

Fig. 6. Comparison of recognition accuracy obtained using stereo-based implementations of RATZ vs. the blind RATZ implementation. Results were obtained using 256 Gaussians and 100 adaptation sentences.

RATZ. The adaptation set consisted of the same 100 sentences that were used in our previous experiments. In the case of RATZ the stereo pairs (clean and noisy) set was used. Not surprisingly, the absence of stereo data is detrimental to recognition accuracy. Nevertheless, even blind RATZ provides considerable benefits when compared to not performing any compensation at all (the CMN case).

4.2.4. Comparisons with FCDCN

The impact of the novel aspects of the RATZ algorithms can be evaluated by comparison to the FCDCN family of algorithms introduced by Acero (1991) and studied further by Liu (Liu et al., 1994). FCDCN can be considered to be a particular case of the RATZ algorithms where a VQ codebook is used to represent the statistics of clean speech and the effect of the environment on this codebook is modeled by shifts in the centroids of the codebook. In FCDCN, additive corrections are associated with each of the centroids in the codebook. All centroids in the codebook are assumed to have the same variance. Compensation is performed on each observation vector by subtracting the correction factor that minimizes VQ distortion.

Fig. 7 compares recognition accuracy using RATZ with the corresponding results obtained using FCDCN, with each system implemented using 256 Gaussians. The same set of sentences were used to learn the statistics or VQ codebook of clean speech and the same 100 stereo sentences were used to learn the correction factors. As can be seen, the RATZ algorithm outperforms FCDCN at all SNRs, and the improvement in recognition accuracy is greater at lower SNRs. The improvement in recognition accuracy obtained using RATZ compared to FCDCN is a consequence of the more detailed statistical representation used by RATZ to express the distributions of clean speech, and the more detailed model used by RATZ to represent the effect of the environment on the distributions of clean speech. While FCDCN models the clean data as a VQ codebook, consequently representing the space with only a set of codewords, RATZ models it as a Gaussian mixture PDF, where the variance of the clean speech is also modeled. Additionally, the effect of the environment on the variance is also modeled, something that is not possible with the representation used by FCDCN.

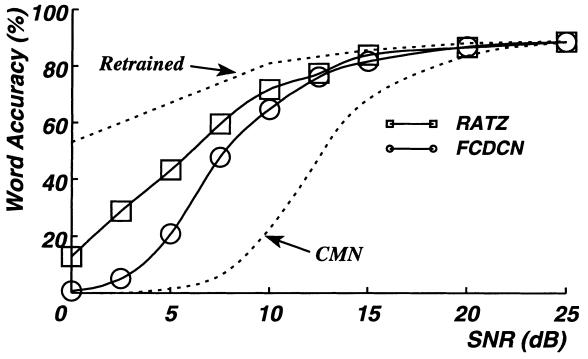

Fig. 7. Comparison of recognition accuracy obtained using the RATZ algorithm with recognition accuracy obtained using the FCDCN algorithm. Both algorithms used a similar configuration with 256 Gaussians learned with the same stereo data. The identity of the environment was also presumed to be known.

5. The STAR family of algorithms

In this section we particularize the general solutions described in Section 3 for the case of the STAtistical Reestimation (STAR) family of algorithms. While the RATZ algorithms apply correction factors to the incoming cepstral vectors of degraded speech, the STAR algorithms modify some of the parameters of the acoustical distributions in the HMM structure.

We present an overview of the STAR algorithm and describe in detail all the steps followed in STAR based compensation. We then provide some experimental results on several databases and environmental conditions.

5.1. Overview of STAR and blind STAR

The concept of data-driven algorithms that adapt HMMs to new environments has been introduced before. For example, the tied-mixture normalization algorithm proposed by Anastasakos et al. (1994) and the Dual Channel Codebook Adaptation (DCCA) algorithm proposed by Liu

(Stern et al., 1994) are similar in spirit to STAR. However, these previous algorithms are based on VQ indices rather than on Gaussian a posteriori probabilities, and they use a weaker model of the effect of the environment on Gaussian distributions in which the covariance matrices are not corrected. Furthermore, they only model the effect of the environment on cepstrum distributions without modeling the effect on the other feature streams such as delta cepstra, double delta cepstra and energy.

The STAR algorithm works in two stages: (1) estimation of the statistics of clean speech, and (2) estimation of the statistics of degraded speech. A formal "compensation" stage is not necessary because the original cepstral vectors of the degraded speech are used for recognition.

Estimation of the statistics of clean speech. The STAR algorithm makes use of the acoustical distributions modeled by the HMMs to represent clean speech. (Strictly speaking this is not a step related to the STAR algorithm.) These distributions, as modeled by HMMs, are mixtures of multivariate Gaussians. Under these assumptions the distribution for clean speech can be written as

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} a_k(t) N_x(\mu_{x,k}, \Sigma_{x,k}), \tag{36}$$

where the $a_k(t)$ term represents the a priori probability of each of the Gaussians of each of the possible states. The $\mu_{x,k}$ and $\Sigma_{x,k}$ terms represent the mean vector and covariance matrix of each multivariate Gaussian mixture element k. These parameters are learned through the well-known Baum–Welch algorithm (Baum, 1972).

Estimation of the statistics of degraded speech. As mentioned in Section 3, we assume that the effect of the environment on the distributions of speech cepstra can be modeled adequately by applying the proper correction factors to the mean vectors and covariance matrices. Therefore, our goal will be to compute these correction factors to estimate the statistics of degraded speech.

If stereo data are not available, we make use of Eqs. (25) and (26) to obtain the compensation parameters. In this case the term $P[s_t(k)|\boldsymbol{y}_t, \phi]$ represent the a posteriori probability of an observation degraded vector $\boldsymbol{y}_t$ being produced by Gaussian k in state $s_t(k)$ given the set of estimated correction parameters ϕ. The solutions are iterative and each iteration guarantees that the likelihood of the adaptation data does not decrease. Notice that in this case the solutions are very similar to the Baum–Welch reestimation solutions commonly used for HMM training (Baum, 1972; Huang et al., 1993).

If stereo data are available we make use of Eqs. (31) and (32). The solutions to these equations are non-iterative. Note that in the stereo case the substitution of $P[s_t(k)|\boldsymbol{y}_t]$ by $P[s_t(k)|\boldsymbol{x}_t]$ assumes implicitly that the a posteriori probabilities do not change due to the environment.

In addition to the cepstra, most HMM-based speech recognition systems use the difference cepstra and the double difference cepstra as parameters. The STAR family of algorithms assumes that all of these streams will be affected by the environment by an additive factor to the means and variances. Even though we do not present evidence to support this assumption, our experimental results as well as results by Gales (1995) support this assumption. We estimate each of the additional correction factors using formulae that are equivalent to those used to estimate the correction factors for the cepstral stream.

Once the correction factors are estimated we can perform recognition using the distributions of degraded speech estimated as distributions of clean speech corrected by the appropriate factors.

5.2. Experimental results

We performed several experiments to evaluate the performance of the STAR family of algorithms. Several dimensions of the algorithm were explored, including the impact of the number of adaptation sentences on speech recognition accuracy and the effect of the availability of stereo data to learn the correction factors. We also compare the recognition accuracy obtained using the STAR algorithms with the corresponding accuracy rates obtained using RATZ. The procedures and databases used in these experiments are the same as had been described in Section 4.2.

5.2.1. Effect of the number of adaptation sentences

Fig. 8 shows the effect of the number of adaptation sentences on the recognition accuracy obtained using STAR as a function of SNR. These experiments were performed using the semi-continuous HMM-based SPHINX-II system with 256 Gaussian mixtures. The STAR algorithm appears to capture all the needed information for environmental adaptation with only 40 sentences.

These results are in agreement with the results we obtained with the RATZ algorithm. Both algorithms require a minimum of 40 utterances to learn the correction factors. This can be explained considering that the number of parameters to be learned is not very large. Only 256 correction factors for the mean vectors and diagonal covariances.

5.2.2. Stereo vs. non-stereo adaptation databases

When stereo data are unavailable, the correction factors must be learned iteratively. We explored different alternatives to bootstrap the iterative learning procedure. Fig. 9 compares recognition accuracy for the original STAR and blind STAR algorithms obtained using 100 adaptation sentences. Ten iterations of the reestimation formulae were used for the blind STAR experiments.

To explore the effect that the initial parameter values have on the blind STAR algorithm, we initialized the reestimation procedure in two ways: (1) with initial correction factors set equal to zero, and (2) with the initial correction factors set equal to the actual optimal factors that are learned for an environment 5 dB higher in SNR using stereo

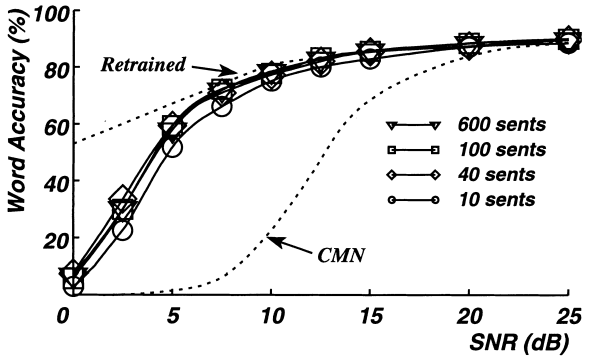

Fig. 8. Effect of the number of adaptation sentences used to learn the correction factors on the recognition accuracy obtained using the STAR algorithm. Stereo data was used to learn the correction factors.

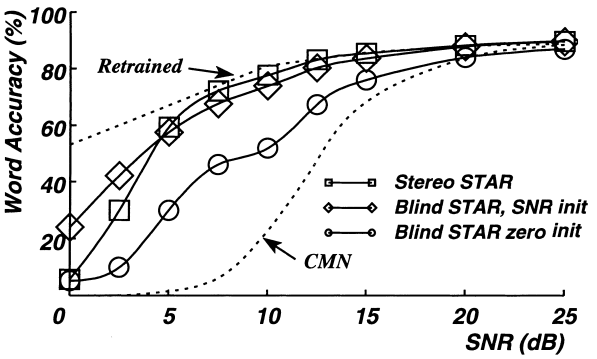

Fig. 9. Comparison of recognition accuracy obtained using the blind STAR, and stereo-based STAR algorithms. The line with star symbols represents the original stereo based STAR algorithm. The line with diamond symbols represents the blind STAR algorithm bootstrapped from the distributions that were closest with respect to SNR to the SNR of the test environment.

data. For example, to learn the correction factors for 25 dB data we would use the correction factors learned with stereo data at 30 dB. This process is repeated up to 0 dB. The goal of this experiment on different initializations was to explore the sensitivity of blind STAR to initial conditions. We observe that the performance of the blind STAR algorithm depends a great deal on which distributions are chosen initially. While the zero initialization gives us some degree of improvement in recognition accuracy compared to CMN, a better initial guess of the correction factors improves the performance of blind STAR to almost the level of stereo STAR. In fact, at the lowest SNRs the blind-STAR actually outperforms the stereo-based STAR, perhaps because the assumption that the a posteriori probabilities $P[s_t(k)|\boldsymbol{y}_t]$ can be replaced by $P[s_t(k)|\boldsymbol{x}_t]$ is not valid at lower SNRs. This dependency on initial conditions is a serious limitation of the blind STAR algorithm that needs to be addressed.

5.2.3. Comparison of STAR with RATZ

Fig. 10 compares the recognition accuracy obtained using the STAR and RATZ algorithms each in both the blind and stereo-based implementations. These results were all obtained using 100 adaptation sentences, an equivalent availability of stereo data to learn the correction factors, and comparable recognition systems with 256 Gaussian mixtures. In general, the STAR algorithms always outperform the RATZ algorithms. The STAR algorithm is able to produce almost the same performance of a fully retrained system for SNRs as low as 5 dB. For lower SNRs the a posteriori invariance assumption made by the algorithm is not appropriate and recognition accuracy suffers.

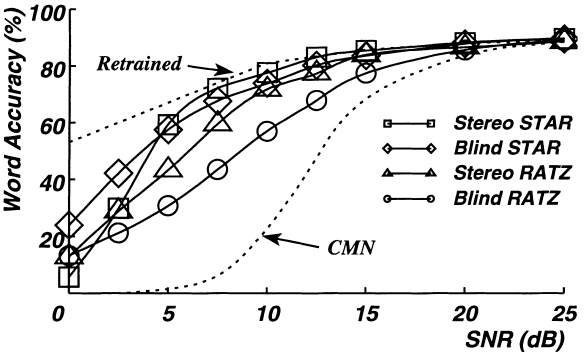

Fig. 10. Comparison of recognition accuracy obtained using blind and stereo configurations of the STAR and RATZ algorithms. The same 100 sentences were used for all adaptation algorithms.

It can be shown that approaches that adapt the HMMs to the environmental conditions of the speech in the testing environment approach the structure of an optimal classifier and should provide improved recognition accuracy performance for any class of techniques. Several authors (Moreno, 1996; Mokbel and Chollet, 1995) provide explanations for this phenomenon. Hence, we expect the STAR technique, which modifies the classifier, to outperform the RATZ technique, which modifies the data. Given equivalent experimental conditions, experimental results bear this hypothesis out.

6. Summary and conclusions

This paper addresses the problem of data-driven environmental robustness algorithms. Starting with a study of the effects of the environment on speech distributions we proposed a mathematical framework based on the EM algorithm for environment compensation. Two generic data-driven approaches have been proposed. The first approach modifies incoming cepstral vectors while the second one modifies the mean vectors and covariances matrices of the acoustical distributions of the statistical representation of the cepstra of clean speech developed by the HMMs. We have shown how both approaches can be derived within a common unified mathematical framework.

We performed a series of simulations using artificially-corrupted data to study in a controlled manner to study the effects of the environment on speech-like log spectral distributions. From these simulations we observed that while the distributions of log spectra of speech are no longer Gaussian when submitted to additive noise and linear channel distortions the changes can be well captured by simple additive correction terms in

the means and variances. Motivated by these observations we modeled the effects of the environment on Gaussian speech distributions as correction factors to be applied to the mean vectors and covariance matrices.

We developed two families of algorithms for data-driven environmental compensation. The first set of algorithms (referred to as the RATZ algorithms) specifies the correction factors to be applied to the incoming vector representation of degraded speech. The second family of algorithms (referred to as the STAR algorithms) provided corrections to the speech distributions that represent degraded speech in the HMM classifier. The values of the correction factors were learned in two different ways: using simultaneously-recorded clean and degraded speech databases ("stereo" databases) to learn correction factors directly from data (stereo RATZ and STAR), and iteratively learning the correction factors directly from the degraded data alone (blind RATZ and STAR). We presented a unified framework for the RATZ and STAR algorithms, showing that techniques that attempt to modify the incoming cepstra vectors and techniques that modify the parameters of the distributions of the HMMs can be described by the same theory. The STAR techniques that modify the parameters of the distributions that internally represent speech generally outperform the RATZ techniques that modify the incoming cepstral vectors of degraded speech, given equivalent experimental conditions.

We have also shown that these data-driven compensation techniques perform quite well even with only ten sentences of adaptation data. When comparing the proposed algorithms with the performance of a fully retrained system we observe that they can provide very similar recognition accuracy for SNRs as low as 15 dB for the RATZ family of algorithms and SNRs as low as 5 dB for the STAR family of algorithms.

Appendix A. EM solutions for the correction factors $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$

In this appendix, we provide detailed solutions for the correction terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$. Given a log likelihood function $L(Y)$:

$$L(\boldsymbol{Y}) = \log(l(\boldsymbol{Y})) = \sum_{t=1}^{T} \log(p(y_t)) = \sum_{t=1}^{T} \log\left(\sum_k a_k(t) N_y(\boldsymbol{r}_k + \mu_{x,k}, \boldsymbol{R}_k + \Sigma_{x,k})\right). \tag{A.1}$$

Our goal is to find the complete set of K terms $\boldsymbol{r_k}$ and $\boldsymbol{R_k}$ that maximize the likelihood (or log likelihood). As it turns out there is no direct solution to this problem and some indirect method is necessary. The Expectation-Maximization (EM) algorithm is one of this methods.

The EM algorithm defines a new auxiliary function $Q(\phi, \overline{\phi})$ as

$$Q(\phi, \overline{\phi}) = E[L(\boldsymbol{Y}, S|\overline{\phi})|\boldsymbol{Y}, \phi], \tag{A.2}$$

where the ($\boldsymbol{Y}$,S) pair represent the *complete* data, composed of the *observed* data $\boldsymbol{Y}$ (the noisy vectors) and the *unobserved* data S (indicating which Gaussian/state produced an observed data vector). This equation can be easily related to the Baum–Welch equations used in Hidden Markov Modelling. The ϕ symbol represents the set of parameters (K correction vectors and K correction matrices) that maximize the observed data

$$\phi = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_k\}. \tag{A.3}$$

The $\overline{\phi}$ symbol represents the same set of parameters as ϕ but with different values. The basis of the EM algorithm lies in the fact that given two sets of parameters, ϕ and $\overline{\phi}$, if $Q(\phi, \overline{\phi}) \geqslant Q(\phi, \phi)$, then $L(Y, \overline{\phi}) \geqslant L(Y, \phi)$. In other words, maximizing $Q(\phi, \overline{\phi})$ with respect to the parameters ϕ is guaranteed not to decrease the likelihood $L(\boldsymbol{Y}, \overline{\phi})$.

If the unobserved data S are represented by the mixture index K, Eq. (A.2) can be expanded as

$$Q(\phi, \overline{\phi}) = E[L(\boldsymbol{Y}, S|\overline{\phi})|\boldsymbol{Y}, \phi] = \sum_{t=1}^{T}\sum_{k=1}^{K} \frac{p(\boldsymbol{y}_t, s_t(k)|\phi)}{p(\boldsymbol{y}_t|\phi)} \log(p(\boldsymbol{y}_t, s_t(k)|\overline{\phi})); \tag{A.4}$$

hence

$$
\begin{aligned}
&Q(\phi,\overline{\phi})\\
&=\sum_{t=1}^{T}\sum_{k=1}^{K}P[s_t(k)|\boldsymbol{y}_t,\phi]\{\log a_k(t)\\
&\quad-\frac{D}{2}\log(2\pi)-\frac{D}{2}\log|\overline{\boldsymbol{R}_k}+\Sigma_{x,k}|-\frac{1}{2}(\boldsymbol{y}_t-\mu_{x,k}\\
&\quad-\overline{\boldsymbol{r}_k})^{\mathrm{T}}(\Sigma_{x,k}+\overline{\boldsymbol{R}_k})^{-1}(y_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})\},
\end{aligned}
\tag{A.5}
$$

where D is the dimensionality of the cepstrum vector. The expression can be further simplified to:

$$
\begin{aligned}
&Q(\phi,\overline{\phi})=\text{constant}\\
&\quad+\sum_{t=1}^{T}\sum_{k=1}^{K}P[s_t(k)|\boldsymbol{y}_t,\phi]\Big\{-\frac{D}{2}\log|\overline{\boldsymbol{R}_k}+\Sigma_{x,k}|\\
&\quad-\frac{1}{2}(\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})(\Sigma_{x,k}+\overline{\boldsymbol{R}_k})^{-1}(\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})\Big\}.
\end{aligned}
\tag{A.6}
$$

To find the ϕ parameters we simply take derivatives and set equal to zero. After some manipulation we obtain

$$
\begin{aligned}
\nabla_{\overline{\boldsymbol{r}_k}}Q(\phi,\overline{\phi})&=\sum_{t=1}^{T}P[s_t(k)|\boldsymbol{y}_t,\phi](\Sigma_{x,k}\\
&\quad+\overline{\boldsymbol{R}_k})^{-1}(\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})=0,\\
&\quad\nabla_{(\Sigma_{x,k}+\overline{\boldsymbol{R}_k})^{-1}}Q(\phi,\overline{\phi})\\
&=\sum_{t=1}^{T}P(s_t(k)|\boldsymbol{y}_t,\phi)\Big\{(\Sigma_{x,k}+\overline{\boldsymbol{R}_k})-(\boldsymbol{y}_t-\mu_{x,k}\\
&\quad-\overline{\boldsymbol{r}_k})(\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})^{\mathrm{T}}\Big\};
\end{aligned}
\tag{A.7}
$$

hence

$$
\overline{\boldsymbol{r}_k}=\frac{\sum_{t=1}^{T}P[s_t(k)|\boldsymbol{y}_t,\phi]\boldsymbol{y}_t}{\sum_{t=1}^{T}P[s_t(k)|\boldsymbol{y}_t,\phi]}-\mu_{x,k},
\tag{A.8}
$$

$$
\begin{aligned}
&\overline{\boldsymbol{R}_k}\\
&=\frac{\sum_{t=1}^{T}P[s_t(k)|\boldsymbol{y}_t,\phi]((\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})(\boldsymbol{y}_t-\mu_{x,k}-\overline{\boldsymbol{r}_k})^{\mathrm{T}})}{\sum_{t=1}^{T}P[s_t(k)|\boldsymbol{y}_t,\phi]}\\
&\quad-\Sigma_{x,k}
\end{aligned}
\tag{A.9}
$$

Eqs. (A.8) and (A.9) for the basis for an iterative algorithm. The EM algorithm guarantees that each iteration does not decrease the likelihood of the observed data.

References

Acero, A., 1991. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Press, Boston, MA.

Acero, A., Stern, R.M., 1990. Environmental robustness in automatic speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, 1990, pp. 849–852.

Anastasakos, A., Kubala, F., Makhoul, J., Schwartz, R., 1994. Adaptation to new microphones using tied-mixtures normalization. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Adelaide, Australia, pp. 433–436.

Baum, L., 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. Inequalities 3, 1–8.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. Ser. B 39, 1–38.

Ephraim, Y., 1992. Statistical-model-based speech enhancement systems. Proc. IEEE 80, 1526–1555.

Flanagan, J., Johnston, J., Zahn, R., Elko, G., 1985. Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. 78, 1508–1518.

Gales, M.J.F., 1995. Model-Based Techniques for Noise Robust Speech Recognition. Ph.D. Thesis, Engineering Department, Cambridge University, Cambridge, UK.

Gales, M.J.F., Young, S.J., 1993. Cepstral parameter compensation for HMM recognition in noise. Speech Commun. 12, 231–239.

Ghitza, O., 1986. Auditory nerve representation as a front-end for speech recognition in a noisy environment. Comput. Speech Language 1, 109–130.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87, 1738–1752.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Huang, X., Ariki, Y., Jack, M.A., 1993. Hidden Markov Models for Speech Recognition. Edinburgh University Press, Edinburgh, UK.

Huang, X., Lee, K.-F., Hon, H.-W., Hwang, M.-Y., 1991. Improved acoustic modeling with the SPHINX recognition system. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Toronto, Canada, pp. 235–238.

Juang, B., 1991. Speech recognition in adverse environments. Comput. Speech Language 5, 275–294.

Junqua, J.-C., Haton, J.-P., 1996. Robustness in Automatic Speech Recognition. Kluwer Academic Press, Boston, MA.

Lee, K.-F., Hon, H.-W., Reddy, R., 1990. An overview of the SPHINX speech recognition. IEEE Trans. Acoust. Speech Signal Process. 38 (1), 35–45.

Leggeter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Language 9 (2), 171–185.

Liu, F.-H., Stern, R.M., Acero, A., Moreno, P.J., 1994. Environment normalization for robust speech recognition using direct cepstral comparison. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Adelaide, Australia, pp. 61–64.

Mokbel, C., Chollet, G.F.A., 1995. Automatic word recognition in cars. IEEE Trans. Speech Audio Process. 3(5), 346–356.

Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.

Moreno, P.J., Raj, B., Gouvea, E., Stern, R.M., 1995a. Multivariate Gaussian-based cepstral normalization. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Detroit, MI, pp. 137–140.

Moreno, P.J., Raj, B., Stern, R.M., 1995b. A unified approach to robust speech recognition. In: Proceedings of Eurospeech 1995, vol. 1, Madrid, Spain, pp. 481–484.

Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Atlanta, GA, pp. 733–736.

Neumeyer, L., Weintraub, M., 1994. Probabilistic optimum filtering for robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, Adelaide, Australia, pp. 417–420.

Paul, D., Baker, J., 1992. The design of the wall street journal-based CSR corpus. In: Proceedings of ARPA Speech and Natural Language Workshop, pp. 357–362.

Sankar, A., Lee, C.-H., 1995. Robust speech recognition based on stochastic matching. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Detroit, MI, pp. 121–124.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. J. Phonetics 16, 55–76.

Stern, R.M., Liu, F.-H., Moreno, P.J., Acero, A., 1994. Signal processing for robust speech recognition. In: Proceedings of International Conference on Spoken Language Processing, vol. 3, pp. 1027–1030.

Stern, R.M., Acero, A., Liu, F.-H., Ohshima, Y., 1996. Signal processing for robust speech recognition. In: Lee, C.-H., Soong, F., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, Boston, MA, pp. 351–378.

Sullivan, T.M., Stern, R.M., 1993. Multi-microphone correlation-based processing for robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Minneapolis, MN, pp. 91–94.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, Albuquerque, NM, pp. 845–848.