FACTOR ANALYSIS INVARIANT TO LINEAR TRANSFORMATIONS OF DATA

R. A. Gopinath, B. Ramabhadran and S. Dharanipragada

IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email:rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

Modeling data with Gaussian distributions is an important statistical problem. To obtain robust models one imposes constraints the means and covariances of these distributions [6, 4, 10, 8]. Constrained ML modeling implies the existence of optimal feature spaces where the constraints are more valid [2, 3]. This paper introduces one such constrained ML modeling technique called *factor analysis invariant to linear transformations* (FACILT) which is essentially factor analysis in optimal feature spaces. FACILT is a generalization of several existing methods for modeling covariances. This paper presents an EM algorithm for FACILT modeling.

1. INTRODUCTION

In Gaussian Modeling the model parameters (means and covariances) are usually estimated using the Maximum Likelihood (ML) principle. In many applications due to data-insufficiency, computational and/or storage considerations, one has to constrain the means and covariances so that there are fewer parameters to estimate (e.g., diagonal covariances, reduced-rank means, shared covariances etc.). In such cases it is desirable to model in a feature space in which the constraints are maximally satisfied. This paper introduces one such constrained ML modeling technique called factor analyzed covariances invariant to linear transformations (FACILT). FACILT is a direct generalization of both standard factor analysis and semi-tied covariance modeling. This paper presents an EM algorithm for FACILT parameter estimation and gives results of speech recognition experiments on two Large Vocabulary Continuous Speech Recognition tasks: the first a dictation task on an IBM internal database and the second on a Voicemail transcription task on a Voicemail database distributed by LDC.

2. DATA MODELING

The basic problem considered here is obtaining a statistical model of observed data where each sample is independent of any other and drawn from a finite set S of Gaussian distributions. Each sample is associated with a single Gaussian or state $s \in S$ viz., the observations are conditionally independent given the state sequence. If $x_1, x_2, \ldots, x_T \stackrel{\Delta}{=} \mathbf{x}$, $x_t \in \mathbb{R}^d$, and $s_1, s_2, \ldots, s_T \stackrel{\Delta}{=} \mathbf{s}$ are the observation and state sequences, this corresponds to the following model for the data:

$$p(\mathbf{x}, \mathbf{s}) = p(\mathbf{s})p(\mathbf{x}|\mathbf{s}) = p(\mathbf{s}) \prod_{t=1}^{T} p(x_t|s_t).$$

Such models occur often in practice. For example, if the state-sequence is iid this corresponds to a Gaussian mixture model of the data; if the state sequence is Markovian this corresponds to a Hidden Markov Model (HMM) of the data with Gaussian observations. This description also includes HMMs with Gaussian mixture observations if the underlying state-sequence is the Gaussian mixture component sequence. Whatever be the underlying $p(\mathbf{s})$, the $p(\mathbf{x}, \mathbf{s})$ is then completely described by state means (μ_s) and covariances (Σ_s). In practical applications the estimation of the means and covariances are constrained. Well-known examples are Linear Discriminant Analysis (where means are constrained to lie in a lower dimensional space viz., $Span(\{\mu_s\}) = k \leq d$) and Diagonal Covariance (DC) Modeling (where covariances are constrained to be diagonal viz., $\Sigma_s = D_s$, D_s diagonal). The constraints on the covariance are typically used because using a Full Covariance (FC) Model is often not warranted. Another example is the DCILT (diagonal covariances invariant to linear transformations) Model or semi-tied covariances model where the covariances are constrained to be of the form $\Sigma = A_s D_s A_s'$, with A_s typically shared by several Gaussians [4, 2, 3]. This corresponds to transforming the data to an optimal feature space (in a state-dependent manner) using A_s and modeling using the diagonal covariance D_s in this space.

3. FACTOR ANALYSIS

Often in speech recognition the DC model is used even though it is known a priori that the dimensions of the sample x_t are known to be correlated (e.g., cepstral features). Factor analysis is one approach to add more flexibility to the DC Model to capture these correlations with few parameters. The basic idea in a Factor Analyzed Covariance (FAC) Model is to estimate covariances of the form

$$\Sigma = \Lambda\Lambda' + \Psi,$$

where Ψ is a diagonal matrix and Λ is a rectangular matrix with typically much fewer columns than rows. Λ is referred to as the *factor loading matrix* and each column of Λ is referred to as a *factor*. If Λ is zero the FAC Model reduces to a DC Model. If Λ has k factors then the FAC Model roughly tries to model the off-diagonal terms in Σ using a rank k matrix ($\Lambda\Lambda'$). The FAC model also corresponds to an additive decomposition of the data into independent components - a *communality* component and a *uniqueness (independent variance)* component: $x = c + u$, where c has zero mean and covariance $\Lambda\Lambda'$, while u is distributed $N(\mu, \Psi)$. The hope is that with very few factors ($k \ll d$) a very good model of the covariance of the underlying Gaussian distribution is obtained. The communality component can be viewed as being generated by an underlying zero-mean unit covariance process. In other words,

$$x_t = \Lambda_{s_t} z_t + u_{s_t},$$

where z_t is distributed $N(0, I)$ and u_s is distributed $N(\mu_s, \Psi_s)$. This viewpoint and the realization that the z_t's

can be considered latent variables leads one to an EM algorithm for FAC Model parameter estimation [1, 7, 8]. The FAC Model was recently applied for modeling the HMM states for speech recognition [9, 10].

4. FACILT MODEL

Factor analysis is a special case of constrained ML modeling with Gaussian distributions. Wherever there are constraints it is important check if the constraints are invariant to linear transformation of the data. If the constrains are *not* invariant to linear transformations, then one can find an optimal linear transformation of the data so that the constraints are optimally satisfied to the extent possible in the transformed space [2]. Clearly the FAC Model is not invariant to linear transformations of the data since after a linear transformation the uniqueness component will not have a diagonal covariance. The FACILT model corresponds to optimally transforming the data (possibly in a state dependent manner) prior to modeling using a FAC Model.

In FACILT covariances are constrained to be of the form:

$$\Sigma = \Lambda\Lambda' + A^{-1}\Psi(A^{-1})',$$

where Λ is the factor loading matrix, Ψ is the diagonal uniqueness matrix, and A is the feature transformation matrix. This corresponds to having a FAC Model of linearly transformed data. Indeed if data from each state is transformed using a matrix A_s then the FAC model corresponds to the decomposition:

$$A_{s_t}x_t = \Lambda_{s_t}z_t + u_{s_t},$$

or equivalently (by redefining Λ_s to be $A_s^{-1}\Lambda_s$, $s \in S$) modeling the original data using the model

$$x_t = \Lambda_{s_t}z_t + A_{s_t}^{-1}u_{s_t},$$

which gives rise to covariances of the form $\Sigma_s = \Lambda_s\Lambda_s' + A_s^{-1}\Psi_s A_s^{-T}$. Thus FACILT is a generalization of both factor analysis (where $A_s = I$) and semi-tied covariance modeling (where $\Lambda_s = 0$).

It turns out that in FACILT modeling, data drawn from a set of Gaussians with covariances constrained as above can be represented in the following fashion

$$x_t = \Lambda_{s_t}z_t + \mu_s + A_{s_t}u_{s_t},$$

where x_t comes from Gaussian s_t at time t and z_t is an unobserved Gaussian random variable distributed as $N(0, I)$ and u_{s_t} is $N(0, \Psi_{s_t})$. In a FACILT Model the conditional distribution of x_t given $s_t \in S$ and z_t, is

$$p(x_t|z_t, s_t) = \frac{e^{-\frac{1}{2}\left[(x_t-\Lambda_{s_t}z_t-\mu_{s_t})'A_{s_t}'\Psi_{s_t}^{-1}A_{s_t}(x_t-\Lambda_{s_t}z_t-\mu_{s_t})\right]}}{\sqrt{(2\pi)^d\left|A_{s_t}^{-1}\Psi_{s_t}(A_{s_t}^{-1})'\right|}} \tag{1}$$

Considering s_t and z_t as latent variables (for all t) we obtain an EM algorithm for ML estimation of FACILT parameters viz. $(\Lambda_s, \Psi_s, \mu_s, A_s)$.

If each Gaussian has its own (Λ_s, Ψ_s, A_s), then the ML estimates are trivially seen to be $(0, E_s, V_s')$ where $V_sE_sV_s'$ is the eigendecomposition of the sample covariance of data from state s. However, if $A_s = I$, then there exists non-trivial ML solutions for Λ_s and Ψ_s - this is akin to standard factor analysis for a single Gaussian variable. More generally, if A_s is shared then there is no trivial solution. The main contribution of this paper is an EM algorithm for this form of covariance modeling allowing for the general case where Λ_s, Ψ_s and A_s are shared independently by arbitrary disjoint collections of Gaussians or states. For the rest of the paper we assume that $S = \cup_l \mathcal{L}_l = \cup_p \mathcal{P}_p = \cup_a \mathcal{A}_a$ are *independent* partitions of S corresponding to the sharing of $\{\Lambda_s\}$, $\{\Psi_s\}$ and $\{A_s\}$ respectively. That is, $\Lambda_s = \Lambda_l$ if $s \in \mathcal{L}_l$, $\Psi_s = \Psi_p$ if $s \in \mathcal{P}_p$ and $A_s = A_a$ if $s \in \mathcal{A}_a$.

5. THE FACILT EM ALGORITHM

The goal is to maximize the likelihood of the data viz., $p(\mathbf{x})$, with respect to $(\mu_s, \Lambda_s, \Psi_s, A_s)$, the parameters in the model. The complete data for this problem is given by the triplet **(x,z,s)**; the data, the hidden factors and the underlying state-sequence. Using conditional independence of observations given hidden variables

$$p(\mathbf{x}, \mathbf{z}, \mathbf{s}) = p(\mathbf{s})p(\mathbf{z}|\mathbf{s})p(\mathbf{x}|\mathbf{z}, \mathbf{s}) = p(\mathbf{s})p(\mathbf{z})\prod_{t=1}^{T} p(x_t|z_t, s_t).$$

The Gaussian parameters depend only on $\prod_{t=1}^{T} p(x_t|z_t, s_t)$, while parameters modeling the state-sequence process are in $p(\mathbf{s})$. The posterior distribution of the latent variables is

$$p(\mathbf{z}, \mathbf{s}|\mathbf{x}) = p(\mathbf{s}|\mathbf{x})p(\mathbf{z}|\mathbf{s}, \mathbf{x}).$$

In EM one computes the Q function which is the expected value of the log likelihood of the complete data with respect to the posterior distribution on the hidden variables.

As for the Gaussian parameters it suffices to consider $p(\mathbf{x}|\mathbf{z}, \mathbf{s})$ instead of $p(\mathbf{x}, \mathbf{z}, \mathbf{s})$. If $\hat{\theta} = \left\{\hat{\mu}_s, \hat{\Lambda}_s, \hat{\Psi}_s, \hat{A}_s\right\}$ and $\theta = \{\mu_s, \Lambda_s, \Psi_s, A_s\}$ are the current and new (to be estimated) values of the parameters, then, from Eqn. 1,

$$\begin{aligned}
Q(\theta, \hat{\theta}) &= E_{\hat{\theta}}\left[\log\prod_{t=1}^{T} p_\theta(x_t|z_t, s_t)\right] \qquad (2)\\
&= E_{\hat{\theta}}\left[\log\prod_{t=1}^{T}\prod_{s\in S}\left[p_\theta(x_t|z_t, s)\right]^{\delta(s,s_t)}\right]\\
&= \sum_{t=1}^{T}\sum_{s\in S} E_{\hat{\theta}}\left[\delta(s, s_t)\log p_\theta(x_t|z_t, s)\right]\\
&= \sum_{t=1}^{T}\sum_{s\in S} E_{\hat{\theta}}\left[\delta(s, s_t)\right] E_{\hat{\theta}}\left[\log p_\theta(x_t|z_t, s)\right]\\
&= \sum_{t=1}^{T}\sum_{s\in S} \gamma_s(t) E_{\hat{\theta}}\left[\log p_\theta(x_t|z_t, s)\right],
\end{aligned}$$

where $\gamma_s(t) = p(s_t = s|\mathbf{x})$ is the posterior probability of being in state s at time t given the old value of the parameters.

EM re-estimation formulae for the parameters are obtained by setting the derivative of $Q(\theta, \hat{\theta})$ with respect to the parameters (θ) equal to zero. Solving the resulting set of simultaneous non-linear equations gives new values of the parameters. In order to express the re-estimation formulae we introduce several convenient variables. Firstly note that $E_{\hat{\theta}}\left[\log p(x_t|z_t, s)\right]$ depends on $E_{\hat{\theta}}[z_t|x_t, s]$ and $E_{\hat{\theta}}[z_tz_t'|x_t, s]$, which are given respectively by

$$E_{\hat{\theta}}[z_t|x_t, s] = \beta_s(x_t - \hat{\mu}_s), \tag{3}$$

and

$$E_{\hat{\theta}}[z_t z_t' | x_t, s] = I - \beta_s \hat{\Lambda}_s + E_{\hat{\theta}}[z_t | x_t, s] E_{\hat{\theta}}[z_t | x_t, s]', \quad (4)$$

where $\beta_s = \hat{\Lambda}_s' (\Psi_s + \hat{\Lambda}_s \hat{\Lambda}_s')^{-1}$. Now define the following statistics:

$$\langle 1_s \rangle = \sum_{t=1}^{T} \gamma_s(t). \quad (5)$$

$$\langle x_s \rangle = \sum_{t=1}^{T} \gamma_s(t) x_t. \quad (6)$$

$$\langle x_s x_s' \rangle = \sum_{t=1}^{T} \gamma_s(t) x_t x_t'. \quad (7)$$

$$\langle z_s \rangle = \sum_{t=1}^{T} \gamma_s(t) E_{\hat{\theta}}[z_t | x_t, s]. \quad (8)$$

$$\langle z_s z_s' \rangle = \sum_{t=1}^{T} \gamma_s(t) E_{\hat{\theta}}[z_t z_t' | x_t, s]. \quad (9)$$

$$\langle x_s z_s' \rangle = \sum_{t=1}^{T} \gamma_s(t) x_t E_{\hat{\theta}}[z_t' | x_t, s]. \quad (10)$$

$$\langle \frac{z_s z_s'}{\Psi_s(i,i)} \rangle = \sum_{t=1}^{T} \gamma_s(t) \frac{1}{\Psi_s(i,i)} E_{\hat{\theta}}[z_t z_t' | x_t, s]. \quad (11)$$

$$\langle \Psi_s^{-1} A_s z_s z_s' \rangle = \sum_{t=1}^{T} \gamma_s(t) \Psi_s^{-1} A_s E_{\hat{\theta}}[z_t z_t' | x_t, s]. \quad (12)$$

$$\langle A_s x_s \rangle = \sum_{t=1}^{T} \gamma_s(t) A_s x_t. \quad (13)$$

$$\langle (A_s x_s)^2 \rangle = \sum_{t=1}^{T} \gamma_s(t) (A_s x_t)^2. \quad (14)$$

$$Var(A_s x_s) = \frac{\langle (A_s x_s)^2 \rangle}{\langle 1_s \rangle} - \left(\frac{\langle A_s x_s \rangle}{\langle 1_s \rangle} \right)^2. \quad (15)$$

$$Cov(x_s) = \frac{\langle x_s x_s' \rangle}{\langle 1_s \rangle} - \frac{\langle x_s \rangle}{\langle 1_s \rangle} \frac{\langle x_s \rangle'}{\langle 1_s \rangle} \quad (16)$$

$$Cov(z_s) = \frac{\langle z_s z_s' \rangle}{\langle 1_s \rangle} - \frac{\langle z_s \rangle}{\langle 1_s \rangle} \frac{\langle z_s \rangle'}{\langle 1_s \rangle} \quad (17)$$

$$Cov(x_s z_s') = \frac{\langle x_s z_s' \rangle}{\langle 1_s \rangle} - \frac{\langle x_s \rangle}{\langle 1_s \rangle} \frac{\langle z_s \rangle'}{\langle 1_s \rangle} \quad (18)$$

$$G_a^{(i)} = \sum_{s \in \mathcal{A}_a} \frac{1}{\Psi_s(i,i)} \left[\langle 1_s \rangle (Cov(x_s) + \Lambda_s Cov(z_s) \Lambda_s') - \langle x_s z_s' \rangle \Lambda_s' - \Lambda_s \langle x_s z_s' \rangle' \right]. \quad (19)$$

The re-estimation formulae are as follows:

Λ_l estimation - Case 1 If Λ_s are not shared among states or are shared at a "lower level" than either the A's or Ψ's i.e., if for all $s \in \mathcal{L}_l$, $A_s = A_r$ for some r and $\Psi_s = \Psi_p$ for some p then

$$\Lambda_l = \left(\sum_{s \in \mathcal{L}_l} (\langle x_s z_s' \rangle - \mu_s \langle z_s' \rangle) \right) \left(\sum_{s \in \mathcal{L}_l} \langle z_s z_s' \rangle \right)^{-1}. \quad (20)$$

Λ_l estimation - Case 2 If for all $s \in \mathcal{L}_l$, $A_s = A_r$ for some r then the i^{th} row of Λ_l is given by the i^{th} row of

$$\left(\sum_{s \in \mathcal{L}_l} \langle \Psi_s^{-1} A_s x_s z_s' \rangle - \Psi_s^{-1} A_s \mu_s \langle z_s \rangle' \right) \left(\sum_{s \in \mathcal{L}_l} \left\langle \frac{z_s z_s'}{\Psi_s(i,i)} \right\rangle \right)^{-1}. \quad (21)$$

Mean (μ_s) estimation

$$\mu_s = \frac{\langle x_s \rangle - \Lambda_s \langle z_s \rangle}{\langle 1_s \rangle}. \quad (22)$$

Diagonal Variance (Ψ_s) estimation

$$\Psi_s = Var(A_s x_s) + diag(A_s Cov(z_s) A_s') - 2 diag(A_s \Lambda_s Cov(x_s z_s') A_s').$$

Optimal Feature Space (A_a) estimation The rows of A_a are obtained in an iterative fashion using the following formula:

$$\mathbf{a_i'} = \mathbf{c_i'} G_a^{(i)-1} \sqrt{\frac{\sum_{s \in \mathcal{A}_a} \langle 1_s \rangle}{\mathbf{c_i'} G_a^{(i)-1} \mathbf{c_i}}}, \quad (23)$$

where $\mathbf{c_i}$ are the cofactors of the i^{th} row of A_a.

5.1. Sufficient Statistics

Since we are estimating Gaussians the sufficient statistics for the estimation are the zeroth, first and second order statistics of the data for each Gaussian. In other words it suffices to have the statistics in Eqn. 5-Eqn. 7. If these statistics are available, then FAC Model can be readily be obtained using EM. In some problems where storage of the second order statistics is prohibitive (e.g., speech recognition), one can compute the alternate statistics Eqn. 8-Eqn. 19. However, with these statistics only one iteration of EM can be performed since the z statistics change with each iteration. Therefore, repeated alignments of the training data is required in this mode of computation. There is an evident tradeoff between computational and storage requirements.

5.2. Likelihood Computation

The likelihood computation for FAC Modeling is as computationally intensive as FC Modeling since the inverse covariance (which is full) is required in both cases. However, in a FAC Model the storage requirements can be reduced by using the Sherman-Morrison-Woodbury formula for inversion of rank-one updates of a matrix (recall the FAC covariance is a rank k update of Ψ). Indeed for a FAC Model we have

$$\Sigma_s^{-1} = \Psi_s^{-1} - \Psi_s^{-1} \Lambda_s (I + \Lambda_s' \Psi_s^{-1} \Lambda_s)^{-1} \Lambda_s' \Psi_s^{-1} \Lambda_s'. \quad (24)$$

FACILT model likelihood computation is identical but for the fact that the data has to be transformed in a class-dependent fashion prior to likelihood computation. Specifically, in the global ILT case (where there is a single global transformation) the data can be transformed a priori and a standard FAC Model used.

Sharing	k=0	k=2	k=3	k=4	k=6
Phone		12.0	12.1	12.3	12.3
Arc		12.3			
HMM State		12.3			
Baseline	12.9				

Table 1: FAC Model Word Error Rate on Read Speech Test Data: 32K Gaussians, 2.7K HMM states

Sharing	k=0	k=2	k=4	k=6
Phone		39.2	39.4	40.0
HMM State		39.6		
Baseline	48.0			

Table 2: FAC Model Word Error Rate on Spontaneous Speech Test Data: 70K Gaussians, 3K HMM states

5.3. Initialization of Parameters

The EM algorithm leads to a local maximum of likelihood and as such the solution depends on the initial value of the parameters. If we start with a DC Model a good starting point - one might suppose - is to initialize the μ_s's and Ψ_s's with the means and diagonal covariances of the DC Model and to initialize the Λ_s's to zero. Unfortunately, this is a local minimum of the likelihood function and hence a stationary point. Therefore, typically Λ_s are randomly initialized with small values. One possibility is to use Joreskog's method as suggested in [6] for standard factor analysis.

6. EXPERIMENTAL RESULTS

The first set of experiments were run on an internal test database of about 1200 words each from 10 speakers. The training data for this database is from an internal database of read speech and the acoustic models used 52 phones, 156 sub-phonetic units, about 2.7K HMM states and about 32K Gaussians. The goal was to study the effect of sharing factors at various levels of tying: phone-level, sub-phonetic unit (arc) level, and context-dependent sub-phonetic unit (or HMM state) level. The results shown in Table 6 indicate that there may be an over-training problem. The best results are obtained for the two-factor case with the least number of parameters - i.e., with maximal sharing. With sharing fixed at the phone level the number of factors were changed to 3, 4, and 6 and the results seem to degrade with number of factors - again perhaps indicating over-training. All these systems were obtained by boot-strapping from a diagonal covariance system.

To study the effectiveness of FAC Modeling on spontaneous speech data a series of experiments were conducted on a Voicemail Test Data. The training data consists of about 20 hours of Voicemail data and the test data consists of about 43 Voicemail messages each about 20 seconds duration. The acoustic models in these experiments used an augmented phone set (two additional phones added to model clicks and disfluencies), 162 sub-phonetic units, 3K HMM states and 73K Gaussians. The results are shown in Table 6.

Currently experiments are underway to study the effects of using A_s's shared at the global, phone and sub-phonetic unit levels. Also, it is fairly straightforward to decide on the number of factors in a data-driven fashion based on the relative gain in likelihood.

7. CONCLUSION

This paper introduces FACILT - a new covariance modeling technique for Gaussian distributions - and presents and EM algorithm for estimating the model parameters in FACILT. FACILT is factor analysis in optimal feature spaces and hence is a generalization of classical factor analysis. Semi-tied covariance modeling and diagonal covariance modeling can be seen as a special cases of FACILT.

8. REFERENCES

1. A. P. Dempster et al., "Maximum Likelihood from Incomplete data via the EM Algorithm", Journal of the Royal Statistical Society, 39:1-38, 1977.
2. R. A. Gopinath, "Constrained Maximum Likelihood Modeling With Gaussian Distributions", Proceedings of the DARPA Speech Recognition Workshop, Lansdowne, VA, Feb 1998. *Also available at http://www.nist.gov/speech.*
3. R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions for Classification", Proceedings of ICASSP 1998.
4. M. J. F. Gales, "Semi-tied Full-covariance matrices for hidden Markov Models", Tech. Report, CUED/FINFENG/TR287, Cambridge Univ., 1997.
5. M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Tech. Report, CUED/FINFENG/TR291, Cambridge Univ., 1997.
6. K. V. Mardia, J. T. Kent and J. M. Bibby, "Multivariate Analysis", Academic Press, 1979.
7. D. B. Rubin and D. T. Thayer, "EM Algorithms for ML Factor Analysis". Psychometrika, Vol 47, No. 1, March, 1982.
8. Z. Gharamani and G. E. Hinton, "The EM Algorithm for Mixtures of Factor Analyzers", CRG-TR-96-1, May 1997, University of Toronto.
9. M. Rahim and L. Saul, "Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Proceedings of 1997 IEEE ASRU Workshop, Santa Barbara, CA, Dec 1997.
10. L. Saul and M. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Submitted to Trans. in SP, 1997. Also an AT&T Tech. Report available from the authors: lsaul,mazin@research.att.com.
11. M. Padmanabhan et al., "Transcription of new speaking styles - Voicemail", Proceedings of the DARPA Speech Recognition Workshop, Lansdowne, VA, Feb 1998. *Also available at http://www.nist.gov/speech.*