AN EVALUATION OF A NONLINEAR FEATURE TRANSFORMATION FOR CONVERSATIONAL SPEECH RECOGNITION

Mohamed Kamal Omar *

University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Brian Kingsbury

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA

ABSTRACT

We test the nonlinear symplectic maximum-likelihood transformation (SMLT) on two large-vocabulary, conversational speech recognition tasks: IBM's Superhuman test and the DARPA 2003 Rich Transcription (RT03) test. Features in these tests are computed via linear discriminant analysis (LDA) on spliced MFCC features and subsequent transformation of the projected features using either a maximum-likelihood linear transformation (MLLT), an SMLT, or both. In contrast to previous tests of the SMLT on TIMIT phone recognition with static and delta MFCCs, these tests use a more difficult task and very different features. The four results of this work are that both LDA+MLLT and LDA+SMLT systems outperform an LDA-only system; the LDA+MLLT system outperforms the LDA+SMLT system (but the MLLT has 20 times more parameters than the SMLT); small improvements over an LDA+MLLT system are obtained with an LDA+MLLT+SMLT system on well-matched material; and no improvements are obtained using two class-dependent SMLTs in an LDA+MLLT+SMLT system.

1. INTRODUCTION

Most automatic speech recognition (ASR) systems use hidden Markov models (HMMs) with mixtures of diagonal-covariance Gaussians as the state-conditional observation densities. The use of diagonal-covariance Gaussians, which is typically justified for computational reasons or concerns about data sparsity, can degrade recognition performance [1]. Inaccurate modeling of continuous sources of inter-feature correlation is one cause of this degradation. Methods that attempt to address this problem can be divided into two major classes. The first class uses Gaussian mixture components with a richer covariance structure than diagonal covariances and reduces the number of parameters that must be estimated from training data by tying some parameters across models. An example is semi-tied covariance matrices [2]. The second class transforms the recognition features to better satisfy the constraints of the diagonal-covariance models. Examples include state-specific principal component analysis [1] and the MLLT [3].

All previous approaches assume that independent or decorrelated components mix linearly to generate the observed data. However, this assumption is unjustified for most acoustic features used in ASR. For example, in the Mel-frequency cepstral coefficients (MFCC) representation, coarticulation effects and additive noise combine nonlinearly with the information about vocal tract shape that is important for recognition. Thus, nonlinear feature transforms are a promising approach to improved acoustic modeling.

In [4], a unified feature transformation framework that estimates the parameters of a nonlinear transform and the probabilistic model that jointly minimize the relative entropy between the true likelihood and its estimate based on the model was introduced. An iterative algorithm to jointly estimate the parameters of a class of volume preserving transforms — namely reflecting symplectic maps — and the parameters of the model is described also in [4]. This algorithm is applied to TIMIT phone recognition.

In this paper, we test the performance of the SMLT on two large-vocabulary, conversational speech recognition tasks: IBM's Superhuman test [5] and the DARPA 2003 Rich Transcription (RT03) test. Conversational speech recognition is a significantly more difficult task than TIMIT phone recognition. Also, the current work uses features computed by a linear projection of spliced cepstra, while the earlier work used static and delta MFCC features. As will be discussed further, this is an important difference because the current implementation of SMLT imposes a partition of the input feature space into two half-spaces.

2. PROBLEM FORMULATION

The goal of our work is to search for a map of the features that improves the validity of the diagonal-covariance Gaussian mixture HMM in the new feature space. First, we describe our approach for designing a global transform, then we generalize to class-dependent transforms.

2.1. A Global Volume-Preserving Transform

As shown in the following proposition, the problem may be reduced to maximum likelihood estimation (MLE) of the model and map parameters; however, we first need to define volume-preserving maps in $\Re^n$, where n is an arbitrary positive integer.

Definition: A C^∞ map $f : S_x \rightarrow S_y$ where $S_x \subset \Re^n$ and $S_y \subset \Re^n$ is volume-preserving if and only if $\left| det \left(\frac{\partial f(x)}{\partial x} \right) \right| = 1$ $\forall x \in S_x$.

Proposition: Let $y^t = f(x^t)$ be an arbitrary one-to-one C^∞ volume-preserving map of the random vector X^t at time t in $\Re^n$ to Y^t in $\Re^n$, and let $\hat{P}_\Lambda(y)$ be the estimated likelihood using an HMM, where $y = y^1 \cdots y^t \cdots y^T$, and T is the length of the utterance. The map $f^*(.)$ and the set of HMM parameters Λ^* jointly minimize the relative entropy between the hypothesized and the true likelihoods of Y if and only if they also maximize the expected log likelihood based on the model, $E_{P(Y)}[\log \hat{P}_\Lambda(Y)]$.

*The first author performed the work during a summer internship at the IBM T. J. Watson Research Center

Using the definition of volume-preserving maps, the proof of the proposition is straightforward [4].

2.2. Strong-Sense Class-Dependent Transforms

In weak-sense class dependency, features have observable values for all classes, but the features and some class variables are conditionally independent given a set of classes [6]. This increases the computational and storage requirements of the system, and results in the introduction of meaningless models that degrade the performance of the recognizer. Features are said to be class-dependent in the strong sense if they are assumed to be observable only for one class or cluster of classes, but are undefined for the remaining classes. In the following, the generation of strong-sense class-dependent features using volume-preserving transforms is described.

Let us define a set of functions $\{f_i(.)\}_{i=1}^{J}$ such that $y_i = f_i(x)$ is an arbitrary one-to-one map of the random vector X in $\Re^n$ to Y_i in $\Re^n$. The relation between the joint class-conditional probability of X and Y_i is

$$P_{Y_i|C}(f_i(x)|c_i) = \frac{P_{X|C}(x|c_i)}{|det(\frac{\partial f_i}{\partial x})|}, \quad (1)$$

where $det(\frac{\partial f_i}{\partial x})$ is the determinant of the Jacobian matrix of the map $f_i(.)$ [7].

Therefore, the Bayesian classification rule for the classifiers that use a set of class-dependent features, $\{y_i\}_{i=1}^{J}$ becomes

$$\hat{c} = \arg \max_{c \in \{1,\cdots,J\}} P_{Y_i|C}(f_i(x)|c)P(c)|det(\frac{\partial f_i}{\partial x})|. \quad (2)$$

Equation 2 shows that we can design strong-sense class-dependent features for any statistical recognition or classification system by accounting for the determinant of the Jacobian matrix in the decision rule [8]. Therefore, the Bayesian classification rule for the classifiers that use a set of class-dependent features, $\{y_i\}_{i=1}^{J}$ generated using a set of volume-preserving maps becomes

$$\hat{c} = \arg \max_{c \in \{1,\cdots,J\}} P_{Y_i|C_i}(f_i(x)|c_i)P(c_i). \quad (3)$$

This means that the decoding is unaffected by using class-dependent volume-preserving transforms. To train the parameters of these class-dependent transforms, the following lemma generalizes the previous proposition for the case of strong-sense class-dependent features.

Lemma: Let $y_i^t = f_i(x^t)$ for $i = 1, \cdots, J$ be arbitrary one-to-one C^∞ volume-preserving maps of the random vector X^t at time t in $\Re^n$ to Y_i^t in $\Re^n$, and let $y^t = y_i^t$ if $c^t = c_i$, $y = y^1 \cdots y^t \cdots y^T$, T is the utterance length in frames, and $\hat{P}_\Lambda(y)$ be the estimated likelihood using an HMM, where $\Lambda = \{\Lambda_i\}_{i=1}^{J}$. The set of maps $\{f_i^*(.)\}_{i=1}^{J}$ and the set of parameters $\{\Lambda_i^*\}_{i=1}^{J}$ jointly minimize the relative entropy between the hypothesized and the true likelihood of Y if and only if they also maximize the expected log likelihood based on the model, $E_{P(Y)}[\log \hat{P}_\Lambda(y)]$.

3. IMPLEMENTATION OF THE MAXIMUM LIKELIHOOD APPROACH

In the previous section, we showed that by using a volume-preserving map, the problem is reduced to maximizing the likelihood of the training data in the new feature space. In this section, we use a symplectic map to generate the new set of features.

3.1. Symplectic Maps

Symplectic maps are volume-preserving maps that can be represented by scalar functions. This interesting result allows us to jointly optimize the parameters of the symplectic map and the model parameters using the EM algorithm or one of its incremental forms [9].

Let $x = (x_1, x_2)$, and $y = (y_1, y_2)$, with $x_1, x_2, y_1, y_2 \in \Re^{\frac{n}{2}}$, then any reflecting symplectic map can be represented by

$$y_1 = x_1 - \frac{\partial V(x_2)}{\partial x_2}, \quad (4)$$

$$y_2 = x_2 - \frac{\partial T(y_1)}{\partial y_1}, \quad (5)$$

where $V(\cdot)$ and $T(\cdot)$ are two arbitrary scalar functions [10]. We parameterize these scalar functions with three-layer feed-forward neural networks

$$V(u, A, C) = \sum_{j=1}^{M} c_j S(a_j u), \quad (6)$$

$$T(u, B, D) = \sum_{j=1}^{M} d_j S(b_j u), \quad (7)$$

where $S(.)$ is a nonlinear function such as the sigmoid or hyperbolic tangent, a_j is the jth row of the $M \times n$ matrix A, c_j is the jth element of the $M \times 1$ vector C, b_j is the jth row of the $M \times n$ matrix B, and d_j is the jth element of the $M \times 1$ vector D. The parameters of the neural networks and the parameters of the model are jointly optimized to maximize the likelihood of the training data.

3.2. Joint Optimization of The Map and Model Parameters

Using the EM algorithm, the auxiliary function [9] to be maximized is

$$Q(\Phi^k, \Phi^{k+1}) = E_\xi[\log P(y, \zeta|\Phi^{k+1})|y, \Phi^k], \quad (8)$$

where $\zeta \in \xi$ is the state sequence corresponding to the sequence of observations $x \in \Re^{n \times T}$ that are transformed to the sequence $y \in \Re^{n \times T}$, T is the sequence length in frames, and $\Phi^k = (\Lambda^k, W^k)$ is the set of the recognizer parameters and symplectic parameters at iteration k of the algorithm. The update equations for the HMM parameters are unaffected by the introduction of the feature transform, and therefore are not given here.

We assume that the recognizer models the conditional probability density function (PDF) of the observation as a mixture of diagonal-covariance Gaussians, and therefore

$$\frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_j} = \sum_{i=1}^{N} \sum_{m=1}^{K} \frac{P(y^i, m|\Phi^k)}{P(y^i|\Phi^k)} \frac{(\mu_{mj} - y_j^i)}{\sigma_{mj}^2}, \quad (9)$$

where μ_{mj}, and σ_{mj}^2 are the mean and the variance of the jth element of the mth PDF respectively, N is the number of frames in the training data, and K is the total number of Gaussian models.

Let the nonlinearity, $S(.)$, in the neural networks be the hyperbolic tangent. Starting with A and B, to update the values of the symplectic parameters a_{qr} and b_{qr} for $q = 1, 2, \cdots, M$, and for $r = 1, 2, \cdots, \frac{n}{2}$, we have to calculate the partial derivative of

the auxiliary function with respect to these parameters using the following relations

$$\frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial a_{qr}} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{1j}} \frac{\partial y_{1j}}{\partial a_{qr}} + \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial a_{qr}}, \quad (10)$$

and

$$\frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial b_{qr}} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial b_{qr}}, \quad (11)$$

where

$$\frac{\partial y_{1j}}{\partial a_{qr}} = \begin{cases} 2x_{2r}\sum_{h=1}^{M}\left(c_h a_{hj} S(a_h x_2)[1-S^2(a_h x_2)]\right) & \text{for } r \neq j \\ 2x_{2r}\sum_{h=1}^{M}\left(c_h a_{hj} S(a_h x_2)[1-S^2(a_h x_2)]\right) - c_q[1-S^2(a_q x_2)] & \text{for } r = j \end{cases} \quad (12)$$

$$\frac{\partial y_{2j}}{\partial a_{qr}} = \sum_{k=1}^{\frac{n}{2}} \frac{\partial y_{1k}}{\partial a_{qr}} \frac{\partial y_{2j}}{\partial y_{1k}}, \quad (13)$$

$$\frac{\partial y_{2j}}{\partial y_{1k}} = -\sum_{h=1}^{M}\left(d_h b_{hj} b_{hk} S(b_h y_1)[1-S^2(b_h y_1)]\right), \quad (14)$$

and

$$\frac{\partial y_{2j}}{\partial b_{qr}} = \begin{cases} 2y_{1r}\sum_{h=1}^{M}\left(c_h b_{hj} S(b_h y_1)[1-S^2(b_h x_2)]\right) & \text{for } r \neq j \\ 2y_{1r}\sum_{h=1}^{M}\left(c_h b_{hj} S(b_h y_1)[1-S^2(b_h x_2)]\right) - d_q[1-S^2(b_q y_1)] & \text{for } r = j \end{cases} \quad (15)$$

For C and D, we have the following relations

$$\frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial c_q} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{1j}} \frac{\partial y_{1j}}{\partial c_q} + \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial c_q}, \quad (16)$$

and

$$\frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial d_q} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k,\Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial d_q}, \quad (17)$$

where

$$\frac{\partial y_{1j}}{\partial c_q} = a_{qj}[1-S^2(a_q x_2)], \quad (18)$$

$$\frac{\partial y_{2j}}{\partial c_q} = \sum_{k=1}^{\frac{n}{2}} \frac{\partial y_{1k}}{\partial c_q} \frac{\partial y_{2j}}{\partial y_{1k}}, \quad (19)$$

and

$$\frac{\partial y_{2j}}{\partial d_q} = b_{qj}[1-S^2(b_q y_1)]. \quad (20)$$

Using Equations 9 to 20, the values of the symplectic map parameters can be updated in each iteration using any gradient-based optimization algorithm [4].

4. EXPERIMENTS

We tested the SMLT in a number of configurations on two large-vocabulary, conversational tasks: IBM's Superhuman test [5] and the DARPA 2003 Rich Transcription (RT03) test. The Superhuman test comprises data from five sources of conversational American English, namely the Switchboard portion of the 1998 Hub 5e test (*swb98*), one meeting from the ICSI meeting corpus [11] (*mtg*), two collections of call center data (*cc1* and *cc2*), and the test set from the IBM Voicemail corpus [12] (*vm*). The RT-03 test material is two-party telephone conversations, like the *swb98* portion of the Superhuman test, but some of the material was collected more recently, and it is about three times longer than *swb98*.

The raw features for the recognition system used in the tests were 18-dimensional MFCC features computed every 10 ms. from 25-ms. frames with a Mel filter bank that spanned 0.125–3.8 kHz. The recognition features were computed from the raw features by splicing together nine frames of raw features (± 4 frames around the current frame), projecting the 162-dim. spliced features to 60 dimensions using an LDA projection, and then optionally transforming the 60-dim. projected features with one or more transforms intended to reduce the mismatch between the statistics of the final features and the constraints of the diagonal-covariance Gaussian mixtures that model the HMM observation densities. We tested five different configurations of the LDA projection and subsequent transforms:

L the LDA projection alone;

L+S the LDA projection followed by a nonlinear SMLT;

L+M the LDA projection followed by a linear MLLT;

L+M+S the LDA projection, then a linear MLLT, then a nonlinear SMLT; and

L+M+S2 the LDA projection, then a linear MLLT, then two class-dependent SMLTs, one for speech states and one for non-speech states.

We tested these configurations in order to answer three questions. First, because the SMLT is a nonlinear transform, will an LDA+SMLT cascade match or improve on the performance of an LDA+MLLT cascade? Second, will an LDA+MLLT+SMLT cascade outperform an LDA+MLLT cascade, given the flexibility that the nonlinear SMLT offers? Finally, can we obtain additional improvements in recognition performance with multiple, class-dependent SMLTs, taking full advantage of the SMLT's volume-preserving property?

The acoustic model training data were 315 hours of material from the Switchboard, Switchboard Cellular, and Callhome English corpora. For all five feature sets, an acoustic model comprising 4807 context-dependent states and 156K diagonal-covariance

transform	Superhuman swb98	mtg	cc1	cc2	vm	all	RT03
L	47.6	58.0	68.1	48.5	40.2	52.5	–
L+S	46.9	57.4	68.2	47.9	39.3	51.9	–
L+M	43.8	51.7	65.6	41.7	35.4	47.6	39.9
L+M+S	43.5	51.8	65.6	41.4	35.4	47.5	39.7
L+M+S2	43.5	51.9	65.4	41.4	35.3	47.5	39.7

Table 1. Word error rates (%) on the IBM Superhuman test and the RT-03 test for features generated with an LDA transform (L), LDA+SMLT transform (L+S), LDA+MLLT transform (L+M), LDA+MLLT+SMLT transform (L+M+S) or with an LDA+MLLT transform followed by one of two class-dependent SMLTs (L+M+S2).

Gaussian mixtures was used. The states were clustered using decision trees that could ask questions about phone identity within the current word in a ± 5-phone window. The number of Gaussian mixtures assigned to a state was chosen by maximizing the Bayesian Information Criterion (BIC). The decision trees and allocation of mixture components to states were based on the **L+M** feature space.

The Superhuman test was run using an interpolation of four back-off trigram language models (LMs) using modified Kneser-Ney smoothing. The data used to train the four component LMs were 3M words from Switchboard, 160M words from Broadcast News, 1M words from Voicemail, and 600K words of call center data [5]. The RT03 test was run using an interpolation of four back-off 4-gram LMs using modified Kneser-Ney smoothing. The component LMs were trained on 3M words of Switchboard, 58M words of web data collected and distributed by the University of Washington, 3M words of Broadcast News relevant to Switchboard topics, and 7M words from the English Gigaword corpus [13]. Decoding was done using a Viterbi decoder operating on a statically compiled decoding graph and employing a hierarchical Gaussian acoustic model [13].

5. RESULTS AND DISCUSSION

The results for our tests of the various transform configurations on the Superhuman and RT03 tests are presented in Table 1. A comparison of the **L**, **L+S**, and **L+M** results shows that in almost all cases, the use of an SMLT or MLLT transform improves performance over using only the LDA projection, and that the LDA+MLLT cascade consistently outperforms the LDA+SMLT cascade. This can be partially attributed to the fact that MLLT has roughly 20 times more parameters than the current implementation of SMLT. It should also be noted that the LDA solution is invariant to full-rank linear transforms such as the MLLT, but that no such invariance exists for nonlinear transforms such as the SMLT. A comparison of the **L+M** and **L+M+S** results shows a small advantage for the LDA+MLLT+SMLT cascade, especially on the *swb98* and RT03 tasks — tasks that are well matched to the training data. A number of factors may account for the relatively small improvement obtained with the SMLT: (1) the limited number of parameters in the SMLT, (2) the lack of a natural partition of the LDA+MLLT feature space into two half-spaces (recall that the implementation used for the reflecting symplectic transform imposes a partition of the feature space), and (3) optimization of the decision trees and mixture allocation to the LDA+MLLT feature space. Finally, we see no significant improvement with the two class-dependent SMLTs over the **L+M+S** results. This result is consistent with results on the TIMIT database reported in [8] for the SMLT and results reported for class-dependent MLLTs in [3]. We argue that transforms trained using MLE on observations corresponding to specific classes are less likely to reduce recognition error compared to MLE global transforms and a discriminative criterion should be used to estimate the class-dependent transforms.

Further investigation of the effect of the type of the input features and the structure used to implement the symplectic map on recognition performance will be our main goal in future research.

6. REFERENCES

[1] A. Ljolje, "The importance of cepstral parameter correlations in speech recognition," *Computer Speech and Language*, vol. 8, pp. 223–232, 1994.

[2] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[3] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, 1998.

[4] M. K. Omar and M. Hasegawa-Johnson, "A non-linear maximum likelihood transform for speech recognition," in *Proc. Eurospeech*, 2003.

[5] B. Kingsbury, L. Mangu, G. Saon, G. Zweig, S. Axelrod, V. Goel, K. Visweswariah, and M. Picheny, "Toward domain-independent conversational speech recognition," in *Proc. Eurospeech*, 2003.

[6] A. Bailey, *Class-dependent features and multicategory classification*, Ph.D. thesis, University of Southampton, 2001.

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1991.

[8] M. K. Omar and M. Hasegawa-Johnson, "Strong-sense class-dependent features for statistical recognition," in *Proc. IEEE Statistical Signal Processing Workshop*, 2003.

[9] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed., pp. 355–368. Kluwer Academic Publishers, 1998.

[10] L. C. Parra, "Symplectic nonlinear component analysis," in *Advances in Neural Information Processing Systems*. 1996, vol. 8, pp. 437–443, MIT Press.

[11] A. Janin *et al.*, "The ICSI Meeting corpus," in *Proc. ICASSP*, 2003.

[12] M. Padmanabhan, G. Saon, J. Huang, B. Kingsbury, and L. Mangu, "Automatic speech recognition performance on a voicemail transcription task," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 433–442, October 2002.

[13] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, , and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. Eurospeech*, 2003.