Non-Linear Maximum Likelihood Feature Transformation For Speech Recognition

Mohamed Kamal Omar, Mark Hasegawa-Johnson

Department of Electrical And Computer Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801. omar,jhasegaw@uiuc.edu

Abstract

Most automatic speech recognition systems (ASR) use Hidden Markov model (HMM) with a diagonal-covariance Gaussian mixture model for the state-conditional probability density function. The diagonal-covariance Gaussian mixture can model discrete sources of variability like speaker variations, gender variations, or local dialect, but can not model continuous types of variability that account for correlation between the elements of the feature vector. In this paper, we present a transformation of the acoustic feature vector that minimize an empirical estimate of the relative entropy between the likelihood based on the diagonal-covriance Gaussian mixture HMM model and the true likelihood. We show that this minimization is equivalent to maximizing the likelihood in the original feature space.

Based on this formulation, we provide a computationally efficient solution to the problem based on volume-preserving maps; existing linear feature transform designs are shown to be special cases of the proposed solution. Since most of the acoustic features used in ASR are not linear functions of the sources of correlation in the speech signal, we use a non-linear transformation of the features to minimize this objective function. We describe an iterative algorithm to estimate the parameters of both the volume-preserving feature transformation and the hidden Markov models (HMM) that jointly optimize the objective function for an HMM-based speech recognizer. Using this algorithm, we achieved 2% improvement in phoneme recognition accuracy compared to the original system that uses the original Mel-frequency cepstral coeeficients (MFCC) acoustic features. Our approach is compared also to previous similar linear approaches like MLLT and ICA.

1. Introduction

An important goal for designers of ASR systems is to achieve a high level of performance while minimizing the number of parameters used by the system. Not only because it increases the computational load and the storage requirements, but also because it increases the size of the training data required to estimate the parameters. One way of controlling the number of parameters is to adjust the structure of the conditional joint PDF used by the recognizer. For example, the dimensionality of the acoustic feature vectors in Gaussian mixture HMM is too large for their conditional joint PDFs to have full covariance matrices. On the other hand, approximating the conditional PDF by a diagonal covariance matrix Gaussian PDF degrades the performance of the recognizer [**?**], as the acoustic features used in ASR systems are neither decorrelated nor independent given the Gaussian component index. The mixture of Gaussian components can model discrete sources of variability like speaker variations, gender variations, or local dialect, but can not model continuous types of variability that account for correlation between the elements of the feature vector like coarticulation effects and background noise.

Recent approaches to this problem that offer new alternatives can be classified into two major categories. The first category try to decrease the number of parameters required for full covariance matrices. This category include a variety of choices for covariance structure other than diagonal or full. Two examples that can be used in ASR systems are block-diagonal [**?**] and banded-diagonal matrices. Another method often used by ASR systems is tying, where certain parameters are shared amongst a number of different models. For example, the semi-tied covariance matrices approach that estimates a transform in a maximum likelihood fashion given the current model parameters is described in [**?**] . Factor analysis also was used in [**?**] to model the covariance matrix of each Gaussian component of the Gaussian mixture used within each state of the HMM recognizer.

The second category choose to transform the original feature space to a new feature space that satisfies the diagonal-covariance models better. This is achieved by optimizing the transform based on a criterion that measures the validity of the assumption. An example is a state-specific principal component analysis (PCA) approach that was introduced in [**?**]. Another example is independent component analysis (ICA) that was used in developing features for speaker recognition [**?**] and speech recognition [**?**], [**?**], [**?**]. The maximum likelihood linear transform (MLLT) introduced in [**?**] is also an example of feature-based solutions.

All previous approaches assume that independent or decorrelated components are mixed linearly to generate the observation data. However, for most acoustic features used in ASR, this assumption is unjustified or unacceptable. An example is cepstral features like MFCC and PLPCC; In the cepstral domain, coarticulation effects and additive noise are examples of independent sources in the speech signal that are nonlinearly combined with the information about the vocal tract shape that is important for recognition. The source-filter model proposes that the excitation signal and the vocal tract filter are linearly combined in the cepstral domain, but the source-filter model is unrealistic in many cases, especially for consonants. Time-varying filters and filter-dependent sources result in nonlinear source-filter combination in the cepstral domain [**?**].

In [**?**], we formulated the problem as a non-linear independent component analysis (NICA) problem. We showed that using the features generated using NICA in speech recognition increased the phoneme recognition accuracy compared to linear feature transforms like ICA [**?**], linear discriminant analysis (LDA) [**?**], and MLLT. However, using PCA or ICA approaches is justified only if a different feature transform is designed for each Gaussian component in the model, as it assumes that the probabilistic model imposes independence or decorrelation on the features.

In this work, we will introduce a unified information-theoretic approach to feature transformation that makes no assumptions about the true probability density function of the original features and can be applied for any probabilistic model with arbitrary constraints. It estimates a nonlinear transform and the parameters of the probabilistic model that jointly minimize the relative entropy between the true likelihood and its estimation based on the model. Unlike previous approaches, this formulation justify using a single transform for observations generated by different classes. In the next section, an information-theoretic formulation of the problem is described and a solution based on volume-preserving maps is introduced. An iterative algorithm is described in section 4 to jointly estimate the parameters of the transform of the features and the parameters of the model. Then, experiments based on an efficient implementation of this algorithm are described in section 5. Finally, section 6 provides discussion of the results and a summary of this work.

2. Problem Formulation

We will take here a different approach to the problem, motivated by the discussion of the previous section. Instead of focusing on specific model assumptions, we will choose any hypothesized parametric family of distributions to be used in our probabilistic model, and search for a map of the features that improves the validity of our model. To do that, we will need the following proposition.

Proposition: Let $y = f(x)$ be an arbitrary one-to-one map of the features random vector X in $\Re^n$ to Y in $\Re^n$, and let $\hat{P}_\Lambda(y)$ be the likelihood of the new features using HMM. The map $f^*(.)$ and the set of parameters Λ^* minimize the relative entropy between the hypothesized and the true likelihoods of Y if and only if they also maximize the objective function

$$L \quad = \quad E_{P(Y)}\left[\log\left(\left|det\left(\frac{\partial f}{\partial x}\right)\right|\right) + \log \hat{P}_\Lambda(Y)\right], \quad (1)$$

where $[\frac{\partial f}{\partial x}]$ is the Jacobian matrix of the map $f(.)$.

This can be shown by writing the expression for the relative entropy after an arbitrary transformation, $y = f(x)$, of the input random vector X in $\Re^n$, as

$$R(P(Y), \hat{P}(Y)) \quad = \quad -H(P(Y)) - E_{P(Y)}\left[\log\left(\hat{P}(Y)\right)\right], \quad (2)$$

where $H(P(Y))$ is the differential entropy of the random vector Y based on its true PDF $P(Y)$.

The relation between the output differential entropy and the input differential entropy is in general [**?**],

$$\begin{aligned} H(P(Y)) \quad \leq \quad & H(P(X)) \\ & + \int_{\Re^n} P(x) \log\left(\left|det\left(\frac{\partial f(x)}{\partial x}\right)\right|\right) dx, \end{aligned} \quad (3)$$

where $P(x)$ is the probability density function of the random vector X, for an arbitrary transformation, $y = f(x)$, of the random vector X in $\Re^n$, with equality if $f(x)$ is invertible.

Therefore the relative entropy can be written as

$$\begin{aligned} R(P(Y), \hat{P}(Y)) \quad = \quad & -H(P(X)) \\ & -E_{P(X)}\left[\log\left(\left|det\left(\frac{\partial f(x)}{\partial x}\right)\right|\right)\right] \\ & -E_{P(Y)}\left[\log \hat{P}(Y)\right], \end{aligned} \quad (4)$$

for an invertible map $y = f(x)$.

The expectation of a function $g(x)$ for an arbitrary one-to-one map $y = f(x)$ can be written as [**?**],

$$E_{P(X)}[g(x)] \quad = \quad E_{P(Y)}\left[g(f^{-1}(y))\right], \quad (5)$$

where $f^{-1}(.)$ is the inverse map.

Therefore

$$\begin{aligned} R(P(Y), \hat{P}(Y)) \quad = \quad & -H(P(X)) \\ & -E_{P(Y)}\left[\log\left(\left|det\left(\frac{\partial f(x)}{\partial x}\right)\right|\right)\right] \\ & -E_{P(Y)}\left[\log \hat{P}(Y)\right]. \end{aligned} \quad (6)$$

Equation 6 proves the proposition.

The proposition states that minimizing the relative entropy is equivalent to maximizing the likelihood in the original feature space, but with the new features are modeled by HMM instead of the original features.

2.1. A Maximum Likelihood Approach

An important special case that reduces the problem to maximum likelihood estimation (MLE) of the model and map parameters is given in the following lemma, but first we 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 said to be volume-preserving if and only if $\left|det\left(\frac{\partial f(x)}{\partial x}\right)\right| = 1 \; \forall x \in S_x$.

Lemma: Let $y = f(x)$ be an arbitrary one-to-one C^∞ volume-preserving map of the random vector X in $\Re^n$ to Y in $\Re^n$, and let $\hat{P}_\Lambda(y)$ be the estimated likelihood using HMM. The map $f^*(.)$ and the set of 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 hypothesized PDF.

Using the definition of the volume-preserving maps, the proof of the lemma is straightforward. By reducing the problem to MLE problem, efficient algorithms based on the incremental EM algorithm can be designed [**?**].

2.2. Generality of The Approach

Our approach generalizes previous approaches to feature transform for speech recognition in two ways. First, transforms can be designed to satisfy arbitrary constraints on the model, not necessarily those that impose an independence or decorrelation constraint on the features. Second, it can also be applied to any parameterized probabilistic model not necessarily Gaussian. Therefore, it can be used to design a single transform of the observations, if the whole HMM recognizer is taken as our probabilistic model, and it can be used to design state-dependent or phoneme-dependent transforms, if the state or the phoneme probabilistic models in the recognizer are used respectively. To show the generality of our approach and its wide range of applications, we relate it with previous methods.

PCA may be viewed as a special case of the proposition under two equivalent constraints. First, if the transform is constrained to be linear and the model PDF is constrained to be a diagonal-covariance Gaussian, then the propositionreduces to PCA. Equivalently, if the true feature PDF is assumed to be Gaussian, and the model PDF is constrained to be a diagonal-covariance Gaussian, the propositionreduces to PCA.

ICA also can be shown as a special case of proposition when the hypothesized model assumes statistical independence of the transformed features and the transform is constrained to be linear. Nonlinear ICA removes the constraint that the transform must be linear. Factor analysis is also a special case of the proposition by assuming that the hypothesized joint PDF is Gaussian with special covariance structure.

MLLT is a special case of the proposition by using a linear volume-preserving map of the features and assuming the hypothesized joint PDF is Gaussian or a mixture of Gaussins. The two assumptions of linearity and Gaussianity together are equivalent to the assumption that the original features are Gaussian.

It should be noted that all linear maps designed to improve the satisfaction of the features of a given model are special cases of the lemma, as any linear map is equivalent to a linear volume-preserving map multiplied by a scalar.

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 output components. 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 very 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 [**?**].

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

$$y_1 = x_1 - \frac{\partial V(x_2)}{\partial x_2}, \quad (7)$$

$$y_2 = x_2 - \frac{\partial T(y_1)}{\partial y_1}, \quad (8)$$

where $V(\cdot)$ and $T(\cdot)$ are two arbitrary scalar functions [**?**]. We use two multi-layer feed-forward neural networks to get a good approximation of these scalar functions [**?**].

$$V(u, A, C) = \sum_{j=1}^{M} c_j S(a_j u), \quad (9)$$

$$T(u, B, D) = \sum_{j=1}^{M} d_j S(b_j u), \quad (10)$$

where $S(.)$ is a nonlinear function like sigmoid or hyperbolic tangent, a_j is the jth row of the $M \times n$ matrix A, and 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 these two 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

We will explain in this section, how the parameters of the volume-preserving map and the probabilistic model can be jointly optimized to maximize the likelihood of the estimated features. We will assume that the system is HMM-based recognizer [**?**]. However, this approach can be applied to any statistical classification, detection, or recognition systems. We will assume also that the scalar functions in the symplectic map are represented by three-layer feed forward neural networks (NN) with the nonlinearity in the NNs represented by hyperbolic tangent functions. The derivation for any other non-linear function is a straightforward replication of the derivation provided here.

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

$$Q(\Phi^k, \Phi^{k+1}) = E_\xi[\log P(y, \zeta|\Phi^{k+1})|y, \Phi^k], \quad (11)$$

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, $\Phi^k = (\Lambda^k, W^k)$ is the set of the recognizer parameters and the symplectic parameters at iteration k of the algorithm.

The updating equations for the HMM parameters are the same as mentioned in [**?**], and therefore will not be given here.

We will assume that the recognizer models the conditional PDF of the observation as a mixture of diagonal-covariance Gaussians and therefore

$$\frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_j} = \sum_{i=1}^{N}\sum_{m=1}^{K} \frac{P(y^i, m|\Phi^k)}{P(y^i|\Phi^k)} \frac{(\mu_{mj} - y_j^i)}{\sigma_{mj}^2}, \quad (12)$$

where μ_{mj}, and σ_{mj}^2 are the mean and the variance of the jth element of the mth PDF respectively.

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 this parameters. These partial derivatives are related to the partial derivatives of the auxiliary function with respect to the features by the following relation

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

and

$$\frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial b_{qr}} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial b_{qr}}, \quad (14)$$

where

$$\frac{\partial y_{1j}}{\partial a_{qr}} = \begin{cases} 2x_{2r}\sum_{h=1}^{M}\left(c_h a_{hj} S(a_h x_2)[1-S^2(a_h x_2)]\right) \\ \qquad\qquad\qquad \text{for } r \neq j \\ 2x_{2r}\sum_{h=1}^{M}\left(c_h a_{hj} S(a_h x_2)[1-S^2(a_h x_2)]\right) \\ -c_q[1-S^2(a_q x_2)] \qquad \text{for } r = j \end{cases} \tag{15}$$

$$\frac{\partial y_{2j}}{\partial a_{qr}} = \sum_{k=1}^{\frac{n}{2}} \frac{\partial y_{1k}}{\partial a_{qr}} \frac{\partial y_{2j}}{\partial y_{1k}}, \tag{16}$$

$$\frac{\partial y_{2j}}{\partial y_{1k}} = -\sum_{h=1}^{M}\left(d_h b_{hj} b_{hk} S(b_h y_1)[1-S^2(b_h y_1)]\right), \tag{17}$$

and

$$\frac{\partial y_{2j}}{\partial b_{qr}} = \begin{cases} 2y_{1r}\sum_{h=1}^{M}\left(c_h b_{hj} S(b_h y_1)[1-S^2(b_h x_2)]\right) \\ \qquad\qquad\qquad \text{for } r \neq j \\ 2y_{1r}\sum_{h=1}^{M}\left(c_h b_{hj} S(b_h y_1)[1-S^2(b_h x_2)]\right) \\ -d_q[1-S^2(b_q y_1)] \qquad \text{for } r = j \end{cases} \tag{18}$$

For C and D, to updated values of the symplectic parameter c_q and d_q for $q = 1, 2, \cdots, M$, we have to calculate the partial derivative of the auxiliary function with respect to these parameters. These partial derivatives are related to the partial derivatives of the auxiliary function with respect to the features by the following relation

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

and

$$\frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial d_q} = \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial d_q}, \tag{20}$$

where

$$\frac{\partial y_{1j}}{\partial c_q} = a_{qj}[1-S^2(a_q x_2)], \tag{21}$$

$$\frac{\partial y_{2j}}{\partial c_q} = \sum_{k=1}^{\frac{n}{2}} \frac{\partial y_{1k}}{\partial c_q} \frac{\partial y_{2j}}{\partial y_{1k}}, \tag{22}$$

and

$$\frac{\partial y_{2j}}{\partial d_q} = b_{qj}[1-S^2(b_q y_1)]. \tag{23}$$

Using Equations from 12 to 23, the values of the symplectic map parameters can be updated in each iteration using any gradient-based optimization algorithm.

4. EXPERIMENTS AND RESULTS

The symplectic maximum likelihood algorithm described in section 3 is used to study the optimal feature space for diagonal-covariance Gaussian mixture HMM modeling of the TIMIT database. We used the conjugate-gradient algorithm to update the values of the symplectic map parameters in each iteration.

The Mel-frequency Cepstrum Coefficients are calculated for 4500 utterances from the TIMIT database. The overall 26-feature vector consists of 12 MFCC coefficients, energy and their deltas.

In each iteration, the new feature vector is calculated using the current symplectic transformation parameters by using the symplectic mapping equation, then the maximum likelihood estimates of the HMM model parameters are calculated. Then, the maximum likelihood estimates of the symplectic map parameters are estimated using the conjugate-gradient algorithm.. After the iterative algorithm converges to a set of locally optimal HMM and symplectic parameters, the training data are transformed by the symplectic map yielding the final symplectic maximum likelihodd tranform (SMLT) feature vector. The new features are compared to LDA, linear ICA, and MLLT in their phoneme recognition accuracy.

In our experiments, the 61 phonemes defined in the TIMIT database are mapped to 48 phoneme labels for each frame of speech as described in [**?**]. These 48 phonemes are collapsed to 39 phoneme for testing purposes as in [**?**]. A three-state left-to-right model for each triphone is trained using the EM algorithm. The number of mixtures per state was fixed to four. After training the overall system and obtaining the symplectic map parameters, the approximately independent output coefficients of the symplectic map are used as the input acoustic features to a Gaussian mixture hidden Markov model speech recognizer [**?**]. The parameters of the recognizer are trained using the training portion of the TIMIT database. The parameters of the triphone models are then tied together using the same approach as in [**?**].

To compare the performance of the proposed algorithm with other approaches, we generated acoustic features using LDA, linear ICA, and MLLT. We used the maximum likelihood approach to LDA [**?**] and kept the dimensions of the output of LDA the same as the input. We used also the maximum likelihood approach to linear ICA as described in [**?**] and briefly overviewed in section 2. Finally we implemented MLLT as described in [**?**] and briefly overviewed in section 2. All these techniques used a feature vector that consists of twelve MFCC coefficients, the energy, and their deltas as their input.

Testing this recognizer, using the test data in the TIMIT database, we get the phoneme recognition results in table 1. These results are obtained by using a bigram phoneme language model and by keeping the insertion error around 10% as in [**?**]. The table compares these recognition results to the ones obtained by MFCC, LDA, linear ICA and MLLT.

Table 1: Phoneme Recognition Accuracy

Acoustic Features	**Recognition Accuracy**
MFCC	73.7%
Linear ICA	73.5%
LDA	73.8%
MLLT	74.6%
SMLT	75.5%

5. DISCUSSION

In this work, we described a framework for feature transformation for speech recognition. We introduced a nonlinear symplectic maximum likelihood feature transform algorithm. This can be attributed to the ability of the algorithm to find a better representation of the acoustic clues of different phonemes. The improvement due to this different representation over the input MFCC features that have the same amount of information about phonemes, is due to the approximate independence property of the new features that allow a more efficient probabilistic modeling of the conditional probabilities with the same model complexity.

6. ACKNOWLEDGMENT

This work was supported by NSF award number 0132900.