

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 (ASR) systems 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 minimizes an empirical estimate of the relative entropy between the likelihood based on the diagonal-covariance Gaussian mixture HMM model and the true likelihood.

Based on this formulation, we provide a solution to the problem using 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 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 baseline system. Our approach shows also improvement in recognition accuracy compared to previous linear approaches like linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and independent component analysis (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 a large number of parameters 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 probability density function (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 Gaussian mixture degrades the performance of the recognizer [1]. This is due to 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. Another method often used by ASR systems is tying, where certain parameters are shared amongst a number of different models like the semi-tied covariance matrices approach in [2]. The second category chooses to transform the original feature space to a new feature space that satisfies the diagonal-covariance models better. Examples are the state-specific principal component analysis (PCA) [1], ICA [3], and MLLT [4].

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; 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 [5].

In [6], we formulated the problem as a non-linear independent component analysis (NICA) problem. We showed an increase in phoneme recognition accuracy compared to linear feature transforms like ICA, LDA [7], and MLLT. However, NICA approach (like PCA and ICA) is not rigorously justified, if a single global transform of the features is designed.

In this work, we 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 the parameters of both a nonlinear transform and 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 allows using a non-linear global transform for observations generated by different classes. In the next section, the problem is formulated and a solution based on volume-preserving maps is introduced. An iterative algorithm is described in section 3 to jointly estimate the parameters of the feature transform and the parameters of the model. Then, recognition experiments are described in section 4. Finally, section 5 provides discussion of the results and a summary of this work.

2. Problem Formulation

Instead of focusing on specific model assumptions, we will choose any hypothesized parametric 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 \mathfrak{R}^n to Y in \mathfrak{R}^n , and let $\hat{P}_\Lambda(y)$ be the likelihood of the new features using HMM. The map $f^*(\cdot)$ 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

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

where $\left[\frac{\partial f}{\partial x} \right]$ is the Jacobian matrix of the map $f(\cdot)$, and $P(y)$ is the true likelihood.

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 \mathfrak{R}^n , as

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

where $H(P(Y))$ is the differential entropy of the random vector Y [8].

The relation between the output differential entropy and the input differential entropy is in general [9],

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

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 \mathfrak{R}^n , with equality if $f(x)$ is invertible.

Therefore the relative entropy can be written as

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

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 [9],

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

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

Therefore

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

Equation 6 proves the proposition.

It can be shown that maximizing Equation 1 is equivalent to maximizing the likelihood in the original feature space given the parametric model in the new feature space.

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 \mathfrak{R}^n , where n is an arbitrary positive integer.

Definition: A C^∞ map $f : S_x \rightarrow S_y$ where $S_x \subset \mathfrak{R}^n$ and $S_y \subset \mathfrak{R}^n$ 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 \mathfrak{R}^n to Y in \mathfrak{R}^n , and let $\hat{P}_\Lambda(y)$ be the estimated likelihood using HMM. The map $f^*(\cdot)$ 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 model, $E_{P(Y)} [\log \hat{P}_\Lambda(Y)]$.

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 [10].

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 global 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.

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 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 [10].

Let $x = (x_1, x_2)$, and $y = (y_1, y_2)$, with $x_1, x_2, y_1, y_2 \in \mathfrak{R}^{\frac{n}{2}}$, 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 [11]. We use two three-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 $S(\cdot)$ is a nonlinear function like sigmoid or hyperbolic tangent, a_j is the j th row of the $M \times n$ matrix A , and c_j is the j th element of the $M \times 1$ vector C , b_j is the j th row of the $M \times n$ matrix B , and d_j is the j th 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

Using the EM algorithm, the auxiliary function [10] 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 $x \in \mathbb{R}^{n \times T}$ that are transformed to the sequence $y \in \mathbb{R}^{n \times T}$, 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 not affected by the introduction of the feature transform, 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

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

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

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

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

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

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

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

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

and

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

For C and D , to update values of the symplectic parameters c_q and d_q for $q = 1, 2, \dots, M$, 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

$$\begin{aligned} \frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial c_q} &= \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_{1j}} \frac{\partial y_{1j}}{\partial c_q} \\ &+ \sum_{j=1}^{\frac{n}{2}} \frac{\partial Q(\Phi^k, \Phi^{k+1})}{\partial y_{2j}} \frac{\partial y_{2j}}{\partial c_q}, \end{aligned} \quad (19)$$

and

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

where

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

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

and

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

Using Equations 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.

The baseline 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, then the maximum likelihood estimates of the HMM model parameters are calculated. Then, the maximum likelihood estimates of the symplectic map parameters are calculated 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 likelihood transform (SMLT) feature vector.

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 [12]. These 48 phonemes are collapsed to 39 phoneme for testing purposes as in [12]. A three-state left-to-right model for each triphone is trained. The number of mixtures per state was fixed to four. The parameters of the recognizer and the symplectic map 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 [13].

To compare the performance of the proposed algorithm with other approaches, we generated acoustic features using LDA, linear ICA, and MLLT. We kept the dimensions of the output of LDA the same as the input. We used also the linear ICA approach described in [3]. Finally we implemented MLLT as described in [4].

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 [12]. The table compares SMLT 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	75.6%

5. DISCUSSION

In this work, we described a framework for feature transformation for speech recognition. This framework is an extension of current approaches to both non-linear and global transforms. We introduced also a nonlinear symplectic maximum likelihood feature transform algorithm. Phoneme recognition experiments using features generated by this algorithm show significant improvement compared to previous linear transforms like LDA, MLLT, and ICA.

6. ACKNOWLEDGMENT

This work was supported by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

7. References

- [1] A. Ljolje, "The importance of cepstral parameter correlations in speech recognition," *Computer, Speech, and Language*, vol. 8, pp. 223-232, 1994.
- [2] Mark 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] Jong-Hwan Lee, Ho-Young Jung, and Te-Won Lee, "Speech Feature Extraction Using Independent Component Analysis," *IEEE Proceedings of ICASSP*, Vol. 3, pp. 1631-1634, Istanbul, Turkey, 2000.
- [4] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions For Classification," *IEEE Proceedings of ICASSP*, Seattle, Washington, 1998.
- [5] Thomas F. Quateri, *Discrete-Time Speech Signal Processing Principles And Practice* Prentice Hall, Upper Saddle River, NJ, 2002.
- [6] Mohamed Kamal Omar, and Mark Hasegawa-Johnson, "Approximately Independent Factors of Speech Using Symplectic Maps," *IEEE Trans. on Speech and Audio Processing*, In Press.
- [7] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley, New York, NY, 2000.
- [8] Thomas M. Cover, and Joy A. Thomas, *Elements of Information Theory*, Wiley, New York, NY, 1997.
- [9] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1991.
- [10] R. Neal and G. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants," *Learning in Graphical Models*, Kluwer Academics, 1998.
- [11] L. C. Parra, *Symplectic nonlinear component analysis*, In *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, Cambridge, MA., pp. 437-443, 1996.
- [12] Kai-Fu Lee, and Hsiao-Wuen Hon, "Speaker Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 37, pp. 1641-1648, November 1989.
- [13] S. Young, and P. Woodland, "State Clustering in hidden Markov model continuous speech recognition," *Computer, Speech, and Language*, vol. 8, pp. 369-383, October 1994.