DERIVATIVE-AUGMENTED FEATURES AS A DYNAMIC MODEL FOR TIME-SERIES

Paul M Baggenstoss

Naval Undersea Warfare Center
Newport RI, 02841
and Fraunhofer FKIE
53343 Wachtberg, Germany

ABSTRACT

In the field of automatic speech recognition (ASR), it is common practice to augment features with time-derivatives, which we call derivative-augmented features (DAF). Although the method is effective for modeling the dynamic behavior of features and produces significantly lower classification error, it violates the assumption of conditional independence of the observations. The traditional approach is to ignore the problem (simply apply the mathematical approach that assumes independence). In this paper, we take an alternative approach in which we still use the same mathematical approach as before, but calculate a correction factor by integrating out the redundant dimensions. This makes it possible to compare and combine a DAF PDF and a non-DAF PDF. We conduct experiments to demonstrate the usefulness of the approach.

Index Terms— PDF estimation, feature derivatives, HMM

1. INTRODUCTION

1.1. Background and Motivation

The hidden Markov model (HMM), although having many benefits for modeling human speech, models the data using discrete states. It can only model continuous feature variations using a large number of states. This problem is generally solved by augmenting the features with time-derivatives [1]. Despite new probabilistic models that address the dynamic behavior of features such as segmental HMMs [2], and a wider class of graphical models [3], the derivative augmented feature (DAF) combined with hidden Markov model (DAF-HMM) remains the most widely-used method of modeling the dynamic behavior of features. Unfortunately, the DAF feature vector is of higher dimension with built-in redundancy. As a result, the assumption of conditional independence of the observations is violated. The probability density function (PDF), or likelihood function (LF) of DAF cannot be compared to the PDF of the original (un-augmented) features. Being able to do this could enable new quantitative means of evaluating dynamic models based on augmentation and comparing with those not based on augmentation and allow classifiers with "mixed" models, taking advantage of DAF when necessary and using un-augmented features when not.

To this end, we derive an analytic expression for the integral of DAF-HMM model with respect to the un-differenced input data, allowing it to be normalized so that it integrates to one. The computational complexity of our method is order $O(M^T)$ where M is the number of Markov states and T is the length of the feature stream. But, the correction term reaches a steady-state at low values of T, allowing an efficient means to compensate PDFs for large T.

2. MATHEMATICAL PRELIMINARIES

2.1. DAF

Consider the feature stream $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_T\}$, where $\mathbf{x}_t \in \mathcal{R}^D$. For simplicity, we assume that the first derivatives are obtained by the first-order difference: $\mathbf{d}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$ and define the DAF as

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{d}_t \end{bmatrix} = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_t - \mathbf{x}_{t-1} \end{bmatrix}.$$

Note that we are forced to eliminate one time step, thus $\mathbf{Z} = \{\mathbf{z}_2, \mathbf{z}_3 \dots \mathbf{z}_T\}$. For analysis, it is more convenient to work with the equivalent *history* form of the DAF defined by $\mathbf{Y} = \{\mathbf{y}_2, \mathbf{y}_3 \dots \mathbf{y}_T\}$, where

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{x}_{t-1} \\ \mathbf{x}_t \end{bmatrix}. \tag{1}$$

From a theoretical point of view, features $\mathbf{y}_t$ and $\mathbf{z}_t$ are equivalent since we can obtain $\mathbf{z}$ from $\mathbf{y}$ by linear transformation $\mathbf{T}$ with determinant 1.

2.2. DAF-HMM

The M-state HMM model parameters consist of $\mathbf{\Lambda} = [\{\pi_m\}, \{A_{i,j}\}, \{b_i(\mathbf{y})\}]$, where $\pi_m,\ 1 \le m \le M$ are the prior probabilities, $A_{i,j},\ 1 \le i \le M,\ 1 \le j \le M$ are the state transition probabilities, and $b_i(\mathbf{y}),\ 1 \le i \le M$

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

are the state observation probability densities. The well-known *forward procedure* [4] computes the likelihood function or joint probability density function (PDF) $L_y(\mathbf{Y}) = p(\mathbf{y}_2, \mathbf{y}_3 \dots \mathbf{y}_T; \mathbf{\Lambda})$. To convert $L_y(\mathbf{Y})$ into a PDF on $\mathbf{X}$, we require the integral $K = \int_{\mathbf{X}} L_y(\mathbf{D}(\mathbf{X}))\ \mathrm{d}\mathbf{X}$, where $\mathbf{Y} = \mathbf{D}(\mathbf{X})$ is the DAF transformation that implements (1).

3. THE DAF INTEGRAL.

The desired integral is

$$K_T = \int_{\mathbf{x}_1} \cdots \int_{\mathbf{x}_T} L_y(\mathbf{D}(\mathbf{X}))\ \mathrm{d}\mathbf{x}_1\ \mathrm{d}\mathbf{x}_2\ \cdots \mathrm{d}\mathbf{x}_T. \quad (2)$$

We first expand $L(\mathbf{Y}) = \sum_{\mathbf{q}\in\mathbf{Q}}\ p(\mathbf{q})\ L(\mathbf{Y}|\mathbf{q})$. where $\mathbf{q}$ is a particular length-T Markov state sequence $\mathbf{q} = \{i, j, k, l \dots p, q, r\}$ with *apriori* probability

$$p(\mathbf{q}) = \pi_i\ A_{i,j}\ A_{j,k} \cdots\ A_{p,q}\ A_{q,r}.$$

In what follows, the indexes $\{i, j, k, l \dots p, q, r\}$ will always stand for the assumed states at times $1, 2, 3, 4 \dots T-2, T-1, T\}$, respectively. Using conditional independence,

$$L(\mathbf{Y}|\mathbf{q}) = b_i(\mathbf{y}_2)\ b_j(\mathbf{y}_3)\ \cdots\ b_q(\mathbf{y}_{T-1})\ b_r(\mathbf{y}_T). \quad (3)$$

Thus, we have

$$\begin{aligned} L(\mathbf{Y}) \quad = \quad & \sum_{i=1}^{N}\sum_{j=1}^{N}\cdots\sum_{q=1}^{N}\sum_{r=1}^{N} \pi_i\ A_{i,j} \cdots\ A_{q,r} \\ & \cdot b_i(\mathbf{y}_2)\ b_j(\mathbf{y}_3)\ \cdots\ b_q(\mathbf{y}_{T-1})\ b_r(\mathbf{y}_T) \end{aligned}$$

For tractability, we assume the state observation PDFs $b_k(\mathbf{y})$ are Gaussian. This assumption does not limit this discussion since an HMM with Gaussian mixture state PDFs can be represented as an HMM with Gaussian state PDFs by expanding the individual mixture kernels as separate Markov states. We assume a special form for the means and covariances of $b_k(\mathbf{y})$:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^a \\ \\ \boldsymbol{\mu}_k^b \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{\Sigma}_k^{aa} & \mathbf{\Sigma}_k^{ab} \\ \\ \mathbf{\Sigma}_k^{ba} & \mathbf{\Sigma}_k^{bb} \end{bmatrix}, \quad (4)$$

where superscripts a and b refer to the partitions of $\mathbf{y}_t$ corresponding to $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, respectively (thus, a,b are in order of increasing time). Note that the marginal PDFs are easily found, for example $b_k^b(\mathbf{x})$ has mean $\boldsymbol{\mu}_k^b$, and covariance $\mathbf{\Sigma}_k^{bb}$.

The only term in (3) that depends on $\mathbf{x}_1$ is $b_i(\mathbf{y}_2)$, which integrated over $\mathbf{x}_1$ is

$$\int_{\mathbf{x}_1} b_i(\mathbf{y}_2) = b_i^b(\mathbf{x}_2)\ \mathrm{d}\mathbf{x}_1,$$

so

$$\int_{\mathbf{x}_1} L(\mathbf{Y}|\mathbf{q}) = b_i^b(\mathbf{x}_2)\ b_j(\mathbf{y}_3)\ \cdots\ b_q(\mathbf{y}_{T-1})\ b_r(\mathbf{y}_T). \quad (5)$$

We now proceed to integrate (5) over $\mathbf{x}_2$. The only terms that depend on $\mathbf{x}_2$ are $b_i^b(\mathbf{x}_2)$ and $b_j(\mathbf{y}_3)$. We have

$$\begin{aligned} \int_{\mathbf{x}_1,\mathbf{x}_2} L(\mathbf{Y}|\mathbf{q}) &= \left\{\int_{\mathbf{x}_2} b_i^b(\mathbf{x}_2)\ b_j(\mathbf{x}_2|\mathbf{x}_3)\ \mathrm{d}\mathbf{x}_2\right\}\ b_j^b(\mathbf{x}_3) \\ \cdot b_k(\mathbf{y}_4)\cdots b_r(\mathbf{y}_T) &= \int_{\mathbf{x}_2} \mathcal{N}(\mathbf{x}_2 - \boldsymbol{\mu}_i^b, \mathbf{\Sigma}_i^{bb})\ b_j(\mathbf{x}_2|\mathbf{x}_3) \\ & \mathrm{d}\mathbf{x}_2\ b_j^b(\mathbf{x}_3)\ b_k(\mathbf{y}_4)\cdots b_q(\mathbf{y}_{T-1})\ b_r(\mathbf{y}_T) \end{aligned} \quad (6)$$

Using (4) and standard identities for the conditional distribution,

$$b_j(\mathbf{x}_2|\mathbf{x}_3) = \mathcal{N}\left(\mathbf{x}_2 - \boldsymbol{\mu}_c(\mathbf{x}_3), \mathbf{\Sigma}_c\right)$$

, where

$$\boldsymbol{\mu}_c(\mathbf{x}_3) = \boldsymbol{\mu}_j^a + \mathbf{\Sigma}_j^{ab}\ (\mathbf{\Sigma}_j^{bb})^{-1}\ \left(\mathbf{x}_3 - \boldsymbol{\mu}_j^b\right),$$

and

$$\mathbf{\Sigma}_c = \mathbf{\Sigma}_j^{aa} - \mathbf{\Sigma}_j^{ab}\ (\mathbf{\Sigma}_j^{bb})^{-1}\mathbf{\Sigma}_j^{ab'}$$

. Then, using the standard identity for the product of two Gaussians,

$$\begin{aligned} \mathcal{N}(\mathbf{x}_2 - \boldsymbol{\mu}_i^b, \mathbf{\Sigma}_i^{bb})\ b_j(\mathbf{x}_2|\mathbf{x}_3) &= \mathcal{N}(\mathbf{x}_2 - \boldsymbol{\mu}^d, \mathbf{\Sigma}^d) \\ \cdot\mathcal{N}\left(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_c(\mathbf{x}_3), \mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_c\right) & \end{aligned}$$

where

$$\begin{aligned} \mathbf{\Sigma}^d &= \left((\mathbf{\Sigma}_i^{bb})^{-1} + \mathbf{\Sigma}_c^{-1}\right)^{-1}, \\ \boldsymbol{\mu}^d &= \mathbf{\Sigma}^d\left((\mathbf{\Sigma}_i^{bb})^{-1}\boldsymbol{\mu}_i^b + \mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c(\mathbf{x}_3)\right) \end{aligned}$$

Integrating over $\mathbf{x}_2$ leaves us with

$$\begin{aligned} \int_{\mathbf{x}_2} b_i^b(\mathbf{x}_2)\ b_j(\mathbf{x}_2|\mathbf{x}_3)\ \mathrm{d}\mathbf{x}_2 &= \mathcal{N}(\mathbf{x}_2 - \boldsymbol{\mu}^d, \mathbf{\Sigma}^d) \\ \cdot\int_{\mathbf{x}_2} \mathcal{N}\left(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_c(\mathbf{x}_3), \mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_c\right) & \\ = \mathcal{N}\left(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_c(\mathbf{x}_3), \mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_c\right). & \end{aligned}$$

We can convert this into a density of $\mathbf{x}_3$ using the fact that for any invertible matrix $\mathbf{A}$,

$$\mathcal{N}(\mathbf{x}, \mathbf{\Sigma}) = \frac{\mathcal{N}(\mathbf{A}^{-1}\mathbf{x}, \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-1\prime})}{|\det(\mathbf{A})|} \quad (7)$$

Define $\mathbf{A}_j = \mathbf{\Sigma}_j^{ab}\ (\mathbf{\Sigma}_j^{bb})^{-1}$. We have

$$\mathcal{N}\left(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_c(\mathbf{x}_3), \mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_c\right) =$$

$$\frac{\mathcal{N}\left(\mathbf{A}_j^{-1}(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_c(\mathbf{x}_3)), \mathbf{A}_j^{-1}(\mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_c)\mathbf{A}_j^{-1}\right)}{|\det(\mathbf{A}_j)|}$$

$$= \frac{\mathcal{N}\left(\mathbf{A}_j^{-1}(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_j^a) - \mathbf{x}_3 + \boldsymbol{\mu}_j^b, \mathbf{A}_j^{-1}(\mathbf{\Sigma}_i^{bb} + \mathbf{\Sigma}_j^{aa})\mathbf{A}_j^{-1} - \mathbf{\Sigma}_j^{bb}\right)}{|\det(\mathbf{A}_j)|}.$$

So we have

$$\int_{\mathbf{x}_1,\mathbf{x}_2} L(\mathbf{Y}|\mathbf{q}) = \tfrac{1}{|\det(\mathbf{A}_j)|}\,\mathcal{N}\left(\mathbf{x}_3-\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right) \cdot b_j^b(\mathbf{x}_3)\; b_k(\mathbf{y}_4)\cdots b_q(\mathbf{y}_{T-1})\; b_r(\mathbf{y}_T), \tag{8}$$

where

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_j^b + \mathbf{A}_j^{-1}(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_j^a),$$
$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}_j^{-1}(\boldsymbol{\Sigma}_i^{bb} + \boldsymbol{\Sigma}_j^{aa})\mathbf{A}_j^{-1} - \boldsymbol{\Sigma}_j^{bb}. \tag{9}$$

We now proceed to integrate over $\mathbf{x}_3$. We re-write the product $\mathcal{N}\left(\mathbf{x}_3-\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\; b_j^b(\mathbf{x}_3)$ as

$$\mathcal{N}\left(\mathbf{x}_3-\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\;\mathcal{N}\left(\mathbf{x}_3-\boldsymbol{\mu}_j^b,\boldsymbol{\Sigma}_j^{bb}\right)$$
$$= \mathcal{N}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_j^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_j^{bb}\right)\;\mathcal{N}\left(\mathbf{x}_3-\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}}\right),$$

where

$$\hat{\hat{\boldsymbol{\Sigma}}} = \left[\hat{\boldsymbol{\Sigma}}^{-1} + (\boldsymbol{\Sigma}_j^{bb})^{-1}\right]^{-1},$$
$$\hat{\hat{\boldsymbol{\mu}}} = \hat{\hat{\boldsymbol{\Sigma}}}\left[\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + (\boldsymbol{\Sigma}_j^{bb})^{-1}\boldsymbol{\mu}_j^b\right]. \tag{10}$$

Collecting results and integrating over $\mathbf{x}_3$,

$$\int_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3} L(\mathbf{Y}|\mathbf{q}) = \tfrac{\mathcal{N}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_j^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_j^{bb})}{|\det(\mathbf{A}_j)|}$$
$$\cdot\int_{\mathbf{x}_3}\left\{\mathcal{N}\left(\mathbf{x}_3-\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}}\right)\; b_k(\mathbf{y}_4)\right\}\; b_l(\mathbf{y}_5)\cdots,$$
$$= \tfrac{\mathcal{N}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_j^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_j^{bb})}{|\det(\mathbf{A}_j)|}$$
$$\cdot\int_{\mathbf{x}_3}\left\{\mathcal{N}\left(\mathbf{x}_3-\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}}\right)\; b_k(\mathbf{x}_3|\mathbf{x}_4)\right\}\; b_k^b(\mathbf{x}_4)\; b_l(\mathbf{y}_5)\cdots, \tag{11}$$

Define

$$Q(\mathbf{x}_{t+1}\ldots\mathbf{x}_T;\{j,k,l\ldots p,q,r\},\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq$$
$$\int_{\mathbf{x}_t}\{\mathcal{N}(\mathbf{x}_t-\boldsymbol{\mu},\boldsymbol{\Sigma})\; b_j(\mathbf{x}_t|\mathbf{x}_{t+1})\}\; b_j^b(\mathbf{x}_{t+1})\; b_k(\mathbf{y}_{t+2})$$
$$\cdots b_q(\mathbf{y}_{T-1})\; b_r(\mathbf{y}_T),$$

then we may re-write (6) and (11) as

$$\int_{\mathbf{x}_1,\mathbf{x}_2} L(\mathbf{Y}|\mathbf{q}) = Q(\mathbf{x}_3\ldots\mathbf{x}_T;\{j,k\ldots p,q,r\},\boldsymbol{\mu}_i^b,\boldsymbol{\Sigma}_i^{bb})$$

and

$$\int_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3} L(\mathbf{Y}|\mathbf{q}) = \tfrac{\mathcal{N}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_j^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_j^{bb})}{|\det(\mathbf{A}_j)|}$$
$$Q(\mathbf{x}_4\ldots\mathbf{x}_T;\{k\ldots p,q,r\},\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}}). \tag{12}$$

Comparing the above equations, we can see a recursion. Because we have previously identified indexes $\{i,j,k,l\cdots p,q,r\}$ with fixed time indexes, to make a general expression for the recursion, we need to define the free indexes m,n representing the assumed Markov states at the arbitrary times t,$t+1$, respectively. The recursion is

$$\int_{\mathbf{x}_{t+1}} Q(\mathbf{x}_t\ldots\mathbf{x}_T;\{m,n\ldots p,q,r\},\boldsymbol{\mu},\boldsymbol{\Sigma}) =$$
$$\tfrac{\mathcal{N}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_m^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_m^{bb})}{|\det(\mathbf{A}_m)|}Q(\mathbf{x}_{t+1}\ldots\mathbf{x}_T;\{n\ldots p,q,r\},\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}}),$$

where

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_m^b + \mathbf{A}_m^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_m^a),$$
$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}_m^{-1}(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_m^{aa})\mathbf{A}_m^{-1} - \boldsymbol{\Sigma}_m^{bb} \tag{13}$$

and

$$\hat{\hat{\boldsymbol{\Sigma}}} = \left[\hat{\boldsymbol{\Sigma}}^{-1} + (\boldsymbol{\Sigma}_m^{bb})^{-1}\right]^{-1},$$
$$\hat{\hat{\boldsymbol{\mu}}} = \hat{\hat{\boldsymbol{\Sigma}}}\left[\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} + (\boldsymbol{\Sigma}_m^{bb})^{-1}\boldsymbol{\mu}_m^b\right]. \tag{14}$$

The recursion starts by integrating (12) over $\mathbf{x}_4$ and ends with $Q(\;,\{\;\},\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq 1$. It can be seen that the full integral

$$K(\mathbf{q}) = \int_{\mathbf{x}_1}\int_{\mathbf{x}_2}\cdots\int_{\mathbf{x}_T} L(\mathbf{Y}|\mathbf{q})$$

is obtained by the product

$$K(\mathbf{q}) = \prod_{m=j,k,\ldots p,q,r} \frac{\mathcal{N}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_m^b,\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_m^{bb}\right)}{|\det(\mathbf{A}_m)|}. \tag{15}$$

Finally, the desired integral (2) is given by

$$K_T = \sum_{\mathbf{q}\in\mathbf{Q}} p(\mathbf{q})\; K(\mathbf{q}) \tag{16}$$

Since there are M^T elements in $\mathbf{Q}$, the computation is of order $O(M^T)$, but the terms in (15) converge to a limiting distribution, since the ratio

$$\frac{Q(\mathbf{x}_{t+1}\ldots\mathbf{x}_T;\{j,k\ldots n\},\hat{\hat{\boldsymbol{\mu}}},\hat{\hat{\boldsymbol{\Sigma}}})}{Q(\mathbf{x}_t\ldots\mathbf{x}_T;\{i,j,k,l\ldots p,q,r\},\boldsymbol{\mu},\boldsymbol{\Sigma})} \to C$$

quickly converges to a constant C. This convergence is related to the property of limiting distributions for Markov chains [5] and is fortunate because $L(\mathbf{Y})$ needs only be calculated for a few values of T, then the constant C stored.

We tested the expression for K_T by comparing to the numerically-integrated PDF. We created samples of $\mathbf{X}$ by selecting the first D MFCC coeffients extracted from some arbitrary samples of speech data and trained an HMM on samples of $\mathbf{Y}$. With HMM parameters held fixed, we evaluated $L_y(\mathbf{D}(\mathbf{X}))$ using the forward procedure on a fine grid spanning the DT-dimensional space of $\mathbf{X}$. In theory the

D	T	Numerical result	K_T	K_T/K_{T-1}
1	2	0.999999	1.000000	1
1	3	0.412523	0.412307	0.412307
1	4	0.191555	0.191275	0.463914
1	5	0.092301	0.092048	0.481233
1	6		0.044915	0.487951
1	7		0.022039	0.490682
1	8		0.010839	0.491809
1	9		0.005335	0.492204
1	10		0.002628	0.492442
1	11		0.001294	0.492506
1	12		0.000637	0.492526
1	13		0.000314	0.492529
2	2	0.99999	1.000000	1
2	3	0.06426	0.063756	

Table 1. Comparison of numerically integrated likelihood function with equation (16) over feature dimension D and length T. The number of Markov states was $N = 2$.

integral equals 1.0 for $T = 2$ since in this case, $\mathbf{X}$ and $\mathbf{Y}$ are equivalent. For $D = 1$, we were able to carry out the numerical integration up to $T = 5$. For $D = 2$, the numerical integration could be carried out only up to $T = 3$. Table 1 shows the comparison of K_T with numerical integration as a function of T. Note the close agreement with K_T from equation (16). The accuracy was limited by the grid sampling used in the numerical integration since it greatly affected the computation time. The ratio K_T/K_{T-1} is shown to converge quite rapidly. Therefore the values K_T can be extrapolated to much higher T with no additional calculations.

4. EXPERIMENTS

Now that we are able to correct the DAF-HMM likelihood function so that it is a true PDF on the un-augmented features $\mathbf{X}$, we can make quantitative evaluation of the effects of feature augmentation.

4.1. Data sets

To illustrate the effect of feature augmentation, we chose two data sets with different amount of dynamic information.

1. **dyphthongs.** This data set consisted of three *dyphthongs* (phonemes with time-varying formants) from the TIMIT corpus [6]. We extracted examples of the phonemes AY, EY, and OW. An example of AY is shown in Figure 1 (left). The total number of samples were 3196 for "AY", 3030 for "EY", and 2858 for "OW". We joined all available utterances of the phonemes from both the training and testing subsets, then divided them into two sets for 2-fold holdout.

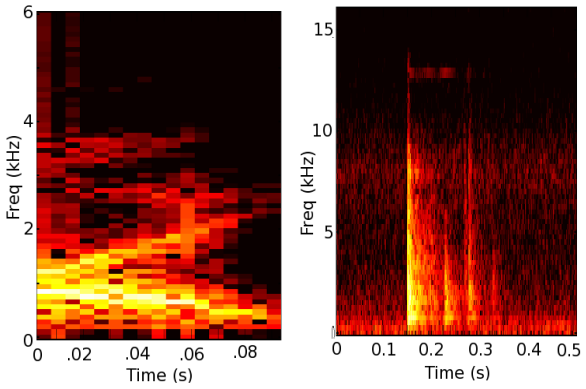

Fig. 1. Sample spectrograms. Left: dyphthong (AY). Right: office sounds (penny). Note the gradually changing spectral content of dyphthong "AY" in contrast to the abrupt character of "penny".

2. **Office sounds.** The Office Sounds database [7] contains twenty-four signal classes of 102 samples each created by dropping common objects or operating office tools such as scissors or staplers. All time-series are 16128 samples long (1/2 second in duration at 32000 Hz). We chose three classes with abrupt temporal character: *penny*, *quart*, *skit*. An example of "penny" is shown in Figure 1 (right).

4.2. Features

We extracted features by 2/3 overlapped hanning-weighted MEL frequency cepstral coefficient (MFCC) feature analysis [8]. For the TIMIT data, which is sampled at 16 KHz, we first downsampled the data to 12 KHz, then used 288-sample windows (24 milliseconds). For the office sounds data, which is sampled at 32 KHz, we used 288-sample windows (18 milliseconds). For both data sets, we used 24 Hanning-shaped MEL bands (including the zero and Nyquist bands), and no DCT truncation, producing a 24-dimensional feature.

4.3. PDF estimation

All PDFs were modeled as an HMM with Gaussian state PDFs (single-component Gaussian mixture) in accordance with the method of Rabiner [4]. We used $M = 7$ Markov states for HMM. We used fewer Markov states ($M = 5$) for DAF-HMM. These numbers were chosen by trial and error to provide the best classification performance. It makes intuitive sense also. The additional derivative information inherent in DAF permits modeling dynamic behavior with fewer discrete states. Also, the increased feature dimension of DAF makes it wise to reduce the number of states, or risk overparameterization. Each PDF was estimated from training data using five trials in which the initial parameters were randomly initialized. The PDF parameters achieving the highest

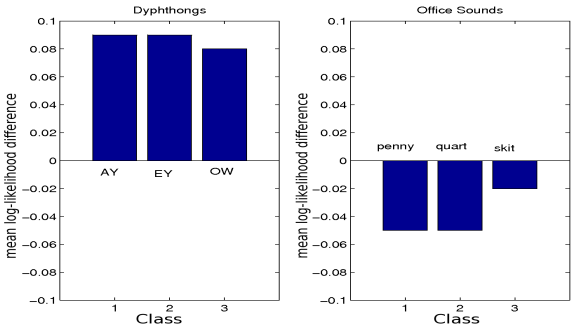

Fig. 2. Difference in μ_m for DAF-HMM-corrected and HMM for Dyphthongs (left) and Office sounds (right).

log-likelihood after convergence was chosen.

4.4. Experimental Procedure

We evaluated both likelihood function types (a) straight HMM on the un-augmented features, and (b) DAF-HMM that had been corrected by K_T, on each data set. For each data set and likelihood function type, we measured mean log-likelihood and classification error rate. Let

$$\boldsymbol{\mu}_m = \frac{1}{N_m} \sum_{k=1}^{N_m} \frac{\log L_m(\mathbf{X}_k)}{T_k},$$

where $L_m(\mathbf{X})$ is a likelihood function for class m, N_m is the number of testing samples for class m, and T_k is the length of the feature stream for sample k. We only evaluated a likelihood function on data from it's own class. We assume that the N_m testing samples have been separated from the training data used to train $L_m(\mathbf{X})$. To separate the data, we trained on half of the available samples, then determined $\boldsymbol{\mu}_m$ on the other half. We then switched the halves and avaraged the results. We also evaluated the classification error rate in percent for each likelihood function type, using the same data separation.

4.5. Results

Figure 2 shows the results of the mean log-likelihood experiment. The vertical scale of the bar-graph equals $\boldsymbol{\mu}_{m(DAF)} - \boldsymbol{\mu}_{m(HMM)}$. In the dyphthong data, the values are positive, indicating that DAF-HMM produces a higher log-likelihood. For office sounds data, DAF-HMM produces a lower log-likelihood. This indicates that the spectral content of the dyphthongs data was smoothly changing and feature time-derivatives are more meaningful from a statistical modeling point of view. For the abrupt sounds in the office sounds data, the feature time difference were random and not predictable. Thus, augmenting the features did more harm than good.

Below we list the results of the classification performance experiment.

Classification Error (percent)			
Dyphthongs		Office Sounds	
HMM	DAF-HMM	HMM	DAF-HMM
9.6%	7.5%	0.66%	1.64%

The results are consistent with the mean log-likelihood experiment. They indicate that augmenting the feature is helpful for the Dyphthong data, but detrimental for the office sounds data.

5. CONCLUSIONS

We have derived an expression for the integral of the DAF-HMM likelihood function with respect to the un-augmented features. This allows normalizing the DAF-HMM likelihood function so that it can be compared with likelihood functions based on the un-augmented features. We demonstrated the use of the method on two data sets. In particular, we have shown that appending feature time derivatives achieves lower classification error as well as higher average log-likelihood for data with slowly-varying spectral character. For data with abrupt spectral character, the opposite was observed. This indicates the possibility of a "litmus test" for the use of DAF. It suggests the possibility of using DAF-HMM along with HMM togeter in a single classifier.

REFERENCES

[1] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans on ASSP*, vol. 34, no. 1, pp. 52–59, 1986.

[2] Kannan Achan, Sam Roweis, and Brendan Frey, "A segmental hmm for speech waveforms," *University of Toronto Technical Report*, Jan 2004.

[3] Jeffrey A. Bilmes, "Graphical models and automatic speech recognition," *Mathematical Foundations of Speech and Language Processing*, 2003.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[5] Dani Gamerman, *Markov Chain Monte Carlo*, Chapman and Hall, 1997.

[6] John S. Garofolo, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[7] P. Baggenstoss, "Office sounds database," *http://class-specific.com/os*.

[8] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, p. 374388, 1976.