Improving Automatic Speech Recognition Using Tangent Distance

W. Macherey, D. Keysers, J. Dahmen, and H. Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
{w.macherey, keysers, dahmen, ney}@informatik.rwth-aachen.de

Abstract

In this paper we present a new approach to variance modelling in automatic speech recognition (ASR) that is based on *tangent distance* (TD). Using TD, classifiers can be made invariant w.r.t. small transformations of the data. Such transformations generate a manifold in a high dimensional feature space when applied to an observation vector. While conventional classifiers determine the distance between an observation and a prototype vector, TD approximates the minimum distance between their manifolds, resulting in classification that is invariant w.r.t. the underlying transformation. Recently, this approach was successfully applied in image object recognition. In this paper we describe how TD can be incorporated into ASR systems based on Gaussian mixture densities (GMD). The proposed method is embedded into a probabilistic framework. Experiments performed on the *SieTill* corpus for telephone line recorded German digit strings show a significant improvement in comparison with a conventional GMD approach using a comparable amount of model parameters.

1. Introduction

The design of a classifier that is invariant w.r.t. certain transformations is an important aspect in pattern recognition. Many approaches to invariant pattern recognition are known [1], among them an invariant distance measure called *tangent distance*. TD was proposed in [2, 3] and proved to be very effective in the domain of optical character recognition. Distance measures like the Euclidean distance and related ones are very sensitive to small transformations, even though these transformations do not affect class membership. In contrast to that, TD is able to partially compensate the effect of such transformations. The approach has been successfully applied in different image object recognition tasks [4]. In this paper we demonstrate, how TD can successfully be incorporated into ASR systems that are based on Gaussian mixture densities. For this, TD is embedded into a statistical framework. In section 2 we motivate TD on the basis of a comparison with the Euclidean distance. Section 2.1 presents a probabilistic interpretation of TD and describes the effect on the Mahalanobis distance. Section 3 deals with the incorporation of TD into ASR systems based on GMDs. A discussion of the experimental results obtained on the *SieTill* corpus for continuous digit strings concludes the paper.

2. Overview of tangent distance

Let $x \in I\!R^D$ be a pattern and $f(x, \alpha)$ denote a transformation that depends on a parameter L-tuple $\alpha \in I\!R^L$. Then the set of points of all transformations of the pattern x is a manifold $\mathcal{M}_x$

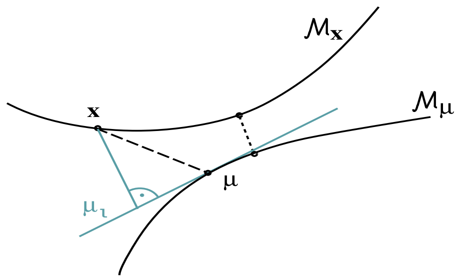

Figure 1: *Illustration of the Euclidean distance between an observation x and a reference vector μ (dashed line) in comparison with the minimal distance between the corresponding manifolds (dotted line). The tangent approximation is depicted by the light gray lines.*

of at most dimension L in pattern space:

$$\mathcal{M}_x = \{f(x, \alpha) \,:\, \alpha \in I\!R^L\} \subset I\!R^D \tag{1}$$

Consider f with the property that (small) transformations of the pattern do not affect class membership. If the discriminant function for a class c is based on e.g. the Euclidean distance, $d(x, c)$ and $d(f(x, \alpha), c)$ may no longer be equal for certain α which could lead to misclassification. In contrast to that a classifier would be invariant w.r.t. f, if the discriminant function was based on the minimum distance between the manifold $\mathcal{M}_x$ of a pattern x and the manifold $\mathcal{M}_\mu$ of a class specific prototype vector μ (cf. Figure 1):

$$d_{\text{Manifold}}(x, \mu) = \min_{\alpha, \beta \in I\!R^L} \{||f(x, \alpha) - f(\mu, \beta)||^2\} \tag{2}$$

However, distance calculation between manifolds is a hard nonlinear optimization problem. Moreover, a manifold does not have a closed expression in general, so it cannot be handled in an analytical way. To overcome these problems the manifolds can be approximated by a *tangent subspace* $\widehat{\mathcal{M}}$. The *tangent vectors* x_l that span the tangent subspace are defined as the partial derivatives of a transformation f w.r.t. to its parameters α_l ($l = 1, \ldots, L$):

$$x_l = \partial f(x, \alpha)/\partial \alpha_l \tag{3}$$

Using this definition the transformation $f(x, \alpha)$ can be approximated as a Taylor expansion around $\alpha = 0$:

$$f(x, \alpha) = x + \textstyle\sum_l \alpha_l x_l + \mathcal{O}(\alpha_l^2) \tag{4}$$

The set of points consisting of all linear combinations of the pattern x with the tangent vectors x_l forms the tangent subspace

$\widehat{\mathcal{M}}_x$, which is a first-order approximation of $\mathcal{M}_x$:

$$\widehat{\mathcal{M}}_x = \left\{x + \textstyle\sum_l \alpha_l x_l \;:\; \alpha \in I\!R^L\right\} \subset I\!R^D \tag{5}$$

The definition of $\widehat{\mathcal{M}}_x$ has the advantage that it is a linear approximation of the manifold $\mathcal{M}_x$ and thus easy to use in distance calculations. A drawback is that the distance measure is no longer globally invariant w.r.t. f, but only locally invariant. On the other hand global invariance may not be necessary since sometimes, large transformations of a pattern do not respect class membership. Using the squared Euclidean norm TD is then defined as:

$$d_{2S}(x,\mu) = \min_{\alpha,\beta}\left\{||(x + \textstyle\sum_l \alpha_l x_l) - (\mu + \sum_l \beta_l \mu_l)||^2\right\} \tag{6}$$

Eq. (6) is also known as *two-sided* tangent distance (2S) [5]. In order to reduce the effort for determining $d_{2S}(x,\mu)$ it is convenient to restrict the calculation of the tangent subspaces to prototype vectors. The resulting distance measure is called *one-sided* tangent distance (1S).

$$d_{1S}(x,\mu) = \min_{\alpha}\left\{||x - (\mu + \textstyle\sum_l \alpha_l \mu_l)||^2\right\} \tag{7}$$

Even though the new distance measure has been introduced using the Euclidean distance, the same applies as well for the Mahalanobis distance, as will be shown in the next section.

2.1. A probabilistic framework for tangent distance

For the purpose of embedding TD into a statistical framework we will focus on the consideration of one-sided TD, assuming that only the references are subject to variations. A detailed overview including the two-sided TD can be found in [6].

For the moment we assume that the tangent vectors μ_l are known. The observations x shall be normal distributed with expectation μ and covariance matrix Σ. In order to simplify the notation, class indices are omitted. Using the first-order approximation of the manifold $\mathcal{M}_\mu$ for a mean vector μ one obtains the probability density function (pdf) for the observations x:

$$p(x \,|\, \mu, \alpha, \Sigma) = \mathcal{N}(x \,|\, \mu + \textstyle\sum_l \alpha_l \mu_l,\, \Sigma) \tag{8}$$

The integral of the joint distribution $p(x, \alpha \,|\, \mu, \Sigma)$ over the unknown transformation parameters α leads to the following distribution:

$$\begin{aligned} p(x \,|\, \mu, \Sigma) &= \int p(x, \alpha \,|\, \mu, \Sigma)\, d\alpha \\ &= \int p(\alpha \,|\, \mu, \Sigma) \cdot p(x \,|\, \mu, \alpha, \Sigma)\, d\alpha \\ &= \int p(\alpha) \cdot p(x \,|\, \mu, \alpha, \Sigma)\, d\alpha \end{aligned} \tag{9}$$

Note that we assume α is independent of μ and Σ. Thus, $p(\alpha \,|\, \mu, \Sigma) \equiv p(\alpha)$ applies. The α_l are assumed to be normal distributed with mean 0 and a covariance matrix $\gamma^2 I$, i.e.

$$p(\alpha) = \mathcal{N}(\alpha \,|\, 0, \gamma^2 I), \tag{10}$$

where I denotes the identity matrix and γ^2 is an empirical parameter. W.l.o.g., the tangent vectors of the pdf in Eq. (8) can be assumed as stochastically independent since they form a basis of the tangent subspace. Hence, it is always possible to decorrelate the tangent vectors using e.g. a singular value decomposition. The evaluation of the integral in Eq. (9) leads to the following expression [6]:

$$\begin{aligned} p(x \,|\, \mu, \Sigma) = \Big|\Sigma + \gamma^2 \sum_{l=1}^{L} \mu_l \mu_l^T\Big|^{-\frac{1}{2}} \cdot \\ \exp\Bigg\{ -\frac{1}{2}\Big[(x-\mu)^T\Big(\Sigma^{-1} - \\ \sum_{l=1}^{L} \frac{[\mu_l^T \Sigma^{-1}]^T [\mu_l^T \Sigma^{-1}]}{1/\gamma^2 + \mu_l^T \Sigma^{-1} \mu_l}\Big)(x-\mu)\Big]\Bigg\} \end{aligned} \tag{11}$$

Note that the exponent in Eq. (11) leads to conventional Mahalanobis distance for $\gamma \to 0$ and TD for $\gamma \to \infty$. Thus, the incorporation of tangent vectors adds a corrective term to the Mahalanobis distance that only affects the covariance matrix which can be interpreted as structuring Σ [7].

3. Incorporating TD into ASR

In the last section the assumption was made that the transformations for which invariance is desired are known. However, in contrast to most image object recognition tasks, the transformations to be selected are not obvious in ASR and often there is no prior knowledge available. In order to circumvent this difficulty, the tangent vectors can be learned from the training data. As there is class specific variation in the data, we obtain a suitable approximation of the tangent vectors by estimating the class specific variance and determining its derivatives. The estimation of the tangent vectors can be formulated within a maximum likelihood approach.

For this let the training data be given by $n = 1, \ldots, N$ training utterances, each consisting of a sequence of acoustic observation vectors $x_{n,1}, x_{n,2}, \ldots, x_{n,T_n}$. The HMM state to which an acoustic observation is aligned to during the training phase shall be denoted with $s(n,t)$. ϑ shall comprise all distribution parameters, i.e. class specific means, variances, mixture weights, and tangent vectors. In the following the class indices will be identified with HMM states. Assuming that the number L of tangent vectors is known (note that L can be determined automatically [8]) the objective function that has to be maximized over all training samples is given by:

$$F_\vartheta = \sum_s \sum_{n=1}^{N} \sum_{t=1}^{T_n} \delta_{s,s(n,t)} \log p(x_{n,t} \,|\, \mu_s, \Sigma) \stackrel{!}{=} \max \tag{12}$$

Here, δ denotes the Kronecker delta. W.l.o.g. we can assume that the vectors $(\Sigma^{-1/2})^T \cdot \mu_{sl}$ are orthonormalized, i.e.:

$$\mu_{sl}^T \Sigma^{-1} \mu_{sm} = \mu_{sl}^T \Sigma^{-1/2} (\Sigma^{-1/2})^T \mu_{sm} = \delta_{lm}, \tag{13}$$

where $\Sigma^{-1/2}$ is defined as $A \cdot \Omega^{-1/2}$ with $A := [v_1, \ldots, v_D]$ and $\Omega := \mathrm{diag}(\omega_1, \ldots, \omega_D)$, where v_d is the eigenvector of the eigenvalue problem $\Sigma \cdot v_d = \omega_d \cdot v_d$ for $d = 1, \ldots, D$. Now, the normalization term in Eq. (11) is a constant in s and thus, Eq. (12) leads to the following expression (constant terms have been dropped):

$$\begin{aligned} \sum_s \sum_{n=1}^{N} \sum_{t=1}^{T_n} \delta_{s,s(n,t)} \cdot \Big[(x_{n,t} - \mu_s)^T \Sigma^{-1} (x_{n,t} - \mu_s) - \\ \sum_{l=1}^{L} \frac{[(x_{n,t} - \mu_s)^T \Sigma^{-1} \mu_{sl}]^2}{1/\gamma^2 + \mu_{sl}^T \Sigma^{-1} \mu_{sl}}\Big] \stackrel{!}{=} \min \end{aligned} \tag{14}$$

For a fixed state s this is equivalent to the maximization of

$$\sum_{l=1}^{L} \frac{\mu_{sl}^T (\Sigma^{-1})^T V_s \Sigma^{-1} \mu_{sl}}{1/\gamma^2 + \mu_{sl}^T \Sigma^{-1} \mu_{sl}} \stackrel{!}{=} \max \qquad (15)$$

with

$$V_s = 1/N_s \cdot \sum_n \sum_t \delta_{s,s(n,t)} (x_{n,t} - \mu_s)(x_{n,t} - \mu_s)^T$$

as the state specific empirical covariance matrix (N_s denotes the number of training vectors that are aligned to the state s). Σ and V_s can be regarded as covariance matrices of two competing models. Their concrete form is discussed in section 3.1. Taking the constraints of orthonormality of the tangent vectors w.r.t. $\Sigma^{-1/2}$ into account, the objective function is modified using state specific Lagrange multipliers λ_{slm}:

$$\mathcal{F}_\vartheta(\{\lambda_{slm}\}) = F_\vartheta - \sum_s \sum_{l=1}^{L} \sum_{m=1}^{L} \lambda_{slm} \cdot \left(\mu_{sl}^T \Sigma^{-1/2} (\Sigma^{-1/2})^T \mu_{sm} - \delta_{lm}\right) \qquad (16)$$

The derivation of Eq. (16) w.r.t. $\widehat{\mu}_{sl} := (\Sigma^{-1/2})^T \cdot \mu_{sl}$ and equating the result to zero yields [9]:

$$\partial \mathcal{F}_\vartheta / \partial \widehat{\mu}_{sl} \stackrel{!}{=} 0 \quad \Longleftrightarrow \quad V_s \Sigma^{-1} \mu_{sl} = \lambda_{sll} (1/\gamma^2 + 1) \mu_{sl}$$

Thus, the maximization of $\mathcal{F}_\vartheta$ is equivalent to the solution of a generalized symmetric eigenvalue problem, where the eigenvalues correspond with the Lagrange multipliers λ_{sll}. The state specific tangent vectors μ_{sl} maximizing Eq. (15) are those eigenvectors with the largest corresponding eigenvalues.

3.1. Models for the covariance matrices

As mentioned in the previous section two different models have to be determined for the covariance matrices Σ and V_s. While V_s is defined as a state specific covariance matrix, a globally pooled covariance matrix proved to be a suitable choice for Σ. Using these models the effect of incorporating TD into Mahalanobis distance is equivalent to performing a global whitening transformation of feature space and then employing the L state specific eigenvectors with the largest eigenvalues. This eliminates those directions of class specific variation that cause the greatest reconstruction error towards Σ. Because of this, TD has the advantage that it also works very well in combination with globally operating feature transformations as for instance a linear discriminant analysis (LDA), since Σ can obviously be assumed as a global covariance matrix of an LDA transformed feature space.

3.2. Adjusting γ^2

Preliminary experiments on TD have shown that the covariance matrix in Eq. (11) tends to become singular, especially if more than one tangent vector is used. In order to compensate this effect, γ is adjusted mixture specifically in the following way. For a state index s γ_{sl} is chosen as the l-th eigenvalue λ_{sll} scaled by a mixture specific factor η_s which ensures that the resulting matrix is positive definite. For the experiments described in section 4 each η_s was initially chosen as a power of 2. Then η_s was as often halved until the resulting covariance matrix was no longer singular.

4. Experimental results

Experiments were performed on the *SieTill* corpus [10] for telephone line recorded German continuous digit strings. The corpus consists of approximately 43k spoken digits in 13k sentences for both training and test set. In Table 1 some information on corpus statistics is summarized.

Table 1: *Corpus statistics for the SieTill corpus.*

corpus	female		male	
	sent.	digits	sent.	digits
test	6176	20205	6938	22881
train	6113	20115	6835	22463

The recognition system is based on whole word HMMs using continuous emission densities. The baseline system is characterized as follows:

- vocabulary of 11 German digits including '*zwo*',
- gender-dependent whole-word HMMs, with every two subsequent states being identical,
- for each gender 214 distinct states plus one for silence
- Gaussian mixture emission distributions,
- one globally pooled diagonal covariance matrix Σ,
- 12 cepstral features plus first derivatives and the second derivative of the first feature component.

The baseline recognizer applies ML training using the Viterbi approximation in combination with an optional LDA. A detailed description of the baseline system can be found in [11]. The word error rates obtained with the baseline system for the combined recognition of both genders are summarized in Table 2 (0 tangent vectors (tv) per mixture (mix)). The V_s were trained as state specific full covariance matrices. Note that the V_s are only necessary in the training phase.
For single densities the incorporation of TD improved the word error rate by 18.1% relative for one tangent vector and 21.6% relative using four tangent vectors per state. In combination with LDA transformed features the relative improvement was 13.8% for the incorporation of one tangent vector and increased

Table 2: *Word error rates (WER) on the SieTill corpus obtained with tangent distance. In column 'tv/mix' the number of used tangent vectors per mixture is given. A value of* 0 *means that the conventional Mahalanobis distance is used. 'dns/mix' gives the average number of densities per mixture.*

LDA	dns/mix	tv/mix	error rates [%]		
			del - ins	WER	SER
no	1	0	1.17-0.83	4.59	11.34
		1	1.17-0.52	3.76	9.22
		4	0.69-1.07	3.60	9.10
	16	0	0.59-0.83	2.67	6.92
		1	0.54-0.58	2.49	6.56
		4	0.46-0.80	2.60	6.76
	128	0	0.52-0.54	2.24	5.87
		1	0.50-0.48	2.12	5.75
		4	0.55-0.49	2.13	5.71
yes	1	0	0.71-0.63	3.78	9.74
		1	0.97-0.49	3.26	8.46
		5	**0.48-0.88**	**2.70**	**7.18**
	16	0	0.44-0.68	2.28	5.92
		1	0.58-0.40	1.97	5.06
		4	0.38-0.55	1.97	5.35
	128	0	0.45-0.39	1.85	4.94
		1	**0.42-0.34**	**1.67**	**4.50**
		4	0.39-0.41	1.76	4.81

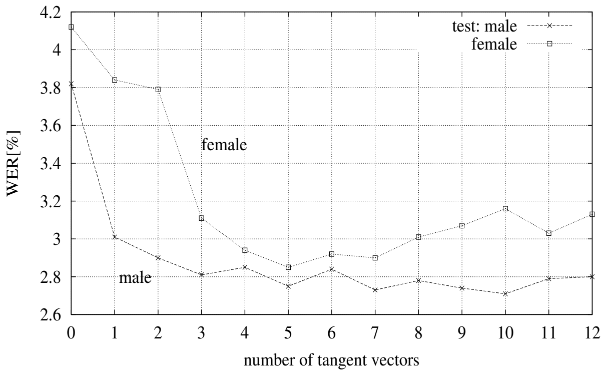

Figure 2: *Evolution of word error rates on the SieTill test corpus for single densities using ML training on LDA transformed features for different numbers of tangent vectors.*

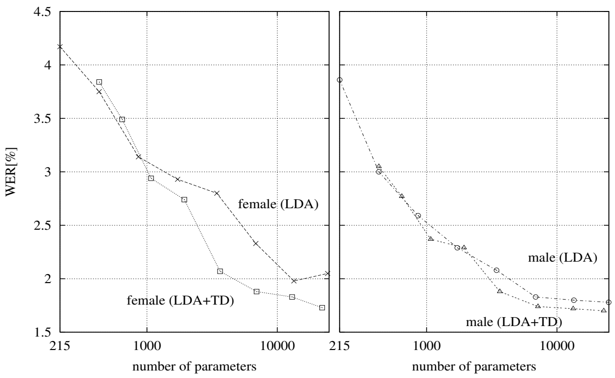

Figure 3: *Comparison of WER for mixture densities on the SieTill test corpus using equal overall parameter numbers.*

to 28.6% for five tangent vectors per state. Figure 2 depicts the evolution of the word error rates on the *SieTill* test corpus for different numbers of tangent vectors using single densities that were trained on LDA transformed features. For this setting the optimal choice for gender dependent trained references was five tangent vectors per state.

Using mixture densities the performance gain in word error rate decreased but was still significant. Thus the relative improvement between the baseline result and TD was 6.7% (16 densities plus one tangent vector per mixture) for untransformed features and 13.6% for LDA transformed features (16 dns/mix, 1 tv/mix). The same applies for the optimal number of tangent vectors which was found at one tangent vector per mixture. Consequently, a larger number of densities is able to partially compensate for the error that is made in the case that the covariance matrix is estimated using the conventional method. The best result was obtained using 128 densities per mixture in combination with LDA transformed features and the incorporation of one tangent vector per state. Using this setting the word error rate decreased from 1.85% to 1.67% which is a relative improvement of 5%. Figure 3 depicts the evolution of word error rates for conventional training in comparison with TD using equal numbers of parameters. Even though the incorporation of tangent vectors into the Mahalanobis distance increases the number of parameters that are necessary to modify the globally pooled variance the overall gain in performance justifies the higher expense.

5. Conclusion

In this paper we presented a new approach for modelling variances in automatic speech recognition based on tangent distance (TD). For that purpose TD was embedded into a probabilistic framework. In accordance with the theory, the new model proved to be very effective in combination with globally operating feature transformations as the linear discriminant analysis. Comparative experiments were performed on the *SieTill* corpus for continuous German digit strings. Using one-sided TD, a relative improvement in word error rate of approximately 20% was achieved for single densities. For mixture densities we could gain a relative improvement of up to 13.6% in word error rate. Incorporating TD we were able to reduce the word error rate of our best recognition result based on ML trained references from 1.85% to 1.67%.

6. References

[1] J. Wood, “Invariant pattern recognition: A review.,” *Pattern Recognition*, vol. 29, pp. 1–17, Jan. 1996.

[2] P. Simard, B. Victorri, Y. Le Cun, and J. Denker, “Tangent prop – A formalism for specifying selected invariances in an adaptive network,” in *Advances in Neural Information Proc. Systems*, vol. 4, pp. 895–903, Morgan Kaufmann Publishers, Inc., 1992.

[3] P. Simard, Y. Le Cun, and J. Denker, “Efficient pattern recognition using a new transformation distance,” in *Advances in Neural Information Proc. Systems*, vol. 5, (San Mateo, CA), pp. 50–58, Morgan Kaufmann, 1993.

[4] D. Keysers, J. Dahmen, T. Theiner, and H. Ney, “Experiments with an extended tangent distance,” in *Proceedings 15th International Conference on Pattern Recognition*, vol. 2, (Barcelona, Spain), pp. 38–42, Sept. 2000.

[5] R. Duda, P. Hart, and D. G. Stork, *Pattern Classification*, pp. 191–192. John Wiley & Sons, 2nd ed., 2000.

[6] D. Keysers, J. Dahmen, and H. Ney, “A probabilistic view on tangent distance,” in *22. DAGM Symposium Mustererkennung 2000*, (Kiel, Germany), pp. 107–114, Springer, Sept. 2000.

[7] J. Dahmen, D. Keysers, M. Pitz, and H. Ney, “Structured covariance matrices for statistical image object recognition,” in *22. DAGM Symposium Mustererkennung 2000*, (Kiel, Germany), pp. 99–106, Springer, Sept. 2000.

[8] C. M. Bishop, “Bayesian PCA,” in *Advances in Neural Information Processing Systems*, vol. 11, pp. 382–388, Morgan Kaufmann Publishers, Inc., 1998.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, pp. 400–405. San Diego, CA: Computer Science and Scientific Computing Academic Press Inc., 2nd ed., 1990.

[10] T. Eisele, R. Haeb-Umbach, and D. Langmann, “A comparative study of linear feature transformation techniques for automatic speech recognition,” in *Proc. of Int. Conf. on Spoken Language Processing*, vol. I, (Philadelphia, PA), pp. 252–255, Oct. 1996.

[11] L. Welling, H. Ney, A. Eiden, and C. Forbrig, “Connected digit recognition using statistical template matching,” in *1995 Europ. Conf. on Speech Communication and Technology*, vol. 2, (Madrid, Spain), pp. 1483–1486, Sept. 1995.