OUTLIER CORRECTION FOR LOCAL DISTANCE MEASURES IN EXAMPLE BASED SPEECH RECOGNITION

Mathias De Wachter, Kris Demuynck and Dirk Van Compernolle

Katholieke Universiteit Leuven –Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven
{mathias.dewachter,kris.demuynck,dirk.vancompernolle}@esat.kuleuven.be

ABSTRACT

Example based speech recognition is critically dependent on the quality of the acoustic distance measure between input and reference vectors. In the past, the commonly used Euclidean distance has been refined to take into account the covariance of the different sounds, resulting in a class dependent distance measure. However, using the same measure for the whole class is still too crude: vectors in the tails of the distribution (outliers) are unduly considered equally representative of the class as those in the centre.

In this paper, we derive two techniques inspired by non-parametric density estimation that explicitly adjust the distance measure based on the position of the reference vector in its class. Experiments on three low-level acoustic tasks show that "data sharpening" results in a substantial improvement, while "adaptive kernels" have minimal effect.

Index Terms— Example based recognition, DTW, Adaptive kernels, Non-parametric density estimates

1. INTRODUCTION

Speech recognition research has been dominated by Hidden Markov Models (HMMs) for over a quarter century. This enormous research effort has produced many valuable extensions to the basic framework of modeling speech as a succession of (hidden) independent stable states. These extensions, in conjunction with the good scalability of the HMM framework, have allowed the successful deployment of state-of-the-art HMM recognizers in commercial applications. Still, there is a growing research community that is frustrated by the limitations that lie at the heart of the HMM framework [1].

Over the past few years we have been working on a revival of example based speech recognition —a.k.a. dynamic time warping (DTW), template based recognition or episodic modeling— as an alternative to HMMs [2, 3, 4, 5]. Example based recognition offers a natural solution to some of the problems HMMs face, especially those related to long-term sequential modeling [5]. Since the two approaches are quite similar, many of the engineering solutions from the HMM framework can be ported to example based recognition.

In HMMs the acoustic scores in essence are a complex (based on e.g. Gaussian mixtures) weighted distance between the test vectors and the state mean vector, while in DTW the distance between the test vectors and their nearest neighbours in the reference database is relevant. Hence outlier reference vectors will have a much larger impact in a DTW based system than in HMMs as each individual outlier is considered a fully acceptable representation of the class.

In [2] we showed that a class based weighted distance measure improves significantly over the more traditional Euclidean distance metric. In this paper we refine this approach with some ideas borrowed from the domain of non-parametric density estimation that explicitly compensate for the position of a reference vector in its class.

2. LOCAL DISTANCE MEASURES

2.1. Between-frame distances: a probabilistic view

The local distance measure proposed in [2] extends the Euclidean distance with a local scaling (covariance matrix):

$$d(\mathbf{x};\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \hat{\mathbf{\Sigma}}_c^{-1} (\mathbf{x}-\mathbf{y}) + \log|\hat{\mathbf{\Sigma}}_c|, \qquad (1)$$

with $\mathbf{x}$ an input frame, $\mathbf{y}$ a reference frame and $\hat{\mathbf{\Sigma}}_c$ an estimate of the covariance matrix of the class c to which the reference frame belongs. In [2] the mapping $\mathbf{y} \mapsto c$ was obtained from a state-level database segmentation by tree-clustered context dependent HMMs, since that information was readily available. Entirely example based methods to estimate the data covariance, such as the one used in [6], could be used as well.

For the remainder of this discussion, it is more appropriate to use a probabilistic view of the scaled local distance:

Definition 1 *The (M-variate)* **Gaussian kernel function** $\kappa(\mathbf{z};\mathbf{\Sigma})$ *with scaling matrix* $\mathbf{\Sigma}$ *is given by*

$$\kappa(\mathbf{z};\mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{\Sigma}^{-1} \mathbf{z}\right). \qquad (2)$$

Definition 2 *The* **local Mahalanobis distance** *between an input* $\mathbf{x}$ *and a database vector* $\mathbf{y}$ *belonging to class* c *is*

$$d_{LM}(\mathbf{x};\mathbf{y}) \quad = \quad -\log\left(\kappa(\mathbf{x}-\mathbf{y};\hat{\mathbf{\Sigma}}_c)\right). \qquad (3)$$

The above definition of the local Mahalanobis distance adds a constant scaling factor and additive term compared to equation 1. This is of no importance in our recognizer as both the offset and the scale factor are identical for all templates.

The explicit change of terminology is useful when discussing example based recognition in the light of related statistical modeling techniques. Especially comparisons with non-parametric density estimation, where it is common to view each sample as a kernel centered around the sample, provided valuable inspiration.

This research was funded by the Fund for Scientific Research Flanders (FWO-project G.0249.03) and by the IWT in the GBOU programme (project number 020192).

Definition 3 *The* **Parzen density estimator** *or* **kernel density estimator** $\hat{f}(\mathbf{x})$ *with an arbitrary kernel function* K *and* **bandwidth** *or* **window width** h *based on* n *samples* $\mathbf{y}_1^n$ *is*

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^M} \sum_{i=1}^{n} K\left(\frac{1}{h}(\mathbf{x} - \mathbf{y}_i)\right). \tag{4}$$

For our DTW system, the kernel function was always set to the Gaussian kernel $\kappa(\mathbf{z}; \mathbf{\Sigma}_c)$.

It should be stressed that, although in the following we will discuss techniques from Parzen density estimation and perform some basic classification experiments, the final DTW problem differs significantly from both Parzen density estimation and nearest neighbour classification. We are looking for neither the average kernel score (Parzen) nor the maximal kernel score (nearest neighbour classification). In DTW recognition, each frame is part of a template, and the ultimate recognition is based on template scores while the local distance is defined between frames. Averaging (cf. forward pass score) or maximizing (cf. Viterbi score) is only done at the template level and final recognition at the sentence level.

2.2. Adaptive kernels

In Parzen density estimation —in fact in most smoothing estimators— there is a well-known trade-off between correctly modeling the tails (or in general any low-density regions) of a distribution and capturing sufficient detail in the main part of the distribution [7]. Since the distance between kernels is larger in the tails, a large bandwidth is needed to "hide" the individual observations in the tails which cause a noisy estimate. However, large bandwidths will oversmooth the main part of the distribution.

Breiman et al. [8] proposed adaptive kernel estimates as a solution to this problem in the context of Parzen density estimation. In an adaptive kernel estimator, each kernel has a different bandwidth that is proportional to some local density estimate called the pilot density. Breiman et al. used k nearest neighbours (kNN) estimators for the pilot density, but they also showed that the method is robust with respect to the exact pilot density.

A general algorithm for adaptive kernel estimation is given in [7]:

Step 1 Find a **pilot estimate** $\tilde{f}$ that satisfies $\tilde{f}(\mathbf{y}_i) > 0$ for all reference vectors $\mathbf{y}_i$.

Step 2 Define **local bandwidth factors** λ_i by

$$\lambda_i = \left\{\tilde{f}(\mathbf{y}_i)/g\right\}^{-\alpha} \tag{5}$$

where g is the geometric mean of the $\tilde{f}(\mathbf{y}_i)$ values:

$$g = \left(\prod_{i=1}^{n} \tilde{f}(\mathbf{y}_i)\right)^{1/n}$$

and α is the **sensitivity parameter** ($0 \leq \alpha \leq 1$).

Step 3 Define the **adaptive kernel estimate** $\hat{f}(\mathbf{x})$ by

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(h\lambda_i)^M} K\left(\frac{1}{h\lambda_i}(\mathbf{x} - \mathbf{y}_i)\right). \tag{6}$$

In high-dimensional spaces most samples are located in the tails of the distribution [7]. Since acoustic vectors in speech recognition are typically of length 20 or more, a DTW setup will suffer from problems related to heavy-tailed distributions. Hence we expect that making the kernels adaptive in the local distance measure will alleviate some of these problems. In our experiments with real speech data, a kNN density estimate was used as a pilot distribution. Definition 2 is thus extended to

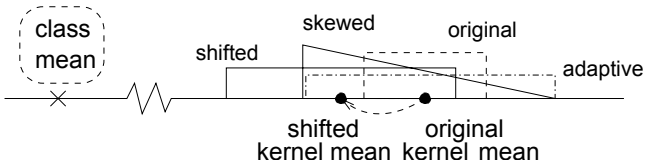

Fig. 1. Schematic motivation for data sharpening using rectangular kernels: the original, adaptive, skewed and shifted kernels are drawn.

Definition 4 *The* **adaptive kernel local Mahalanobis distance** *between an input* $\mathbf{x}$ *and a database vector* $\mathbf{y}$ *of class* c *is defined as*

$$d_{AKLM}(\mathbf{x}; \mathbf{y}) \quad = \quad -\log\left(\kappa(\frac{\mathbf{x} - \mathbf{y}}{\lambda_y}; \mathbf{\Sigma}_c)\right), \tag{7}$$

with λ_y *the local bandwidth factor for database vector* $\mathbf{y}$.

2.3. Data sharpening: adjusting the kernel means

The use of adaptive kernels results in wide kernels in low density areas, meaning the distance to these kernels varies more slowly. However, the (symmetric) kernels in the tail are smoothed both towards the center (the more likely region) and away from the center, where there are very few or no data points. As a result, the region where the maximal kernel density for a class is non-negligible is enlarged. Having a larger likely area may mean higher confusability/overlap. Therefore, adaptiveness can easily overshoot the intended smoothing of the tail estimates.

Figure 1 sketches the problem. A possible remedy would be to skew the kernel function towards the class mean (schematically shown as the full-line triangle in the figure). However, as this implies giving up the symmetry of the kernels, such a solution would be complex.

Instead, we propose a shift of the kernel *means* towards the center of the class to shrink the likely area. To this end, each kernel mean is transformed by

$$\hat{\psi}(\mathbf{y}_i) = \frac{1}{k} \sum_{j=0}^{k-1} \mathbf{y}_{(j)}, \tag{8}$$

with $\mathbf{y}_{(j)}$ the jth nearest neighbour (in local Mahalanobis sense) of $\mathbf{y}_i$ within the same class, and $\mathbf{y}_{(0)}$ being $\mathbf{y}_i$ itself. For all our experiments, we used the same k nearest neighbours as for the calculation of the pilot estimate.

Similar methods can be found in the literature. Fukunaga and Hostetler [9] were the first to use an iterative version of this idea in cluster analysis. Their approach, called *mean shift* was later formalized and shown to be a general method which includes, for example, the popular k-means clustering algorithm [10]. Choi and Hall [11] used the same idea as a preprocessing step for non-parametric density estimation, calling it *data sharpening*.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed extensions, the Euclidean, local Mahalonobis and adaptive kernel local Mahalanobis

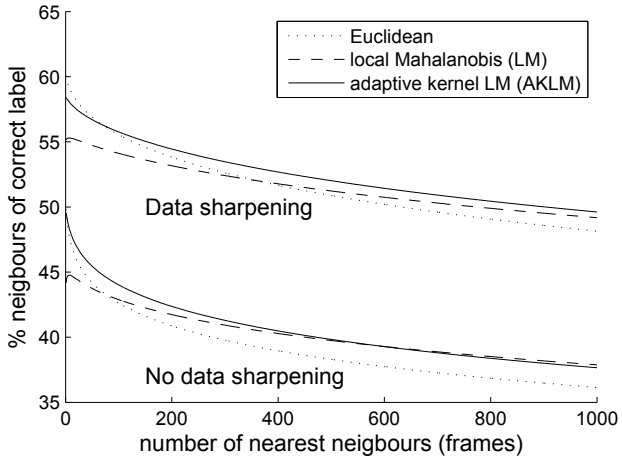

Fig. 2. Percentage of nearest neighbour frames of the correct phone identity for an increasing number of neighbours.

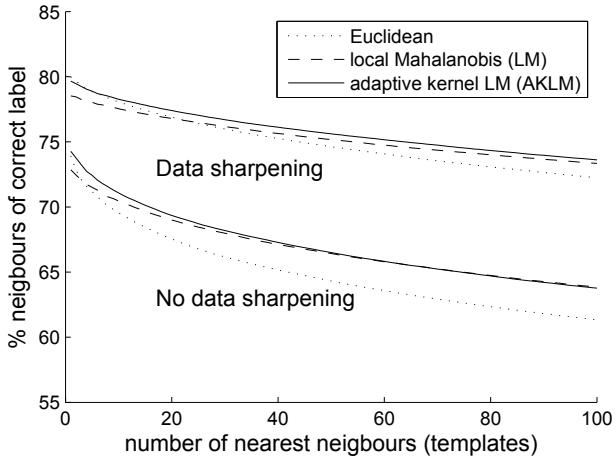

Fig. 3. Percentage of nearest neighbour templates of the correct phone identity for an increasing number of neighbours.

distances, each with and without data sharpening, are compared on three different low-level acoustic tasks: frame-based phone classification, template-based phone classification and phone string recognition.

3.1. Experimental setup

All experiments are evaluated on the November 92 5k WSJ benchmark. The SI-84 WSJ0 database, which contains about 15 hours of noise-free read speech, is used for training. The training database contains about 4.5 million speech frames (excluding silence frames), and the test set about 175000. Both the training database and the test set were segmented using the same HMM system. This HMM uses 1078 cross-word context-dependent tied states which share a pool of 17932 diagonal covariance Gaussians. 25 dimensional feature vectors were obtained by means of a mutual information based discriminant linear transformation (mida) on 24 MEL spectra and their first and second order time derivatives [12].

We used a set of 43 phones, based on the cmu v0.6d phonetic lexicon. All experiments use diagonal covariance matrices $\hat{\mathbf{\Sigma}}_c$, where the class assignment is based on the state segmentation from the reference HMM system. The data sharpening step is performed using the same class assignment. The number of nearest neighbours k used for both the pilot distribution and the data sharpening was fixed at $\lceil \sqrt{n} \rceil$, with n the number of samples in a class.

3.2. Frame classification

A first experiment looks at the nearest database frames for each non-silence frame in the test set (silence frames are classified nearly 100% correctly, hence results would be strongly dependent on the cutoff length of the recordings). Figure 2 shows the percentage of those neighbours that have the correct phone label for an increasing number of neighbours.

The results clearly show that data sharpening is extremely beneficial for classification. Contrary to the behaviour in continuous speech recognition the Euclidean distance performs best for a small number of nearest neighbours. However, as the distances get larger, the local Mahalanobis scaling increasingly outperforms the Euclidean distance. Adaptive kernels have a beneficial effect for the smaller distances, but their performance gets asymptotically closer to non-adaptive kernels as the distances get larger.

Single best frame classification correctly identifies 60% of the frames. In comparison, HMM based frame classification (using the sum of all state likelihoods of a phone and the phone occurence in the training database as a prior) achieves 63% correct classification. Note that if frame classification was the primary goal, better results could be obtained by means of kNN voting or real Parzen density estimation. However, those results would not be relevant for DTW based continuous speech recognition, since the basic unit is the template, whose score is based on the individual kernel scores of all its member frames.

3.3. Phone classification

The phone classification task uses small DTW alignments between each non-silence phone segment of the test set and all phone segments of the database. This task is more relevant to example based recognition, since now the temporal dependencies imposed by the templates become relevant.

Figure 3 shows the percentage of correct neighbours for an increasing number of neighbours. The results are very similar to those of the frame classification experiment, although the advantage of the scaled distances over the Euclidean distance is more pronounced.

3.4. Phone string recognition

In the phone string recognition experiment, the complete test sentence is used as input, and the recognizer finds the most likely template string. The search space is limited by bottom-up template selection, as described in [3]. The recognizer uses context dependency based on template concatenation costs as in [3], as well as a 3-gram phone transition model.

Table 1 shows the phone error rate for these experiments. Again, the difference between the results before and after data sharpening are highly significant. When no data sharpening is performed, the scaled distances perform better than the Euclidean distance measure. However, this advantage vanishes almost completely after data sharpening.

No data sharpening			Data sharpening		
Eucl.	LM	AKLM	Eucl.	LM	AKLM
21.6	18.7	18.1	14.4	14.3	14.0

Table 1. Phone error rates for the different distance measures, without and with data sharpening.

To place the DTW phone recognition results in perspective, we also compared the results with those obtained by classical HMM systems. The HMM system used to segment the training database (see section 3.1) achieved a phone error rate of 16.7% on the same task. A more complex HMM (36-dimensional features and 1818 tied states) improves that score to 14.9%. While these results show the DTW system is certainly competitive, there are some caveats: the DTW system is gender-dependent (both in the acoustic scaling and through extra template concatenation costs [3]), while the HMM system doesn't use explicit gender information. Furthermore, the DTW system uses multiple successive phone templates from the original recordings as often as possible, which to some extent corresponds to longer-span phone transition models.

4. DISCUSSION AND FUTURE RESEARCH

The most remarkable experimental result is the very large improvement caused by data sharpening, which is confirmed in all three tasks. Data sharpening causes large shifts towards the class mean for outliers, while vectors in the main part of the distribution stay near their original location. The resulting distribution is therefore compacted. The degree of compaction can be controlled by varying k in equation 8. $k = 1$ causes no sharpening, while $k = n$ maps all kernels to the class mean. In all experiments reported on in this paper, k was set to $\lceil \sqrt{n} \rceil$, which we deemed to be a reasonable first guess. Future work will focus on establishing an optimal k, either theoretically or experimentally.

Extra research is also necessary to determine the main cause of the strong beneficial effect of the data sharpening. A combination of two factors seems most likely: First, badly labeled data in the tails are removed in the classical "outlier removal" sense. Secondly, there may be an inherent overrepresentation of the tails in the DTW setup.

Another noteworthy result is that most of the effect of the local distance scaling is lost when using data sharpening. Based on the results from the classification experiments, it seems plausible that the locally scaled distance measures only have an advantage over the Euclidean distance in the case of relatively large distances. The fact that data sharpening compacts the distribution –hence making the average distances within the class smaller– is consistent with this view in explaining the diminished effect of local scaling after data sharpening.

An interesting statistic in this context is the fact that in the phone string recognition experiment, the templates in the chosen phone string hypothesis are on average only about the 100^{th} nearest neighbour template for the chunk of input data they explain. This is because of the influence of template concatenation costs and the phone transition model. Adding further constraints to the task (e.g. lexical constraints in CSR) causes the inclusion of even worse-matching templates in the hypotheses, necessitating better "long-range" distance measures. Therefore, in CSR there might be a clearer advantage of the local Mahalanobis distances over the Euclidean, even after data sharpening.

5. CONCLUSIONS

We introduced two extensions to the scaled local distance measure in example based speech recognition based on equivalents in non-parametric density estimation. We tested the influence of the extensions in three low-level acoustic tasks. Adaptive kernels, a widely used technique in density estimation, caused only a small improvement. On the other hand, a data sharpening approach loosely based on the popular mean shift algorithm produced a large improvement. For phone string recognition, the DTW system using the best available local distance measure performed better than a state-of-the-art HMM system.

6. REFERENCES

[1] R. K. Moore, "Towards a unified theory of spoken language processing," in *Proceedings of the 4th IEEE International Conference on Cognitive Informatics*, Irvine, CA, USA, August 2005.

[2] M. De Wachter, K. Demuynck, P. Wambacq, and D. Van Compernolle, "A locally weighted distance measure for example based speech recognition," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 181–184.

[3] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven example based continuous speech recognition," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 1133–1136.

[4] M. Matton, M. De Wachter, D. Van Compernolle, and R. Cools, "A discriminative locally weighted distance measure for speaker independent template based speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, Oct. 2004, vol. I, pp. 429–432.

[5] M. De Wachter, K. Demuynck, and D. Van Compernolle, "Boosting HMM performance with a memory upgrade," in *Proc. ICSLP*, Pittsburgh, U.S.A., September 2006, pp. 1730–1733.

[6] E. L. Bocchieri and G. R. Doddington, "Speaker independent digit recognition with reference frame-specific distance measures," in *Proc. ICASSP*, Tokyo, Apr. 1986, vol. IV, pp. 2699–2702.

[7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

[8] L. Breiman, W. Meisel, and E. Purcell, "Variable kernel estimates of multivariate densities," *Technometrics*, vol. 19, no. 2, pp. 135–144, May 1977.

[9] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. on IT*, vol. 21, pp. 32–40, 1975.

[10] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. on PAMI*, vol. 17, no. 8, pp. 790–799, Aug. 1995.

[11] E. Choi and P. Hall, "Data sharpening as a prelude to density estimation," *Biometrika*, vol. 86, no. 4, pp. 941–947, 1999.

[12] K. Demuynck, J. Duchateau, and D. Van Compernolle, "Optimal feature sub-space selection based on discriminant analysis," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, vol. III, pp. 1311–1314.