Comparing Gaussian and Polynomial Classification in SCHMM-Based Recognition Systems

Alfred Kaltenmeier, Jürgen Franke
Daimler Benz AG, Research Institute, Wilhelm Runge Str. 11, D-89081 Ulm Germany
e-mail: kaltenmeier@dbag.ulm.DaimlerBenz.COM

1 Abstract

Semi-continuous Hidden Markov Models (SCHMM) with gaussian distributions are often used in continuous speech or handwriting recognition systems. Our paper compares gaussian and tree-structured polynomial classifiers which have been successfully used in pattern recognition since many years. In our system the binary classifier tree is generated by clustering HMM states using an entropy measure. For handwriting recognition, gaussians are clearly outperformed by polynomial classification. However, for speech recognition, polynomial classification currently performs slightly worse because some system parameters are not yet optimized.

2 Introduction

Semi-continuous Hidden Markov Models with *gaussian distributions* are often used in *continuous speech* (CSR) or *unrestricted handwriting* recognition systems (UHR) [1],[2]. Our paper describes an almost real-time speaker-independent continuous speech recognizer implemented in the German Verbmobil demonstrator [3]. With different input features the same system is used for UHR. Further, we compare our current system to another based on tree-structured polynomial classifiers (TPC) [4] instead of gaussians. Polynomial classifiers have been successfully used in pattern recognition [5], but not in speech recognition. They offer some advantages for SCHMMs:

- no assumptions on underlying probability distributions are neccessary.
- adaptation is based on linear regression, i.e. algorithmically straightforward and simple.
- regression yields individual coefficient ranking orders of polynomials, i.e. only most significant coefficients must be evaluated for classification.
- polynomial coefficients can be coarsely quantized.
- the tree structure allows an efficient pruning strategy for classification because only promising branches of the tree must be evaluated.
- due to the tree structure, a larger recognition system can be easily shrinked to a smaller target system according to given memory or computational constraints. System retraining is not necessary.

The outline of the paper is as follows: Section 3 gives a short overview of our current gaussian-based recognition system and its main characteristics. Section 4 describes the system adaptation including HMM training, clustering and generation of the tree-structured polynomial classifier. Section 5 compares the recognition performances of both systems also for UHR and for CSR.

3 System Description

The acoustic front end of our baseline system has following characteristics [7]:

- *Mel-based signal analysis* with 12 cepstral coefficients $\boldsymbol{c}$ and one companded energy value e.
- *Vector quantization* (VQ) of primary feature vectors $\boldsymbol{x}$=[$\boldsymbol{c}$,e] with four independent codebooks for $\boldsymbol{c}$, $\Delta\boldsymbol{c}$, $\Delta\Delta\boldsymbol{c}$ and $[e, \Delta e, \Delta\Delta e]$. Each codebook includes 256 gaussians with full covariance matrices. These codebooks are only used for the *first* pass of HMM-training.
- *LDA transform* of 9 succeeding feature vectors $\boldsymbol{x}$ yielding an intermediate vector $\boldsymbol{v}$ with $9 \cdot 13 = 117$ components which is then transformed into a secondary vector $\boldsymbol{z}$ with 32 components.
- *Vector quantization* of $\boldsymbol{z}$ with one codebook and 1024 classes – either with full-covariance gaussians or tree-structured polynomial classifiers. LDA transform and one-codebook VQ are only used in the *second* pass HMM training and the final recognition system.
- The *HMM model pool* includes more models than our previous system [7]: context-independent phones, biphones, triphones, crossword-triphones for functional phrases, and whole word models for digits, the alphabet, and several types of noises and non-speech events.

The handwriting front end [2] transforms binary images into sequences of feature vectors comprising credibilities for several kinds of geometrical information. Except for the feature extraction module, the num-

ber of HMMs and appropriate model topologies our speech and handwriting systems are identical both for training and recognition.

4 System adaptation

4.1 First Pass System Training

HMMs are trained with the forward-backward algorithm. First context-independent HMMs are trained which are then used as initial seeds for context-dependent models. At the end of the first-pass training, a model state s_i of an HMM is characterized by its transition probability vector $\boldsymbol{a}_i$ and its four-parts emission probability vector $\boldsymbol{b}_i$ according to the four VQ codebooks.

After the first training pass all training utterances are segmented by forced *Viterbi-alignment* yielding corresponding pairs of states and feature vectors $[q_t = s_i, \boldsymbol{v}_t$, at time $t]$. With all these pairs about 5000 state/class-dependent means and full covariances are computed.

In order to reduce the number of classes, entropy clustering is performed on emission probabilities [6]. Given the emission probability vector $\boldsymbol{b}_i$, the entropy of state s_i is defined as

$$H_i = -\sum_{k=1}^{1024} b_{ik} \log_2 b_{ik}. \tag{1}$$

With equation (1) the distance between two states s_i and s_j is computed as

$$\Delta_{ij} = (N_i + N_j) H_{ij} - N_i H_i - N_j H_j \tag{2}$$

where N_i and N_j are the counts of s_i and s_j, resp. and H_{ij} is the entropy of the merged state s_{ij} with averaged emission probabilities $\{\boldsymbol{b}_i, \boldsymbol{b}_j\}$. The clustering procedure computes the distances $\{\Delta_{ij}\}$ between all mergeable, i.e. allowed pairs of states, and iteratively merges those states with the smallest distance Δ_{ij}. In contrast to the Kullback divergence clustering described in [7], which assumes one gaussian distribution for each state, entropy clustering does not make such an assumption and performes much better in our system.

In the first clustering pass, the number of classes is reduced to 1024. State-dependent means and covariances are merged according to clustered emission probabilities. The reduced set of classes is used to compute the LDA-transform which is the same for both gaussian and TPC-based SCHMMs. For the gaussian system, the 1024 merged means and full covariances are further transformed by the LDA-matrix yielding the final [1024,32] VQ codebook used for the second training pass and recognition.

4.2 Classifier Adaptation

For the TPC system, entropy clustering of emissions is further performed until two final clusters are reached. The cluster history, i.e the information which emissions are merged in a single cluster step, is kept separately. This history defines the nodes and branches of the classification tree. Although clustering is completely data-driven, it agrees well to phonetic knowledge, e.g. the two final sets include noises, pauses, non speech-events, and closure parts of plosives in the first set and all other speech events in the second one.

The adaptation of polynomial classifiers is completely described in [5, chapt. 6], thus it will be only roughly sketched here. The corresponding pairs of $[q_t = s_i, \boldsymbol{v}_t$, at time $t]$ of state indices (or clustered sets of indices) and feature vectors are the basis of TPC adaptation. According to $q_t = s_i$ we define target vectors $\boldsymbol{y}_t$ with $y_{it} = 1$ and $y_{jt} = 0, j \neq i$. The goal of classifier adaptation is to approximate target vectors $\{\boldsymbol{y}_t\}$ in a minimum mean-squared error sense. With

$$d_t(w) = C^T w(z_t) \tag{3}$$

the optimization criterion is

$$S^2 = E[|y_t - d_t(w)|^2] = \min \tag{4}$$

with respect to the coefficient matrix $\boldsymbol{C}$. Here $\boldsymbol{z}_t$ is the LDA-transformed vector of the intermediate feature vector $\boldsymbol{v}_t$, and $\boldsymbol{w}$ is the quadratic expansion of $\boldsymbol{z}_t$. For practical reasons we are limited to quadratic polynomials, Fig. 1. In our experiments we used complete quadratic polynomials for LDA-transformed feature vectors $\boldsymbol{z}_t$ with 32 dimensions. Thus the expanded polynomial vector $\boldsymbol{w}$ has 560 enhanced components. Thus complete quadratic polynomials have the same number of parameters as full-covariance gaussians.

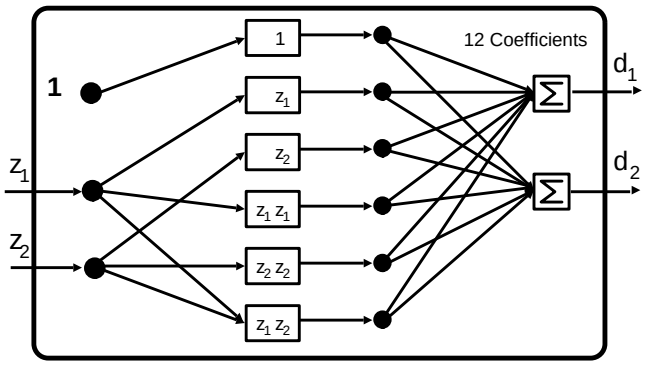

Figure 1: Structure of a two-classes quadratic polynomial classifier with two input features z_1, z_2.

Although quadratic terms are used in the classifier, above equations state a linear optimization problem. It can be effectively solved by well-known matrix equation techniques. In contrast to neural networks used as HMM-labelers [8] which have to be trained iteratively, our solution is algorithmically straightforward and fast.

For each node in our cluster tree we have to compute one classifier which has to separate two classes or two super-classes (clusters of classes). Fig. 2 shows a simple classifier tree for handwritten digits. The leaves of the tree are the ten terminal classes {0,...,9}; the non-terminal nodes {C$_i$} are the nine classifiers.

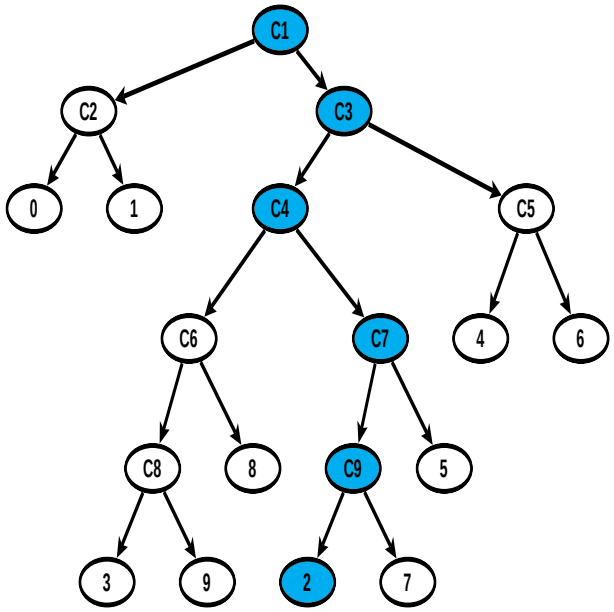

Figure 2: Classifier tree for handwritten digits.

Due to the property [5]:

$$\sum_{k=1}^{K} d_k = 1 \qquad (5)$$

each classifier has to evaluate only one equation because $k = 2$ in our case. Thus the coefficient matrix $\boldsymbol{C}$ is simply a vector $\boldsymbol{c}$ for each classifier.

TPC operates very similar to a decision tree. The main difference is that no hard decisions are forwarded along the tree branches but rather conditional probabilities which are repeatedly multiplied. Thus at tree leaves (and at intermediate nodes) estimations for a-posteriori probabilities are gained [5]. Hence TPC offers the advantage of evaluating only those branches where a-posteriori probabilities are high enough. In each active node the current decision is analyzed, and if one of both estimations is higher than a threshold (e.g. 0.75 in our experiments) only the associated branch will be further alive, otherwise both branches.

Another problem arises because classifier decisions are only estimations of probabilities, hence their values may be larger than 1.0 or less than 0.0. This behavior is typical for polynomial classifiers. Estimations deviating too much from the [0,1] interval indicate that the classifier is forced to classify an object *not* belonging to the classes it was trained with. For such cases another threshold is used in order to deactivate corresponding nodes and branches.

This twofold pruning strategy significantly reduces the number of computations: on an average only 8% of nodes had to be evaluated in our recognition experiments. Fig. 2 shows classification for a sample of the handwritten digit 2: bright nodes/leaves were not evaluated, grey nodes show the activated classifier branch and the correct recognition result.

5 Results and Conclusions

Our system was evaluated for both CSR and UHR. The CSR system was trained with the official German Verbmobil training data (CD-ROMs 1 to 5, 7, and 12). The test set includes the evaluation data of year 95 (EVAL95, 343 sentences, 7204 words) and the data of year 96 (EVAL96, 305 sentences, 5417 words). The test lexicon includes 5394 entries. It should be mentioned that our evaluation system is the same as our demonstrator system which runs close to real-time for a 2500 word vocabulary.

The USR system is used for *offline* single-word postal address reading on letters. It was trained with 13200 US city-names manually labeled according to writing style (small or capital, block or cursive). The test set consists of 1265 US city-names different from the trainig set. UHR is performed in two steps. First, a ZIP-code recognizer reduces the US postal lexicon from more than 60000 entries to about 400 on average, i.e. the recognition lexicon is different for each address. Then SCHMM recognition is performed using the reduced lexicon.

Entropy clustering described by equations (1) and (2) may depend on the state counts $\{N_i\}$. Thus following clustering strategies are investigated:

- CLU1 – unchanged state counts.
- CLU2 – clipped state counts $\{N_i \leq N_{max}\}$.
- CLU3 – combination of top-down decision tree clustering and bottom-up entropy clustering (results will be presented at the conference).

Strategie CLU2 yields best-balanced trees for TPC generation. The tree depth, i.e. the length of the longest path through the tree, was shorter compared to CLU1. Therefore CLU2 reduces computational requirements for TPC classification

Polynomial classifiers have another degree of freedom, namely the weights/counts of classes (means and covariances) used for adaptation. Here we investigated to strategies:

- WEI1 – unchanged class counts $\{N_i\}$.
- WEI2 – averaged class counts $\{N_i = N_{avr}\}$.

Tables 1 and 2 show UHR results for US city-names. It is obvious that the system is well adapted with 13200 handwritten names; the results for the (randomly chosen) test set are even better than the results for the learn set. Further TPC adapted with strategies CLU2

and WEI2 clearly outperforms gaussian classification (96.1% vs. 94.1% for the test set and 94.5% vs. 92.7% for the learn set). This is not only valid for the forced recognition rate (accuracy) but also for the reject and acceptance rates, Tab. 3. The latter rates are important for real applications where unsafely recognized words have to be rejected.

Type	Classes	Cluster	Weight	Accuracy
Gauss.	280	CLU1	—	93.7%
Gauss.	300	—	—	94.1%
TPC	300	CLU1	WEI1	93.1%
TPC	300	CLU2	WEI1	94.5%
TPC	300	CLU2	WEI2	96.1%

Table 1: Word accuracy for UHR of the US city-names *test set* (confidence level = ±1.2%).

Type	Classes	Cluster	Weight	Accuracy
Gauss.	300	—	—	92.7%
TPC	300	CLU2	WEI2	94.5%

Table 2: Word accuracy for UHR of the US city-names *learn set* (confidence level = ±0.4%).

Type	Reject	Error	Accept
Gauss.	10.0%	2.1%	87.9%
	15.9%	1.1%	84.0%
TPC	6.5%	1.1%	92.4%
	14.0%	0.3%	85.7%

Table 3: Reject, error and acceptance rates for UHR of the US city-names *test set* (confidence level = ±1.2%).

Table 4 shows CSR word accuracies for the Verbmobil test sets EVAL95 and EVAL96. TPC cluster and adaptation strategies CLU2/WEI2 perform again better than strategies CLU1/WEI1. The reason is that pauses, breath, coughs and other noises occur very frequently in spontaneous speech. Occurrence-based clustering and weighting (CLU1/WEI1) thus over-represent such classes; hence the recognizer performes better for non-speech events, but worse for speech itself.

For handwriting, TPC performs significantly better than gaussian classification. For speech, however, gaussian classifiers still perform about 1.5% better than our best TPC. Possible reasons could be:

- LDA-transformed credibilities used for handwriting recognition can be better modeled by TPC than LDA-transformed mel-cepstral coefficients used for speech recognition?
- For handwriting, initial clustering reduced the number of classes from 360 to 300. For speech, initial clustering reduced the number from about 5000 to 1024. Perhaps the number of polynomials is too small for speech?
- TPC parameters optimized for pattern recognition may not be optimal for speech recognition; e.g. the branching factor was fixed to 0.75.

Type	Cluster	Weight	EVAL95	EVAL96
Gauss.	—	—	73.1%	78.2%
TPC	CLU1	WEI1	66.7%	71.0%
TPC	CLU2	WEI1	67.7%	73.7%
TPC	CLU2	WEI2	71.6%	76.3%

Table 4: Word accuracy for CSR of speech data EVAL95 and EVAL96, (confidence level = ±1.1%).

Nevertheless, our first results look promising. TPC seems to be an attractive alternative to gaussian or neural network classification in SCHMM recognition systems. Currently we work on following topics:

- Optimization and adaptation of parameters.
- Combined tree-based and entropy-based clustering, i.e. a combination of top-down and bottom-up strategies.
- Clustering of classes based on polynomial separability instead of entropies (for bottom-up clustering).

6 References

[1] *From Sphinx-TT to Whisper*, Advanced Topics in Speech and Speaker Recognition, Kluwer Publisher, 1996

[2] T. Caesar, J. Gloger, A. Kaltenmeier, E. Mandler:, *Sophisticated Topology of Hidden Markov Models for Cursive Script Recognition*, Proc. ICDAR 93, Tsukuba, Japan, Oct. 93.

[3] T. Bub, J. Schwinn: *VERBMOBIL: The Evaluation of a Complex Large Speech-to-Speech Translation System*, Proc. ICSLP 96, Philadelphia, Oct. 96

[4] J. Franke: *Isolated Handprinted Digit Recognition*, Handbook on Optical Character Recognition and Document Image Analysis, ed. by H. Bunke and P.S.P. Wang, World Scientific Publishing Company, 1996

[5] J. Schuermann: Pattern Classification, J. Wiley & Sons, Inc., New York, 1996

[6] M-Y. Hwang, X. Huang, F.A. Alleva: Predicting Unseen Triphones with Senones, IEEE Trans. on Speech and Audio Processing, vol. 4, num. 6, Nov. 96

[7] F. Class, A. Kaltenmeier, P. Regel: *Optimization of an HMM-based Continuous Speech Recognizer,* Proc. EUROSPEECH 93, Berlin, Sept. 93.

[8] P. Le Cerf, W. Ma, D. Van Compernolle: *Multilayer Perceptrons as Labelers for Hidden Markov Models*, IEEE Trans. on Speech and Audio Processing, vol. 2, num. 1, Jan. 94