ANALYSIS OF PHONE POSTERIOR FEATURE SPACE EXPLOITING CLASS-SPECIFIC SPARSITY AND MLP-BASED SIMILARITY MEASURE

Afsaneh Asaei[1,2], Benjamin Picart[3], Hervé Bourlard[1,2]

[1]IDIAP Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]TCTS Lab, Faculté Polytechnique, Université de Mons (UMons), Belgium

ABSTRACT

Class posterior distributions have recently been used quite successfully in Automatic Speech Recognition (ASR), either for frame or phone level classification or as acoustic features, which can be further exploited (usually after some "ad hoc" transformations) in different classifiers (e.g., in Gaussian Mixture based HMMs). In the present paper, we show preliminary results showing that it may be possible to perform speech recognition without explicit subword unit (phone) classification or likelihood estimation, simply answering the question whether two acoustic (posterior) vectors belong to the same subword unit class or not. In this paper, we first exhibit specific properties of the posterior acoustic space before showing how those properties can be exploited to reach very high performance in deciding (based on an appropriate, trained, distance metric, and hypothesis testing approaches) whether two posterior vectors belong to the same class or not. Performance as high as 90% correct decision rates are reported on the TIMIT database, before reporting kNN phone classification rates.

Index Terms— Posterior feature space, posterior-based metrics, posterior space properties, kNN classifier.

1. INTRODUCTION

Posterior probabilities are currently often used as additional acoustic features to improve Automatic Speech Recognition (ASR) systems [1,2]. These features are usually extracted by a Multilayer Perceptron (MLP) using spectral-based features such as MFCC or PLP as input. In this approach, each output unit of the MLP is associated with a particular phone (or subword unit) and is trained to generate a posteriori probability distributions over the output classes conditioned on the input acoustic observation sequence [3]. While allowing for discriminant training, such an approach also has the advantage of accommodating acoustic context by providing several frames at the MLP input. The MLP in the case of posterior features performs a nonlinear discriminant transformation, which projects the input feature space onto a nonlinear sub-space of maximum possible sound class discriminatory information. This projection is expected (and has been shown) to be able to suppress (to some extent) the non-relevant variability (including noise), while preserving the speech discriminatory information, thus resulting in a set of features highly rich in contextual and phonetic information which could be considered as "optimal phone detectors".

These appealing properties make posterior probabilities powerful features for ASR systems, and are now part of many state-of-the-art systems. However, these features also exhibit very specific statistical properties (like much squeezed, non-Gaussian, distributions) and metric spaces, which do not make them very suitable to Gaussian Mixture Model (GMM)-based HMM. In this case, the usual "solution" is to "gaussianize and decorrelate" the posterior features by applying a log function, followed by a PCA transform [2]. However, recent studies have used posterior features directly in Kullback-Leibler (KL)-based HMM models where the reference KL-HMMs are parameterized by multinomial distributions [1].

The goal of the present work is to better understand the properties of the posterior acoustic space (assumed to be close enough to a binary space) and see how new families of ASR systems could emerge from those properties. Following our intuition, and encouraged by previous studies on the properties of binary spaces [5] showing that, the probability that two vectors are orthogonal is equal to 1 for a large enough dimensional binary space, the first goal of this paper is thus to show evidence that inter-class posterior vectors are indeed often orthogonal. Based on this fact, we then studied several approaches to see how accurately we could predict whether two posterior vectors belong or not to the same "class" (phone, phone in context, etc). This problem was approached by using standard hypothesis testing using different metrics well suited to capture the orthogonality/sparsity properties of the space. In this paper, this problem will be referred to as "pairwise classification" (likelihood that two vectors belong to the same class), as opposed to "full classification" (likelihood that a specific vector belong to one of the K possible classes). While our conclusions open up new perspectives towards novel ASR approaches, we also provide here "full classification" rates by simply incorporating our distance metrics into kNN classifiers.

2. POSTERIOR-BASED DISTANCE MEASURES

2.1 Angle-based Similarity Measure

To gain insight into the orthogonality and sparsity properties of (high dimensional) posterior feature spaces, we started by measuring the relative angle between multiple pairs of posterior feature vectors, belonging or not to the same phonetic class. This angular distance between two *K*-dimensional posterior vectors x_i and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jK})^T$ was defined as:

$$\theta(x_i, x_j) = \cos^{-1} \frac{\sum_{k=1}^{K} x_{ik} x_{jk}}{\sqrt{\left(\sum_{k=1}^{K} x_{ik}^2\right)\left(\sum_{k=1}^{K} x_{jk}^2\right)}} \qquad (1)$$

where $x_{ik} = p(\omega_k|y_i)$, the a posteriori probability that the correct class associated with y_i is ω_k, from the set of all possible subword unit (phonetic) classes, $k = 1, \ldots, K$. These posterior probabilities are estimated on a wide training database by a large MLP where we have regular (e.g., MFCC, PLP) *d*-dimensional features y_i at the input and *K* output classes.

As exploited in Section 3, angles between a large number of training vector pairs belonging or not to the same class (known during training) were calculated according to (1). The distribution of those angles was then approximated by an histogram, similar to the one described in Section 3. This histogram clearly displayed two separate distributions with very small overlap, one for the "same class angles" and one for the "different class angles". Interestingly, the intersection of those two distributions was around 80-85° (hence close to 90°), thus confirming that inter-class posteriors are often very close to orthogonality. In other words, posterior features associated to a class are nearly orthogonal to any given posterior features belonging to the other classes. This conclusion can also be interpreted as "class-specific sparsity".

To exploit this property we investigated a geometric extension of the kNN classifier. First, the distance of the test sample to the training data is computed as the cosine function of the feature vectors. Then, the nearest neighbors are defined as the samples which lie in a hypercone defined by a specified angle originated from the test data. This angle could be interpreted as a look-angle into the space. Subsequently, the classification is performed based on voting to find the dominating class of training data which are geometrically in the nearby samples. By specifying the look-angle, we fix the maximum value for the relative angle up to which the labels will be counted in kNN majority voting. This procedure eliminates the need to find the optimal k values by cross-validation. Instead, the look-angle is specified based on how restrictively the orthogonality assumption is exploited. We call this extension of the NN classifier the Geometric NN (GNN) classifier.

For some of the test vectors, and for a given look angle, there are not enough neighbors to take a reliable decision. By specifying a minimum value of k under which the decision is considered unreliable, the GNN also provides an interesting tool to distinguish vectors whose classification is not as reliable as the other ones. These vectors are labeled as "unreliable". This capability opens new research opportunities to further improve the classification accuracy (by means of post-processing of these uncertainties in the feature space). This latter point being outside our present scope, we do not exploit further this particular capability of the GNN in this paper.

2.2 MLP-based Similarity Measure

As a possible "universal" distance estimator, we also considered using an MLP as a similarity measure and to estimate the probability that two posterior vectors are part of the same (phonetic) class or not [4].

As illustrated in Figure 1, the parameters of this "pair similarity MLP" (referred to as MLP-s) are trained over a very large set of feature vector pairs. For each input pair, the target output of the MLP-s is fixed to 1 when the two vectors belong to the same class and to 0 otherwise.

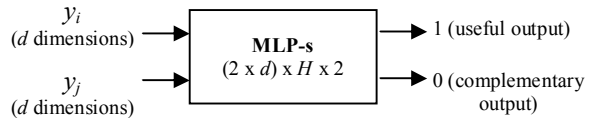

Figure 1: The MLP-s is composed of 2xd input units, H hidden units (optimized on a cross-validation database) and 2 output units (although one could be enough, we observed that training was faster with two output units)

However, after having trained and tested (on an independent test set) different MLP topologies (but always yielding very good performances), we were prompted to prove that the MLP-s optimal output, trained to minimize the average Mean Squared Error (MSE), was nothing else but the scalar product of the two posterior vectors associated with the input vector pairs $\{(y_i,y_j)\}$, whatever they were (PLP, MFCC, posterior features, etc). It is indeed extensively discussed and theoretically proved in [6] that minimization (in the MLP parameters space, hence the MLP weights) of

$$E = \sum_{i,j} \left(g(y_i, y_j) - t(y_i, y_j)\right)^2 \cdot p(y_i, y_j) \qquad (2)$$

over all the training data pairs $\{(y_i,y_j)\}$ will, at best, result in the optimal MLP output:

$$g^{opt}(y_i, y_j) = \sum_{k=1}^{K} p(\omega_k | y_i) \cdot p(\omega_k | y_j) \qquad (3)$$

where $g(y_i,y_j)$ is the observed output (given input feature vectors y_i and y_j), $p(\omega_k|y_i)$ are the "theoretical" posterior probabilities, and $t(y_i,y_j)$ is the 1/0 target output, depending whether the two vectors belong to the same class or not. On

TIMIT, as used here, but also in general, this information is known for the training data (e.g., from a preliminary training and resulting segmentation). Taking into account the definition of posterior feature vectors given in Section 2.1, with the caveat that our posterior features x_i are *estimates* of the true posteriors $p(\omega_k|y_i)$, (3) simply becomes:

$$g^{opt}(y_i, y_j) = x_i \bullet x_j \qquad (4)$$

Given (4) it is now obvious that it is not even necessary to (expensively) train (or run during testing) MLP-s and that the optimal solution (according to MSE cost function) is simply obtained by the Posterior Scalar Product (PSP) of the best posterior estimates that can be obtained from the two input vectors (whatever they are). The comparisons and uses of this will be performed in our hypothesis tests (explained in Section 3) to decide whether they belong to the same class or not. As in Section 2.1, those "best" posterior estimates are assumed to be given by an MLP with y_i at the input (usually within 9 frames of acoustic context), and yielding the posterior vector estimate x_i at its output.

Of course, both the angle distance defined in Section 2.1 and this PSP are measures of "orthogonality" of the inter-class posterior (close to binary) features. These two intra/inter-class distance measures will now be used in a standard hypothesis testing approach to decide whether two posterior vectors belong to the same class or not. In the experimental section, these two distance measures will be compared to more classical ones like Euclidian, Kullback-Leibler, and Bhattacharyya distances.

3. HYPOTHESIS TESTING

Now that we have defined proper distance measures, we can address the following question: "Given two feature vectors, what is the probability that these belong to the same (phonetic) class or not, whatever the class ω_k (k=1,…,K) is?". This is a standard hypothesis testing problem where, given two vectors x_i and x_j and a distance measure $d_{ij}=d(x_i,x_j)$ between them, we have to decide whether they belong to the same class or not. We thus aim at classifying vector pairs as belonging to class S ("same class") or class D ("different class").

As illustrated in Figure 2, we first estimate (on the training or cross-validation data) the continuous distributions $p(d_{ij}|S)$ and $p(d_{ij}|D)$ by two histograms [7,8], and then determine the "optimal" decision threshold τ (minimizing Bayes' risk) by their intersection. In this case, $\tau = 0.06$.

In our work, the decision threshold τ was picked by hand, but it will be estimated automatically in the future. In the test phase, classification is then simply performed by comparing d_{ij} with that decision threshold τ. Final accuracy is assessed using the test pairs made from a set of test vectors independently drawn from the same population. Of course, as shown below, the choice of the correct distance

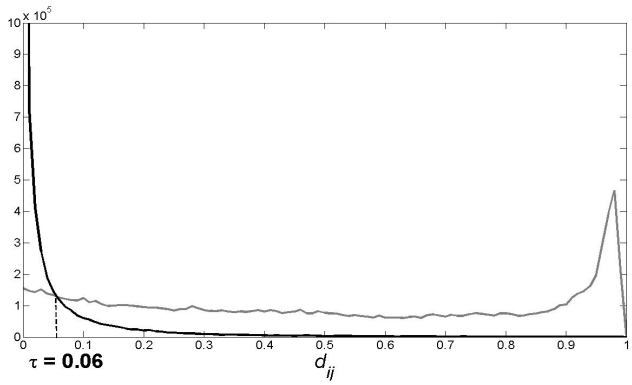

Figure 2: Histogram of the "Scalar Product"-based similarity d_{ij} between same-class (grey) and different-class (black) posterior feature vectors. Decision threshold τ at 0.06.

metric is critical, and depends on the topological properties of the feature space. Therefore, in addition to our angle-based and PSP distances, experiments have also been conducted with other distance/divergence measures, such as Euclidian, Kullback-Leiber (KL), and Bhattacharyya (BT).

4. EXPERIMENTS

4.1. Acoustic Front-End

All experiments were performed on the TIMIT database. A three layered MLP was trained on TIMIT and used to estimate the phone posterior probabilities $x_{ik} = p(\omega_k|y_i)$, where

- k=1, …, K (K=40 in this case)
- y_i are standard d-dimensional features (13 PLP coefficients, together with their delta and delta-delta parameters, resulting in 39-dimensional acoustic features).

The data consisted of 3,000 training utterances from 375 speakers, 696 cross-validation utterances from 87 speakers, and 1,344 test utterances from 168 speakers. This MLP had 351 input nodes corresponding to the concatenation of nine frames of 39 dimensional acoustic vectors, one hidden layer with 2,000 units, and 40 output units, each of them being associated with one of the K phone classes.

4.2. Histogram-based Hypothesis Tests

The classification accuracies obtained over training and test pair sets of posterior feature vectors, for the different metrics are summarized in Table 1. The tuning of the optimal decision point was performed using 20,000,000 training pairs, while the test accuracy was computed using 4,000,000 test pairs. In [6], we discuss in details how the training data set was built to compensate for the obvious strong bias towards the D class (which has a much higher prior probability) without modifying the phonetic class priors $p(\omega_k)$.

Table 1: Pairwise classification accuracies over 4 millions test posterior feature pairs, using several distances: Euclidian (Eucl), Kullback-Leibler (KL), Bhattacharyya (BT), Angle, Posterior Scalar Product (PSP) and MLP-s

Dist / Pairs	Eucl (%)	KL (%)	BT (%)	Angle (%)	PSP (%)	MLP-s (%)
Trai-ning	84.4	88.4	89.3	89.7	90.2	89.9
Test	78.8	85.4	86.6	87.5	88.5	85.4

From Table 1, it is clear that the PSP metric performs better than all the others, while the angle-based distance is the second best (which was expected since the latter is simply the PSP divided by the product of the norms of the two vectors belonging to the pair). Thus, starting from the "simple" Euclidian distance, improved performance is achieved when using metrics exploiting the properties of the a posteriori probability space, like KL, BT, Angle and, finally, PSP (as also supported by theory)[1].

4.3. kNN Classification

The k-Nearest Neighbor (k-NN) classifier is a simple but effective classification method which associates an unknown sample to the class the most frequently represented within its *k* nearest neighbors. Since kNN is non-parametric, there is no need to assume any knowledge about the underlying statistical distribution. Assuming that enough training data is available and using the "proper metric", kNN will minimize Bayes risk, while providing good estimates of a posteriori distribution [8].

To further validate the above key results (which are at the center of the present paper), we also performed kNN-based classification. This was done over 410,920 test vectors, and tuning of the "optimum" k value on a cross-validation set of 204,657 vectors. The classification accuracies obtained over the test posterior feature vectors[2], for the different metrics, are given in Table 2. When using GNN, the minimum number of neighbors to make a decision was chosen equal to 40. GNN accuracy is computed on reliable samples, "unreliable" ones being left undecided. As it could be seen, a small fraction of unreliable samples significantly contributes to degrade the overall classification performance, resulting in approximately the same "full classification" performance as the classical approaches.

Actually, the main goal of this section was to show that whatever metric is used, and in spite of the very high "pairwise classification" accuracies that can be reached, it is still very difficult to beat the best "full classification" accuracies obtained with modern methods.

[1] We also have applied the Euclidian, Angle and Scalar Product metrics to the PLP features space. The performance was much worse (around 10% absolute decrease in accuracy).

[2] Again, the classification of PLP features (instead of posterior features) provided much worse results (50% accuracy only for the Euclidian distance).

Table 2: kNN classification accuracies for the posterior feature vectors

Dist / Test vector set	Eucl (%)	KL (%)	BT (%)	PSP (%)	Look-Angle (°)	GNN (%)	Un-reliable Samples (%)
Test vectors	68.3	68.5	68.2	68.3	0.5	82.3	22
					1	78.5	15
Optimum k value	260	200	20	5	1.5	76.2	11
					17	68.6	0

5. CONCLUSIONS

In this paper, we provided evidence that posterior features have interesting orthogonality and sparsity properties, which could be exploited to reach high "pairwise classification" accuracies. In this context, the Posterior Scalar Product (PSP), which has been shown here to be the optimal solution that can be achieved when training an MLP on such task, yields the best performance, just above the angle-based distance. Using these distances in kNN to perform "full classification", we see that we reach again the same performance as reported in the literature. As a conclusion, this work led us to start developing new ASR approaches exploiting the high performances of posterior vector/sequence "pairwise classification" and our distance metrics (instead of the explicit classification used in the state-of-the-art ASR systems).

6. ACKNOWLEDGEMENTS

This work was supported by a European ERASMUS exchange program, complemented by the EU-FP6 AMIDA training program, and by the EU Marie-Curie SCALE project.

7. REFERENCES

[1] G. Aradilla, J. Vepa, H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features", ICASSP 2007.

[2] H. Hermansky, D. Ellis, S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems", Proceedings of the ICASSP 2000.

[3] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Press, 1994.

[4] M. Gerber, T. Kaufmann, B. Pfister, "Perceptron-based Class Verification", NOLISP, Paris, 2007.

[5] P. Kanerva, *Sparse Distributed Memory*, The MIT Press, 1988.

[6] B. Picart, "Improved Posterior Estimation Through kNN and MLP-based Similarity", Idiap Research Report RR-18-2009, http://publications.idiap.ch/index.php/publications/show/1682

[7] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[8] P.A. Devijver, and J. Kittler, *Pattern Recognition: A Statistical Pattern Approach*, Prentice Hall Intl., 1982.