SUBSPACE MODELS FOR SPEECH TRANSITIONS USING PRINCIPAL CURVES

K Reinhard (1), M Niranjan (1)

(1) University of Cambridge, Engineering Department, Cambridge CB2 1PZ, UK

1. INTRODUCTION

The temporal evolution of the short time spectrum is an important characteristic of speech signals. This time variation is caused by the movement of the vocal tract and is a rich source of information not only of the phonetic content of what is spoken, but also other information, such as the speaker. State of the art statistical models make crude approximations to the temporal variation, essentially by a piecewise constant approximation that is inherent in the hidden Markov model. Small extensions to this approximation, such as the inclusion of *delta* and *delta-delta* parameters has become common practice, but one is immediately faced with the problem of reliability in parameter estimation caused by the expansion in dimensionality. The use of Recurrent Neural Networks is seen as one plausible mechanism to capture such transitional information. An alternative approach is the use of segmental models that model the time evolution of feature vectors within a segment. Typically, these approaches use the phone as the unit of segmentation [8]. We start from a slightly different premise that attempts to focus on the transition between phones. The spectral trajectory into the vowel [i:], for example is quite different in the CV transition /bee/ than in that of /gee/. Clearly, a phone model for the vowel [i:] derived from all contexts would be noisy. Hence we focus on diphone units, defining the diphone as half of one phone followed by half of the next phone. While the number of segments to model increases rapidly, the hope is that one has a greater chance of capturing the transitional information explicitly. An adaptation of a modified Principal Component Analysis (PCA) approach is used to compute projections onto a two dimensional subspace of the diphone transitions. The two dimensional projections were a starting point for this work, enabling the visualisation of the trajectory in the projected space, but the method itself is not restricted to two dimensions. The work described in this paper shows that much of the discriminatory information is retained in the projection we propose. This is illustrated on a simple problem involving the discrimination of /b/, /d/ & /g/, on the ISOLET [3] and TIMIT [5] database. Receiver Operating Characteristic (ROC) curve is used to present the compromise between detection of a transition and false alarms.

2. SUBSPACE MODEL

Finding a low-dimensional projection of the diphone data describing transitions in the spectral domain is needed. The number of parameters required to perform a projection onto 2D is $2 * (n + p)$, where p is the dimensionality of the parametric spectral representation and n is the number of spectral frames in the diphone. Information of the temporal ordering of the data frames is captured by introducing a constrained PCA, described below.

2.1 Time-constraint PCA (TC-PCA)

A very popular unsupervised technique for dimensionality reduction is the well known principal component analysis or *Karhunen-Loéve-Transform* (for e.g. see [1, 4]), the principal directions being given by the eigenvectors of the data covariance matrix. Given a data set $\mathcal{T}$ which consists of D sequences of N p-dimensional points $\mathcal{T} = \mathbf{T_1}, \cdots, \mathbf{T_D}$ with $\mathbf{T_k} = \mathbf{t_{k1}}, \cdots, \mathbf{t_{kN}}$, it is the temporal evolution of these vectors that is of interest.

In order to preserve the temporal sequence information, the expansion of the dimensionality of the data by one is performed, using $\mathbf{t_{k*}} = \tau * (\mathbf{1}, \cdots, \mathbf{N})$. Hence $\mathbf{T}_* = \mathbf{t_{k*}}, \mathbf{t_{k1}}, \cdots, \mathbf{t_{kN}}$, the extra dimension representing a scalable frame ordering as time constraint. The scale factor τ is introduced to control the weighting imposed by this extra time dimensionality. Tuning this parameter is achieved by an exhaustive search to determine the most discriminant subspace among all models during performance tests. The subspace definition using TC-PCA can be described by solving the covariance matrix of the set of temporal extended vectors and is given by:

$$\Sigma^\tau = \sum_{\mathbf{k=1}}^{\mathbf{D}} \left[\sum_{\mathbf{i=1}}^{\mathbf{N+1}} (\mathbf{t_{ki}} - \bar{\mathbf{t}})(\mathbf{t_{ki}} - \bar{\mathbf{t}})^{\mathbf{T}} \right] \tag{1}$$

the solution of the minimisation problem with respect to the choice of basis vectors $\mathbf{u_i^\tau}$ leads to the equation $\Sigma^\tau \mathbf{u_i^\tau} = \lambda_i^\tau \mathbf{u_i^\tau}$ which is satisfied by $\mathbf{u_i^\tau}$ being the eigenvectors of the covariants matrix. Optimal dimensionality reduction can be performed in terms of projecting the data onto the eigenvectors corresponding to the largest eigenvalues λ_i^τ. Defining our 2-dimensional subspace which is dependent on the time constraint τ introduced above the transformation matrix is characterised by the following equation assuming that $\lambda_1^\tau \geq \lambda_2^\tau \geq \cdots \geq \lambda_{N+1}^\tau$.

$$\mathcal{P}_\tau = \begin{bmatrix} \mathbf{u}_1^{1^\tau} & \mathbf{u}_2^{1^\tau} \\ \vdots & \vdots \\ \mathbf{u}_1^{p^\tau} & \mathbf{u}_2^{p^\tau} \end{bmatrix} \tag{2}$$

3. TRAJECTORY MODEL

The aim is to extract quantitative non-linear dynamics from the training data to form a trajectory model. In the time constrained planes described above strong indicator for typical trajectories for the different CV syllables were found. The trajectories can be represented as a sequence of points in low dimensional space which are found using principal curves.

The algorithm for principal curves by Hastie/Stuetzle [6] describes a method of extracting a smooth one-dimensional curve that passes through the "middle" of a p-dimensional data set providing a non-linear summary of the data. The algorithm for constructing principal curves starts with an initial guess, originally the first principal line, to define the initial ordering of the data and hence the neighbourhood relation. This line is then smoothly bent according to the actual data representation, by locally averaging p-dimensional points and iteratively minimising the orthogonal distances to the new curve. Below, a brief description of the principal curves idea is given. For details, the reader is referred to Hastie/Stuetzle [6].

Given $\mathbf{T} = \{\mathbf{t_1} \cdots \mathbf{t_n}\}$ random vectors in $\mathcal{R}^{\mathbf{p}}$. Then principal curves is defined by $\mathbf{f}$, where $\mathbf{f}$ denote a smooth curve in $\mathcal{R}^{\mathbf{p}}$ parameterised over $\Phi \subseteq \mathcal{R}^{\mathbf{1}}$, a closed interval, that does not intersect itself ($\phi_1 \neq \phi_2 \Longrightarrow \mathbf{f}(\phi_1) \neq \mathbf{f}(\phi_2)$). The projection index $\phi_{\mathbf{f}} : \mathcal{R}^{\mathbf{p}} \longrightarrow \mathcal{R}^{\mathbf{1}}$ is defined as:

$$\phi_{\mathbf{f}}(\mathbf{t_i}) = \sup_{\phi} \{\phi : \|\mathbf{t_i} - \mathbf{f}(\phi)\| = \inf_{\mu} \|\mathbf{t_i} - \mathbf{f}(\mu)\|\}. \tag{3}$$

Equation (6) is read as follows: A search for each data point $\mathbf{t_i}$ is performed over all points on the principal curve $\mathbf{f}$ parameterised over μ. For the closest point on the curve its parameter μ is assigned to ϕ as the best fit. The projection index $\phi_{\mathbf{f}}(\mathbf{t_i})$ is the value of ϕ for which $\mathbf{f}(\phi)$ is closest to $\mathbf{t_i}$. If there are several values, the largest is picked. The definition can be interpreted as starting with an initial guess to provide an ordering and collecting for any particular parameter value ϕ all observations that have $\mathbf{f}(\phi)$ as their closest point on the curve. If $\mathbf{f}(\phi)$ is the average of those observations, and if this holds for all ϕ, then $\mathbf{f}$

is called a principal curve. Because there is in general only one observation $\mathbf{t}_i$ related to a certain ϕ_i the observations projecting into a neighbourhood are locally averaged.

The principal curve algorithm iterates through *Projection-Expectation* steps until the relative change in the distance from the data points to its projections is below a certain threshold (see Figure 1(b)). The initial guess is very important because it should already represent temporal neighbourhood. Therefore principal curve starts as a line segment which is generated by connecting each frame average by a straight line. The initial projection step of the data is hence calculated by projecting the data onto the line segments which defines a neighbourhood relationship using the parameterisation ϕ of the projected data points along the one-dimensional curve $\mathbf{f}(\phi)$. This carefully chosen initial line delivers a good first guess for the temporal ordering of the data (see Figure 1(a)).

$$\frac{|D^2(\mathbf{T},\mathbf{f}^{i-1}) - D^2(\mathbf{T},\mathbf{f}^i)|}{D^2(\mathbf{T},\mathbf{f}^{i-1})} < \text{threshold.} \tag{4}$$

with

$$D^2(\mathbf{T},\mathbf{f}^i) = \sum_{i=1}^{n} ||\mathbf{t}^i - \mathbf{f}^i(\phi_i)||^2. \tag{5}$$

Expectation

For the expectation step, a locally weighted running line smoother is employed to estimate a new $\mathbf{f}(\phi)$ which will be used to calculate the new projected points and hence the new distance from the data points to its projections. In principle one considers a number of data points which project onto the neighbourhood of each projected point which is defined by a kernel function over ϕ to perform a linear regression. The linear regression delivers a new projected point which results in a new principal curve. A new projecting step is then performed to define a new temporal ordering. The algorithm for robust locally weighted regression was used suggested by Cleveland [2]. The parameters using linear regression were calculated, minimising the following expression, such that $\hat{\beta}_j(\phi_i)$ are the values of β_j:

$$\sum_{k=1}^{n} w_k(\phi_i)(\mathbf{f}(\phi_k) - \beta_0 - \beta_1\phi_k)^2 \quad \text{with} \quad w_k(\phi_i) = W(h_i^{-1}(\phi_k - \phi_i)). \tag{6}$$

and

$$\begin{aligned} W(x) &= (1-|x|^3)^3 \quad \text{for} \quad |x| < 1 \\ &= 0 \qquad\qquad\;\; \text{for} \quad |x| \geq 1. \end{aligned} \tag{7}$$

For each ϕ_i, weights, $w_k(\phi_i)$, are defined for all ϕ_k, $k = 1, \ldots, n$ using the weight function W. This is done by centering W at ϕ_i and scaling it so that the point at which W first becomes zero is at the r^{th} nearest neighbour of ϕ_i. For each i let h_i be the distance from ϕ_i to the r^{th} nearest neighbour of ϕ_i. That is, h_i is the r^{th} smallest number among $|\phi_i - \phi_j|$, for $j = 1, \ldots, n$. This procedure for computing the initial fitted values is referred to as locally weighted linear regression, where $\hat{\mathbf{f}}(\phi_i) = \hat{\beta}_0(\phi_i) + \hat{\beta}_1(\phi_i)\phi_i$.

Projection

The aim of the projection step is to reorder the data according to its distance relationship to the new generated curve. Starting from the curve represented by n points $\mathbf{f}^{(\mathbf{j})}(\cdot)$ for each $\mathbf{t_i}$ in the sample the value $\phi_i = \phi_{\mathbf{f}^{(\mathbf{j})}}(\mathbf{t_i})$ which represents the temporal ordering is computed. First the closest point on the line segment joining each pair $(\mathbf{f}(\phi_k), \mathbf{f}(\phi_{k+1}))$ is found. The closest point to the curve is then either the projection onto a line segment or one of the $\mathbf{f}(\phi_k)$. Using these values to represent the curve, ϕ_i is replaced by the arc length from $\mathbf{f_1^{(j)}}$ to $\mathbf{f_i^{(j)}}$. Potential problems with this projection step can be found where the expected values of the observations project onto $\mathbf{f}(\phi_{min})$ or $\mathbf{f}(\phi_{max})$. To circumvent this problem the corresponding data points generate a new $\mathbf{f}(\phi_{min})$ or $\mathbf{f}(\phi_{max})$ by projecting the data point onto the first or last line segment extending the original principal curve.

This algorithm doesn't use frame specific information but adjust the curve according to the density distribution of the training data. Hence it is very important that the initial guess supports the time correlation of the data points. To find a suitable trajectory model $\mathcal{M}_{PC}$, one has to find the closest point on the principal curve for each average frame point. This data driven approach adjusts the frame specific average points to the underlying data distribution.

$$\begin{aligned} \mathcal{M}_{PC} &= \{\mathbf{f}(\arg\min_{\mu} \|\hat{\mathbf{t}}_1^{av} - \mathbf{f}(\mu)\|) \cdots \mathbf{f}(\arg\min_{\mu} \|\hat{\mathbf{t}}_N^{av} - \mathbf{f}(\mu)\|)\} \\ &= \{\hat{\mathbf{t}}_1^{pc}, \cdots, \hat{\mathbf{t}}_N^{pc}\}. \end{aligned} \tag{8}$$

4. TRAJECTORY MAPPING

To use the idea discussed above in a classification setting we impose a smoothing of the test data by fitting a constrained natural spline through it before projecting onto the subspace. This trajectory comparison method is motivated from the observation that the speech signal tends to follow certain paths corresponding to the underlying phonemic units. This was utilised by Sun [13] who modelled the target positions for phonetic units. He found that it is less important to model accurately the path of the intermediate positions that are results of the coarticulation process.

Distance Measure for Classification

In the subspace representation a similarity measurement has to be found which takes time evolution as well as geometrical position of the sequence of observations into account. Therefore a distance measure which compares individual frames geometrically, normalising its overall distance by its arc length [7] was calculated.

The initial operation for the subspace method is a projection of a vector. This is performed by $\tilde{\mathbf{t}}_P = (\tilde{\mathbf{t}})^T \mathbf{P}_\tau$ where the projection matrix $\mathbf{P}_\tau$ is the found subspace matrix defined by the training data and a certain time constraint factor τ. Using the trajectory model $\mathcal{M}_{PC}$ the computation of a normed squared orthogonal distance $d_{sub}(\hat{\mathbf{t}}, \tilde{\mathbf{t}})$ from our trajectory model $\hat{\mathbf{t}}_P = (\hat{t})^T \mathbf{P}_\tau$ can be performed:

$$d_{sub}(\hat{\mathbf{t}}_P, \tilde{\mathbf{t}}_P) = \sum_{i=1}^{N} d_{sub}(\hat{\mathbf{t}}_P^i, \tilde{\mathbf{t}}_P^i) = \frac{\sum_{i=1}^{N} \|\hat{\mathbf{t}}_P^i - \tilde{\mathbf{t}}_P^i\|^2}{arclength(\tilde{\mathbf{t}}_P)}. \tag{9}$$

Performing the distance measure for all models, classification result can be obtained by finding the diphone template with the minimum distance.

5. EXPERIMENTAL ILLUSTRATIONS

A subset of the ISOLET [3], an isolated speech, alphabet database to illustrate the idea was used, extracting the isolated spoken characters /B/, /D/ and /G/ to obtain the diphones /bee/, /dee/ and /gee/. The available data was split into a training and test set. 80% of the data was used for training (ISOLET1-4), 20% for tests (ISOLET5), as recommended by the originators of the dataset. Further experiments were conducted using TIMIT database [5] to show the influence of continuous speech. TIMIT is a convenient database to demonstrate the approach of using diphones as speech segments because of TIMITs phonetic labelling which makes it possible to extract all occurring diphones. With the best recognition accuracy of

Database	**Accuracy**			
	BEE	DEE	GEE	Average
ISOLET	73.3%	80.0%	90.0%	81.1%
TIMIT	72.5%	72.4%	85.6%	76.8%

Table 1: Obtained results for our trajectory model from diphones /bee/,/dee/ and /gee/.

81.1% the result produced with the subspace model using ISOLET is similar to the results for a baseline HMM using one mixture and a diagonal covariance matrix on a BTL E-SET giving 81.2 % accuracy. The results obtained in diphone accuracy for TIMIT with the best recognition accuracy of 76.8% were far better from what one has expected because of the influences of realistic continuous speech to the acoustic transitions in comparison to isolated spoken letters. However, what is important is to note that the representation here is very simplistic, namely, a projection onto two dimensions. In comparison to 720 parameters per model of the baseline HMM system [14], the subspace trajectory approach within ISOLET uses only $2 * (6 + 9 + 9) = 48$ parameters (see [10] for more detailed information).

6. DISCUSSION

In this study, a new method of modeling speech transitions with a subspace model was proposed. It was shown that temporal transitions in speech can be visualised and modelled in a low dimensional space using the principal curve technique. This approach has the advantages that the memory requirements for our subspace model is much less demanding in comparison with models involving context-dependent speech units, which furthermore leads to a model which is easier trainable with the presents of a limited amount of training data which is true for diphones in speech.

The results are encouraging to further investigate the use of subspace models for speech transitions [10], which could be used as compensational models in respect to the inter-segment independent assumption used in state-of-the-art recognition systems. Future work will concentrate on the usefulness of the obtained information whether modelling transitions provides one with orthogonal information in comparison with the information obtained by standard HMM systems. Extension of the experimental work is on the way, introducing multiple trajectory concepts per diphone while investigating more suitable speech representations, trajectory mapping methods and subspace projections. Alternatively a N-best rescoring scheme is proposed [11, 12, 9] to incorporate the subspace mode into a phone-based system. The rescoring mechanism can be used to emphasise paths in the lattice of hypothesis using transitional models which might avoid pruning out the correct sequence of phones. The modelled inter-phone characteristics, which are captured by diphones, should complement baseline systems and should lead to better performances.

7. REFERENCES

[1] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[2] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, December 1979.

[3] R. Cole, Y. Muthusamy, and M. Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, Oregon Graduate Institute, 1994.

[4] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* New York, Wiley, 1973.

[5] J.S. Garofolo. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST), 1988.

[6] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[7] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies Press, Letchworth, U.K., 1983.

[8] M. Ostendorf, V.V. Digalakis, and O.A. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transaction on Speech and Audio Processing*, 4(5):360–378, 1996.

[9] M. Rayner, D. Carter, V.V. Digalakis, and P. Price. Combining knowledge sources to reorder N-best speech hypothesis lists. *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, 1994.

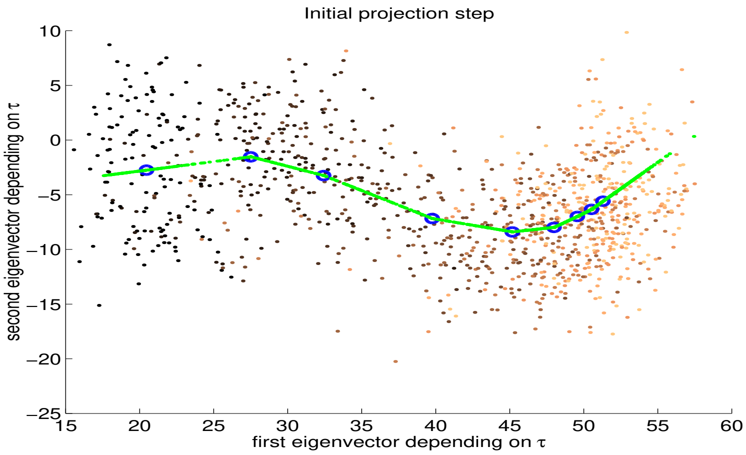
[10] K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. Technical Report CUED/F-INFENG/TR.308, Cambridge University Engineering Department, February 1998.

[11] P. Schmid. *Explicit N-best formant features for segment based speech recognition.* PhD thesis, Oregon Graduate Institute, 1996.

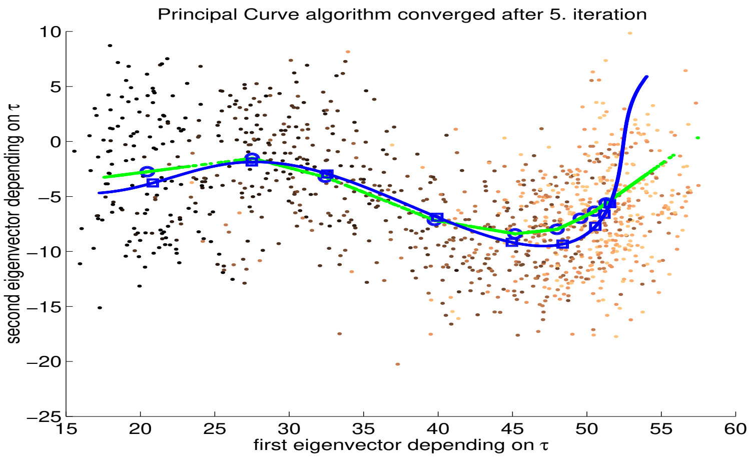
[12] P. Schmid and E. Barnard. Explicit N-best formant features for vowel classification. *Int. Conf. in Acoustics, Speech and Signal Processing*, 2:991–994, 1997.

[13] D.X. Sun. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. *Int. Conf. in Acoustics, Speech and Signal Processing*, 3:1751–1754, 1997.

[14] V. Valtchev. *Discriminative Methods in HMM-based speech recognition.* PhD thesis, Cambridge University Engineering Department, 1995.

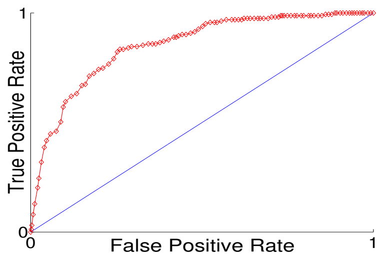


(a) Algorithm after initial temporal ordering

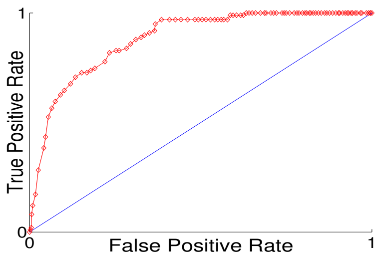


(b) Converged Principal Curve algorithm

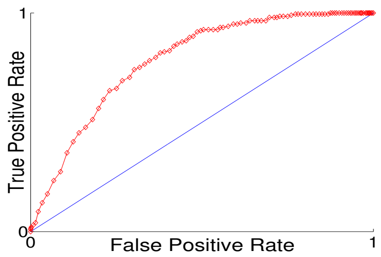
Figure 1: Principal Curves generated for the diphone */dee/* for male speaker extracted from ISOLET database The time-constrained transformation to emphasize temporal ordering is using $\tau = 2.8$. (a) shows the initial guess which represents the first temporal ordering. (b) plots the converged principal curve and the found trajectory model as squares on the curve. Time evolution is shown in the scatterplot of the data in changing gray scales from dark to bright.



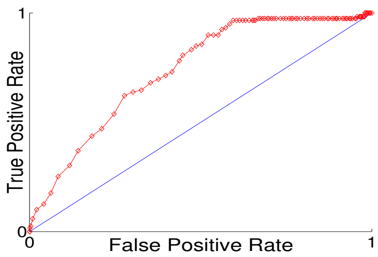
(a) ROC curve for male /bee/



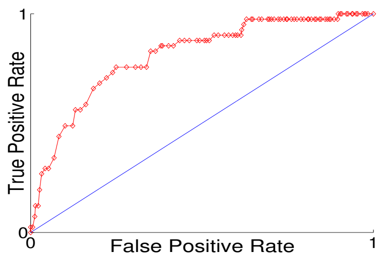
(b) ROC curve for female /bee/



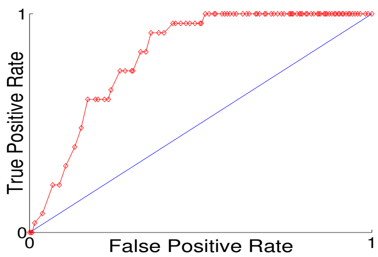
(c) ROC curve for male /dee/



(d) ROC curve for female /dee/



(e) ROC curve for male /gee/



(f) ROC curve for female /gee/

Figure 2: ROC curves obtained for female and male diphone model /bee/, /dee/ and /gee/ using TIMIT. Upper distance threshold was 100 which leads to a more pessimistic curve because all other false-positive examples with a distance score higher than 100 will not appear in this curve.