Color Figures

Reference:

Morrison, G.S., Enzinger, E., Ramos, D., González-Rodríguez, J., Lozano-Díez, A. (2020). Statistical models in forensic voice comparison. In Banks, D.L., Kafadar, K., Kaye, D.H., Tackett, M. (Eds.) *Handbook of Forensic Statistics*. Boca Raton, FL: CRC.

Figure 1. Procedure for the calculation of MFCCs. The numbers in black circles correspond to the numbered steps in the main text. DFT = discrete Fourier transform. DCT = discrete cosine transform.

Figure 2. Feature warping. The original feature value is replaced by the warped feature value. The warping is achieved by mapping from the empirical cumulative probability distribution of the original feature values to a parametric target cumulative probability distribution, in this case the standard cumulative Gaussian distribution.

Figure 3. Examples of the effects of channel and feature warping on the distribution of MFCCs. The first column represents MFCC values extracted from a high-quality audio signal (the same in both rows). The next three columns represent the results of sequentially applying various signal processing techniques to simulate casework conditions. The top row represents a simulated questioned-speaker recording condition and the bottom row represents a simulated known-speaker recording condition. The final column represents the results of applying feature warping to the values represented in the immediately preceding column, applied separately to the simulated questioned-speaker recording condition and to the simulated known-speaker recording condition.

Figure 4. Example of using the EM algorithm to train a two-component GMM. The example is based on artificial data that were created for illustrative purposes. The dotted curve represents the distribution that was used to generate the data. The dashed curve represents the fitted GMM, and the black curve and the white curve represent the two Gaussian components of the fitted GMM. The top panel shows the initial GMM distribution based on a random seed. The data points (the circles on the x axis) are shaded according to their responsibilities with respect to the two components. The second panel shows the fitted GMM distribution after 1 iteration of maximization (along with the responsibilities after that iteration). The third panel shows the results after 20 iterations, and the bottom panel the results after 40 iterations.

Figure 5. Example of using a GMM-UBM model to calculate a likelihood ratio for a single questioned-speaker-recording feature vector. This example is based on artificial two-dimensional data generated for illustrative purposes. The UBM is represented by the surface drawn with the darker mesh and the MAP adapted known-speaker GMM is represented by the surface drawn with the lighter mesh. The *x*-*y* location of the vertical line indicates a questioned-speaker-recording feature vector value at which the likelihoods of the UBM and the known-speaker GMM are being evaluated (the likelihoods are given by the *z* values of the intersections of the vertical line with each of the surfaces).

Figure 6. Example of using PCA to reduce the data from two dimensions to one dimension. The example uses artificial data created for illustrative purposes. Different shaped symbols represent different speakers and the two different intensities of shading represent two different conditions, e.g., a questioned-speaker-recording condition and a known-speaker-recording condition. The PCA dimension (represented by the oblique axis) is in the direction of maximum variance in the original *x*-*y* space.

Figure 7. Example of using CLDF for mismatch compensation and to reduce the data from two dimensions to one dimension. The example uses artificial data created for illustrative purposes (the same data as used for Figures 6). Different shaped symbols represent different speakers and the two different intensities of shading represent two different conditions, e.g., a questioned-speaker-recording condition and a known-speaker-recording condition. The CLDF dimension (represented by the oblique axis) is in the direction in the original *x*-*y* space that has the maximum ratio of between- versus within-speaker variance.

Figure 8. Example of using a PLDA model to calculate a score for a pair of i-vectors. The numerator of the PLDA model is represented by the surface drawn with the lighter mesh and the denominator is represented by the surface drawn with the darker mesh. The *x*-*y* location of the vertical line indicates the values of the i-vectors v_q and v_k (extracted from the questioned- and known -speaker recordings respectively) at which the likelihoods of numerator and denominator of the PLDA model are being evaluated (the likelihoods are given by the *z* values of the intersections of the vertical line with each of the surfaces).

Figure 9. Simplified example of the architecture of a feed-forward DNN consisting of an input layer, two hidden layers, and an output layer. Not all connection between nodes have been drawn, and weights have been indicated for only a few connections.

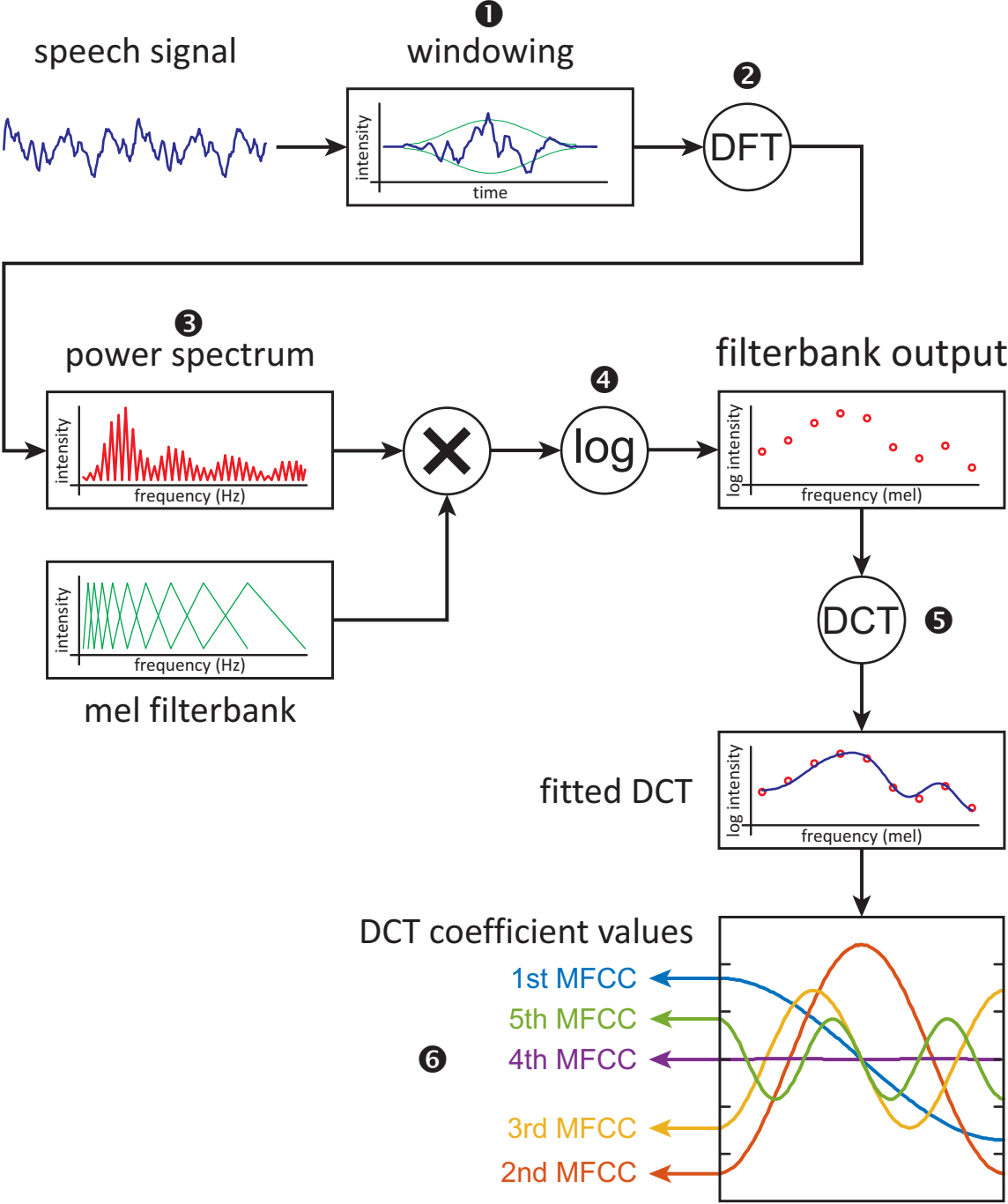
Figure 10. Illustration of the architecture of a DNN embedding system. Only the time-dimension is

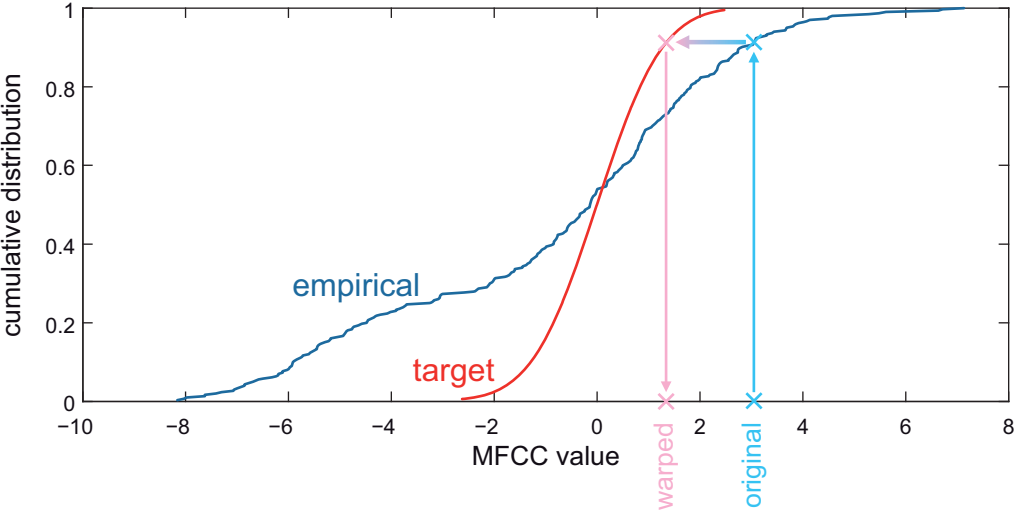
shown for the frame level (the frequency dimension is not shown). The input layer (at the bottom) spans t–7 through t+7. The time dimension is collapsed by the third hidden layer of the frame level. The statistics-pooling layer calculates the mean and standard deviation of each node in the preceding layer when the whole of the speech of interest in a recording is passed by the input layer. The frequency dimension is shown for the segment level.

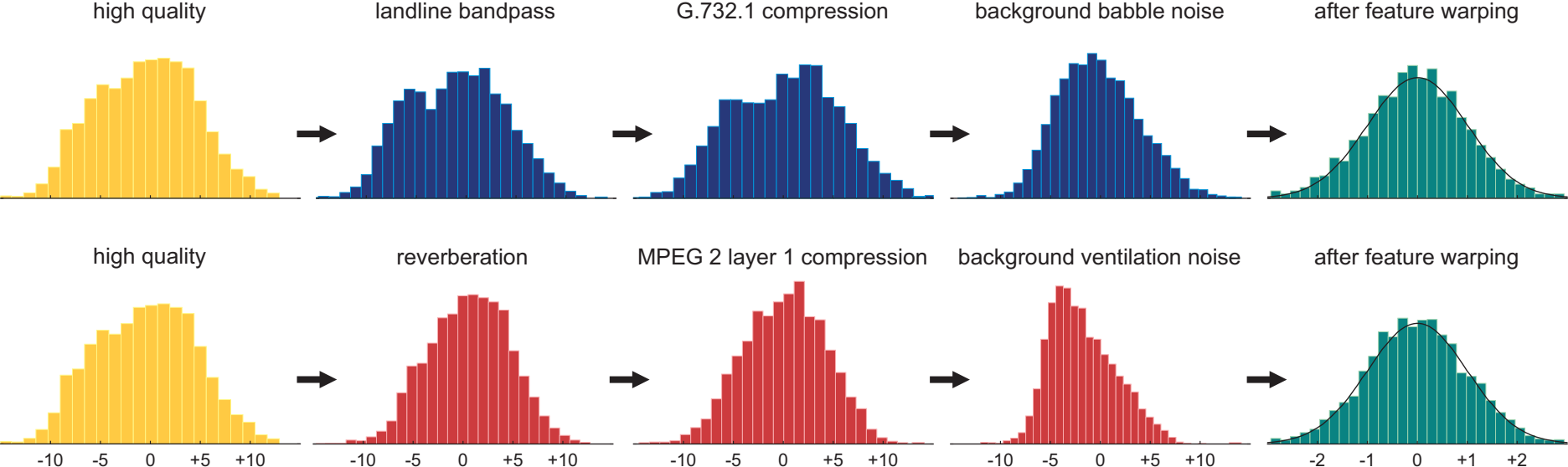
Figure 11. Example of using linear discriminant analysis or logistic regression to convert a score to a likelihood ratio. The example uses artificial data created for illustrative purposes. Different-speaker training scores are shown as grey triangles and same-speaker training scores are shown as white circles. The top panel shows a linear discriminant analysis model fitted to the data. The middle and bottom panels can be derived from the top panel. The middle and bottom panels can also be derived by fitting a logistic regression model to the same data. The vertical line represents a score value that is being converted to a likelihood-ratio value. In the top panel the output is clearly the ratio of two likelihoods. The middle panel shown that the conversion from scores to log-likelihood-ratio values is a linear function. (In reality the plotted training data are illustrative only and the plotted functions show ideal values based on specified parametric distributions.)

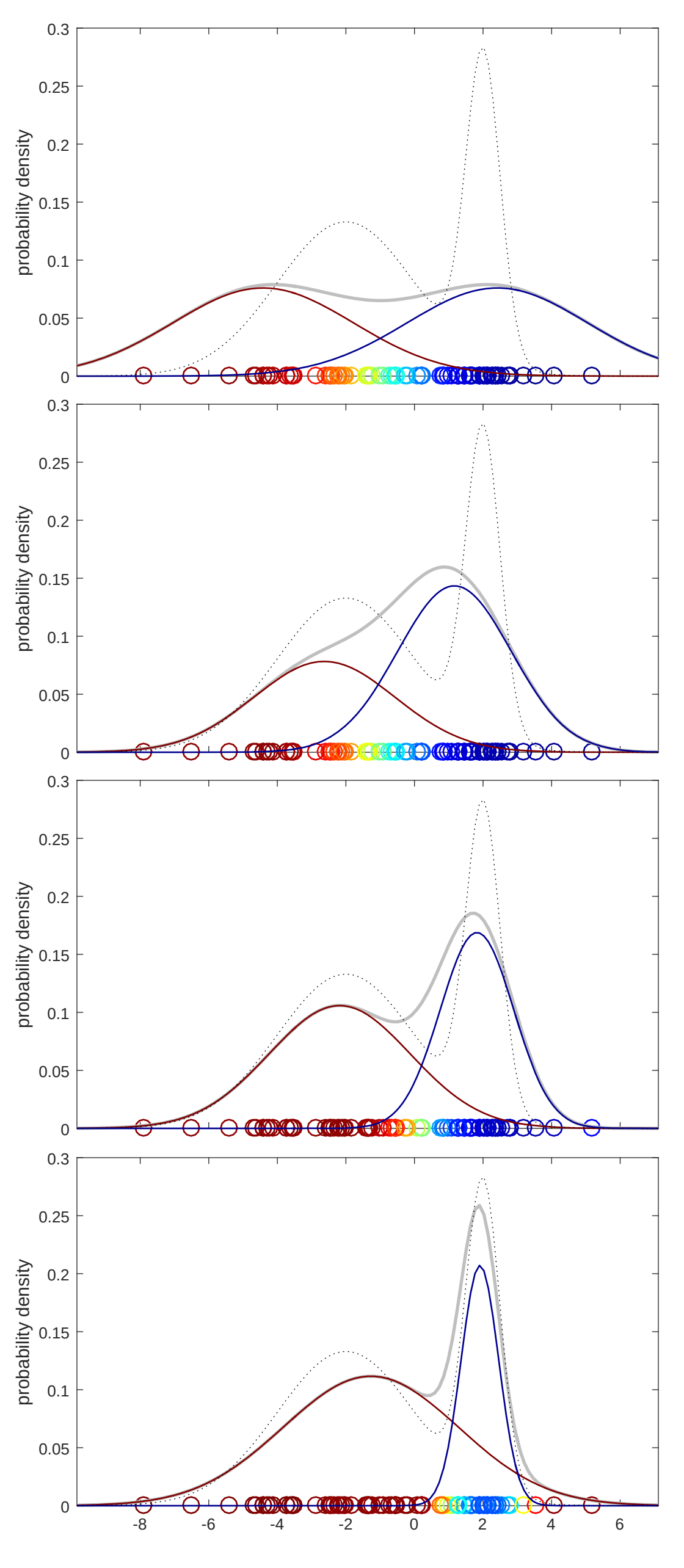
Figure 12. Penalty functions for calculating C_{llr}. The same-speaker and different-speaker curves correspond to the functions within Equation 20's left and right summations respectively.

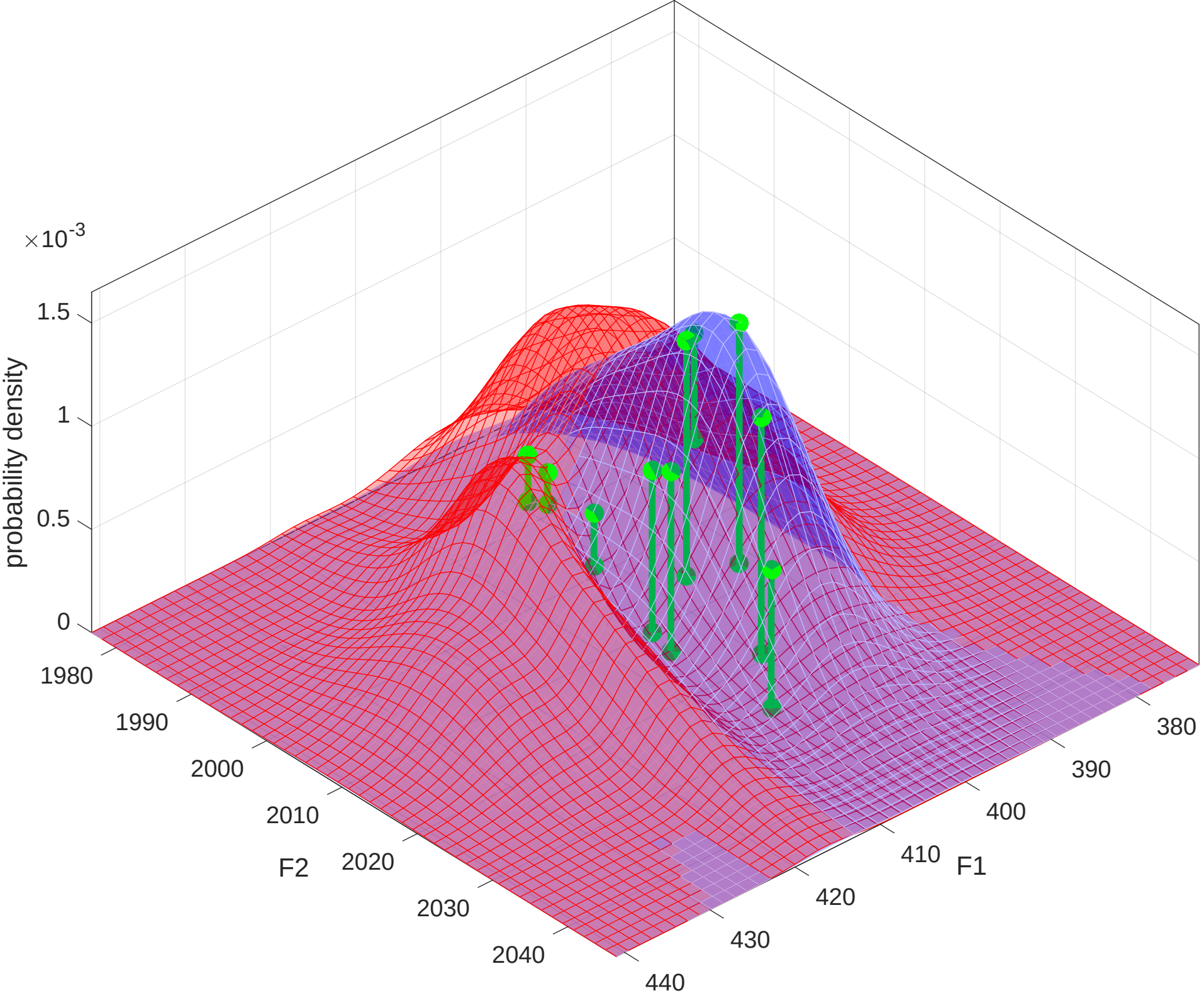
Figure 13. Examples of Tippett plots. The three plots represent three different systems tested on the same test data. The examples are based on artificial data created for illustrative purposes. The data represent 50 same-speaker test pairs and 200 different-speaker test pairs – imbalance in the number of same-speaker and different-speaker test pairs is usual since it is easier to construct different-speaker pairs than same-speaker pairs. The C_{llr} values corresponding to the results shown in the top, middle, and bottom panels are 1.068, 0.698, and 0.307 respectively.

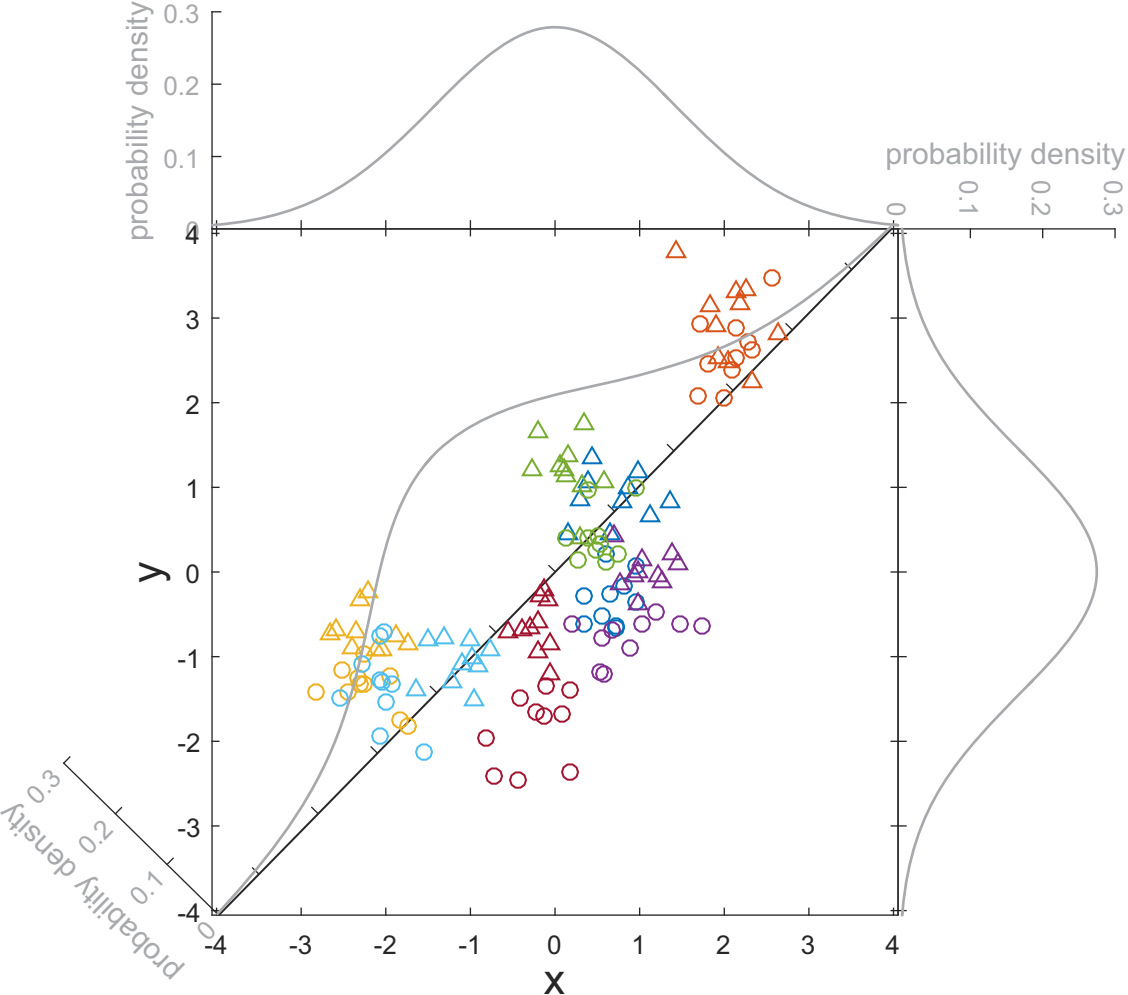

speech signal
1
windowing
intensity
time
2
DFT
3
power spectrum
intensity
frequency (Hz)
4
log
filterbank output
log intensity
frequency (mel)
DCT
5
intensity
frequency (Hz)
mel filterbank
fitted DCT
log intensity
frequency (mel)
DCT coefficient values
6
1st MFCC
5th MFCC
4th MFCC
3rd MFCC
2nd MFCC

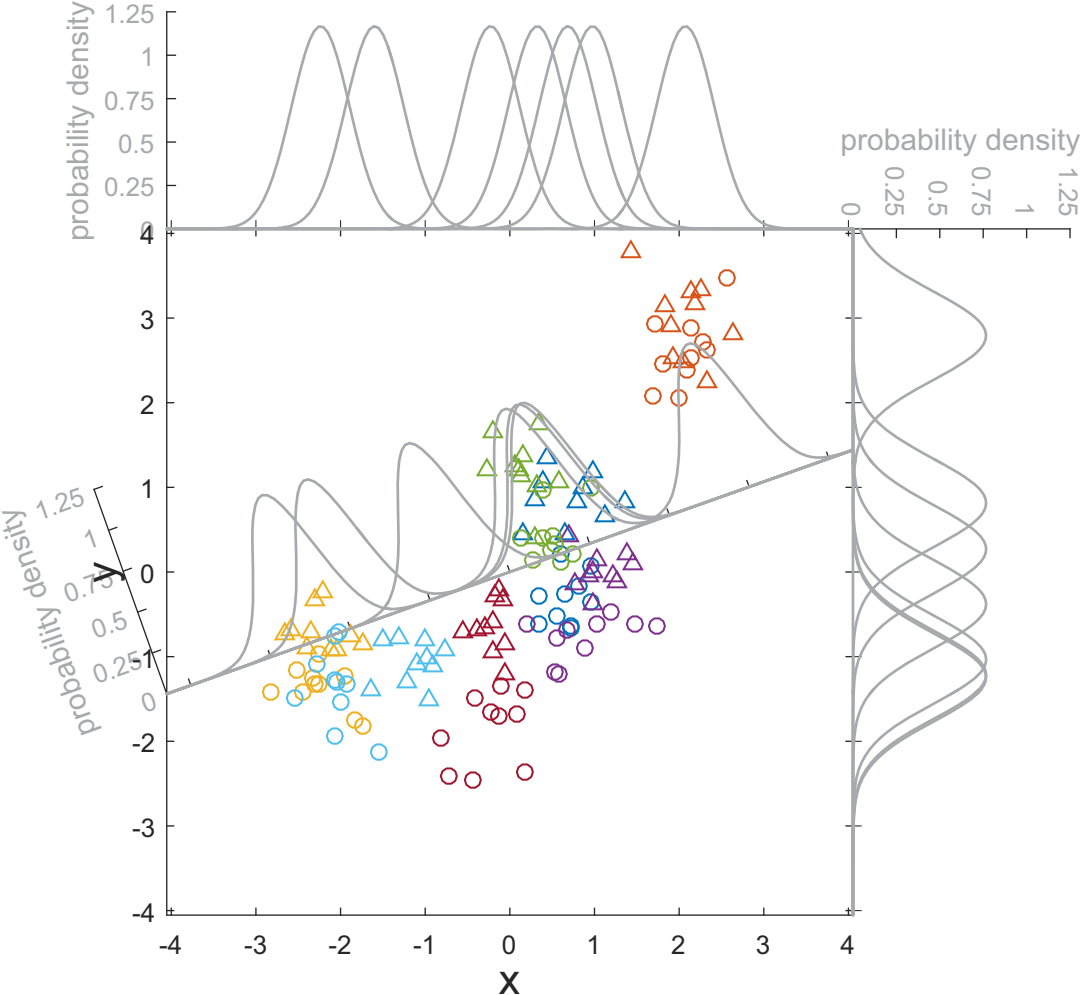

cumulative distribution
1
0.8
0.6
0.4
0.2
0
empirical
target
warped
original
−10
−8
−6
−4
−2
0
2
4
6
8
MFCC value

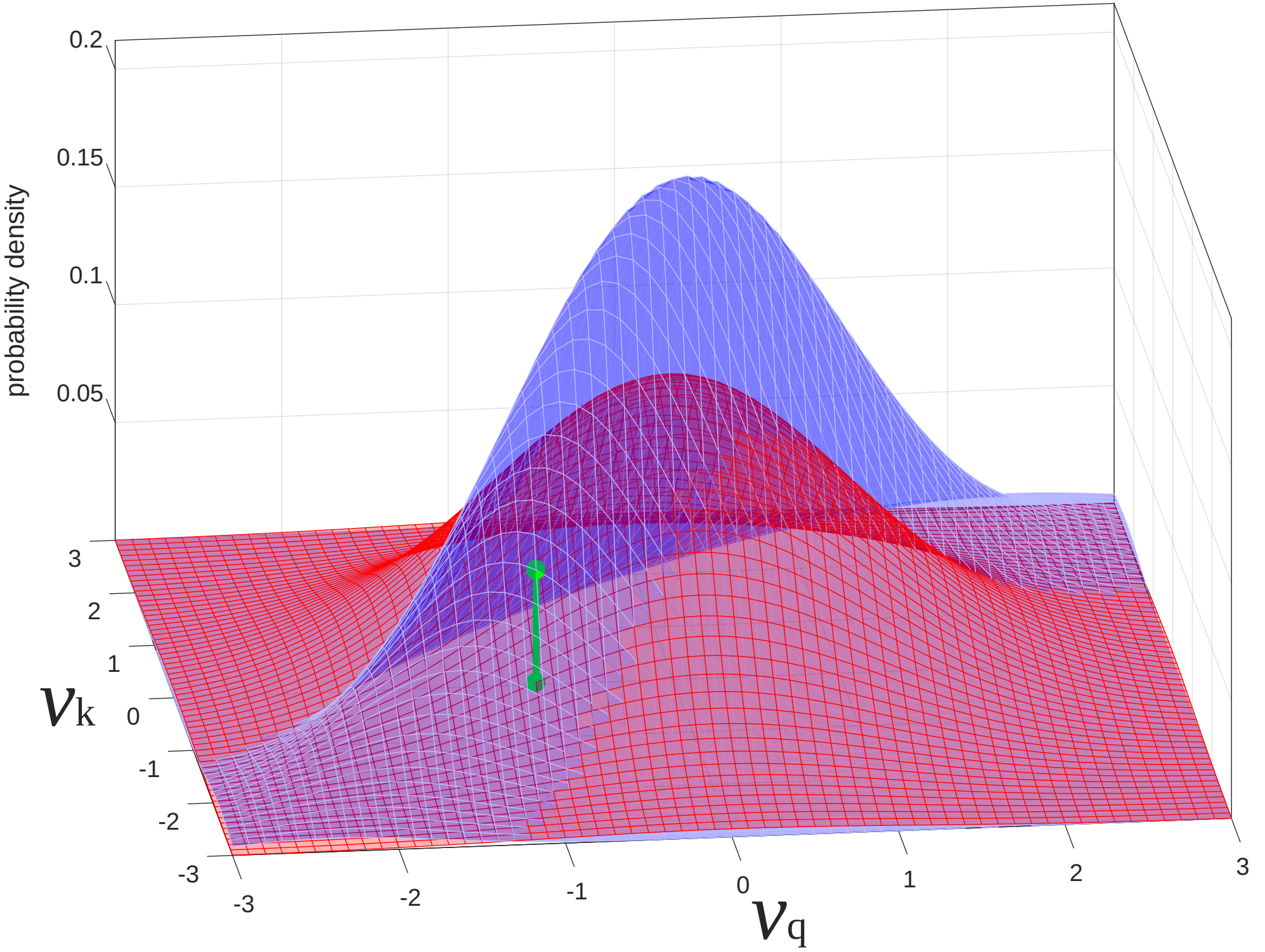

high quality
landline bandpass
G.732.1 compression
background babble noise
after feature warping
high quality
reverberation
MPEG 2 layer 1 compression
background ventilation noise
after feature warping
-10 -5 0 +5 +10
-10 -5 0 +5 +10
-10 -5 0 +5 +10
-10 -5 0 +5 +10
-2 -1 0 +1 +2

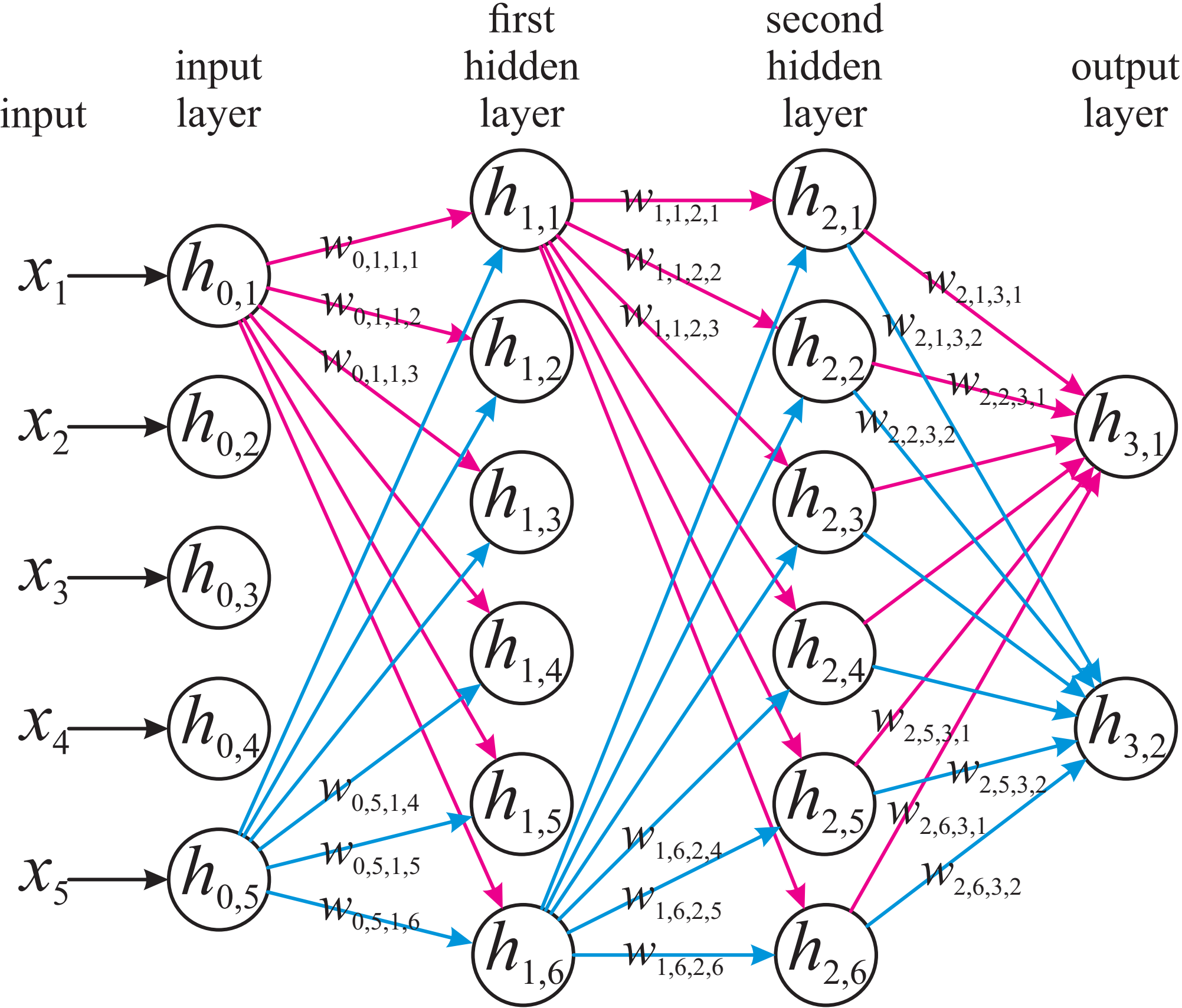

probability density
0.3
0.25
0.2
0.15
0.1
0.05
0
-8
-6
-4
-2
0
2
4
6

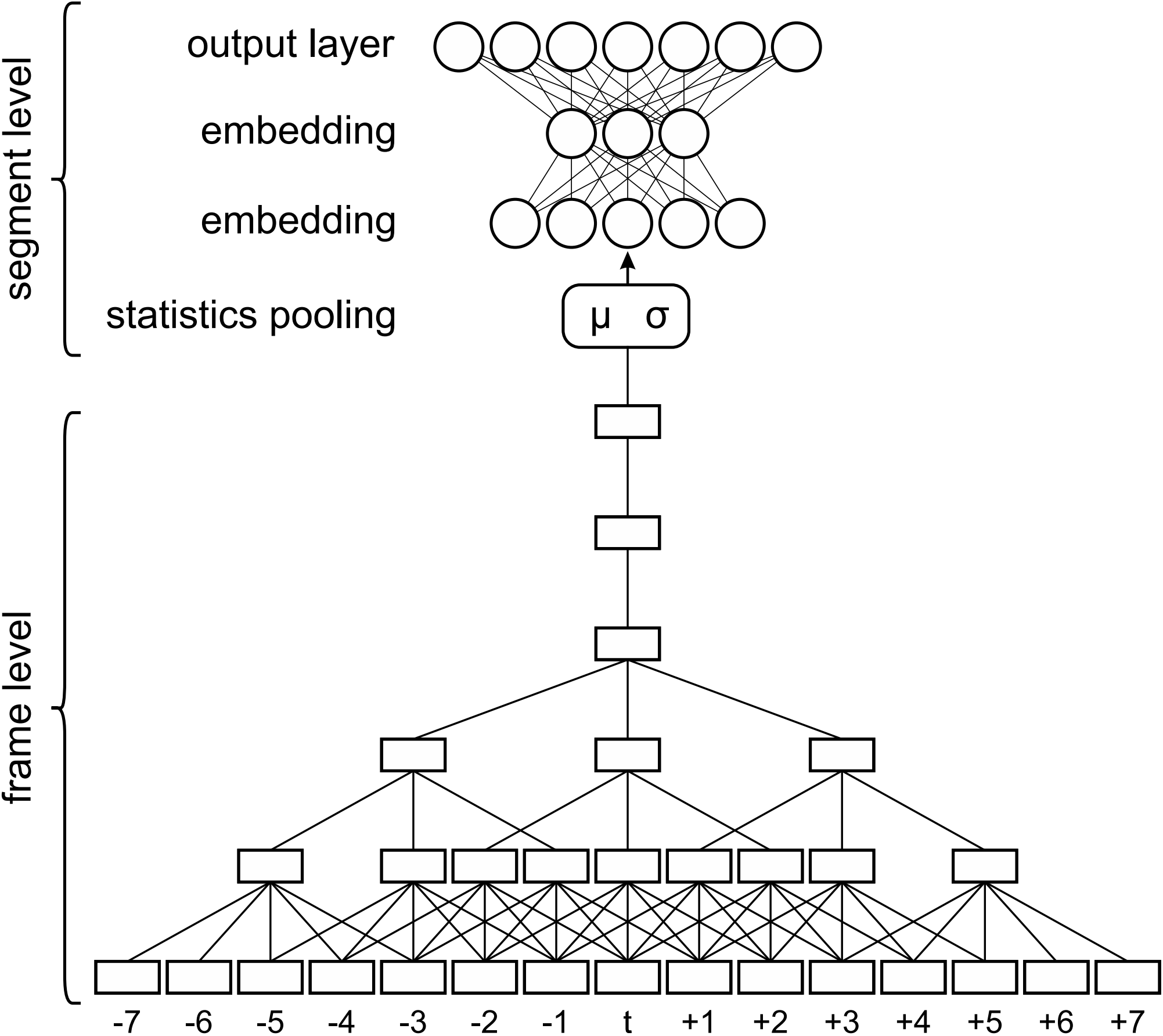

×10-3
1.5
1
0.5
0
probability density
1980
1990
2000
2010
2020
2030
2040
F2
380
390
400
410
420
430
440
F1

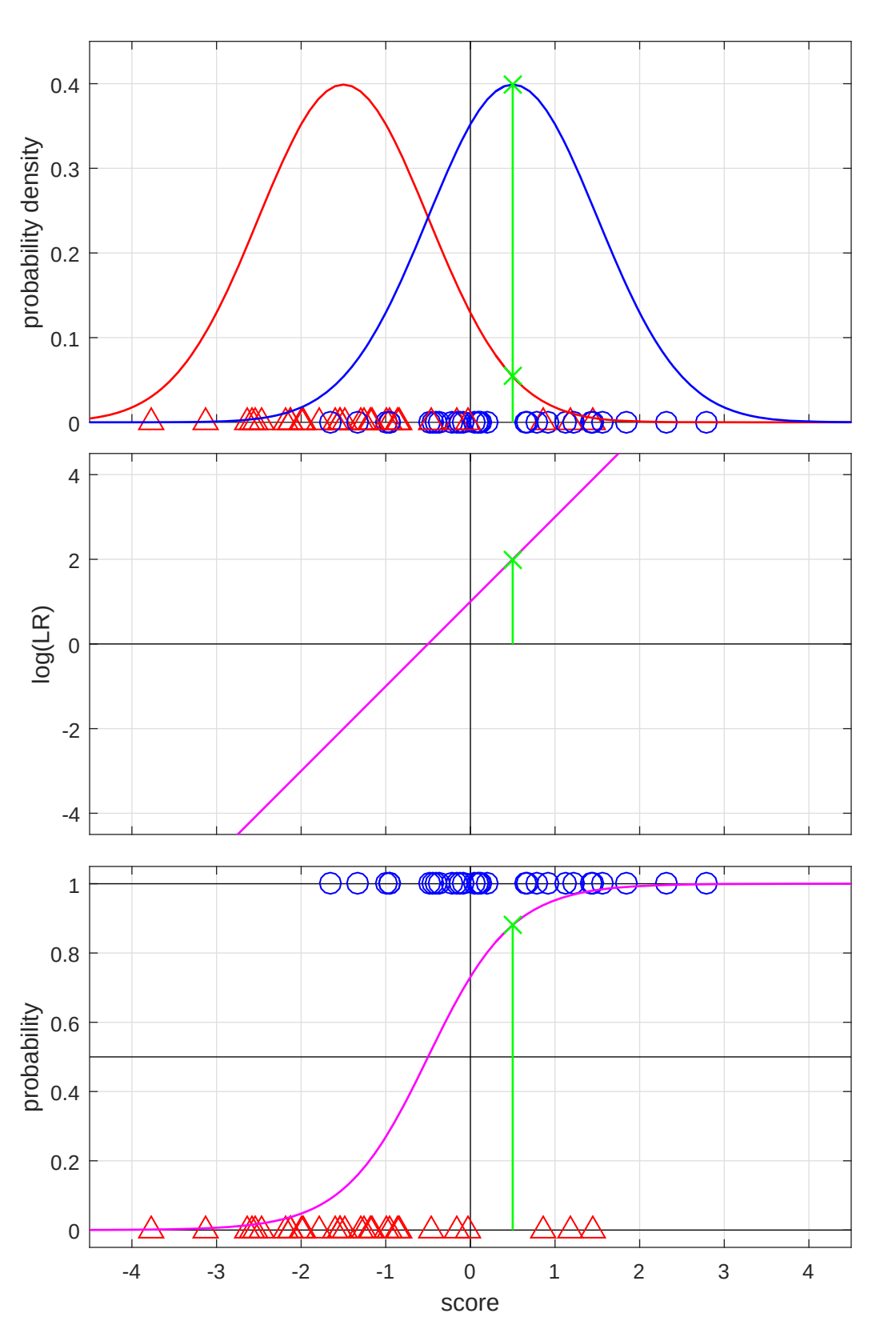

probability density
0.3
0.2
0.1
0
probability density
0
0.1
0.2
0.3
y
x
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
probability density
0
0.1
0.2
0.3

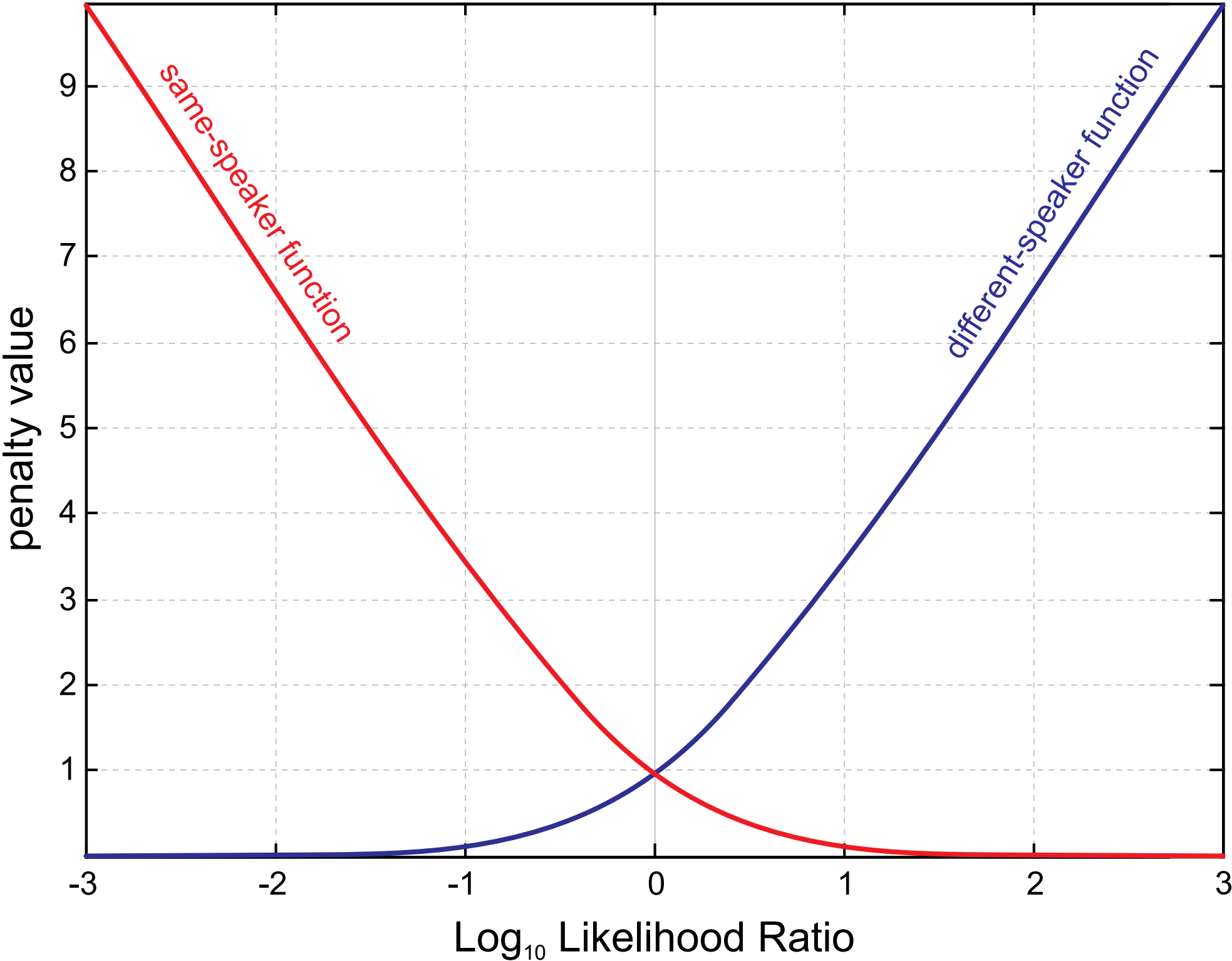

probability density
1.25
1
0.75
0.5
0.25
0
probability density
0
0.25
0.5
0.75
1
1.25
4
3
2
1
0
-1
-2
-3
-4
probability density
1.25
1
0.75
0.5
0.25
0
y
-4
-3
-2
-1
0
1
2
3
4
x

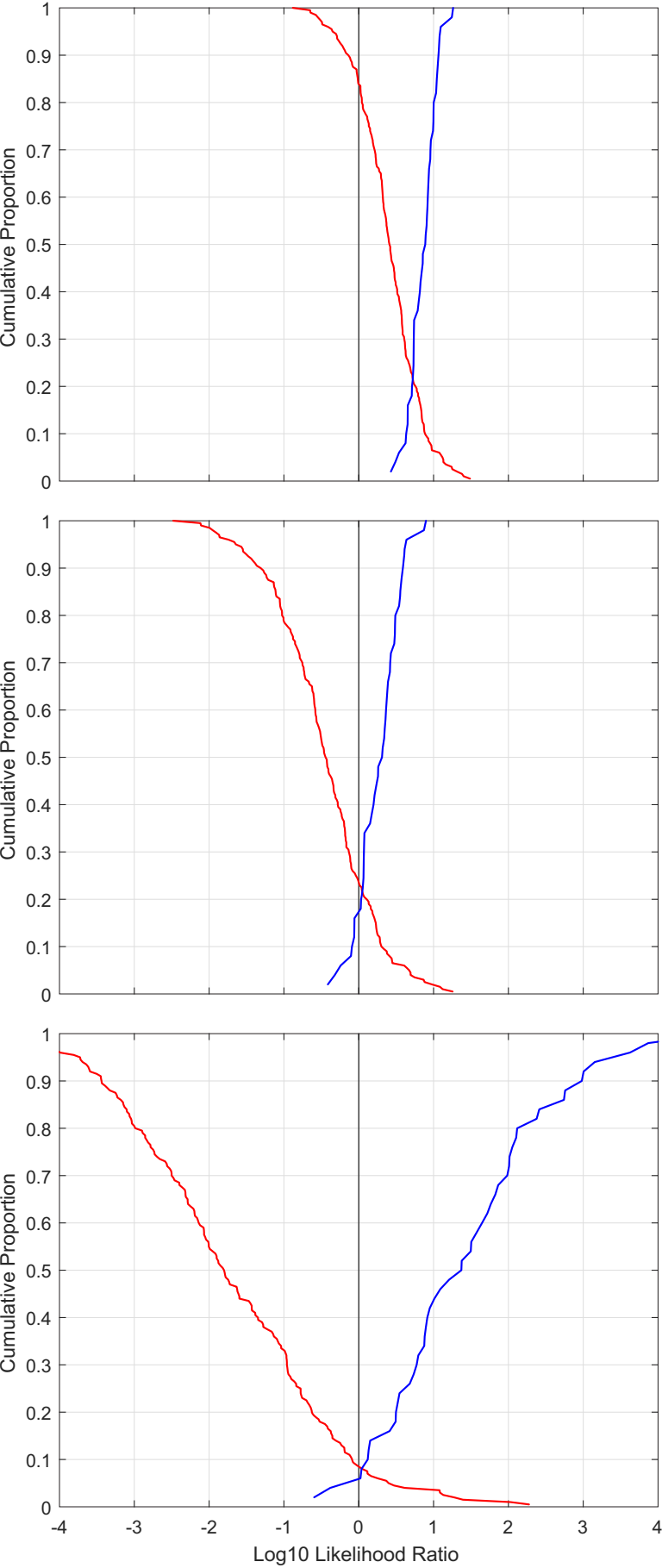

probability density
0.2
0.15
0.1
0.05
3
2
1
0
-1
-2
-3
ν_k
-3
-2
-1
0
1
2
3
ν_q


input
input layer
first hidden layer
second hidden layer
output layer
x_1
x_2
x_3
x_4
x_5
$h_{0,1}$
$h_{0,2}$
$h_{0,3}$
$h_{0,4}$
$h_{0,5}$
$h_{1,1}$
$h_{1,2}$
$h_{1,3}$
$h_{1,4}$
$h_{1,5}$
$h_{1,6}$
$h_{2,1}$
$h_{2,2}$
$h_{2,3}$
$h_{2,4}$
$h_{2,5}$
$h_{2,6}$
$h_{3,1}$
$h_{3,2}$
$w_{0,1,1,1}$
$w_{0,1,1,2}$
$w_{0,1,1,3}$
$w_{0,5,1,4}$
$w_{0,5,1,5}$
$w_{0,5,1,6}$
$w_{1,1,2,1}$
$w_{1,1,2,2}$
$w_{1,1,2,3}$
$w_{1,6,2,4}$
$w_{1,6,2,5}$
$w_{1,6,2,6}$
$w_{2,1,3,1}$
$w_{2,1,3,2}$
$w_{2,2,3,1}$
$w_{2,2,3,2}$
$w_{2,5,3,1}$
$w_{2,5,3,2}$
$w_{2,6,3,1}$
$w_{2,6,3,2}$


segment level
output layer
embedding
embedding
statistics pooling
μ σ
frame level
-7 -6 -5 -4 -3 -2 -1 t +1 +2 +3 +4 +5 +6 +7


probability density
0.4
0.3
0.2
0.1
0
log(LR)
4
2
0
-2
-4
probability
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4
score


same-speaker function
different-speaker function
penalty value
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log10 Likelihood Ratio


Cumulative Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Cumulative Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Cumulative Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-4
-3
-2
-1
0
1
2
3
4
Log10 Likelihood Ratio