Contract between Phonexia, BUT and MV ČR, č. smlouvy: 16002

# Unsupervised Score Calibration for Speaker Verification Systems

February 27, 2018

Authors: Ján Profant, jan.profant@phonexia.com
Pavel Matějka, matejkap@fit.vutbr.cz

# Contents

# Introduction

This report describes analysis of the effect of calibration and a recipe for creating a calibration set from the customer data.

Speaker identification systems can produce „wrong scores" on the data which are very different than the data the system was trained on. Such „wrong scores" can be corrected by the score normalization/calibration. It is a process where we show the data from the new domain/channel to the SID system and the SID system can analyze how it behaves on the new data and correct the output score so that it gives meaningful output. Attached is a paper which analyzes different techniques of score normalization. But the essential step is the correct selection of the data which are exposed to the system – we call it **calibration set**.

All processing is ideally done on the audio files, but most of the steps can be done on ivectors from the SID system.

**Ideal calibration set requirements:**

- Same recording conditions/channels as in the evaluation set
- One recording per speaker
- One speaker in the recording
- Different spoken content of the recording
- Enough net speech – shorter recordings to match testing scenario can be cut from the long ones

## Real Data Collection

A random selection from one-day collection of traffic is a good approximation, but can lead to some pathologic cases. If there are not enough data from the day traffic, there might be for example one speaker with lots of calls, repeated calls, very short calls, operator announcements, recordings with very bad quality, etc.

We propose several steps for selecting a set of recordings from the „big amount of data" from the day traffic. We expect rather bigger amount of data so that we can afford to throw away the data which are not suitable for the calibration set. The selection process contain several steps described in the next section.

# Selection Process

Selection of recordings is crucial for a successful SID calibration. In this chapter, the requirements and approaches for a selection are proposed.

## Recording Quality

This block checks the recording quality – mainly net speech and SNR.

We propose the SNR of the recording is higher than 5dB, because very bad recordings can cause the problem in the calibration process.

The net speech of the recording is at least 60sec. This number is experimentally verified. More speech does not significantly influence the calibration results. If the customer data are generally of shorter duration and there would be a problem in collecting data, then this limit can be lowered. The new limit will be set as 75th percentile of the net speech length (per call) on the daily traffic of the customer data.

## Recording Duplicity

The purpose of this block is to eliminate recordings which are the same or are cuts from the longer ones. Since we are working with ivectors the best solution for this is to use cosine distance between length normalized ivectors. If the cosine distance between two ivectors is bigger than *0.95* we consider that these two recordings come from the same source = are the same recordings or cuts from the same recordings or very close to each other and can cause problems in the calibration set.

## Speaker Clustering

This block ensures that there will be only one recording per speaker and also one speaker in the recording.

1. Keep only recordings shorter than 5 minutes of net speech or take only the first five minutes. Longer recordings have minimal effect on calibration.

2. Run speaker diarization and keep only recordings where there is only one speaker per file.

3. Cut each recording to two – we expect there is only one speaker in the recording which makes these two halves a **target trial**. For verification, compute cosine distance between these two halves and keep only those where cosine distance is lower than *0.3*.

4. Compute PLDA score of target trials.

5. Compute PLDA score of non-target trials - non-target trials are formed by scoring different recordings.

6. Based on 4 and 5, compute linear calibration (LinCal) parameters for recomputing score to be presentable as LLR (log likelihood ratio).

$$mtar = \frac{1}{n}\sum_{i=1}^{n} target\_scores_i$$

$$mnon = \frac{1}{m}\sum_{i=1}^{m} nontarget\_scores_i$$

$$var = \frac{\sum_{i=1}^{n}(target\_scores_i - mtar)^2 + \sum_{i=1}^{m}(nontarget\_scores_i - mnon)^2}{n+m}$$

7. Compute PLDA score between all original full length recordings and recompute them according to:

$$new\_score = \frac{(old\_score - mnon)^2 - (old\_score - mtar)^2}{2 \cdot var}$$

8. Find all pairs of the recordings which have *new_score* higher than a threshold - these recordings come from the same speaker. Value of threshold can be set analytically or experimentally. With this approach, we can group all recordings which come from the same speaker. Ideal threshold would be set to 0, but since our new calibration is biased (we computed the target scores from the same recording and in non-target scores might be target ones), it is better to set the threshold to more positive values. Experimentally it would be set in the way that we have some evaluation set and run this recipe with varying threshold and observe the results on this validation set. The lowest CLLR would correspond to the threshold we want to use for the clustering in deployment.

9. The process can be done once more from point 4 if we already know the recordings which belong to the same speaker from point 8, which makes more target trials and more precise estimation of the mean of these two distributions – but from our experience this step is not mandatory and will not change the results significantly.

10. Select only one recording from each speaker and form calibration set.

There might be objections that speaker clustering is not precise, but for our purpose — eliminating very similar recordings from the same speaker — this process is good enough.

# Final Selection

1. Randomly select 1000 or more recordings in total

   (a) Random selection ensures that the final calibration set would be similar to the target channel characteristics

(b) The amount of data in the final set depends on the technique which is used for the score normalization/calibration. Current normalization is normalization to number of False Alarms (FA). If we want to normalize to 1% of FA, we need to have at least 1000 recordings (10/0.01=1000). If we want to normalize to 0.5% FA, we need to have 10/0.005=5000 recordings.

2. Create a set of these recordings and give it a name. Later, when using this set, point to this name, so that each test can be reproducible.

# Experiments

This chapter verifies the proposed recipe with results and analysis.

## Evaluation Metric

We choose the $C_{llr}$ metric as a main evaluation metric, because it considers how well all scores represent the likelihood ratio and also penalizes for errors in score calibration [1]. As the accuracy of a speaker recognition system gets higher the Cllr tends towards zero. If we calibrate the system on the evaluation data (which is cheating) we obtain perfectly calibrated system hence zero calibration loss - this we call $C^*_{llr}$ or $C^{min}_{llr}$ which reflects only the discriminative power of the speaker ID system. The difference between $C_{llr}$ and $C^*_{llr}$ represent the calibration loss.

The $C_{llr}$ performance measure is defined as:

$$C_{llr} = \frac{1}{2 \times log_2} \times \left( \frac{\sum log(1 + \frac{1}{s})}{N_{TT}} + \frac{\sum log(1 + s)}{N_{NT}} \right) \tag{1}$$

where the first summation is over all target trials $N_{TT}$, the second is over all non-target trials $N_{NT}$, and $s$ represents a trial's likelihood ratio

## Data

For SID calibration, we used FISHER ENGLISH data - 1000 files with 1000 speakers together with NIST SRE 2004 data - 1556 files containing 75 speakers, which makes it in total 2556 files containing 1075 speakers and 17931 target trials and 3232044 non-target trials. Our approach was evaluated on different NIST SRE 2004 dataset with 70 speakers in 1394 files. All recording have English as spoken language. Average length of clean speech in recording is 121 seconds.

## System description

Phonexia vp_extract with XL3 model is used to generate the i-vectors. The i-vector extraction scheme is shown in Figure 1.

In our system, we used only English data for PLDA training - shown as Postprocessing in Figure 1, therefore discriminability and calibration of the used system is worse compared to the official Phonexia system. PLDA for this experiment is trained on 59223 recordings with 5068 speakers.

[1] The reasons for choosing this cost function, and its possible interpretations, are described in detail in the paper "Application- independent evaluation of speaker detection" in Computer Speech & Language, volume 20, issues 2-3, April-July 2006, pages 230- 275, by Niko Brummer and Johan du Preez.

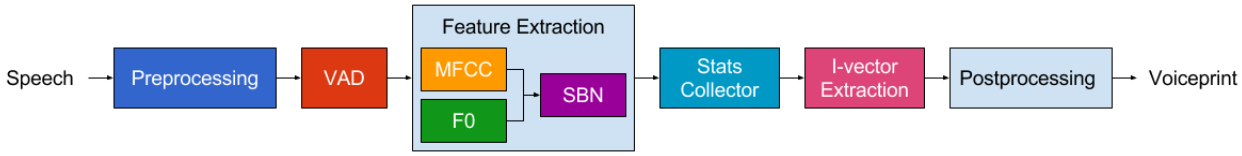


Figure 1: General block scheme of voiceprint(i-vector) extractor.

# Results

Figure 2 shows the score distribution on the evaluation set for raw scores (uncalibrated system) plotted with solid line. The optimal Linear calibration (LinCal)[2] trained on the evaluation set[3] is plotted with dotted line. The figure shows that the score from the original system is far away from the well calibrated system.

Table 1 compares the results of the system without linear calibration (line 1) and with different stages of the linear calibration derived during the speaker clustering process. This calibration is needed there for estimating the number of speakers in the calibration datasets and correct speaker clustering. EER in Table 1 do not change, because linear calibration does not change the discriminability of the system, it only shifts and scales the scores. The main effect is on the CLLR metric which takes into account also the calibration loss.

At first we can see the CLLR without any calibration is 3.4 and EER threshold is 17.27. This means that the system produces very badly calibrated scores. Figure 2 confirms this - it shows how much the score is shifted if we compare it to the ideal case where the calibration parameters are trained on the same data with true labels.

Next, there are results if we apply *STEP 3* from the speaker clustering schema - where we cut the recording to 2 halves to get the target trial. It is followed by *STEPs 4,5,6* where we score it and then plot the distribution of scores (Figure 3). In this scenario, the target scores are very optimistic (shifted to positive values, because they contain only very easy trials which come from the same recording) and non-target scores contain a lot of target trails which come from the same speaker but different recording. The CLLR for this system is 0.81.

Next, we can apply *STEP 8* to find recordings which come from the same speaker. For this, we need to set a threshold which says which recordings belong to the same speaker. In this case it is better to be more on the safe side - set a higher threshold rather than lower. We choose threshold which reflects 1% FA and can be calculated as $-logit(0.01) = 4.59$. This threshold eliminates most of the obvious errors as CLLR drops to 0.43 and EER threshold to -1.37 - see also Figure 4 for plot of score distribution. In an ideal case, we would have some labeled development/evaluation data from the same channel, where we can run our approach the get a right threshold for speaker clustering - in our case, we used the evaluation set (which is cheating) and we run the calibration with different thresholds for speaker clustering in *STEP 8*. We can see from Table 2 that the optimal threshold is around 3.0 where the CLLR is 0.36 while with the labeled data, we have 0.33.

Then, we can continue and select only one recording per speaker to form a calibration set

[2]LinCal is linear calibration, where global scale and shift are trained to transform scores to be interpreted as Likelihood Ratio (LR) - see equations in section on Speaker Clustering. This calibration do not change the discriminability of the system (EER is still same), but changes the interpretability of the system = calibration loss

[3]This is cheating, but can show us how far we are from the ideal case

which will be used for the final calibration to 1% FA. Table 3 describes the effect of calibration on 1% FA (implemented in the current Phonexia system) on the different calibration set. @@@ Describe the results, when the table is done @@@

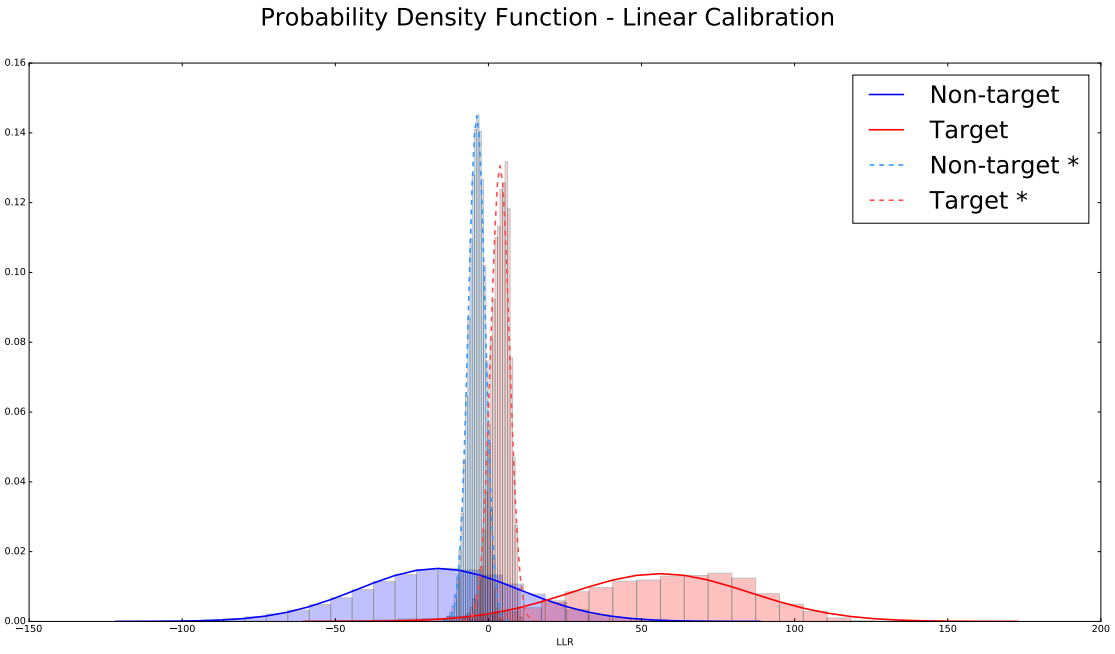


Figure 2: *Distribution of scores not using any score calibration compared to ideal distribution using linear calibration (dotted line).*

Table 1: *Results (CLLR, EER, EER threshold) for different calibration cases.*

| **Calibration** | **CLLR** | **EER** | **EER threshold** |
|---|---|---|---|
| Turned off | **3.40** | 9.33 | 17.27 |
| Cuts | **0.81** | 9.33 | -4.34 |
| Unsupervised | **0.43** | 9.33 | -1.37 |
| Unsupervised/Tuned | **0.36** | 9.33 | 0.37 |
| Supervised (cheating) | **0.33** | 9.33 | -0.23 |

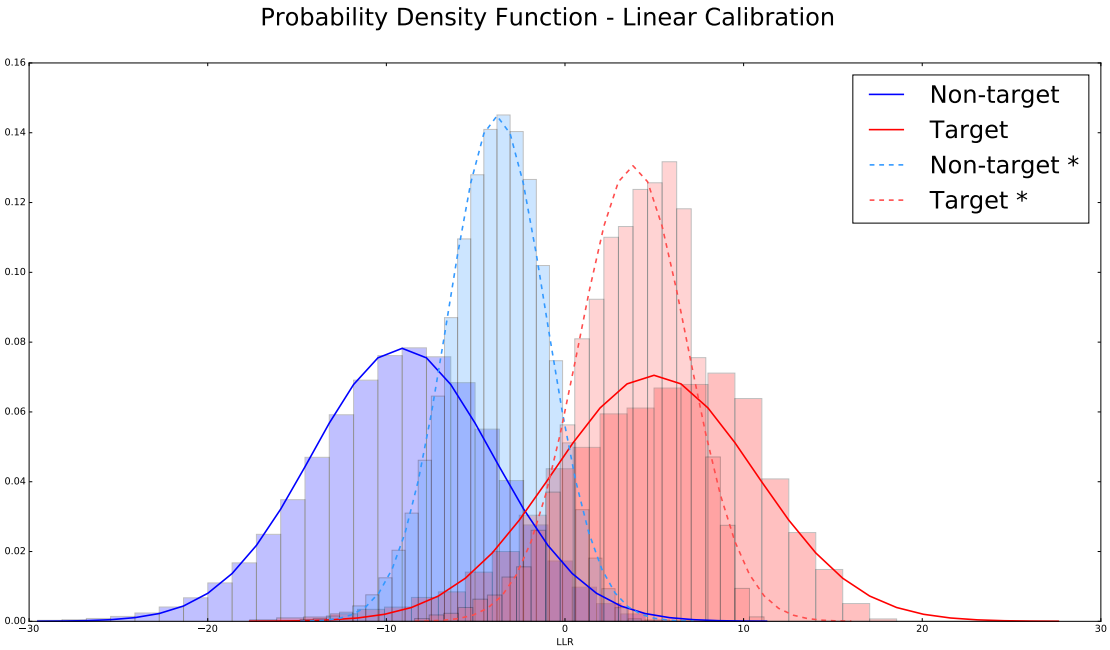


Figure 3: *Distribution of scores after calibration using cuts of the recording to halves compared to ideal distribution using linear calibration.*

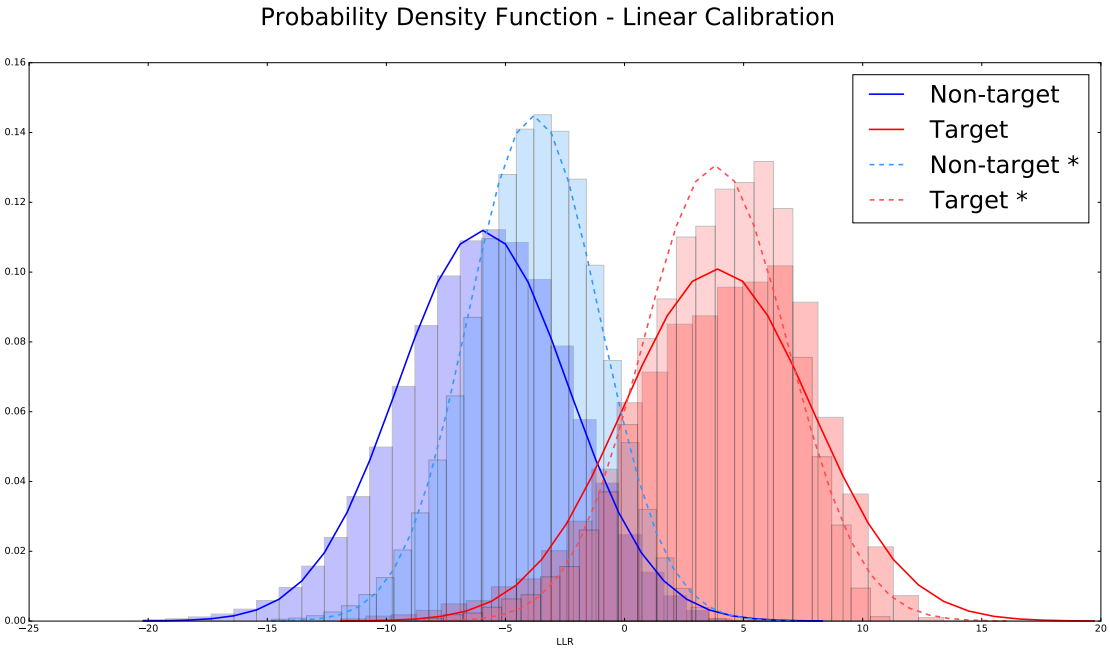


Figure 4: *Distribution of scores after whole unsupervised calibration process compared to ideal distribution using linear calibration.*

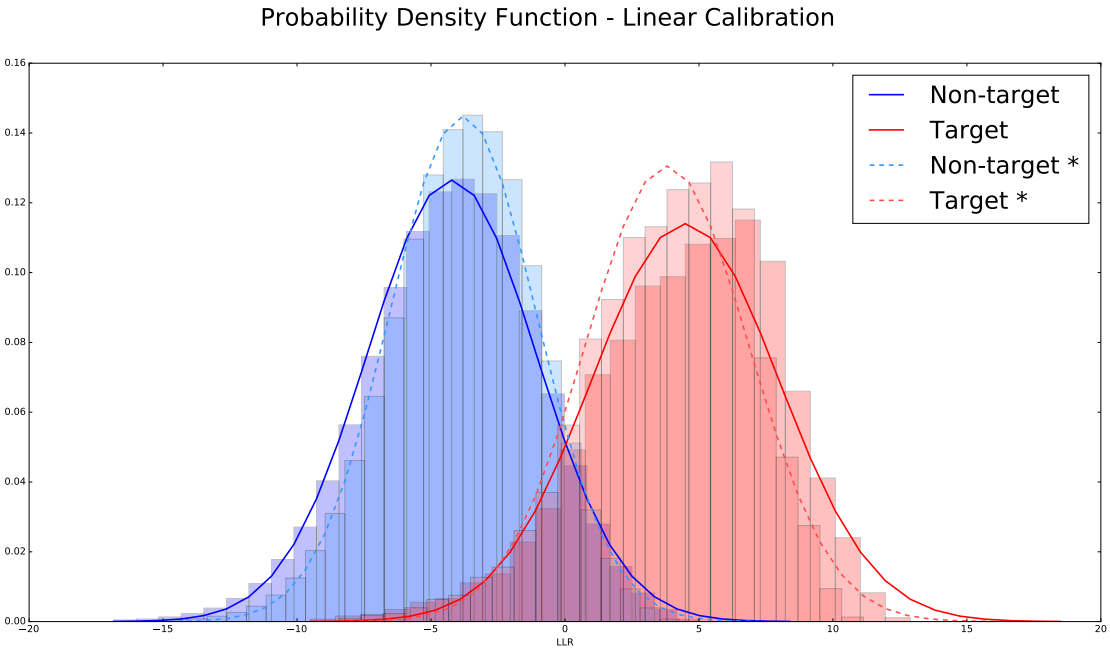


Figure 5: *Distribution of scores after whole unsupervised calibration process with tuning compared to ideal distribution using linear calibration.*

Table 2: *Threshold estimation for speaker clustering.*

| **Threshold** | **CLLR** | **EER threshold** | **Scale** | **Shift** |
|---|---|---|---|---|
| 0.0 | 0.44 | 1.30 | 0.09 | 0.72 |
| 1.0 | 0.43 | 1.20 | 0.09 | 0.65 |
| 2.0 | 0.40 | 0.84 | 0.07 | 0.38 |
| 3.0 | 0.36 | 0.37 | 0.07 | -0.09 |
| 4.0 | 0.40 | -1.30 | 0.12 | -2.05 |
| 5.0 | 0.44 | -1.60 | 0.12 | -2.38 |
| 6.0 | 0.44 | -1.67 | 0.12 | -2.47 |

Table 3: *Results (CLLR, EER, EER threshold) for different calibrations on 1% FA with different selection from the original calibration set. FAST means that quality and duplicity steps were applied on the calibration datasets. Reference means that calibration set was derived based on the assumptions for ideal calibration set based on the true labels. CLLR* is the result with linear calibration trained on the evaluation data = there is no calibration loss*

| **Calibration for 1% FA** | **EER** | **CLLR** | **CLLR*** |
|---|---|---|---|
| No Calibration | | | |
| Calibration original full set | | | |
| Calibration FAST(=Quality+Duplicity) | | | |
| Calibration FAST+Clustering(Unsupervised) | | | |
| Calibration FAST+Clustering(Unsupervised/Tuned) | | | |
| Calibration FAST+Clustering(Supervised) | | | |
| Calibration with reference (cheating) | | | |

# Conclusions

The proposed recipe for selecting the recordings for the calibration set is designed to eliminate negative issues in the usage of such set for the calibration of an SID system on the customer data. Such calibration can improve the results and also shift and scale the output scores to meaningful range. The tests are done on the NIST data with artificially designed set for deriving calibration set. We are currently running tests on the data close to the real customer data.

# Appendix 1

In this appendix, procedure described in the document above is verified on the same set of recordings with one difference - duration of the net speech in recordings. We designed the set so that audio recordings are from 15 to 40 seconds of net speech in uniform distribution. The trends and conclusion are same as before with one difference. The optimal threshold for speaker clustering is 9, which is higher than before and calls for another tests with different dataset.

Table 4: *Results (CLLR, EER, EER threshold) for different calibration cases using short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

| **Calibration** | **CLLR** | **EER** | **EER threshold** |
|---|---|---|---|
| Turned off | **2.16** | 8.30 | 10.53 |
| Cuts | **0.63** | 8.30 | 2.59 |
| Unsupervised | **0.71** | 8.30 | 2.68 |
| Unsupervised/Tuned | **0.39** | 8.30 | 0.43 |
| Supervised (cheating) | **0.32** | 8.30 | -0.09 |

Table 5: *Threshold estimation for speaker clustering using short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

| **Threshold** | **CLLR** | **EER threshold** | **Scale** | **Shift** |
|---|---|---|---|---|
| 0.0 | 0.74 | 2.68 | 0.09 | 1.70 |
| 1.0 | 0.74 | 2.68 | 0.09 | 1.68 |
| 2.0 | 0.73 | 2.68 | 0.10 | 1.66 |
| 3.0 | 0.72 | 2.68 | 0.10 | 1.64 |
| 4.0 | 0.70 | 2.67 | 0.10 | 1.59 |
| 5.0 | 0.63 | 2.57 | 0.12 | 1.34 |
| 6.0 | 0.57 | 2.39 | 0.12 | 1.03 |
| 7.0 | 0.50 | 2.04 | 0.14 | 0.53 |
| 8.0 | 0.41 | 1.07 | 0.17 | -0.73 |
| 9.0 | 0.40 | -0.15 | 0.20 | -2.20 |

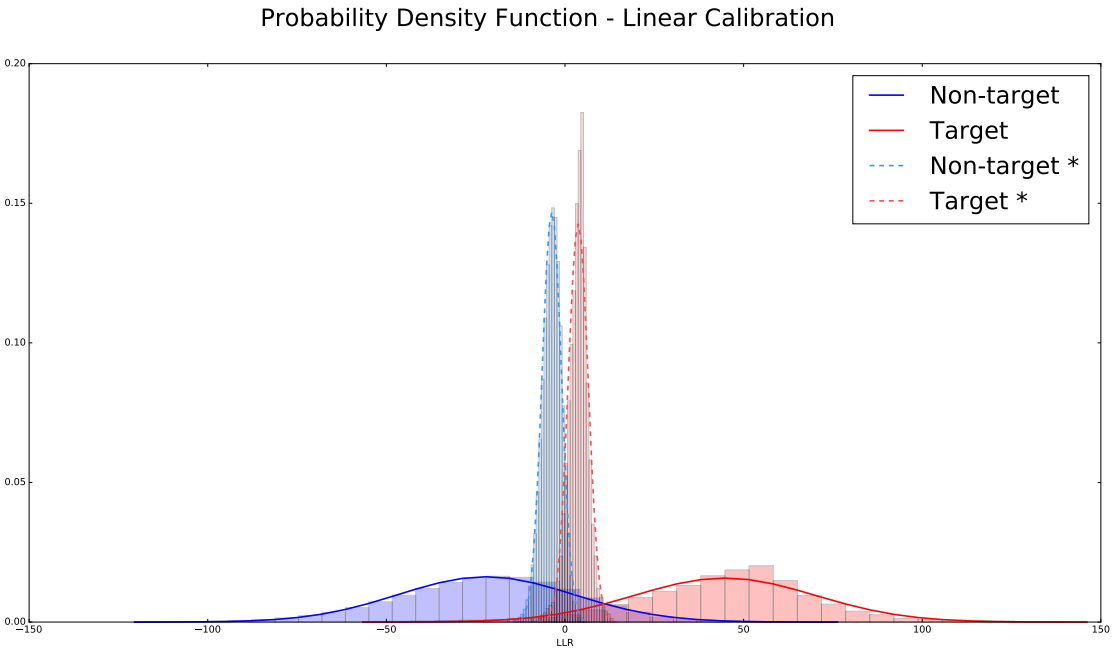


Figure 6: *Distribution of scores not using any score calibration compared to ideal distribution using linear calibration for short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

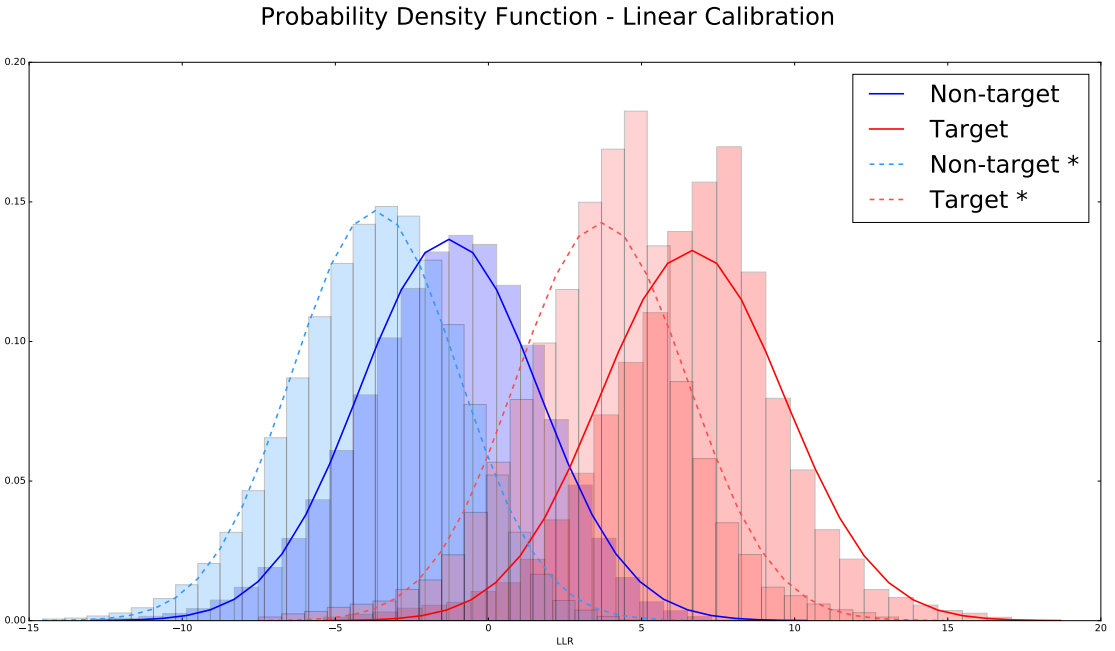


Figure 7: *Distribution of scores after calibration using cuts of the recording to halves compared to ideal distribution using linear calibration for short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

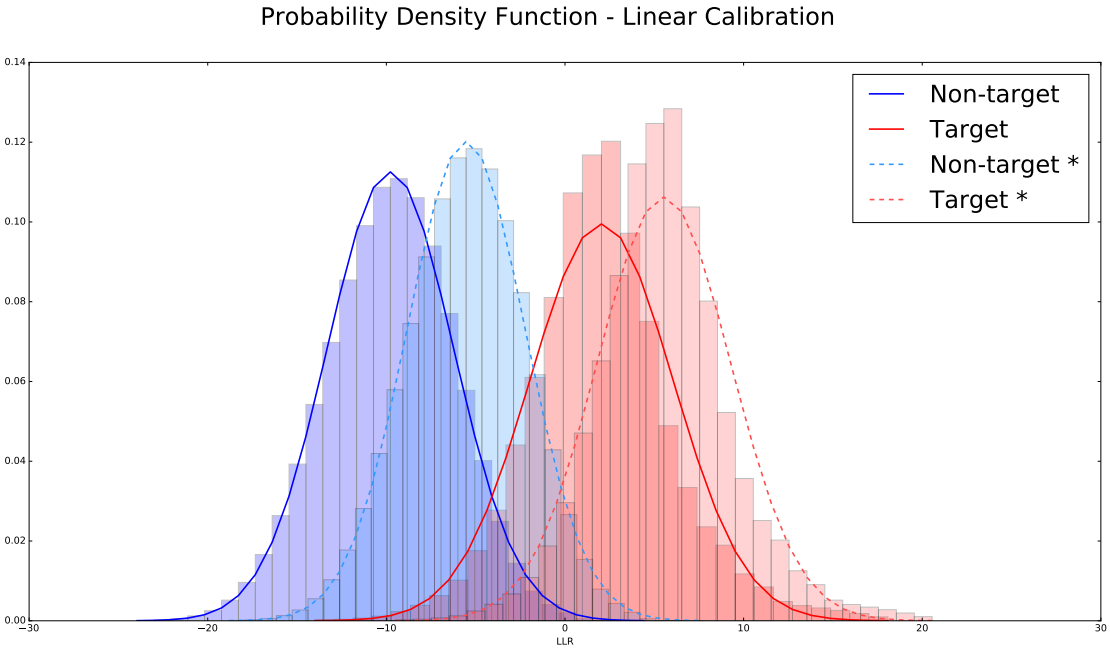


Figure 8: *Distribution of scores after the whole unsupervised calibration process compared to ideal distribution using linear calibration for short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

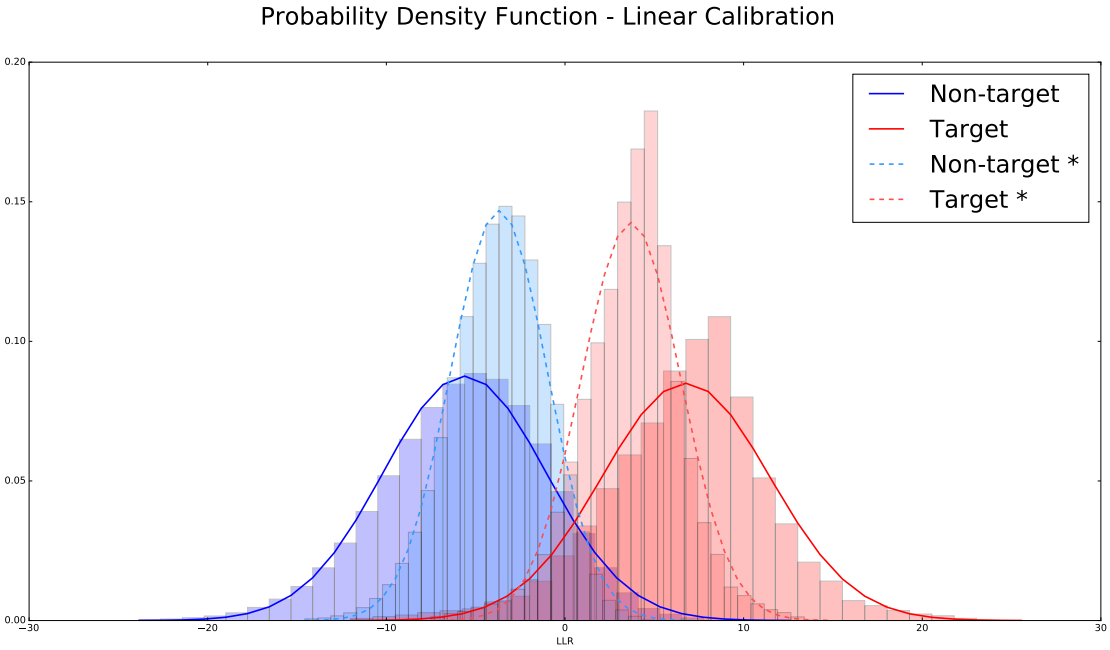


Figure 9: *Distribution of scores after the whole unsupervised calibration with tuning process compared to ideal distribution using linear calibration for short recordings from 15 second to 40 seconds of clean speech in uniform distribution.*

# Bibliography