INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Feature Normalisation Technique for PLLR based Language Identification Systems

Sarith Fernando [1, 2], *Vidhyasaharan Sethu* [1], *Eliathamby Ambikairajah* [1, 2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia

sarith.fernando@student.unsw.edu.au

Abstract

Phone log-likelihood ratio (PLLR) features have been shown to be effective in language identification systems. However, PLLR feature distributions are bounded and this may contradict assumptions of Gaussianity and consequently lead to reduced language recognition rates. In this paper, we propose a feature normalisation technique for the PLLR feature space and demonstrate that it can outperform conventional normalisation and decorrelation techniques such as mean-variance normalisation, feature warping, discrete cosine transform and principal component analysis. Experimental results on the NIST LRE 2007 and the NIST LRE 2015 databases show that the proposed method outperforms other normalisation methods by at least 9.3% in terms of %Cavg. Finally, unlike PCA which needs to be estimated from all the training data, the proposed technique can be applied on each utterance independently.

Index Terms: Spoken Language Recognition, Phone Log-Likelihood Ratios, Feature Transformation, Statistical Normalisation, Gaussian PLDA, i-Vectors.

1. Introduction

Phonotactic features have been a mainstay of automatic Language Identification (LID) systems [1, 2], particularly in combination with acoustic features [3-5]. However, these two approaches have significant differences in the feature extraction processes. Phone-based systems typically employ language models and phone decoders that capture phonotactic information pertaining to sequences or lattices of phonetic tokens such as phonemes. On the other hand, acoustic features such as Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) take a frame based approach resulting in a feature vector corresponding to each frame.

An alternative to the classical phonotactic features are the Phone Log-Likelihood Ratio (PLLR) [6] features which provide a frame level representation of phonetic information. PLLRs are computed from phone posterior probabilities and have shown to be effective in both Language Recognition and Speaker Recognition systems [6, 7].

In this paper, we propose a normalisation technique to mitigate the effects of boundedness and subsequent non Guassian nature of PLLRs as well as to decorrelate the different feature dimensions. The proposed method is compared to standard normalisation and decorrelation techniques such as feature warping, mean-variance normalisation, DCT and PCA and shown to be better suited for language identification systems.

2. PLLR feature space

In [7], it was shown that frame level features that directly represent phone posteriors or phone log-posteriors will fail to perform well due to the highly non-Gaussian nature of these features. In contrast, the distributions of PLLR features are nearly Gaussian along each dimension. PLLR features can be computed from phone posteriors generated by a suitable phone decoder [6] as,

$$LLR_i(t) = log\frac{p_i(t)}{\frac{1}{N-1}\left(1-p_i(t)\right)} \quad ,i = 1, \dots, N \qquad (1)$$

where, $LLR_i(t)$ denotes the log-likelihood ratio of the i^{th} phoneme and $p_i(t)$ denotes the posterior probability of the i^{th} phoneme corresponding to frame t. The N dimensional vectors corresponding to each frame comprising of the log-likelihood ratios for all N phonemes are then referred to as PLLR feature vectors (or PLLRs).

As previously mentioned, the transformation from phone posteriors to PLLRs incorporates an element of Gaussianisation which in turn makes PLLRs more suitable as features for a LID system compared to phone posteriors. However, phone posteriors are highly non-Gaussian [7] and the Gaussianisation implicit in computing PLLRs is not perfect. Since phone posteriors must sum to unity, they are linearly dependent and are constrained to a $N-1$ dimensional subspace. Subsequently, these constrains result in PLLRs being bounded as can be seen from Figure 1 which shows a clear boundary (red dashed line) when visualising 3 dimensions of PLLRs (corresponding to the phonemes *a:, E,* and *O* [8]). This boundedness of PLLRs suggests that assumptions of Gaussianity about PLLR features that are commonly made in LID systems may not be valid and explicit normalisation of PLLRs may be beneficial.

Feature normalisation techniques, which are used to reduce the mismatch between training and test datasets in most LID systems, may also incorporate an additional element of Gaussianisation. However, as described in [6], well-known normalisation techniques, such as feature normalisation [9] and feature warping [10] degrade the performance of PLLR features. In particular, feature warping forces the shape of the distribution to be Gaussian. However, since the original PLLR distributions are bounded, this tends to create a distribution that is not smooth (Figure 2b) which in turn may increase the boundedness of the feature space as shown in Figure 2a and

http://dx.doi.org/10.21437/Interspeech.2016-560

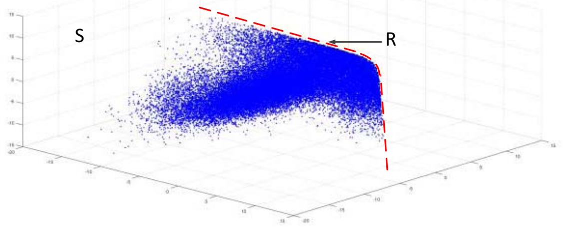

Figure 1: *Distribution of PLLRs for the set of three phones (a:,E,O), computed with the BUT decoder*

consequently degrade performance. Diez et. al. proposed a technique that projects features onto a hyperplane in order to overcome the bounding effect [8]. Our preliminary experiments suggest that the distributions of the projected features [10] are approximately Gaussian.

Diez et. al. have also suggested the use of PCA to decorrelate feature dimensions which may also reduce the bounding effect on original PLLR feature space [8]. The high level of correlation between the feature dimensions of PLLRs can also make the joint distribution harder to characterise and benefits can be expected from decorrelating them.

3. Proposed PLLR Normalisation

We propose a transformation that forces all dimensions of PLLR features to be decorrelated. Further, the proposed transform scales all dimensions to have unit variance. This transformation is applied to utterances as:

$$Y = VD^{-\frac{1}{2}}V^{T}X \tag{2}$$

where, X is an $N \times K$ matrix of N dimensional mean centered PLLR features extracted from an utterance, K is number of frames in that utterance and Y denotes the transformed PLLR feature vectors. The transformation parameters V and D are obtained by Eigen decomposition of the feature covariance matrix C as given by:

$$C = VDV^{-1} \tag{3}$$

where V is the matrix of eigenvectors, D is the diagonal matrix of eigenvalues and C is given by:

$$C = E[X\,X^{T}] \tag{4}$$

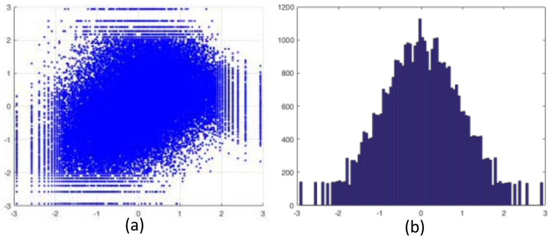

Figure 2*: (a) Distribution of PLLRs followed by feature warping for the set of two phones (a:,E), computed with the BUT decoder and (b) histogram representation of phone (a:)*

It should be noted that if $D = I$ the transformation leaves the data untouched and in general the scaling by $D^{-1/2}$ can be thought of as compressing and stretching data along the directions given by eigenvectors to normalise the variance along these directions. The proposed transform can be considered to comprise of 3 stages, namely, a rotation of the data to the eigenspace, a scaling based on eigenvalues and a rotation back to the original feature space. In Figure 3, these three stages are visualised for the first 2 dimensions of the PLLR feature space corresponding to the phones *'a:'* and *'E'*. Comparing Figure 3a and 3d, it is clear that the sharp boundary in the feature space is smoothed out by the transformation.

Table 1 compares the differences between the proposed transformation and a number of other standard normalisation techniques. It should be noted that unlike PCA, the proposed normalisation matrix is estimated from each utterance independently and there is no requirement to aggregate all training data. The proposed transform normalises the data along the eigenvectors but maps the features back to the original space, while PCA maps the features to a different space in order to decorrelate them. It is expected that PLLRs transformed as per the proposed method will more closely follow a Gaussian distribution compared with the untransformed PLLRs and consequently be more suitable for modelling by diagonal covariance Gaussian mixture models (GMMs) that are used in i-vector based systems.

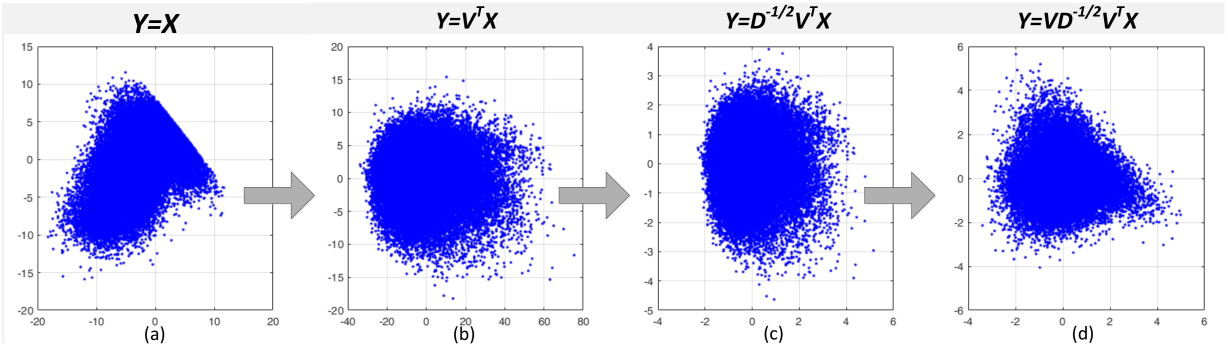

Figure 3: *Distribution of PLLRs for the set of two phones (a:,E), computed with the BUT decoder in the process of Transformation*

Table 1. *Distribution characteristics of different feature transformations applied on PLLRs*

Distribution/ PLLR Systems	Decorrelated	Un-boundedness	Gaussian distributions	Normalized Mean	Normalized Variance
Mean-Variance Normalisation	X	X	X	√	√
Feature Warping	X	X	X	√	√
DCT	√	X	X	X	X
PCA	√	√	√	√	X
Proposed Statistical Transformation	√	√	√	√	√

4. Experimental setup

4.1. Language Identification System

Figure 4 shows the block diagram of the LID system that was used to evaluate the proposed normalisation technique. The overall system follows the well-established total variability factor analysis (i-vector) paradigm [11]. Following frame-based PLLR feature extraction and feature normalisation, i-vectors representing each utterance were estimated based on a Gaussian mixture model (GMM) of the distribution of features adapted from a universal background model (UBM). Length normalisation and linear discriminant analysis (LDA) were then carried out on the i-vectors to further reduce dimensionality and compensate for inter-session variability. Finally, Gaussian PLDA parameters [12] were estimated for target languages and used to compute the scores. In order to compare the proposed normalisation technique to other commonly employed methods of normalisation, the performances of different variations of this system, with each one employing a different normalisation technique were estimated and compared (section 5).

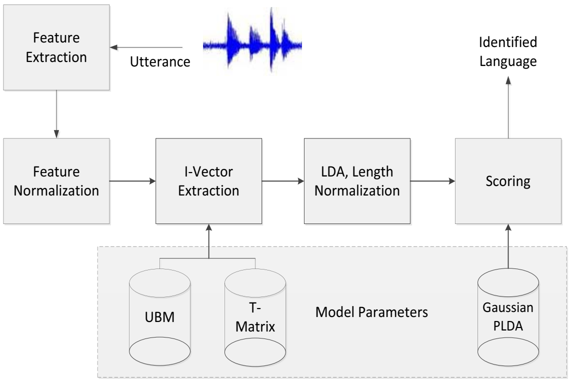

Figure 4: *Block diagram of overall experiment setup*

4.1.1. PLLR feature Extraction

In the front-end, phone posteriors were estimated using the Temporal Patterns Neural Network (TRAPs/NN) phone decoder for Hungarian language, developed by the Brno University of Technology (BUT) [13]. After summing the probabilities of all states in each phonetic unit, 58 phonetic and 3 non-phonetic units were retained. Further, the non-phonetic units were summed and treated as one single model unit to determine the voicing activity.

To determine non speech frames, a PLLR based VAD was used as in [6]. Finally, PLLRs are calculated as per equation (1) and feature vectors where the highest PLLR value corresponded to this integrated non-phonetic unit were removed since these frames correspond to non-speech segments.

4.1.2. I-Vector Gaussian PLDA

Language independent 1024-mixture GMMs with diagonal covariance was used as UBM and trained on half of the training data from all target languages. The i-vector dimensionality was chosen as 400 based on previous experiments conducted on NIST 2007 LRE data [12]. The total variability matrices were estimated as in [14] using data from the target languages only.

4.2. Test data and Evaluation Measures

4.2.1. NIST 2007 LRE

The primary experiments comparing the proposed normalisation technique to other established approaches were conducted on the NIST 2007 LRE [15] which defined a spoken language recognition task for conversational speech across telephone channels, involving 14 target languages. All the training and testing were limited to these 14 target languages. For development purposes, 10 conversations from each language were randomly chosen. The results obtained were computed on 30 second test segments for the closed-set condition, as is standard for NIST 2007 LRE.

4.2.2. NIST 2015 LRE

Following the initial comparison on NIST 2007 LRE data, the performance of the proposed normalisation method is also validated on the NIST 2015 LRE data set as per the evaluation plan [16] involving 20 target languages. The data set contains both conversational telephone speech data and broadcast narrowband speech data. The results obtained were computed on all test segments limited to segments with approximately between 3s, and 30s speech duration. Ten conversations from each language were chosen at random to constitute the development set.

The system performance on both the NIST 2007 and NIST 2015 LRE datasets were compared in terms of average cost performance C_{avg} and log-likelihood ratio cost function C_{llr} as defined in the NIST 2007 and NIST 2015 evaluation plans respectively.

5. Results

5.1. Comparison of Normalisation Techniques

As previously mentioned, the proposed normalisation method was compared to standard normalisation techniques in the context of the language identification system described in section 3 and the results are given in Table 2. The results show that DCT provides a small improvement of (10.1%) and PCA provides an improvement of (22.2%). The proposed transformation outperformed both of these classical decorrelation methods by at least 9.3% in terms of %Cavg for NIST 2007 LRE. Furthermore, the proposed approach also outperforms the combination of PCA and feature warping which brings together decorrelation, mean and variance normalisation and Gaussianisation (Table 1). The overall relative improvement obtained by the proposed transformation (based on comparison with untransformed PLLR features) is 29.4%.

Table 2. *%C_{avg} and C_{llr} performance for the Baseline, applied parameterization, decorrelation and transformed methods for NIST 2007 LRE*

System	%C_{avg} / C_{llr}
PLLR	3.88/ 0.244
PLLR + SDC	3.16/ 0.208
Mean-Variance Normalisation (MVN)	3.78/ 0.241
Feature Warping (FW)	4.25/ 0.260
DCT	3.49/ 0.257
PCA	3.02/ 0.223
PCA + Feature Warping	2.91/ 0.217
Transformed PLLR	2.74/ 0.211
Transformed PLLR + SDC	**1.89/ 0.130**

Finally, transformed PLLRs were concatenated with Shifted Delta Coefficients (SDC) with a 59-1-5-1 configuration and this led to a further increase in performance. The overall relative improvement attained by the combination of the proposed normalisation and SDC is 51.3% in terms of %Cavg compared to un-normalised PLLR features and 40.2% improvement compared to un-normalised PLLR + SDC features.

5.2. Performance on NIST 2015 LRE dataset

The proposed technique was also validated on the NIST 2015 LRE data set and the results are presented in Table 3. Unlike NIST 2007 LRE, the 20 languages in NIST 2015 were grouped into 6 language clusters with a focus on distinguishing languages within each cluster. The results presented in Table 3 describe system performance according to each cluster and finally the overall performance.

The results show that the proposed method outperformed the un-normalised PLLRs by 7.1%. Further, enhancing the transformed PLLRs by concatenating SDCs led to a relative improvement of 9.8% in terms of %Cavg.

Table 3. *%C_{avg} for the Original PLLR, Original PLLR+SDC, Transformed PLLR and Transformed PLLR+SDC features for NIST 2015 LRE*

System/ Cluster	%C_{avg}			
	Original PLLR	Original PLLR + SDC	Transformed PLLR	Tansformed PLLR + SDC
Arabic	29.2	27.7	28.2	**28.0**
English	22.6	18.0	17.3	**17.2**
French	42.5	43.3	41.5	**41.0**
Slavic	9.33	8.47	9.09	**8.56**
Iberian	26.2	26.8	23.7	**23.3**
Chinese	22.4	20.9	21.9	**19.6**
Average	25.4	24.2	23.6	**22.9**

6. Conclusions

In this paper, a normalisation method for PLLR features used in language identification has been proposed, with the explicit aim of reducing the effect of boundedness and non-Gaussianity of the PLLR feature space on the language identification. Experimental results included in the paper suggest that the proposed method outperforms standard approaches to normalisation and decorrelation. In addition, the proposed transform can be estimated and applied on each utterance individually unlike PCA (the standard method that comes closest in terms of performance) which must be estimated on all the training data. Finally, experimental results also show that the use of shifted delta coefficients (which are commonly employed in language identification systems to incorporate longer term temporal information) along with the transformed PLLRs obtained via the proposed method leads to further improvement in performance. A limitation of this approach is that like the other transformations, it is still a linear transformation, which means that the boundedness is not fully removed.

7. References

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: a tutorial," *Circuits and Systems Magazine, IEEE,* vol. 11, pp. 82-108, 2011.

[2] L. F. D'haro Enríquez, O. Glembek, O. Plchot, P. Matějka, M. Soufifar, R. d. Córdoba Herralde*, et al.*, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," 2012.

[3] E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds, A. McCree, F. Richardson, N. Dehak*, et al.*, "The MITLL NIST LRE 2011 language recognition system," in *Odyssey*, 2012, pp. 209-215.

[4] N. Brümmer, S. Cumani, O. Glembek, M. Karafiát, and P. Matejka, "Description and analysis of the Brno276 system for LRE2011," *sign (ℓ i-ℓ j),* vol. 1000, p. 24, 2012.

[5] L. J. Rodrıguez-Fuentes, M. Penagarikano, A. Varona, M. Dıez, G. Bordel, A. Abad*, et al.*, "The BLZ Systems for the 2011 NIST Language Recognition Evaluation," 2012.

[6] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 274-279.

[7] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "Using phone log-likelihood ratios as features for speaker recognition," *NIST 2012 Speaker Recognition Evaluation (SRE),* vol. 3, p. 15, 2013.

[8] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the projection of PLLRs for unbounded feature distributions in spoken language recognition," *Signal Processing Letters, IEEE,* vol. 21, pp. 1073-1077, 2014.

[9] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proceedings of the workshop on Human Language Technology*, 1993, pp. 69-74.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[11] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, 2011, pp. 857-860.

[12] S. Irtza, V. Sethu, P. N. Le, E. Ambikairajah, and H. Li, "Phonemes Frequency Based PLLR Dimensionality Reduction for Language Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/ ,Brno, Czech Republic, 2008.

[14] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy,* pp. 861-864, 2011.

[15] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Odyssey - The Speaker and Language Recognition Workshop*, 2008.

[16] "The 2015 NIST Language Recognition Evaluation Plan (LRE15)," 2015.