Eigenfeatures: An alternative to Shifted Delta Coefficients for Language Identification

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@unsw.edu.au

Abstract

Almost all of the current LID systems use Shifted Delta Coefficients (SDC) as the temporal information features, in addition to spectral based features. However, using normalisation techniques on SDC features to make them more robust to noise or channel effects also tends to distort some of the language specific information. In this paper, we propose Eigenfeatures (EF) as an alternative to SDCs to capture temporal information while being more robust to any distortion caused by normalisation. Experimental results based on NIST LRE 2015 database shows that the Eigenfeatures are a better alternative to SDCs in the context of language recognition.

Index Terms: Shifted Delta Coefficients, Eigenfeatures, Post Log-Likelihood Ratios, Deep Bottleneck Features

1. Introduction

Feature extraction is one of the most important steps in any Language identification system [1]. Typically frame based features such as Mel Frequency Cepstral Coefficients (MFCC) [2] and Perceptual linear predictive coefficients (PLP) [3] are used and more recently, Phone Log likelihood Ratio (PLLR) [4] and deep neural network based Bottleneck Features (BNF) [5] have also been identified as effective features for language identification.

In order to capture longer term temporal context, Shifted Delta Coefficients (SDC) [6] concatenated with acoustic and phonotactic features [7, 8] are commonly employed. These coefficients have been shown to be better than just using the acoustic and phonotactic features on their own or in combination with delta and delta-delta features for language identification systems [1]. Recently, a new ‘local variability feature’ [9] has gained interest in speaker recognition tasks and also in image processing applications [10] and have been used to capture short term temporal information or regional variations. These local variability features have been shown to be comparable to delta and delta-delta coefficients in speaker verification tasks [9].

In this paper, we propose the use of Eigenfeatures to capture short term temporal information in place of SDCs in conjunction with the recently introduced PLLR and BNF features. Experiments carried out on NIST 2007 and NIST 2015 datasets demonstrate that the use of Eigenfeatures is a better alternative to shifted delta coefficients for language identification system.

Generally, in language identification, there is a need for features that represent language specific information which are robust to noise and channel effects. The most common way of achieving this is through suitable feature normalisation techniques [11, 12]. However, the use of normalisation techniques generally tends to make the covariances of all features with utterances of all languages to be more similar to each other [9]. This, in turn, may lead to a loss of some language specific information. Using experimental results, we show that the Eigenfeatures, which are calculated from short term covariance matrices, are less susceptible to normalisation.

2. Proposed Eigenfeature extraction

Eigenfeatures are estimated from frame based features within overlapping windows of multiple frames as shown in Figure 1. The Eigenfeatures are characterised by three main parameters, N-P-K. Where, N represents the dimensionality of the underlying frame based feature vectors, P is the number of frames contained in the window over which the Eigenfeatures are estimated and K is the number of Eigenvectors that are concatenated to form the Eigenfeature vector.

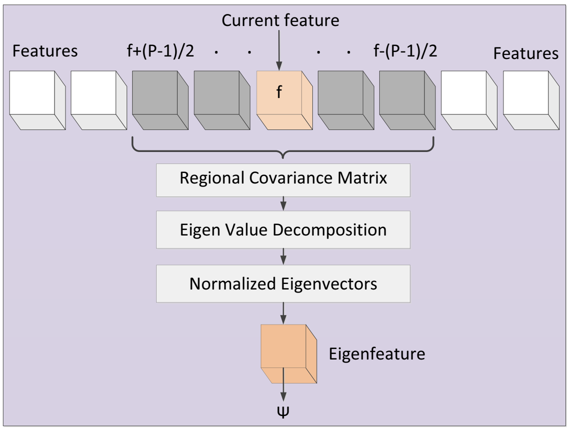

Figure 1: *Plot of Eigenfeature extraction processing starting from root features.*

Let $[X]_{P\times N}$ denote, $N-$ dimensional feature vectors corresponding to P frames in a window R. The regional covariance matrix, C_R, of the features corresponding to this window R can then be computed as,

$$C_R = \frac{1}{P-1}\sum_{t=1}^{P}(x_t - \mu)(x_t - \mu)^T \qquad (1)$$

where, x_t denotes the feature vector corresponding to the t^{th} frame and μ is the mean over the P frames.

Then Eigenvalue decomposition was then performed on this regional covariance C_R as,

$$C_R = VSV^{-1} \qquad (2)$$

where V is the $N \times N$ Eigenvector matrix and S is the $N \times N$ diagonal matrix which consists of the corresponding

Eigenvalues. Finally, each Eigenvector $\boldsymbol{v'}_i$ is normalised as follows,

$$\boldsymbol{v}_i = \frac{\boldsymbol{v'}_i \cdot s_i}{\sum_{j=1}^{N} s_j} \quad (3)$$

where $\boldsymbol{v_i}$ is the normalised Eigenvector. Finally, the $NK-$ dimensional Eigenfeature vector, $\boldsymbol{\Psi}$, is formed by concatenating the eigenvectors correspond to K largest eigenvalues. In our experiments we have used N-5-1 configuration where N was 59 and 42 for Phone Log Likelihood Ratios (PLLR) and Bottleneck Features (BNF) respectively.

3. System Description

A block diagram of the LID system used in our experiments reported in this paper is shown in Figure 2. This system is used to compare the Eigenfeatures (section 2) as an alternative to shifted delta coefficients [1] . The rest of the system is a typical language identification system comprising of an i-vector-GPLDA back-end with either Phone Log-Likelihood Ratio or Bottleneck features employed in the front-end. In addition, length normalisation and LDA were applied on the i-vectors prior to Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) scoring [13]. It should be noted that the i-vectors corresponding to each front-end (PLLR or BNF) are estimated using a UBM and a T-matrix specific to that front-end.

3.1. PLLR feature Extraction

The Phone Log-Likelihood Ratio (PLLR) feature vectors [4] from each frame were computed using phone posteriors estimated from speech utterances using a Hungarian phone decoder developed by the Brno University of Technology [14]. Following the estimation of phone posteriors, a Voice Activity Detector (VAD) [4] was used to remove non-speech frames from each utterance. Finally, the state posteriors corresponding to each phoneme were summed together and the 3 non-phonetic units were combined into one single unit which led to 59 dimensional PLLR values. PLLR features were computed as,

$$LLR_i(t) = log \frac{p_i(t)}{\frac{1}{N-1}(1-p_i(t))} \quad , i = 1, \ldots, N \quad (4)$$

where, $LLR_i(t)$ represents the log-likelihood ratio of the i^{th} phoneme and $p_i(t)$ represents the posterior probability of the i^{th} phoneme corresponding to frame t. The N dimensional vectors (59 dimension in our case) corresponding to each frame were then referred to as PLLR feature vectors (or PLLRs).

3.2. BNF feature extraction

The overall Deep Neural Network (DNN) architecture of Bottleneck feature extraction process is shown in Figure 3. BNF were extracted using a Deep Neural Network [5] trained on MFCC features. The Network has trained on 300 hours of Switchboard 1 data as defined in Kaldi example “tri4a” [15]. DNN consists of 5 layers each with 1024 nodes except at the bottleneck layer at layer four. All of these layers used ‘tanh’ activation function with the exception of the bottleneck layer. The bottleneck layer comprised of 42 nodes using a linear activation function. After extracting Bottleneck Features, Vector Quantization Voice Activity Detection (VQ-VAD) was used.

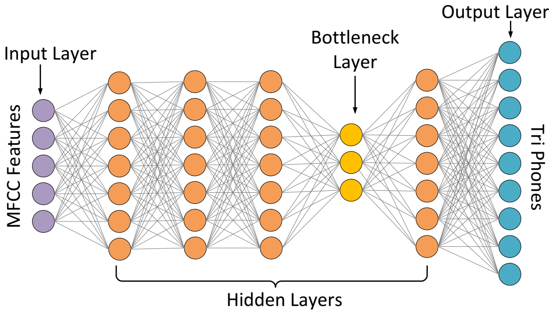

Figure 3: *Block diagram of DNN architecture for BNF extraction process*

3.3. I-Vector Gaussian PLDA

The Total Variability matrix was trained as in [16]. This is based on the language-independent Universal Background model (UBM) consisting of 1024 Gaussian components trained on 118 dimensional PLLR coefficients or on 84 dimensional BNF concatenated with either SDCs or EF separately. Only half of the data available from target languages was used for UBM training and Total Variability matrix was trained on all language data. The i-vector dimension was selected as 400 according to the previous experiments conduct on NIST 2007 dataset [13] and computed

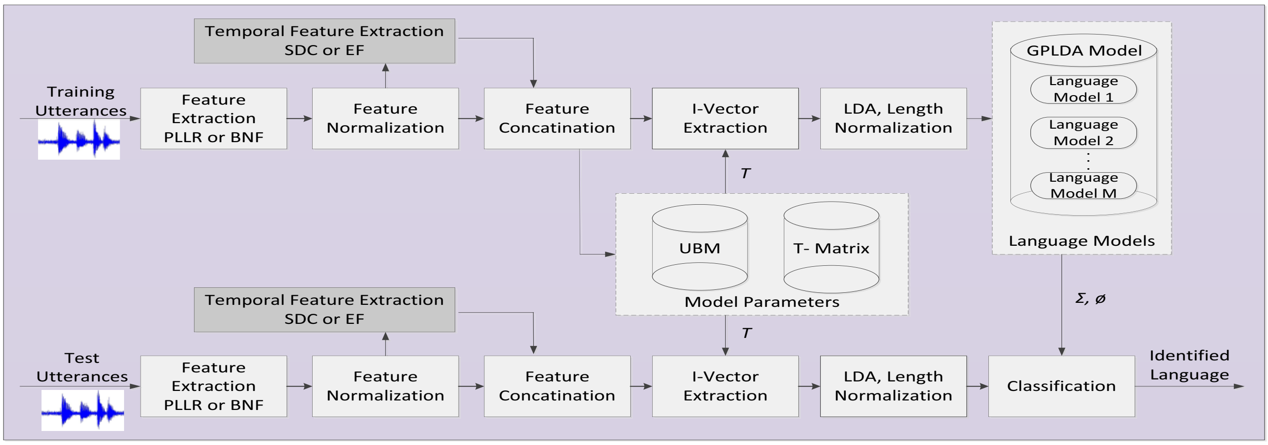

Figure 2: *Block diagram of experimental setup*

as,

$$\boldsymbol{M} = \boldsymbol{m} + T\boldsymbol{w} \quad (5)$$

where $\boldsymbol{M}$ is the utterance dependent GMM mean supervector, $\boldsymbol{m}$ is the language independent mean super vector, T is the total variability matrix and $\boldsymbol{w}$ is the low dimension i-vector.

In this study, a Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) was used as a classifier to make the final decision based on the i-vectors. This can be denoted by "Language models" block in Figure 2. The GPLDA back-end has shown to be effective, in LID tasks [13]. Typically, GPLDA based systems use length normalization of i-vectors to overcome their non-Gaussianity [14]. In the GPLDA approach, the i-vectors are represented by a generative model given as log likelihood ratio between the same $H1$ versus different $H0$ language model hypothesis,

$$s(u, v) = log\frac{p(u, v|H1)}{p(u|H0).p(v|H0)} \quad (6)$$

where u and v are the training and test i-vectors respectively. Finally score level averaging was conducted for each language,

$$w(l, v) = \frac{\sum_{\forall n} s(u, v)}{n} \quad (7)$$

where n is the number of utterances in each language and l is the target language.

3.4. NIST 2007 and 2015 LRE Datasets

The NIST 2007 LRE is based on 14 target languages involving conversational speech across speech channels [17]. We used all 14 target languages for training and testing purposes. 10 conversations from each language were randomly selected for development purposes. In keeping with the structure of NIST 2007, our studies were carried out on 30 seconds speech segments for the closed-set condition.

The NIST 2015 LRE features 20 target languages, subdivided into 6 main clusters as shown in Table 1 involving both conversational telephone speech data and broadcast narrowband speech data recorded in various conditions. The language models should only train on these limited and specified training data according to the evaluation plan [18].

Table 1. *NIST 2015 Language clusters with target languages*

Cluster	Target Languages
Arabic	Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
Chinese	Cantonese, Mandarin, Min, Wu
English	British, General American, Indian
French	West African, Haitian Creole
Slavic	Russian, Polish
Iberian	Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese

There are 7020 utterances for target language modelling and as before, we selected 10 random conversation speech segments from each language for development purposes. Unlike NIST 2007, the results obtained were computed over all test segments, having durations of between 3s and 30s according to their primary task.

4. Results

In this section, experiments are carried out on NIST 2007 and NIST 2015 datasets to demonstrate that the use of Eigenfeatures is a better alternative to the shifted delta coefficients in a language identification system. In order to validate the class separability on training data, we calculated the *J*-measure for both NIST 2007 and NIST 2015 LRE data sets. The *J*-measure is the ratio between inter-class scatter to intra-class scatter as shown below,

$$J = trace\ (S_W^{-1}S_B) \quad (8)$$

where S_W and S_B are the within-class and between class scatter matrix respectively. The larger the value of *J*-measure, the better the discrimination of the classes in the feature space. The results in Table 2 show a better class separability using Eigenfeature than shifted delta coefficient features that leads to a better performance.

Table 2. *J-measure for PLLR and BNF features concatenated with SDC/ EF features*

Features	*J- measure*	
	NIST 2007	NIST 2015
PLLR	10.17	12.03
PLLR_SDC	10.39	12.17
PLLR_EF	**10.45**	**12.25**
BNF	10.76	12.86
BNF_SDC	10.96	13.08
BNF_EF	**10.99**	**13.11**

In our work, all LID systems were compared in terms of average cost performance, (Cavg) and Log likelihood ratio function, (Cllr) provided by the NIST 2007 and NIST 2015 LRE evaluation plans respectively. The results of the experiments using NIST 2007 LRE data set are summarized in Table 3 with the first three rows corresponding to PLLR features and last three corresponding to BNF features. It can be clearly seen that in both cases Eigenfeatures outperformed SDC features having relative improvement of 10.8% in PLLR and 42.3% in BNF.

Table 3. *%Cavg and Cllr performance for the standard PLLR and BNF features compared with concatenated SDC and EF features for NIST 2007 LRE*

System	%C_{avg} (C_{llr})
PLLR	3.88 (0.244)
PLLR_SDC	3.16 (0.208)
PLLR_EF	**2.82 (0.189)**
BNF	1.85 (0.142)
BNF_SDC	1.63 (0.132)
BNF_EF	**0.94 (0.0818)**

To validate the consistency of Eigenfeatures in different databases the above techniques were applied to the NIST 2015 LRE data set. The results presented in Table 4 show a relative improvement of 3.3% on PLLR and 3.9% on BNF in terms of %Cavg using EF compared to SDC features.

Table 4. *%Cavg and Cllr performance for the standard PLLR and BNF features compared with concatenated SDC and EF features for NIST 2015 LRE*

System/ Cluster	%C_{avg} (C_{llr})					
	PLLR	PLLR_SDC	PLLR_EF	BNF	BNF_SDC	BNF_EF
Arabic	29.2 (2.04)	27.7 (1.93)	**27.5 (2.38)**	27.3 (2.30)	26.9 (2.71)	**25.9 (2.34)**
English	22.6 (1.59)	18.0 (1.11)	**17.5 (1.28)**	14.5 (0.97)	12.6 (0.84)	**13.9 (0.89)**
French	42.5 (5.20)	43.3 (5.81)	**42.0 (5.89)**	40.6 (5.25)	41.9 (5.88)	**39.0 (6.37)**
Slavic	9.33 (0.434)	8.47 (0.40)	**8.37 (0.49)**	6.24 (0.36)	6.09 (0.33)	**5.61 (0.32)**
Iberian	26.2 (2.19)	26.8 (1.90)	**23.9 (2.38)**	21.5 (1.87)	21.6 (2.08)	**20.9 (2.22)**
Chinese	22.4 (1.94)	20.9 (1.63)	**20.9 (1.92)**	16.7 (1.47)	15.0 (1.34)	**14.3 (1.35)**
AVERAGE	25.4 (2.23)	24.2 (2.13)	**23.4 (2.39)**	21.1 (2.04)	20.7 (2.19)	**19.9 (2.25)**

In order to improve the system performance further, score fusion was conducted. Fusion parameters were estimated and applied to the development data sets using the Focal toolkit [19]. Table 5 shows the results for fusion of different combinations of systems. The results suggest that the fusion of PLLR_EF system with BNF_EF system outperforms any other combinations for the 2007 and 2015 LRE data sets used in our experiments.

Table 5. *Fusion of LID systems and its performances*

Dataset	Fusion of LID Systems	%C_{avg} (C_{llr})
2007 LRE	(PLLR)+(BNF)	1.03 (0.0377)
	(PLLR_SDC)+(BNF_SDC)	0.55 (0.0213)
	(PLLR_EF)+(BNF_EF)	0.45 (0.018)
2015 LRE	(PLLR)+(BNF)	20.2 (0.529)
	(PLLR_SDC)+(BNF_SDC)	19.5 (0.502)
	(PLLR_EF)+(BNF_EF)	18.9 (0.499)

5. Conclusions

In this paper, we proposed a regional covariance based Eigenfeatures extraction which can capture the short time varying information for language identification tasks. As discussed above, the improvement of results implies that Eigenfeatures contain certain scale and direction invariance over the C_R regions in each language. Furthermore, this directional invariance is increased by choosing Eigenvectors with the highest Eigenvalues.

In contrast to SDC calculation, these Eigenfeatures capture both temporal information as well as language specific information and can be applied to both PLLRs and bottleneck features which are the state-of-the-art phonotactic and acoustic features respectively. The results included in this paper illustrates specifically that concatenating standard features with Eigenfeatures will be more beneficial than concatenating them with SDC features in the front-end of language identification systems.

6. 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] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, pp. 357-366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America,* vol. 87, pp. 1738-1752, 1990.

[4] 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.

[5] F. Richardson, D. Reynolds, and N. Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *INTERSPEECH*, 2002.

[7] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 2005, pp. 1-4.

[8] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "New insight into the use of phone log-likelihood ratios as features for language recognition," in *INTERSPEECH*, 2014, pp. 1841-1845.

[9] M. Sahidullah and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing,* vol. 50, pp. 1-11, 2016.

[10] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Computer Vision–ECCV 2006*, ed: Springer, 2006, pp. 589-600.

[11] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication,* vol. 25, pp. 133-147, 1998.

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

[13] 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.

[14] 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.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel*, et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[16] 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.

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

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

[19] "FoCal, Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers http://sites.google.com/site/nikobrummer/focal," 2008.