IEEE Signal Processing Letters 18.12 (2011): 705 – 708

DOI: http://dx.doi.org/10.1109/LSP.2011.2170566

Von Mises-Fisher Models in the Total Variability Subspace for Language Recognition

Ignacio Lopez-Moreno*, Daniel Ramos, Javier Gonzalez-Dominguez
and Joaquin Gonzalez-Rodriguez

Abstract

This paper proposes a new modelling approach for the Total Variability subspace within a Language Recognition task. Motivated by previous works in directional statistics, von Mises-Fisher distributions are used for assigning language-conditioned probabilities to language data, assumed to be spherically distributed in this subspace. The two proposed methods use Kernel Density Functions or Finite Mixture Models of such distributions. Experiments conducted on NIST LRE 2009 show that the proposed techniques significantly outperform the baseline cosine distance approach in most of the considered experimental conditions, including different speech conditions, durations and the presence of unseen languages.

EDICScategory: SPE-RECO, SPE-LANG

Index Terms

Finite Mixture Models, Kernel Density Function, Language Recognition, Total Variability, Von Mises-Fisher.

I. Introduction

The use of acoustic systems based on Factor Analysis (FA) has become the state of the art in language and speaker recognition due to its strength in dealing with variability, representing either nuisance or useful information [1] [2] [3]. In the initial FA approach, the latent factors that model the undesirable

The authors are with the ATVS Biometric Research Lab., Universidad Autónoma de Madrid, Madrid, Spain, contact: (see http://atvs.ii.uam.es/). Manuscript submitted May 5, 2011; This research was supported by the Ministerio de Ciencia e Innovación under FPI grant TEC2009-14719-C02-01 and cátedra UAM-Telefónica.

acoustic variability are estimated and compensated from the original utterance representation as a super-vector [2]. A type of high-dimensional vector obtained from the parameters of a Gaussian Mixture Model, which is generated using the acoustic features of an utterance (e.g. MFCC). In addition, newer successful FA approaches have shown that the so-considered useful information embedded in the super-vectors, i.e. the variability generated by different languages, can also be represented by a small number of latent variables [1]. Furthermore, they can be thought as derived front-ends that use vectors of latent variables as utterance samples.

Currently, the speaker recognition community is focused on a particular kind of FA-based front-end, known as Total Variability Subspace [1]. In this case, the vectors of latent variables obtained from the super-vectors, which are denoted by identity vectors (i-vectors), do not stand for an specific identified source of information. However all the variability presented in the data is modelled together and regardless of its origin. It may be argued that during this process, part of the useful information could be lost. However, unlike previous FA approaches, a robust estimation of the space of the i-vectors can be achieved using sufficient number of latent variables and a representative set of background data, which does not need to be class labelled beforehand. The result is that Total Variability works as a robust dimensionality reduction model. Target classes (e.g. languages) are spaced out in a further *disentangling* stage using classical discriminant techniques, such as Linear Discriminant Analysis (LDA), typically together with Within Class Covariance Normalization (WCCN) [4] to normalize the distribution of vectors of each class.

In this paper we propose a probabilistic framework for scoring utterances mapped to their own *disentangled* i-vectors. It is motivated by the good performance of the typically used cosine distance score, which overcomes other more complex approaches (e.g. SVM language models) [1]. In our work, the cosine score is understood as an approximation to a univariate directional kernel density function (KDF) that models the language i-vectors [5]. However we argue that it might be unsuitable for learning complex distributions. This is significant in language recognition, where unlike in speaker recognition, it is possible to manage hundreds of utterances for training a single language.

The proposed front-end assigns densities using von Mises-Fisher (vMF) distributions [7], which have been proved to be flexible in modelling directional data in an analogous manner as Gaussians distributions for non-directional data. Using vMF distributions, two approaches are implemented here: vMF Kernel Density Functions (vMF-KDF) and vMF Mixture Models (vMF-MM). Likelihoods generated in either case are normalized to generate scores of language membership. Moreover, this front-end can be generalized to the case of non-directional data via Spherical Normalization [8], which ensures the directionality of the data.

This paper is organized as follows. Sec. II provides a probabilistic interpretation of the typically used cosine distance and compares it with the vMF distribution. Sec. III reviews the formulation of the vMF distribution and its applicability for vMF-KDF and vMF-MM. Experiments and results are presented in Sec. IV. Conclusions and future work can be found in Sec. V.

II. Cosine Kernel Density Function

Consider a Total Variability front-end where utterances are represented by d-dimensional i-vectors [1]. In this domain, a particular language, labelled by c, is typically represented by the centroid $\mathbf{m}_c \in \mathbb{R}^d$, of the n_c available i-vectors for that language as

$$\mathbf{m}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} \mathbf{x}_{i,c}$$

where $\{\mathbf{x}_{i,c} \in \mathbb{R}^d | 1 < i < n_c\}$ is the i-th i-vector of the language c (i-vectors are hereafter referred as samples). Using the popular cosine kernel, the similarity of $\mathbf{m}_c$ with a test sample $\mathbf{x}_t \in \mathbb{R}^d$ is given by the cosine of the angle θ formed between them.

$$K_{cos}(\mathbf{x}_t, m_c) = \cos\theta = \langle \frac{\mathbf{x}_t}{\|\mathbf{x}_t\|}, \frac{\mathbf{m}_c}{\|\mathbf{m}_c\|} \rangle \tag{1}$$

$K_{cos}(\mathbf{x}, \mathbf{m}_c)$ only depends on θ and thus, it can be considered a directional kernel. Samples can be thought as directions that depart from the origin of $\mathbb{R}^d$ and they are frequently expressed in terms of unit vectors ($\mathbf{u}/\|\mathbf{u}\|$) or points on the surface of a hyper-sphere $\mathbb{S}^{d-1}$ (Fig. 1).

Lets denote by $\mathbf{x} \in \mathbb{R}^d$, such that $\|\mathbf{x}\| = 1$, a continuous random variable on the $\mathbb{S}^{d-1}$ hypersphere. Under some slight modifications $K_{cos}(\mathbf{x}, \mathbf{m}_c)$ can be thought as a directional Kernel Density Function (KDF) assigned to the distribution of $\mathbf{x}$ given each class on $\mathbb{S}^{d-1}$ (Fig. 1) [6].

$$\hat{p}(\mathbf{x}|\varphi_c) = \frac{\pi}{4} K_{cos}(\mathbf{x}, \mathbf{m}_c) \mathbf{1}_{\{K_{cos}(\mathbf{x},\mathbf{m}_c)>0\}}$$

where $\mathbf{1}_{\{u\}}$ is the characteristic function, defined as 1 for true values of u and 0 otherwise. φ refers to the model parameters, represented by the class centroid $\varphi_c \equiv \mathbf{m}_c$ for the cosine KDF. Note that $\hat{p}(\mathbf{x}|\varphi_c)$ is a valid probability density function (p.d.f) since $i)\hat{p}(\mathbf{x}|\varphi_c) > 0$ and $ii) \int \hat{p}(\mathbf{x}|\varphi_c) d\mathbf{x} = 1$ [7]. Despite its simplicity and efficiency, the cosine KDF can be non-representative for complexly distributed data. This motivates the study of p.d.f. more fitted to the i-vectors of each language.

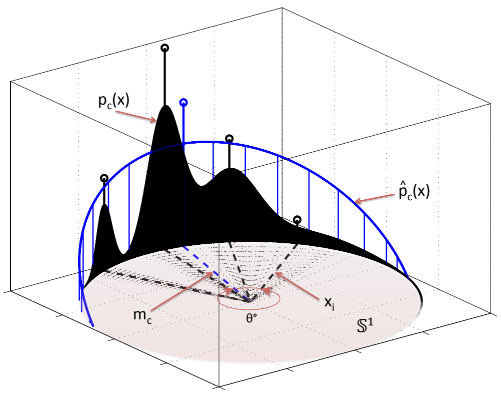

Fig. 1. Example of directional data distribution. The XY-plane shows the concentration of unit vectors $\mathbf{x}_i$ in $\mathbb{S}^1$ for the class c, together with the class centroid $\mathbf{m}_c$. The Z-plane is the multivariate distribution $p(\mathbf{x}|\varphi_c)$ of the class vectors along $\theta \in [0, 2\pi)$, together with the estimated distribution. $\hat{p}(\mathbf{x}|\varphi_c)$ using the cosine kernel density function.

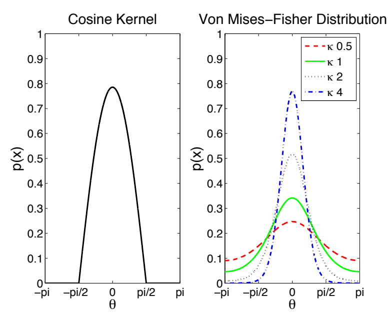

Fig. 2. Estimated value of probability $p(\mathbf{x})$ on $\mathbb{S}^1$ for (left) the cosine distribution and (right) the von Mises-Fisher distribution for different values of the concentration parameter κ.

III. VON MISES-FISHER MODELS

Directional data modelling has been previously developed in the field of directional statistics for its applications in several other fields. Some of them includes text-categorization [9], physics [10] and speaker clustering [11]. For many of these fields, a common successful alternative for the cosine KDF is the usage of von Mises-Fisher distributions (vMF). These two distributions are compared in Fig. 2. In the following subsections we present a more detailed description of the vMF-based p.d.f. assignation methods used in the proposed approaches.

A. The von Mises-Fisher distribution

The d-dimensional von Mises-Fisher distribution of a random variable $\mathbf{x} \in \mathbb{S}^{d-1}$ is given by

$$f(\mathbf{x}, \boldsymbol{\mu}, \kappa) = c_d(\kappa) e^{\kappa \boldsymbol{\mu}^T \mathbf{x}} \tag{2}$$

being $c_d(\kappa)$ a normalization constant denoted by

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$$

I_ν is defined as the modified Bessel function of the first kind and order ν [12]. The parameters $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and $\kappa \in \mathbb{R}$ are denoted respectively as the *mean direction* and the *concentration parameter*. While $\boldsymbol{\mu}$ stands for the direction that maximizes$f(\mathbf{x}, \boldsymbol{\mu}, \kappa)$, κ measures the sparseness degree of the distribution (rotationally symmetric about $\boldsymbol{\mu}$). Particularly, $f(\mathbf{x}, \boldsymbol{\mu}, \kappa)$ is a uniform distribution for $\kappa = 0$ and a point distribution for $\kappa = \infty$ (Fig. 2). Hence, the parameters $\boldsymbol{\mu}$ and κ are analogue to the mean and the inverse of the variance in a Gaussian distribution. Also, likewise the Gaussian distribution in $\mathbb{R}^d$, vMF is the only distribution in $\mathbb{S}^{d-1}$ that maximizes the entropy given the first and second order moments. This means that under some considerations [7], vMF is the closest distribution to any predetermined one on $\mathbb{S}^{d-1}$.

B. Generative Models in Directional Data

In the problem of assigning a probability density function to a random variable assumed to fit to a complex multimodal distribution, most common approaches consider using one of two methods [5]: $i)$ KDF, sum of distributions that assign local values of p.d.f. $\hat{p}(\mathbf{x}|\varphi)$ in the neighborhood of each training sample and $ii)$ finite mixture models, formed by a set of distributions whose parameters are assumed to be learnable from the data, often under a Maximum Likelihood criterion. Concerning directional statistics, the vMF distribution is typically used for both types of assignments of p.d.f. $\hat{p}(\mathbf{x}|\varphi)$ on $\mathbb{S}^{d-1}$. In both cases, the predicted likelihood can be computed as a weighted sum of a set of vMF distributions given by

$$\hat{p}(\mathbf{x}|\varphi) = \sum_{m=1}^{M} w_m f(\mathbf{x}, \boldsymbol{\mu}_m, \kappa_m) \tag{3}$$

This equation can be compactly represented by the collection of tuples $\varphi(w_m, \boldsymbol{\mu}_m, \kappa_m)_{m=1}^{M}$, which defines a vMF model. Following subsections describe the techniques used to obtain φ in each of the proposed approaches. Note that the class model φ_c characterizes the target language given by c by $\hat{p}(\mathbf{x}|\varphi_c)$. In this work, the language membership score of a testing utterance $\mathbf{x}_t$ over the class c is computed according to

$$score(c,t) = \ln \hat{p}(\mathbf{x}_t|\varphi_c) - \ln \sum_{j \neq c} \hat{p}(\mathbf{x}_t|\varphi_j)$$

being j is a positive real valued random variable for the set of J target languages $\{j \in 1 \dots J\}$.

C. The vMF Kernel Density Function

Consider a set of samples $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{S}^{d-1} | 1 \leq i \leq n\}$ and the vMF-based kernel function defined by $K_{vMF}(\mathbf{a}, \mathbf{b}) = f(\mathbf{a}, \mathbf{b}, \kappa)$. The likelihood predicted by a vMF Kernel Density Function (vMF-KDF) with respect to $\mathbf{x}$ is defined by

$$\hat{p}(\mathbf{x}|\varphi) = \sum_{i=1}^{n} w_i K_{vMF}(\mathbf{x}, \mathbf{x}_i)$$

It can be thought as a collection of n unimodal e independent vMF distributions centered at each sample $\mathbf{x}_i$. Individual likelihoods are added up including a common weight term $w_i = n^{-1}$, which normalizes the generated function to a valid p.d.f. The vMF-KDF system can be described as a non-parametric (M variable) vMF model

$$\varphi(w_m, \boldsymbol{\mu}_m, \kappa_m)_{m=1}^{M} = \varphi(n^{-1}, \mathbf{x}_i, \kappa)_{i=1}^{n}$$

D. The vMF Mixture Model

Given a training set $\mathcal{X}$, the Maximum Likelihood (ML) estimation using vMF distributions aims to obtain the vMF model $\varphi(\hat{w}_m, \hat{\boldsymbol{\mu}}_m, \hat{\kappa}_m)_{m=1}^{M}$ that maximizes the log likelihood over $\mathcal{X}$, given by $\ln \hat{P}(\mathbf{x}|\varphi) = \sum_{i=1}^{n} \ln \hat{p}(\mathbf{x}_i|\varphi)$, under the constrains $\|\hat{\boldsymbol{\mu}}\| = 1$ and $\hat{\kappa} \geq 0$ [9]. According to the number of mixtures, two cases can be distinguished: unimodal $(M = 1)$ and multimodal $(M > 1)$ ML estimation.

1) ML in Unimodal vMF-MM: In [9] it is shown that the analytical solution of the unimodal ML estimation exists and can be computed, in terms of $\mathbf{r} = \sum_{i=1}^{n} \mathbf{x}_i$, as follows

$$\hat{\boldsymbol{\mu}} = \mathbf{r} \cdot \|\mathbf{r}\|^{-1} \tag{4}$$

$$I_{d/2}(\hat{\kappa}) \cdot I_{d/2-1}(\hat{\kappa})^{-1} = \bar{r} \tag{5}$$

being $\bar{r} = \|\mathbf{r}\|/n$ for $M = 1$. Since $\hat{\kappa}$ results from the ratio between two Bessel functions of consecutive order, it has no deterministic solution. Several methods exists for its approximation, among which, we found reliable the approximation of small values of d and large values of κ given in [7]

$$\hat{\kappa} \approx \frac{d-1}{2(1-\bar{r})} \tag{6}$$

2) ML in Multimodal vMF-MM: Unlike the previous case, for multimodal vMF-MM models there is not an analytical solution for the ML estimation which needs to be iteratively estimated, commonly using the Expectation Maximization (EM) algorithm [9]. It iterates between E-step and the M-step according to:

(E-step)

- Compute *mixture occupation probabilities* for each mixture m and training sample x_i

$$g(m, \mathbf{x}_i) = \frac{w_m f(\mathbf{x}_i, \boldsymbol{\mu}_m, \kappa_m)}{\sum_{k=1}^{M} w_m f(\mathbf{x}_k, \boldsymbol{\mu}_k, \kappa_k)}$$

(M-step)

- Update $\hat{w}_m$ using the Baum Welch 0-order statistics

$$\hat{w}_m = \frac{1}{n} \sum_{i=1}^{n} g(m, \mathbf{x}_i)$$

- Update $\hat{\boldsymbol{\mu}}_m$ using eq. (4) for $\mathbf{r} = \mathbf{r}_m$ (1^{st} order statistics)

$$\mathbf{r}_m = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i g(m, \mathbf{x}_i)$$

- Update $\hat{\kappa}_m$ as in eq. (5) and (6) for $\bar{r} = \bar{r}_m$

$$\bar{r}_m = \|\mathbf{r}_m\| \cdot (n\hat{w}_m)^{-1}$$

IV. EXPERIMENTS

A. Experimental Setup, Database and Protocol

Language recognition systems based on the approaches presented in the above sections have been tested over the 2009 edition of the biannual Language Recognition Evaluation (LRE'09) carried out by the National Institute of Standard and Technology (NIST) [13]. NIST LRE'09 evaluation data includes test segments of 3, 10 and 30s length, belonging to 40 different languages divided as 23 target and 17 non-target languages. Results presented in this work refer to both closed and open-set conditions. We refer to closed-set as the task when only target languages are included in the test trials set, and to open-set when other non-target languages (unknown to participants) are also included. More detailed information can be found in the LRE'09 evaluation plan [13].

TABLE I

CLOSED-SET RESULTS IN NIST-LRE 09

	min Cavg. x 100			mean EER		
	30s	*10s*	*03s*	*30s*	*10s*	*03s*
Cosine Kernel	4.64	8.76	18.54	4.99	9.07	19.05
Cosine KDF	4.99	9.00	18.53	5.38	9.36	18.96
vMF KDF	3.42	7.74	18.40	3.69	8.03	19.19
vMF MM	**3.32**	**7.61**	**18.40**	**3.59**	**7.84**	**18.70**

Data used with background purposes was obtained from a development set provided by NIST to all LRE'09 participants. It included manually and automatically labelled segments of the two types of audio considered in the LRE'09 evaluation: conversational telephone speech and telephonic speech belonging to internet broadcast news. An average of ~ 400 utterances per language were extracted from long (minimum length 30s.) segments. Later, utterances were mapped into their own i-vectors using a Total Variability Subspace approach with 400 latent variables and an initial UBM model of 1024 Gaussians [1]. Using LDA, i-vectors were finally projected into the $d = 21$ dimensional *disentangled* space used to develop the proposed techniques. More details about the data partition, parametrization and UBM modelling can be found in the authors LRE'09 participation presented in [3].

The algorithm in the vMF-MM system is initialized using binary splitting for assigning vectors to each cluster, until M clusters are found. Distances are computed according to (2) and centroids are computed as in (4) for $\kappa = 16$. Then, a general model is trained using all the available training samples, five EM iterations and adapting mean directions, weights and concentrations parameters. Three additional iterations are used to adapt this general model to each of the target languages. In this case only weights and mean directions were updated, but not the concentration parameters that we found to result in overfitting given the available data. Regarding the systems parameters tuning, the Bayesian Information Criterion (BIC) revealed that the optimal values of M is 1 and 2, for respectively, 17 and 5 of the considered target languages. For the vMF-KDF system, κ was empirically set to 16.

B. Results

A wide set of experiments covering both closed and open-set conditions has been carried out, summarized in Table I and Table II for the two performance metrics considered: $i)$ *min Cavg*, which defines the averaged minimum cost of taking bad decisions and $ii)$ *mean EER*, the averaged Equal Error Rate (EER)

TABLE II

OPEN-SET RESULTS IN NIST-LRE 09

	min Cavg. x 100			mean EER		
	30s	*10s*	*03s*	*30s*	*10s*	*03s*
Cosine Kernel	5.44	9.59	19.25	5.62	9.75	19.55
Cosine KDF	5.68	9.81	**19.19**	5.94	10.00	**19.52**
vMF KDF	**4.90**	9.21	19.44	**4.89**	9.29	20.06
vMF MM	4.96	**9.15**	19.44	4.95	**9.18**	19.63

among the target classes. Both are reliable metrics to evaluate the systems discrimination capabilities.

At a first glance of Table I, two main facts can be observed. First, both systems based on vMF distributions significantly outperform the typically used cosine-based matching presented in [1]. This confirms that vMF distributions are suitable for assigning distributions in the i-vector domain. Notable is the 28% of improvement in the min Cavg and mean ERR achieved for the 30s closed-set condition using the vMF-MM system. Second, since samples derived from utterances longer than 30s were only used in training, it can also be observed that the vMF-based approaches are slightly sensitive to the data mismatch when models are trained with utterances having different lengths than test segments. Despite the relative performance with the cosine-based approaches decreases, the vMF-based approaches still achieves best results for all the closed-set conditions. Particularly, the vMF-MM system outperforms other systems in all cases.

Table II shows the performance improvement of the proposed techniques with the presence of unseen languages. We observe a slight deterioration generated by the further introduced mismatch between training and testing data sets. It could be alleviate using sufficient language variability in the training set. In the worst presented scenario, 3s open-set condition, the cosine-based approaches achieve best performances while small degradation is observed for the vMF-based systems. Moreover, it can be observed that as the mismatching decreases (30s and 10s open-set conditions) the vMF-based system still achieve best results.

V. CONCLUSIONS

We present a new modelling approach for learning distributions of *disentangled* i-vectors for language recognition. The approach is based on the use of von Mises-Fisher distributions. These are directional distributions suited to directional data such as the i-vectors. vMF distributions are exploited using two different approaches: Kernel Density Functions (vMF-KDF) and Finite Mixture Models (vMF-MM). Experiments are carried out over the NIST LRE'09 evaluation and shows that the vMF-MM systems outperforms the typically used cosine distance approach in most of the cases. Particularly, a 28% of improvement is achieved for the 30s closed-set condition.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[2] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 7, pp. 1969 – 1978, 2007.

[3] J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. Toledano, and J. Gonzalez-Rodriguez, "Multilevel and session variability compensated language recognition: Atvs-uam systems at nist lre 2009," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1084–1093, 2010.

[4] A. Hatch, A. Stolcke, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2006

[5] D. Scott and S. Sain, *Multidimensional Density Estimation*. Elsevier, 2005, vol. 24, pp. 229 – 261.

[6] D. Scott, *Multivariate Density Estimation*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1992.

[7] K. Mardia and P. Jupp, *Directional statistics*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[8] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 203–210, 2005.

[9] A. Banerjee, I. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway, "Clustering on the unit hypersphere using von mises-fisher distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[10] T. McGraw, B. Vemuri, B. Yezierski, and T. Mareci, "von mises-fisher mixture model of the diffusion odf," in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, 2006.

[11] H. Tang, S. Chu, and T. Huang, "Generative model-based speaker clustering via mixture of von mises-fisher distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[12] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications, 1965.

[13] "The 2009 nist language recognition evaluation plan," 2009, http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.