SLTU-2014, St. Petersburg, Russia, 14-16 May 2014

On Using Intrinsic Spectral Analysis for Low-resource Languages

Reza Sahraeian[1], *Dirk Van Compernolle*[1], *Febe de Wet*[2]

[1] ESAT, KU Leuven, Belgium
[2]HLT Research Group Meraka Institute, CSIR, South Africa

Abstract

This paper demonstrates the application of Intrinsic Spectral Analysis (ISA) for low-resource Automatic Speech Recognition (ASR). State-of-the-art speech recognition systems that require large amounts of task specific training data fail to reliably model feature distributions in resource impoverished settings. We address this issue by approaching the problem in the front-end, where we can learn an intrinsic subspace that can replace the traditional feature space like mel frequency cepstral coefficients (MFCC). We use ISA features for under-resourced settings to model the acoustic feature distribution with less complexity. We also propose to combine intrinsic features with extrinsic ones to take advantage of both subspaces. Experimental results for a phone recognition task on the Afrikaans language show that a combination of the intrinsic subspace and extrinsic subspaces provides us with improved performance compared to conventional features.

Index Terms: low-resource speech recognition, manifold learning, intrinsic spectral analysis

1. Introduction

Over the past four decades, most efforts in the realm of speech recognition were focused on a very small number of languages spoken by a large number of speakers, and the dominant strategy has relied on the availability of substantial language-specific transcribed speech and text data. However, when forced to deal with limited resources, conventional acoustic modeling techniques perform very poorly. With more widespread use of voice technology, developing ASR systems for under-resourced domains or languages has become of great interest in recent years [1] [2].

An extensive amount of study has been conducted to improve the accuracy of speech recognizers in low-resource conditions, and several strategies have been previously proposed. One group of approaches deal with cross-lingual acoustic modeling to port information from one or more source language systems which are built using larger amounts of training data, in order to build a recognizer for an under-resourced target language [3] [4]. To this end, acoustic modeling techniques capable of exploiting out-of-language data such as Kullback-Leibler divergence based HMM (KL-HMM) [5], Tandem [6] or Subspace Gaussian mixture models (SGMMs) [7] have been developed and extensively used [4]. Moreover, training data-driven feature front-ends, in which a multilayer perceptron (MLP) trained on large amounts of task independent data plays a key role has also been proposed [8].

A less studied alternative approach to accommodating low resource language scenarios moves the focus to feature transformations in the front-end, where we can train more reliable models with less data. Principal components analysis (PCA) and Linear discriminant analysis (LDA) are commonly used linear methods. However, speech production mechanisms imply that our vocalizations are approximately restricted to a low-dimensional manifold embedded in a high-dimensional space [9] [10]. Manifold learning methods have been widely used to learn nonlinear projection maps that recover the underlying configuration space. The applicability of this class of techniques in the speech community was first proposed in [10] by introducing Intrinsic Spectral Analysis (ISA).

Intrinsic Spectral Analysis is the extension of Laplacian Eigenmaps in the framework of unsupervised manifold regularization [11], which naturally deals with out-of-sample data and also results in feature reduction. ISA has been compared with traditional front-ends in high resource speech recognition [12] [13]. The utility of ISA on a completely unsupervised task of spoken term discovery was also investigated in terms of zero resource speech recognition [12]. In the low resource regime, however, the performance of ISA features has not been investigated.

Intrinsic components can discriminate between natural classes of speech sounds [13]. Improved linear separability implies that acoustic modeling may be achieved with less complexity and less training data. In the case of Gaussian mixture monophone or triphone models, improved linear separability may reduce the number of mixture components required. Moreover, an ISA-based classification task showed that the accuracy of a system using ISA features is different from that of a system using cepstral features [14]. This suggests that a combination of both feature styles might be even better. In this paper we address this issue.

The remainder of this paper is structured as follows:

Section 2 contains a brief review of the theoretical background of Intrinsic Spectral Analysis and the data selection method we proposed in [14]. Section 3 describes the utility of ISA features for low-resource settings. Then the database we used is explained in section 4. Section 5 presents experimental results, and finally we have concluding remarks.

2. Background

2.1. Intrinsic Spectral Analysis

Given a set of data points $X = [x_1, x_2, ..., x_n]$ in $\mathcal{R}^H$ sampled from a manifold $\mathcal{M}$, we first construct an undirected adjacency weighted (or binary) graph $G = (X, \mathbf{W})$ with one vertex per data point. We put an edge between node i and j if x_i is among κ nearest neighbors of x_j (or vice versa). $\mathbf{W} \in \mathcal{R}^{n \times n}$ is the similarity matrix whose ijth element, w_{ij}, represents the similarity between x_i and x_j. We use the gaussian similarity function, $w_{ij} = exp(-\|x_i - x_j\|^2/2\tau^2)$, to exploit more structural information [14]. We then define the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is diagonal with $D_{ii} = \sum_{j=1}^{n} w_{ij}$. In this paper we use normalized Laplacian matrix: $\mathbf{L}_{norm} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix, as it has some nice properties [15].

In Laplacian Eigenmaps, the graph Laplacian is used to approximate the intrinsic coordinates for the manifold [16]. However, this method is limited to the eigen functions of the graph and not the entire manifold. Intrinsic Spectral Analysis approaches out-of-sample data by introducing a modified variant of the unsupervised manifold regularization algorithm

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L} \mathbf{f} \tag{1}$$

Where f is a projection to intrinsic bases, $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space (RKHS) for some positive semi-definite $n \times n$ kernel function K, and $\mathbf{L}$ is the graph Laplacian, $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_n)]^T$ is the vector of values of f for the training data. ξ is the parameter which makes the balance between extrinsic and intrinsic smoothness of the functions. The lth component of the solution to this optimization problem, based on the RKHS representer theorem, can be expressed as

$$f_l^*(v) = \sum_{i=1}^{n} a_i^l K(x_i, v) \tag{2}$$

$a^l \in \mathcal{R}^n$ is the lth eigenvector (sorted by eigenvalue) to the following generalized eigenvalue problem

$$(\mathbf{I} + \xi \mathbf{L} K)a = \lambda K a \tag{3}$$

In this paper, we always use a Radial Basis Function (RBF) kernel: $K(y, x) = exp(-\|y - x\|^2/2\sigma^2)$.

2.2. Data Selecetion

Classical manifold learning methods need to store a $n \times n$ Laplacian matrix and to compute an eigendecomposition. This can be problematic for large-scale data (large n). Approximations such as the Nyström method can be used to solve a reduced eigenvalue problem and to approximate the full-size eigenvectors solution [17]. These techniques typically use random subsampling which may lead to choosing a subset of data points that do not represent the underlying structure of the data. We proposed to use quadratic Renyi entropy to find a proper subset being well representative of manifold structure [14]; we will review this approach in this section briefly.

Considering $\mathcal{D}_{full}$ as a full dataset, we seek to find a subset, $\mathcal{D}$, with much smaller number of data points and well representative of the structure of data. To this end, we select a subset of m samples, and then maximize the nonparametric estimation of the quadratic Renyi entropy for the subset using RBF kernel as has been discussed in [18].

$$E(\mathcal{D}, \rho) = -\log \int p(x)^2 dx \approx -\log(\frac{1}{m^2} \mathbf{1}_m^T \mathbf{K} \mathbf{1}_m) \tag{4}$$

Where 1_m is a vector of m ones and $\mathbf{K}$ is the $m \times m$ RBF kernel matrix with parameter ρ. This criterion can be maximized iteratively in a greedy manner as explained in [14]. We use Silverman's rule [19] to find an appropriate kernel parameter:

$$\rho = \delta[\frac{4}{(2H+1)n_c}]^{[1/(H+4)]} \tag{5}$$

Where H is the dimension of data, δ is the sum of diagonal elements in the covariance matrix of data in $\mathcal{D}_{full}$, and n_c is the number of data points in $\mathcal{D}_{full}$.

The greedy method we used in [14] is based on the substitution of one datapoint with the other; however, for the selection of large number of points we can select a small subset in each iteration.

3. ISA for low-resource settings

Short-term spectral-based (typically cepstral) features such as MFCC or PLP are typically non-Gaussian, and are most often modeled by mixtures of Gaussians. Thus, we need many Gaussian mixtures to effectively model the feature distribution. On the other hand, ISA is effective at modeling unknown distributions by recovering the nonlinear articulatory parameter space. The numerical correlation between the distinctive features and intrinsic spectral components is studied in [13]. This implies that individual intrinsic coordinates can be understood according to some broad phonetic class distinction and separate natural classes of speech sounds with no supervision. Although the ISA interpretation cannot be formally developed for all classes of speech sounds, e.g. turbulence-

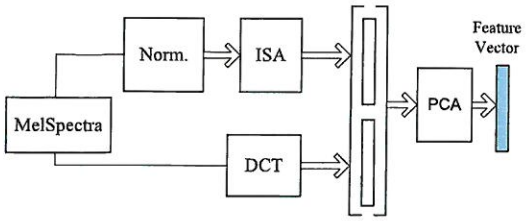

Figure 1: *Schematic diagram of Feature Transformation.*

driven obstruents, it still improves clustering by preserving localities [10]. This separability suggests less complexity to model feature distributions in resource impoverished settings, and this leads to more reliable models.

3.1. Combined ISA features with MFCC features

In addition to intrinsic components there can exist individual extrinsic ones, e.g. frequency bands, that can discriminate some of the natural classes reasonably well. To fully take advantage of the benefit of ISA together with regular features like MFCC, we combine ISA features with MFCC features. It is worth noting that the combination of ISA features with PLP features using the Dempster-Shafer (DS) theory of evidence is addressed for high resource speech recognition in [13]. However, in this paper we show that even a simple concatenation of these feature types improves the recognition in the low resource scenarios.

A simple way to combine MFCC and ISA features is to concatenate them. This can result $39 + 39 = 78$ dimensional feature vectors for example, where we considered 39 as the dimension of both ISA and MFCC features. The resulting feature vector could contain significant redundancies. Thus, different feature reduction approaches such as HLDA or LDA can be applied to discard the insignificant dimensions. In this paper, however, we use Principal Component Analysis (PCA) to find significant dimensions and respect the unsupervised nature of ISA features.

Figure 1 outlines the components of the feature transformation. Starting from a Mel-spectrum, we can extract cepstral features by taking discrete cosine transform (DCT). To obtain ISA features MelSpectra features are first normalized to have zero mean and unit variance. Then, MFCC and ISA features are concatenated to form a long vector. PCA is subsequently applied to reduce dimensionality.

4. Database discription

In this study we use Afrikaans data from the NCHLT corpus [1] [20]. The database consists of 210 speakers, including broadband speech recorded at 16 kHz. The dictionary contains 45 phonemes (including silence); the standard dataset configuration consists of only a training and a test part. The validation set introduced in Table 1 is taken from the training portion. The dataset information including the duration of each part and number of female and male speakers is summarized in Table 1.

Low-resource settings were simulated by using only small amounts of the training data mentioned in Table 1. To this end, we use 18 hours of randomly chosen speech covering all the speakers from the complete train set. We continue to choose smaller amount of data from the new set and keep the balance among speakers. The information regarding these new subsets is summarized in Table 2. In this paper we use these datasets as examples of data from a low-resource language.

Table 1: *Summary of NCHLT Afrikaans dataset. Duration is in hours*

Set	Duration	# male sp.	# female sp.
Train	50.70	98	94
Test	2.55	4	4
Validation	2.70	5	5

Table 2: *Summary of small datasets chosen to represent low-resource settings.*

Set	set1	set2	set3	set4	set5
Duration	1h	4h	8h	12h	18h
# speakers	188	190	191	192	192

5. Experiments

5.1. Feature extraction

For feature extraction, a short-time Fourier analysis is performed with a 30ms Hamming window and a 10ms window shift. Each frame was represented by a 24-dimensional Mel-Spectrum applying triangular shaped filterbank using the full spectrum (24 channels for 16 kHz). To train the intrinsic coordinates, all features were normalized to have zero mean and unit variance. 10k samples were subsequently selected from the training data as explained in section 2.2. Next, the weighted similarity graph is constructed to make the normalized graph Laplacian. After finding the intrinsic coordinates by ISA, we kept only the first 13 ones (skipping the first trivial

[1] Available from the South African Resource Management Agency (http:rma.nwu.ac.za).

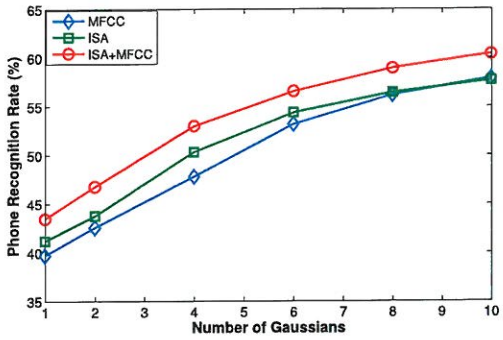

Figure 2: *Comparing phone recognition accuracies with 1 hour of training data for different front-ends.*

one) and add the first and second derivatives (Δ and $\Delta\Delta$) features.

The ISA features together with 39 dimensional MFCC features (13 cepstral $+\Delta+\Delta\Delta$ features) are concatenated to form 78-dimensional feature vectors. Then PCA is applied to reduce the dimensionality to 39 (Figure 1).

5.2. Evaluation

In this section, we analyze the performance of the different features in under resourced settings. Our definition of ISA involves four parameters that must be chosen by the user. To this end, all parameters are jointly optimized on the validation set introduced in Table 1. To save time these parameters are found once using simple monophones trained on the 1 hour dataset (set1 in Table 2), and the resulting optimized parameters are used for the evaluation on the test set in the rest of this section. The suitable parameters are determined as follows: $\kappa = 30, \sigma = 90, \xi = 1, \tau = 0.5$. It is worth mentioning that the efforts to find these parameters using other training sets lead to more or less the same values. This makes an interesting point that setting the parameters shows a good robustness to the amount of training data.

For the first set of experiments, we only used set1 (including 1 hour of data). Since the amount of training data is very low, the standard 3-state left to right HMM architecture to model monophones with a simple phone-loop grammar are trained. Using various numbers of Gaussians per state (1 to 10), phone recognition accuracies for different feature types are shown in Figure 2.

Figure 2 confirms our hypotheses that intrinsic coordinates can separate natural speech sound classes with less model complexity. However, as shown, when the number of Gaussians increases per state, the improvement of ISA over MFCC vanishes. This implies that when adequate data is available to train more complex models, MFCC features contain reasonably good discrimination information. Figure 2 also shows that the combination of ISA and MFCC features (ISA+MFCC) yields the best results in all cases. This suggests that the intrinsic subspace together with the extrinsic one constitutes a suitable feature space for low resource settings.

Table 3: *Comparing phone recognition accuracies (%) using different amounts of data for MFCC and ISA+MFCC.*

Set:	set2	set3	set4	set5
MFCC Features	59.11	64.73	67.60	70.55
ISA Features	58.10	64.68	68.17	69.86
ISA+MFCC Features	61.70	67.34	69.99	71.94

For the second set of experiments, we used more data, (set2,...,set5), to model context-dependent triphones. Triphones were tied at the state level using decision tree clustering, and each tied-state triphone was estimated with 8 Gaussian mixtures per state. The phone recognition accuracy for conventional acoustic features, i.e. MFCCs, compared to the combined feature type based on ISA is shown in Table 3.

As shown, ISA features combined with MFCC features provides a substantial gain over conventional acoustic features in all sets. This demonstrates the usefulness of our approach to use the intrinsic subspace in low-resource settings to generate better features.

6. Conclusions

We have argued that using ISA features combined with MFCC features can improve ASR performance on a low-resource speech recognition task. ISA features provide a data-driven front-end approach to feature extraction that improves discrimination and ease of modeling by recovering a set of intrinsic projections maps that correlate with natural classes of speech sounds. We proposed to combine the intrinsic feature space with the extrinsic one to take full advantage of both. We have conducted phone recognition experiments on Afrikaans language taken from the NCHLT dataset to examine the validity of our proposed features for low-resource conditions. The results showed that the combined feature type outperforms cepstral features not only in very impoverished settings but also for the case of having more training data of about 18 hours.

7. Acknowledgements

This work is based on research supported by the South African National Research Foundation as well as the fund for scientific research of Flanders (FWO) under project AMODA GA122.10N. Thanks also to the Human Language Technology research group (HLT) at the CSIR's Meraka Institute for providing access to

Afrikaans datasets and pronunciation dictionaries.

8. References

[1] P. Fung and T. Schultz, "Multilingual spoken language processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 89–97, 2008.

[2] X. Cui, J. Xue, P. L. Dognin, U. V. Chaudhari, and B. Zhou, "Acoustic modeling with bootstrap and restructuring for low-resourced languages." in *INTERSPEECH*, 2010, pp. 2974–2977.

[3] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual subspace Gaussian mixture models for low-resource speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 22, no. 1, pp. 17–27, 2014.

[4] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Communication*, vol. 56, pp. 142–151, 2014.

[5] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using kl-based acoustic models in a large vocabulary recognition task." in *INTERSPEECH*, 2008, pp. 928–931.

[6] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings.*, vol. 3. IEEE, 2000, pp. 1635–1638.

[7] D. Povey *et al.*, "Subspace Gaussian mixture models for speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),*. IEEE, 2010, pp. 4330–4333.

[8] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," *genre*, vol. 10, no. 15/20, 2005.

[9] M. Nilsson and W. B. Kleijn, "On the estimation of differential entropy from data located on embedded manifolds," *IEEE Transactions on Information Theory,*, vol. 53, no. 7, pp. 2330–2341, 2007.

[10] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2006, pp. 241–244.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[12] A. Jansen, S. Thomas, and H. Hermansky, "Intrinsic spectral analysis for zero and high resource speech recognition." in *INTERSPEECH*, 2012.

[13] A. Jansen and P. Niyogi, "Intrinsic spectral analysis," *IEEE Transactions on Signal Processing*, vol. 61, pp. 1698–1710, April 2013.

[14] R. Sahraeian and D. Van Compernolle, "A study of supervised intrinsic spectral analysis for timit phone classification," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2013, pp. 256–260.

[15] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data represenation," *Neural Computation*, vol. 16, pp. 1373–1396, 2003.

[17] F. Tompkins and P. J. Wolfe, "Approximate intrinsic Fourier analysis of speech." in *INTERSPEECH*, 2009, pp. 120–123.

[18] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, pp. 669–688, 2002.

[19] B. W. Silverman, "Density estimation for statistics and data analysis," Chapman and Hall, London, 1986.

[20] C. van Heerden, M. H. Davel, and E. Barnard, "The semi-automated creation of stratified speech corpora," in *Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2013, pp. 115–119.