Under-resourced Speech Recognition based on the Speech Manifold

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

[1]ESAT, KU Leuven, Belgium
[2]HLT, Meraka Institute, CSIR, South Africa
{Reza.Sahraeian,Dirk.VanCompernolle}@esat.kuleuven.be,fdwet@csir.co.za

Abstract

Conventional acoustic modeling involves estimating many parameters to effectively model feature distributions. The sparseness of speech and text data, however, degrades the reliability of the estimation process and makes speech recognition a challenging task. In this paper, we propose to use a nonlinear feature transformation based on the speech manifold called Intrinsic Spectral Analysis (ISA) for under-resourced speech recognition. First, we investigate the usefulness of ISA features in low resource scenarios for both Gaussian mixture and deep neural network (DNN) acoustic modeling. Moreover, due to the connection of ISA features to the articulatory configuration space, this feature space is potentially less language dependent than other typical spectral-based features, and therefore exploiting out-of-language data in this feature space is beneficial. We demonstrate the positive effect of ISA in the frame work of multilingual DNN systems where Flemish and Afrikaans are used as donor and under-resourced target languages respectively. We compare the performance of ISA with conventional features in both multilingual and under-resourced monolingual conditions.
Index Terms: Under-resourced speech recognition, intrinsic spectral analysis, multilingual deep neural network

1. Introduction

During the past decade, Automatic Speech Recognition (ASR) for under-resourced languages has received much attention in the speech research community [1, 2]. The main issue which makes speech recognition a challenging task in resource constrained settings is that conventional speech recognizers rely heavily on statistically based modeling schemes and need to estimate a large number of parameters to effectively model speech feature distributions. This is mainly due to the non-Gaussian distribution of typical features such as PLPs or MFCCs that are usually used. Finding a feature space in which feature distributions can be modeled with fewer parameters is therefore an interesting challenge. This motivates efforts to investigate the impact of feature transformation in the front end of ASR systems to accommodate low resource language scenarios [3].

Moreover, exploiting out-of-language data, either in the acoustic modeling or feature extraction process, is a very popular approach to improve the performance of under-resourced ASRs [4, 5]. This can be accomplished by porting multilingual information at the Gaussian level in the context of Subspace Gaussian Mixture models (SGMMs) [6]. Furthermore, multilingual data can be used to train deep neural networks assuming that all the language dependent information is concentrated in the neurons which compute the output layer; thus, the hidden layers can be viewed as language independent feature extractors and transferred across languages [7, 8, 9, 10].

The aim of this paper is to address the aforementioned issues by using a nonlinear feature transformation based on the the articulatory configuration of the speech production system. More specifically, we propose to exploit the manifold structure of speech sounds with the aim to discover the intrinsic feature space. The existence of the speech manifold was studied in [11] by introducing a variant of Laplacian eigenmaps named Intrinsic Spectral Analysis (ISA). The use of ISA for different speech recognition tasks has been investigated in many studies such as [12], [13] and [14]. Most of these studies, however, do not show how ISA can be helpful for speech recognition in under-resourced conditions. It has been shown that intrinsic coordinates discriminate between natural classes of speech sounds [13]. This suggests that acoustic modeling can be accomplished with lower complexity and less data. For example, in GMM based models, fewer Gaussian components are required to reliably train phone models.

Moreover, since ISA features are representative of articulatory parameters, we expect them to be less language dependent than extrinsic features [15]. This implies that sharing data of different languages for multilingual learning is more beneficial in the intrinsic feature space than the extrinsic one. Furthermore, many of the sate-of-the-art multilingual techniques such as KL-HMM [4] and Multilingual DNNs [16] involve phoneme mapping which is normally done by merging phones with the same articulatory representation; however, the conventional feature spaces do not directly represent the articulatory space. This motivates us to use ISA features which are linked to articulatory parameters for the phoneme mapping. In this study, we investigate this issue in the frame work of multilingual DNNs where creating the target phoneme set for the multilingual DNN training is done by a knowledge-based mapping to a global phoneme set.

The remainder of this paper is organized as follows: Section 2 explains the speech manifold and the theoretical background of Intrinsic Spectral Analysis. Section 3 describes the utility of ISA features for low resource settings. The databases we used for our experiments are explained in Section 4. Sections 5 presents the experimental results, and finally we have concluding remarks.

2. The Speech manifold

The human speech generation apparatus involves relatively few degrees of freedom. As a result, the number of possible sounds that can be produced are restricted. It has been postulated long ago that regions exist in the articulatory space where small changes lead to relatively large changes in acoustic character, and vice versa [17]. This suggests that the speech production mechanisms define a nonlinear mapping between the configuration space and the acoustic space. With the aim of estimating

an inverse map from the data to recover the articulatory parameters, the manifold learning algorithm has been used for speech signals. The validity of the manifold structure assumption for speech data in the acoustic space was formalized in [11]. Going further, Intrinsic Spectral Analysis (ISA) was proposed as a practical approach in the context of manifold regularization [11, 13]. This technique is explained in the next section.

2.1. Intrinsic Spectral Analysis

Considering a manifold $\mathcal{M}$ embedded in $\mathcal{R}^H$ and a collection of n samples $X = [x_1, x_2, ..., x_n] \subset \mathcal{M}$ that forms a mesh of data points that lie on the manifold, as is typical in manifold learning algorithms, an undirected adjacency weighted (or binary) graph $G = (X, \mathbf{W})$ is constructed with one vertex per data point and the similarity matrix $\mathbf{W} \in \mathcal{R}^{n \times n}$. w_{ij} (the ijth element of $\mathbf{W}$) represents the similarity between x_i and x_j if x_i is one of the κ nearest neighbors of x_j (or vice versa) and 0 otherwise. In this study, we use the gaussian similarity function, $w_{ij} = exp(-\|x_i - x_j\|^2/2\tau^2)$. Then, the so-called graph Laplacian is defined, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal vertex degree matrix with elements $D_{ii} = \sum_{j=1}^{n} w_{ij}$. One can also consider a normalized variant, $\mathbf{L}_{norm} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix. This normalization reduces the effect of large variation in vertex degree arising from sampling sparsity [18].

Conventional Laplacian Eigenmaps regard the graph as a mesh on the manifold and find the basis determined by the graph Laplacian as an approximation to an intrinsic basis for the manifold that the sample was drawn from [19]. However, this method is limited to the eigen functions of the graph and not the entire manifold. Thus, we seek for a projection f to an intrinsic basis on the manifold. In Intrinsic Spectral Analysis out-of-sample data is approximated by learning such a function in the framework of unsupervised manifold regularization:

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

Where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space (RKHS) for some positive semi-definite $n \times n$ kernel function K, $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_n)]^T$ is the vector of values of f for the training data, and $\mathbf{L}$ is the graph Laplacian. 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: $K(y, x) = exp(-\|y - x\|^2/2\sigma^2)$.

3. ISA for under-recourced ASR

Conventional acoustic modeling involves estimating a large number of parameters to effectively model feature distributions. The parameter estimation, however, is often hampered by a lack of data. The simplest solution is to look for a feature space where we need fewer parameters to model the feature distribution properly.

Intrinsic Spectral Analysis is a nonlinear feature transformation yielding features that are expected to be correlated with distinctive features. It has been shown [13] that some of these intrinsic features may have a near binary behavior and directly relate to some broad phonetic class distinction and separate natural classes of speech sounds. Linear separability may therefor be easier in the intrinsic subspace [13], require fewer parameters in the statistical modeling and hence require less training data. It is worth noting that we have used ISA and its combination with MFCC for the low resource ASR in [20]; however, we didn't observe improvement by using ISA features alone.

Moreover, the most common approach to improve an under-resourced speech recognizer is to use out-of-language data for bootstrapping, smoothing or back-off. However, sharing the knowledge across various languages is not a straightforward task because of differences such as different sets of subword units. A common approach towards this is the creation of a universal phoneme set by first pooling the phoneme sets of different languages together and then merging them [21, 22]. The major underlying assumption here is that the articulatory representations of phonemes are similar and their acoustic realization can be assumed language independent. From the very early attempts to perform multilingual acoustic modeling to the most recent ones, phoneme mapping has been popularly used to merge all the monolingual phonemes which share the same symbol, for example in the IPA table [23, 16].

From the above brief overview, it is clear that the language independency assumption for the phonemes with the same articulatory representation is a crucial issue. In the literature, PLP, MFCC and FBANK are the typical feature spaces used where phonemes are merged; however, they do not directly refer to the articulatory features. In this study, we use intrinsic spectral analysis to map speech features into the articulatory configuration space where the amalgamation of the phonemes makes more sense.

4. Experimental setup

In our experiments Afrikaans plays the role of under-resourced target language. Flemish takes on the role of well-resourced donor language. Various mono- and multilingual scenarios are considered. Given the low resource context, all evaluations are done with phoneme recognition experiments, using a phone bigram language model and the ASR performance is reported in phone error rate (PER). The Kaldi ASR toolkit [24] is used for both GMM and DNN based acoustic modeling.

4.1. Database for Afrikaans

The NCHLT corpus[1] [25] is an Afrikaans database consisting of 210 speakers, including broadband speech sampled at 16 kHz. The phoneme set contains 38 phonemes, including silence. All repeated utterances were removed from the original dataset. In our setting, to simulate low resource conditions, we consider one hour of data, five hours of data and the full training set including about 10.7 hours of data as summarized in Table 1. We used the default test and validation sets.

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

Set:	Train1	Train2	Train3	Test	Dev
Duration	1hr	5hr	10.7hr	2.2hr	1.0hr
# speakers	188	192	192	8	10

Table 1: Used Subsets of NCHLT Afrikaans Corpus.

4.2. Database for Flemish

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) is a standard Dutch database that includes speech data collected from adults in the Netherlands and Flanders [26]. This dataset consists of 13 components that correspond to different socio-situational settings. In this study, we used Flemish data (audio recordings of speakers in Flanders) from component-o which contains read speech. The whole dataset includes 38 hours of speech sampled at 16KHz and we have taken 36hr for the training and 2hr for the evaluation. In this work, we used only the training part including 36 hours as donor data produced by 150 speakers. The CGN pronunciation dictionary uses an alphabet of 47 phonemes, for which a mapping to the Afrikaans phoneme set (37 phonemes) is available [27].

4.3. Feature extraction

For feature extraction, a short-time Fourier analysis was performed with a 30ms Hamming window and a 10ms window shift. For FBANK features, each frame was represented by a 24-dimensional log mel-spectrum applying triangular shaped filterbank using the full spectrum (24 channels for 16 kHz). Then, utterance based mean and variance normalization is applied. Afterwards, MFCCs and PLPs were extracted from the FBANK features.

To extract ISA features, the FBANKs were used and 10k samples were randomly selected from the training data to construct the weighted similarity graph and consequently make the normalized graph Laplacian. After finding the intrinsic coordinates by ISA and skipping the first trivial one, we used the next 13 or 24 dimensions as feature vector. Moreover, we need to find proper values for the ISA parameters; our ISA definition involves four parameters which were jointly optimized on the validation set introduced in Table 1. The suitable parameters were determined to be as follows: $\kappa = 5$, $\sigma = 90$, $\xi = 1$, $\tau = 0.5$. It was observed in the experiments that tuning these parameters using different amounts of data or based on different systems, e.g. monophones, triphones or SGMMs, leads to almost the same values.

5. Experimental results

5.1. Monolingual experiments

The first set of experiments provides a monolingual baseline for the Afrikaans data. First a Gaussian mixture model system is built as a reference system on the one hand and for development of the triphone definitions. We trained conventional 3-state left-to-right HMM triphone models. Subsequently, the alignments of the triphone system are used to train SGMMs. SGMMs use mixtures of Gaussians as the underlying state distribution and consider the high dimensional supervector of all the GMM parameters to be constrained to a relatively low dimensional subspace, which is common to all the states [28]. In other words, SGMMs aim to find the underlying manifold at the Gaussian level. Thus, it would be interesting to know if the use of manifold learning in the front end would further improve the performance of SGMMs. In these experiments, 13-dimensional MFCC, PLP and ISA as well as 24-dimensional FBANK features were used. The raw features together with their first and second derivatives were spliced in time taking a context size of 7 frames (i.e.,± 3), followed by decorrelation and dimensionality reduction to 40 using LDA and further decorrelation using MLLT [29].

# of Gaussians	Features			PER(%)
	MFCC	FBANK	ISA	
Total/ Average per tied-state	**4k/9.2**	2.2k/4.8	1.8k/3.9	23.09
	-	**4k/8.8**	3.2k/7.0	22.22
	-	-	**4k/6.6**	21.79

Table 2: PERs(%) for Train1 using different total number of Gaussians for HMM/GMM system.

As explained in section 3, a key motivator is that fewer Gaussians are needed for acoustic modeling of the intrinsic features rather than the extrinsic ones. To investigate this issue, we first consider the Train1 set including 1 hour of data, and then tune the number of Gaussians using the validation set. The PERs for MFCCs, FBANKs and ISAs are 23.09%, 22.22% and 21.79% respectively and are shown in the last column of Table 2. The tuned values for the total number of Gaussians and it's average per tied-state are made bold. As is shown, ISA outperforms MFCC and FBANK features.

This table reveals more interesting trends; first, we can see that fewer Gaussians (on average) are required to model triphones with ISA features than MFCCs and FBANKs. Moreover, for each feature type, the number of Gaussians needed to have the same PERs corresponding to the other features are found. For example, to have a PER equal to 23.09% we need only $\sim$1.8k Gaussians using ISA which is less that half of those required for MFCC. Table 3 summarizes the PERs for the all training sets introduced in Table 1 for MFCC, PLP, FBANK and ISA features. In each scenario the number of Gaussian components is tuned first. It confirms our hypotheses that ISA features outperform the other feature representations since fewer components are needed to effectively train phone models in both the HMM/GMM and SGMM systems.

In addition, since deep neural network is becoming a mainstream technology for speech recognition, it is of interest to contrast ISA with other conventional features as input for DNNs. We trained HMM/DNN systems using a generalized maxout network with p-norm nonlinearity [30]. We compared ISA with FBANK and MFCC features. PLP features were not included in this experiment as we observed that their performance was very similar to MFCCs, and they were never better than FBANK features.

The initial and final learning rates were specified by hand and equal to 0.02 and 0.004 respectively, and we always set $p = 2$. More details about the implementation and parameters are presented in [30]. The number of hidden layers are set based on the amount of training data. The number of units in each hidden layer and the group size are 800 and 5 respectively. Table 4 shows the PERs for various features using HMM/DNN systems. The neural network's inputs were the features being concatenated with 7 left and 7 right neighbor frames; then, an LDA transformation matrix was applied without dimensionality reduction. The number of DNN targets (i.e. context-dependent triphone states) are 505, 1380 and 2281 for Train1, Train2 and Train3 respectively. For Train1 set, we repeated the DNN training three times and report the average of the results. As is shown, DNN performances are worse than SGMM systems, noting that SGMMs reduce the number of parameters by choos-

Systems	Features	Train1	Train2	Train3
HMM/GMM	FBANK	22.22	16.13	13.91
	MFCC	23.09	16.87	14.81
	PLP	23.41	16.77	14.80
	ISA	**21.79**	**15.75**	**13.55**
SGMM	FBANK	21.65	13.07	10.53
	MFCC	22.38	13.83	11.02
	PLP	21.97	13.42	10.83
	ISA	**21.29**	**12.84**	**10.42**

Table 3: PERs(%) using different amounts of data for MFCC, PLP, FBANK and ISA in a monolingual GMM based system.

Set		Train1	Train2	Train3
# of hidden layers		2	3	4
Feature (#dim)	FBANK(24)	23.47	**13.95**	**11.27**
	MFCC(13)	23.30	15.10	12.67
	MFCC(24)	24.06	14.96	12.56
	ISA(13)	**22.39**	14.37	12.14
	ISA(24)	23.62	14.12	11.91

Table 4: PERs(%) for various features and different amounts of training data in a monolingual HMM/DNN system.

ing the Gaussians from a subspace spanned by a background model while we need to train many parameters for DNN systems. Moreover, except for the Train1, FBANK outperforms ISA in the monolingual DNNs systems. This suggests that when the reasonable amount of data exist, it is beneficial to exploit all the nonlinearities in a discriminative manner with DNN rather than applying the unsupervised manifold based transformation before the input layer.

5.2. Multilingual experiments

In this set of experiments, we focused on Train1 including 1hr of Afrikaans as low resource target language and the training part of Flemish dataset including 36 hours as donor data. First, we applied a knowledge-based phoneme mapping where each phoneme from the Flemish phoneme set was mapped to one of the phonemes in the Afrikaans one. To this end, 31 phonemes that share the same symbol in the IPA table were merged. The remaining 16 phonemes in Flemish without any IPA counterpart in Afrikaans were mapped based on linguistic knowledge [27]. It is worth noting that another alternative to train a multilingual DNN is to consider separate output targets for each language and avoid phoneme mapping. In our setup, however, we found that the phoneme mapping improves the results.

Then, we trained multilingual HMM/GMM systems and generate tied-state alignments by using the bi-gram language model trained with 1 hour of Afrikaans. To train ISA coordinates, we randomly chose 5k data points from the Flemish data and 5k data points from Afrikaans. Tuning the ISA parameters yielded the same values as in the monolingual case.

Then, multilingual DNNs were trained by adopting the audio alignments from the multilingual HMM/GMM system. The number of tied-states obtained by using a multilingual decision tree is 4131. The DNNs used a p-norm activation function with $p = 2$ and were trained from 15 consecutive frames like the DNN for the monolingual setting. p-norm input and output dimensionality were empirically set to 1000 and 200 respectively. We used the same learning rates as in the monolingual case. To bootstrap the acoustic model for Afrikaans, the hidden layers of the multilingual DNNs were shared and the softmax layer was retrained with the Afrikaans [8].

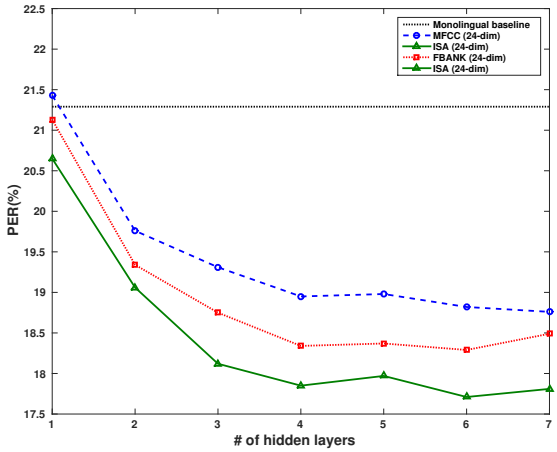

Figure 1: PERs for MFCC, FBANK and ISA features using multilingual DNN with different numbers of hidden layers using 1hr of data in the target language (Train1).

Figure 1 shows the comparisons of PERs obtained by multilingual DNNs with a different number of hidden layers for MFCC, FBANK and ISA; the reported results are for 24-dimensional features as we observed in the experiments that they always outperform 13-dimensional ones. The best monolingual performance achieved in Table 3 using ISA and SGMM is also shown in Figure 1.

Figure 1, moreover, reveals the following trends. First, sharing the hidden layers trained with multilingual data improves the ASR performance for the low resource ASR. Moreover, ISA features consistently outperform MFCC and FBANK features in the multilingual DNN systems. This demonstrates that ISA provides with a better feature space than the conventional extrinsic features for multilingual learning. The ISA improvement in these multilingual experiments is more pronounced by referring to Table 4; where the ISA performance in the monolingual DNN systems is not always better than FBANK features. This can be explained due to the link between ISA and articulatory configuration parameters which makes ISA a less language dependent feature space than MFCC and FBANK.

6. Conclusions

In this paper, we proposed to use ISA features which are representative of speech articulatory configuration parameters in the low-resourced conditions. We successfully showed that this allows acoustic modeling with less number of parameters in the GMM based acoustic modeling. For monolingual DNN based acoustic modeling ISA is only beneficial for 1hr data and in the case of more data DNNs seem to learn everything they need from the FBANKs. We also demonstrated that since ISA feature space is linked to the articulatory configuration space, it is a proper alternative to conventional features such as MFCC and FBANK for multilingual speech recognition systems.

7. Acknowledgements

Part of this work was 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.

8. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[2] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low resource languages: BABEL project research at cued," in *Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014, pp. 16–23.

[3] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *INTERSPEECH*. ISCA, 2012.

[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] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7319–7323.

[6] L. Burget *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4334–4337.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4269–4272.

[8] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7304–7308.

[9] F. Grezl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011, pp. 359–364.

[10] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Workshop on Spoken Language Technology (SLT)*, 2012, pp. 336–341.

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

[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] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[16] N. T. Vu *et al.*, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7639–7643.

[17] G. Fant, "Acoustic theory of speech production," Paris, France, 1970.

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

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

[20] R. Sahraeian, D. Van Compernolle, and F. de Wet, "On using intrinsic spectral analysis for low-resource languages," in *International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU)*, 2014, pp. 61–65.

[21] V. B. Le and L. Besacier, "First steps in fast acoustic modeling for a new target language: Application to Vietnamese," in *ICASSP*, 2005, pp. 821–824.

[22] K. C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2009, pp. 546–551.

[23] J. Köhler, "Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks," in *Multi-Lingual Interoperability in Speech Technology*, 1999, pp. 79–84.

[24] D. Povey *et al.*, "The KALDI speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[25] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, St Peterburg, Russia, May 2014, pp. 194–200.

[26] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation," in *International Conference on Language Resources and Evaluation*, 2000, pp. 887–894.

[27] R. Sahraeian, N. Kleynhans, F. de Wet, and D. Van Compernolle, "Knowledge-based phoneme mapping between flemish and afrikaans," *Internal report PSI-SPCH-14-01*, 2014.

[28] 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.

[29] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, 1999.

[30] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 215–219.