



Staircase Network: structural language identification via hierarchical attentive units

Trung Ngo Trong¹, Ville Hautamäki¹, Kristiina Jokinen²

¹School of Computing, University of Eastern Finland, Finland

²AI Research Center, AIST Tokyo Waterfront, Japan

trung@cs.uef.fi, villeh@cs.uef.fi, kristiina.jokinen@aist.go.jp

Abstract

Language recognition system is typically trained directly to optimize classification error on the target language labels, without using the external, or meta-information in the estimation of the model parameters. However labels are not independent of each other, there is a dependency enforced by, for example, the language family, which affects negatively on classification. The other external information sources (e.g. audio encoding, telephony or video speech) can also decrease classification accuracy. In this paper, we attempt to solve these issues by constructing a deep hierarchical neural network, where different levels of meta-information are encapsulated by attentive prediction units and also embedded into the training progress. The proposed method learns auxiliary tasks to obtain robust internal representation and to construct a variant of attentive units within the hierarchical model. The final result is the structural prediction of the target language and a closely related language family. The algorithm reflects a “staircase” way of learning in both its architecture and training, advancing from the fundamental audio encoding to the language family level and finally to the target language level. This process not only improves generalization but also tackles the issues of imbalanced class priors and channel variability in the deep neural network model. Our experimental findings show that the proposed architecture outperforms the state-of-the-art i-vector approaches on both small and big language corpora by a significant margin.

1. Introduction

The availability of large corpora in speech processing has been one of the major driving forces advancing speech technologies [1, 2]. This has allowed recent state-of-the-art systems to obtain impressive performance in recognizing spoken languages [1, 3, 4]. However, the focus of the recent NIST Language Recognition Evaluation (LRE) has been shifted to closely related languages in LRE’15 [5], then to a mixture of multiple language groups and their dialects in LRE17 [6]. As for the challenges of recognizing more confusing groups of languages, most of the developments in the spoken language recognition field have concentrated on two directions: fine-tuning existing algorithms on the new corpora, and applying recent advances in the field of machine learning in order to obtain improvements in the recognition accuracy [1, 7]. However, the new corpora also provides an interesting test case for more detailed analyses, since the target languages are grouped into language families, and two different types of audio recordings are also introduced.

An effect similar to the language family effect can also be seen in a much smaller setting when we, for instance, look at the regional dialects of the spoken in the Northern parts of Nor-

way, Finland, and Sweden. The Sami languages cover a large geographical area in Northern Europe, and North Sami is the biggest group with about 20000 speakers. It is an official language, but the speakers are at least bilingual and also speak the dominant language of their area, i.e. Finnish, Norwegian, or Swedish. In [8], we have previously shown that the North Sami regional dialects are confounded by multiple hidden factors which significantly influence the performance of an i-vector based system. One of the factors seems to be the variation among the speakers caused by the majority language: Finnish in Finland and Norwegian in Norway. Thus, in the same way as the language family effect is observed in LRE’17, we can observe the majority language effect on the regional dialect recognition in the North Sami case. It should be noticed that North Sami and Finnish belong to the Finno-Ugric language family while Norwegian is an Indo-European language, so a language family effect might play a role here, too; however, the regional variation in North Sami is likely to be caused by the country borders and minority-majority language distinction which have reinforced the North Sami speakers to accommodate their North Sami speaking according to their second language, i.e. the majority language Finnish or Norwegian.

In this paper, we aim at a direct algorithmic approach to tackle the problem of explicitly modeling these confounding factors by addressing the challenges seen in the language family and the audio encoding effects. Since the target languages in LRE17 are grouped into language families [6], we can use them to construct a shared representation of two or more languages. By leveraging a simpler auxiliary task, i.e. identifying the language family, the recognizer efficiently learns how to learn the language characteristic. Similar approaches using multiple domains or tasks to optimize task-specific representation have shown significant improvement on the original objective [9, 10, 11]. In designing multi-task systems, it is important to avoid focusing on unwanted variability, since this could drive the attention of the algorithm away from its main task [12]. Conventional approaches to multi-task learning implicitly model these factors, and are heavily dependent on the assumption that learning them allows to model to learn beneficial shared representations [12]. Thus, we propose a solution for learning the hierarchy of tasks as well as attending to relevant patterns evolved from specific connection among tasks.

Different characteristics of audio recording and utterance are mixed together in a non-linear way, so it remains a challenge to perform speech processing task over very different types of audio collections. As in LRE17, we have a new audio encoding format, which has shown to degrade the LID performance. Prior studies have mainly focused on data augmentation, feature post-processing, score adaptation and transfer learning [1, 13] with

a certain exploration on integrating this task into a learnable classifier. As a result, we present a technique for embedding the effect of audio encoding variation into the training process of a deep neural network.

2. Staircase Methods

In this section, we present our “staircase” strategy for language learning via deep representation. We first introduce the necessary notation and give an overview of our method as well as recent advances in the field. We then delve into the details of the methods.

2.1. Background and notation

Our task is to recognize language $\mathbf{y}_{i \in \mathbb{L}}$ from a set \mathbb{L} different languages. Subsequently, the languages are grouped into language families, \mathbb{F} . Moreover, the utterances were recorded in different sessions by a multitude of recording devices, so we obtain a set of encodings \mathbb{E} , such as VAST and MLS14 in NIST LRE17 which will be introduced in Section 3.1. In general, all of the above information is provided in the training set $\mathbb{D}^t = \{(x_i, y_{i,j,k}) | i \in \mathbb{L}; j \in \mathbb{F}; k \in \mathbb{E}\}_{i=1}^{N^t}$. Also, we have a set of unlabeled data for evaluation $\mathbb{D}^e = \{(x_i)\}_{i=1}^{N^e}$. The goal is to optimize the language classifier by leveraging a series of auxiliary tasks. This process can be seen as collaborative feature learning or multi-task learning since we try to obtain the objective by using a single representation (i.e. deep neural network).

In the context of deep learning, multi-task learning can be achieved by optimizing the same set of parameters under influence of multiple objectives. There exists two common strategies for sharing those parameters. The hard parameter sharing method [10], or the multi-head network [14] shares the hidden layers between all tasks, while creating multiple “heads” (i.e. output layers) specifically for each task as illustrated by Fig 1 (b). As suggested in [10], hard parameter sharing can greatly reduce the risk of overfitting, and at the same time learning beneficial representation from multiple tasks simultaneously. The soft parameter sharing have been proposed in [11], and the design is described in Fig 1 (c). The algorithm maintains multiple instances of the same network for each task. In order to enforce the learning of shared representation, additional regularization is added to minimize the distance of the coordinated parameters among the networks. As a result, the objective becomes overly complicated as the number of tasks exceeds three.

Since both of these techniques implicitly model the multi-task case, meaning that the algorithm is heavily based on the assumption that anything learned from the extra tasks will benefit the main task. As a result, they deliberately ignore the relation among the targets within each task, and the structural relation between the auxiliary tasks and the main task. They introduce the risk of modeling irrelevant pattern as well as lessen the important role of the main objective. Our present work explores solutions to these problems in two directions:

1. Predicting relevant characteristics (i.e. language family) to the main task (Section 2.2).
2. Jointly learning the closely related task (i.e. utterances’ encoding) (Section 2.3).

2.2. Hierarchical attentive units (HAUs)

The intuition behind learning the language families is that if a language belongs to e.g. Arabic languages, two things must be enforced:

- it must not belong to any other family (e.g. Chinese, Iberian, ...),
- the language could only be one of the given languages within the set of all Arabic languages.

However, modeling such correlation requires the network performing reasoning, for instance, if the language is likely to be Arabic, then, it should prioritize predicting 4 different Arabic dialects (e.g. Egyptian, Iraqi, Levantine, Maghrebi). Implementing an “if-then-else” logic in a differentiable system while preserving the end-to-end design of deep learning is problematic.

Inspired by the approach [15], we propose a hierarchical prediction system that is able to attend to a subset of predictive neurons during inference. We tie the connection between the language and its family by adding the families’ logit values to the appropriate languages (Fig 1 (a)). These values weigh up the relevant languages within the chosen group, and weighs down all the other languages.

In order to realize this idea, we design two linear output layers: \mathbf{f}_F for the language families, and \mathbf{f}_L for the target languages. Both layers use softmax activations to predict the language family and target language accordingly for each utterance. The first network procedures the logit values by ℓ_F :

$$p(\mathbf{z}_F | \mathbf{X}, \theta_F) = \text{softmax}(\ell_F)$$

where \mathbf{z}_F refers to the predicted probabilities for each language group and θ_F is the set of all parameters used for \mathbf{f}_F . Subsequently, we weigh the language prediction logits ℓ_L by the following formula

$$p(\mathbf{y}_L | \mathbf{z}_F, \mathbf{X}, \theta_F, \theta_L) = \text{softmax}\left(\bigoplus_{j \in \mathbb{F}} (\ell_L^{(j) \forall i \in \mathbb{L}: F(i)=j} + \ell_F^{(j)})\right), \quad (1)$$

where \bigoplus is a concatenating operator for a sequence of vector, $F(\cdot)$ is the operator extracting the language family of a given language, and θ_L is the set of all parameters used for \mathbf{f}_L . Eq. (1) can be explained as follows: “for each language family, add its logit value to the appropriate logit units of all the languages of the family in the language classifier”.

As it is important to penalize the network more for making a wrong decision on an easy task, we use η to balance the contribution of the two objectives: language grouping and language recognizing. The final training cost is

$$\mathcal{J} = \eta \cdot \mathcal{D}(\mathbf{z}; p(\mathbf{z}_F | \mathbf{X}, \theta_F); \mathbf{W}_F) + (1 - \eta) \cdot \mathcal{D}(\mathbf{y}; p(\mathbf{y}_L | \mathbf{z}_F, \mathbf{X}, \theta_F, \theta_L); \mathbf{W}_L), \quad (2)$$

where $\mathcal{D}(\cdot)$ is the discriminative loss function, $\mathbf{W}_{(\cdot)}$ is a specific weight for each training example initialized to be 1.0 by default. The η value should be greater than 0.5, which emphasizes that a bigger weight is put on clustering a given language to the right language family. It should be noted that the labels of language family are only provided during training, and the network is *self-attentive*, i.e. using its own group prediction for weighting relevant languages during testing.

Additionally, θ_F and θ_L could be an overlapping set of parameters. However, our experiments have shown that a deeper architecture should be used for more complicated task (i.e. predicting the actual language), and a shallower network can be used for simpler task (i.e. predicting language family). As a result, we propose the design Fig 1 (a) for language identification on LRE17.

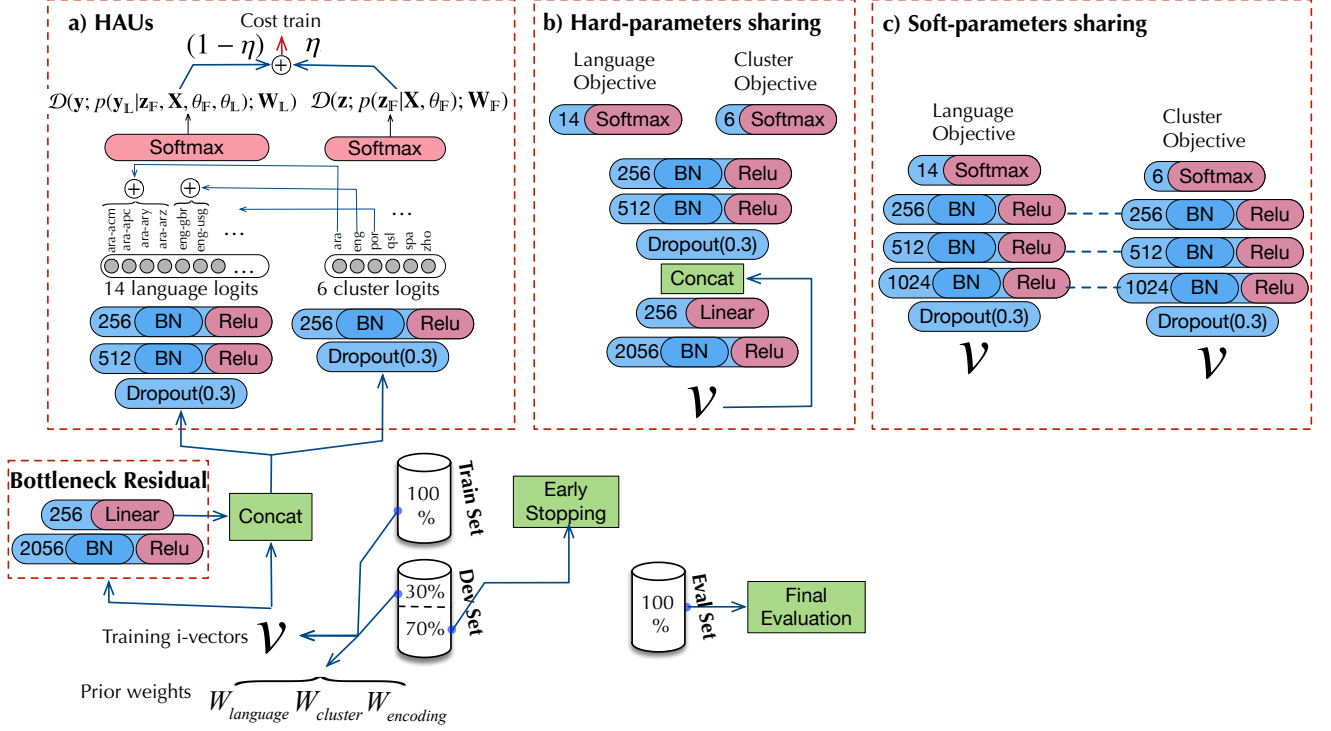


Figure 1: Illustration of our strategy fine-tuned for the NIST LRE17, also present the design of, hierarchical attentive units (HAUs) in comparison to hard parameters sharing and soft parameters sharing approaches.

2.3. Cost adaptive objective

After the training, we observed high variation of the performance among different utterance encodings. The issue can be traced to the imbalance in the utterance distribution between encodings, which has a strong negative impact on the network generalization performance [16, 17]. Specifically, each training step can drive the network to a different sub-optimal solution created by the dominant classes [16]. Since deep learning, in general, can be seen as an automatic feature learning algorithm [18], the network should adapt its representation for modeling the language pattern in all encodings. We introduce a cost-sensitive objective function to let the network learn better adaption among the encodings. The idea here is that if the objective can balance the importance of each encoding internally, the network will be able to focus on all classes equally, and as a result, it could learn crucial language characteristics from all encodings.

The modified version of cross-entropy [17] (Bayesian cross-entropy - BCE) takes into account the prior distribution of the training set, and scales the loss value appropriately for each class

$$\mathcal{D}(\theta|(X, y)) = -\frac{1}{Kn} \sum_{i=1}^n y_i * \frac{\log(f(x_i, \theta))}{\Pr(y_i)},$$

where n is the number of training examples, K is the number of classes, and $\Pr(y_i)$ is the prior probability of class y_i . We then estimate the prior probabilities based on the number of samples from each class in the training set, and use these priors for cal-

culating the weights for each example during training:

$$Pr_i = \frac{n_i}{\sum_{j=1..C} n_j},$$

$$w_i = \frac{\max_{j=1..C} (Pr_j)}{Pr_i},$$
(3)

where C is the number of classes (in LRE17: 14 for language, 6 for language family and 2 for encoding), and n_i is the number of classes within the training set. The w_i is the weight for i^{th} classes, which is the inverted ratio of each class compared to the most dominant label. However, this value could be huge due to some extreme cases, and consequently, makes the network over-concentrated on the minority classes. We rescale the weights using

$$w_{\min} = \min_{j=1..C} (w_j),$$

$$w_{\max} = \max_{j=1..C} (w_j),$$

$$w_i = \frac{(x_{\max} - x_{\min}) * (w_i - w_{\min})}{w_{\max} - w_{\min}} + w_{\min},$$
(4)

where $[x_{\min}, x_{\max}]$ are the lower and upper bounds of the weights and are chosen experimentally. We selected the following values $[0.1, 8.0]$. Using the strategy, we estimate the weights for different audio encoding (VAST and MLS14)- W_{encoding} , language- W_{language} , family- W_{cluster} (Fig 1-bottom). The weight for each task (i.e. recognizing language and language group) is given as the combination of task-specific prior weights and the channel prior weights, as follow

$$\begin{cases} W_F = W_{\text{encoding}} + W_{\text{cluster}} \\ W_L = W_{\text{encoding}} + W_{\text{language}} \end{cases}$$
(5)

As a result, the training weight for each example is selected according to each task. Because the difference between encoding is significant, the network matches the weight of each sample to its appropriate encoding without the need for external labels. It has been shown in our experiments Sec 3.1 that modeling encoding directly causes performance deficit. Since recognizing encoding is not as closely related task as identifying language family, the approach saturates the main objective.

3. Experiments

In this section we validate the presented methods in Section 2 by two experiments:

- First, we leverage the two extensions for improving the classifier backend of LID system on the large NIST corpus.
- Second, we extend the applicability of the method to boost the performance of end-to-end dialect recognition for under-resourced languages using a North Sami corpus. The experiment highlights the crucial role of modeling confounding factors in case the external information is limited and unavailable.

3.1. Improving backend on NIST LRE17

3.1.1. NIST LRE17 corpus

The 2017 NIST language recognition evaluation (LRE17) is the most recent effort to advance the research in LID. The challenge as described in the evaluation plan [6] builds on the history of the LRE campaigns, and it shares many features with the previous challenges. However, there are two major differences that propose challenges to the community.

- The inclusion of *VAST* utterances in the development set and the evaluation set. The audio were extracted from video data using encoding and channel variation which were very different from the traditional telephony speech in *MLS14* corpus.
- Normalized cross-entropy (C_{norm}) is used as performance metrics. The evaluation process calculates C_{norm} for each language under two assumed prior probabilities $P_{true} = 0.5$ and $P_{true} = 1.0$. The final score is the average of all these values and provided in our experiments.

Our strategy for splitting the dataset is presented in Fig 1. The goal is to evaluate the ability of each algorithm in domain adaptation, i.e. match the performance on both *MLS14* and *VAST* utterances. As a result, we limited the amount of *VAST* exposed to the training algorithm by randomly selecting only 30% of development set for the training set. The remaining part, named as *validation set*, is then for early-stopping, hyper-parameters tuning, validation and alternative evaluation. We remind that the evaluation data is untouched and only presented during scoring phase. There are 17425 files for training, 2440 files for validation and 25449 files for evaluation.

3.1.2. Bottleneck features based i-vector

We use state-of-the-art bottleneck-based i-vector extractor [1, 4] as the basis of our LID systems. A bottleneck DNN was trained using the 13-dimensional MFCC features with concatenated delta and delta-delta coefficients extracted from the *Switchboard-1* and *Fisher* corpora (≈ 2000 hours). The features

were then applied per utterance mean and variance normalization, then stacked by 10 past and future frames to form 21 contextual acoustic input. The seven hidden layers have 2048 units each and the bottleneck layer, which has 80 units, is placed 2 layers before the output layer. We use ReLU activation followed by a re-normalization that scales the activations RMSE to 1.0, with the exception of the bottleneck layer, which only has re-normalization nonlinearity. The output layer has 8700 targets corresponding to a set of 8700 senones obtained from the baseline speaker-independent GMM-HMM system. The 80-dimensional bottleneck features are used for extracting the i-vectors. An energy-based voice activity detection (VAD) technique was applied to the raw bottleneck features to exclude the silence frames. The voiced frames were then used to train a universal background model (UBM) with 2048 Gaussians with diagonal covariances. The diagonal UBM was used to train the total variability matrix and extract the 400-dimensional i-vectors.

Within-class covariance normalization (WCCN) [19] was used to compensate unwanted intra-class variation in the total variability space [20]. We then use *linear discriminant analysis* (LDA) to project the i-vectors onto a sub-space where inter-dialect variability is maximized and intra-dialect variability is minimized. This post-processing is repeated for all experiments with the backend to ensure the comparable results.

3.1.3. Baseline backend classifiers

We present the three most common i-vector backends: cosine scoring, support vector machine (SVM) [21], and multi-class logistic regression (MCLR), together with various configuration for the deep classifier.

Given two i-vectors \mathbf{w}_{test} and \mathbf{w}_{target}^l for the language l , cosine similarity score t is computed as follows:

$$t = \frac{\mathbf{w}_{test}^T \mathbf{w}_{target}^d}{\|\mathbf{w}_{test}\| \|\mathbf{w}_{target}^d\|} \quad (6)$$

The representation of l^{th} target language, \mathbf{w}_{target}^d , can be simply obtained by taking the average of the corresponding i-vectors in the training set. This score is calculated for all target languages, and we identify the language by the one with the highest degree of similarity.

SVM as a language classifier involves two issues: the algorithm must handle multi-classification task and deal with the non-linearity of speech and language representation. As a result, we train a multi-class SVM according to a one-vs-one scheme. We empirically selected radial basis function (RBF) kernel after it outperformed other options (linear, polynomial and sigmoid kernel). The optimization process relies upon a maximum margin concept, the focus of SVM is to model the boundary between the positive and the negative classes as opposed to a traditional density estimation approaches.

Direct discriminative training of the languages can be performed via MCLR. The backend leverages gradient descent to minimize the log-likelihood of its predicted probability values to the target language distribution in the training set. We also apply the most recent advances in training deep neural networks to obtain better MCLR solutions. We use mini-batch training, and the adaptive learning rate method (*ADDELTA*) [22] is used instead of the conventional stochastic gradient descent. Implicit regularization [23] via early stopping [24] is used to increase generalizability. The criterion is *generalization loss* [24] and the learning rate decay is 0.96. Whenever the network drops its

validating score we rollback the parameters to the best checkpoint.

3.1.4. Results

Table 1 compares the performance of different classifiers on the LRE17 corpus. All hyper-parameters of each method are fine-tuned using validation set. Our proposed approach (i.e. HAU Section 2.2) outperforms other methods by a significant margin. The performance on both validation set and evaluation set are closely matched. However, HAU performed better on validation set than on the evaluation set. This could be a sign of overfitting or that latent representation has been learned effectively. Since the performance on the evaluation set has also improved, the second explanation is more plausible.

Table 1: Comparison, in terms of the actual cost (C_{primary}), between different backends on the LRE17 corpus.

Methods	Validation set	Evaluation set
Cosine scoring	22.15	21.48
SVM	17.57	17.55
MCLR	18.63	18.52
HAUs (Sec. 2)	16.29	16.88

Table 2: Comparison, in terms of the actual cost (C_{primary}), between other deep neural network architecture variants on the LRE17

Methods	Validation set			Evaluation set		
	mls14	vast	avg	mls14	vast	avg
Single task	15.21	21.01	16.77	17.68	19.61	18.05
Multitask-hard	15.01	19.38	16.19	17.59	18.37	17.66
Multitask-soft	15.25	21.02	17.06	18.37	19.82	18.64
No HAUs	15.19	18.89	16.01	17.61	18.76	17.80
No BCE	18.70	23.82	20.06	19.65	23.25	20.08
HAUs (Sec. 2)	15.12	19.88	16.29	16.65	18.35	16.88

Table 2 delves deeper into the question "why does HAU work?". First, we validate the performance of a fine-tuned single task network instance. The performance gap between validation and evaluation set shows a strong sign of overfitting. The same observation applies to both multi-task systems with either hard or soft parameter sharing, and our design with HAUs objective removed. In the multi-task learning, the loss surface dimensions are exponentially increased with the number of tasks, and as a result, the searching for the state of equilibrium becomes difficult without sufficient amount of training data. Subsequently, the learning of the added tasks could diminish learned representation that benefits the main task. On the other hand, removing HAUs extension while keeping the adaptive objective forces the algorithm to improve VAST on the validation set, but fails to transfer this improvement to the evaluation set.

Finally, omitting the proposed objective (i.e. "No BCE") significantly degrades the performance due to the negative impact of the imbalance in the training set proportions. We see that the performance on VAST is degraded significantly, whereas HAUs keeps the model generalizing well and matching similar performance on both validation and evaluation set. We can see that applying both extensions from Section 2 yields the highest performance single system for the BNF i-vector input.

Although HAUs introduces another hyper-parameters η (Eq. 1), Table 3 highlights the robustness of HAUs. With all

Table 3: The performance of HAUs on evaluation set given different η values (Eq. 1).

η	0.3	0.4	0.5	0.6	0.7	0.8
C_{primary}	16.98	16.98	16.90	16.88	16.96	17.18

selected values of η , HAUs still outperforms other backend classifiers and variants from Table 1 and Table 2. The best choice for η ranges from 0.5 to 0.7 which matches our intuition and suggestion in Sec. 2.2, that the easier task should be penalized more during the training process.

3.1.5. Further analysis

Fig 2 illustrates the differences on the learned representations among the three datasets. The plots are acquired by training PCA on the training set followed by LDA, then applying the trained model on validation and evaluation sets. All data points from the same language family have the same marker style, and each language is represented by different colors. While the first and second representations are highly separable by language family on the training set, they encapsulate strong confusion region marked by the red circle. The region contains data points densely populated within a small area. It is also notable that PCA-LDA failed to generalize to the validation and evaluation sets since the samples are all mixed together regardless of their language group.

Deeper classifier (i.e. "No HAUs") is able to separate the language family within its representation, but the confusion remains high in the center. However, the learned representation by HAUs is spread over a wider area (directions denoted by red arrows), while the algorithm maintains the same margin among language families. This can be understood as HAUs having been able to learn a richer language family space. As a result, it is able to reduce the confusion by not grouping the languages into big clusters only, but also into mini-clusters.

3.2. End-to-end approach for under-resourced corpora

We apply Bayesian cross-entropy objective and HAUs for an end-to-end approach to dialect recognition on under-resourced corpora. However, the objective heavily relies on the assumption that our training set encapsulates the same distribution as the population. The assumption is unlikely to be met in the scenarios of a small dataset, where we only observe an insufficient small number of samples. As a result, we use mini batch statistics to aggregate the prior probability of each class, $\Pr(y_c) = \sum_i^n \mathbb{1}(y_i = c)/n$. This approach has been proven to stabilize the gradients and leads to better results in our experiments.

3.2.1. DigiSami corpora

The DigiSami corpora [25, 26] contain two North Sami corpora recorded in the locations of three villages in Finland: Utsjoki, Inari and Ivalo, and two villages in Norway: Kautokeino and Karasjoki. The participants were invited to take part in two different tasks: discussion and reading aloud existing Wikipedia texts, and taking part in a free conversation. Table 4 shows the distribution of the two collected corpora. There are 10 men and 18 women. Their age ranges from 16 to 65 years: 17 were 16-21 years old, five 30-44 years old, and six 49-65 years old. The participants are native speakers of North Sami and are also bilingual with Finnish or Norwegian depending on their home

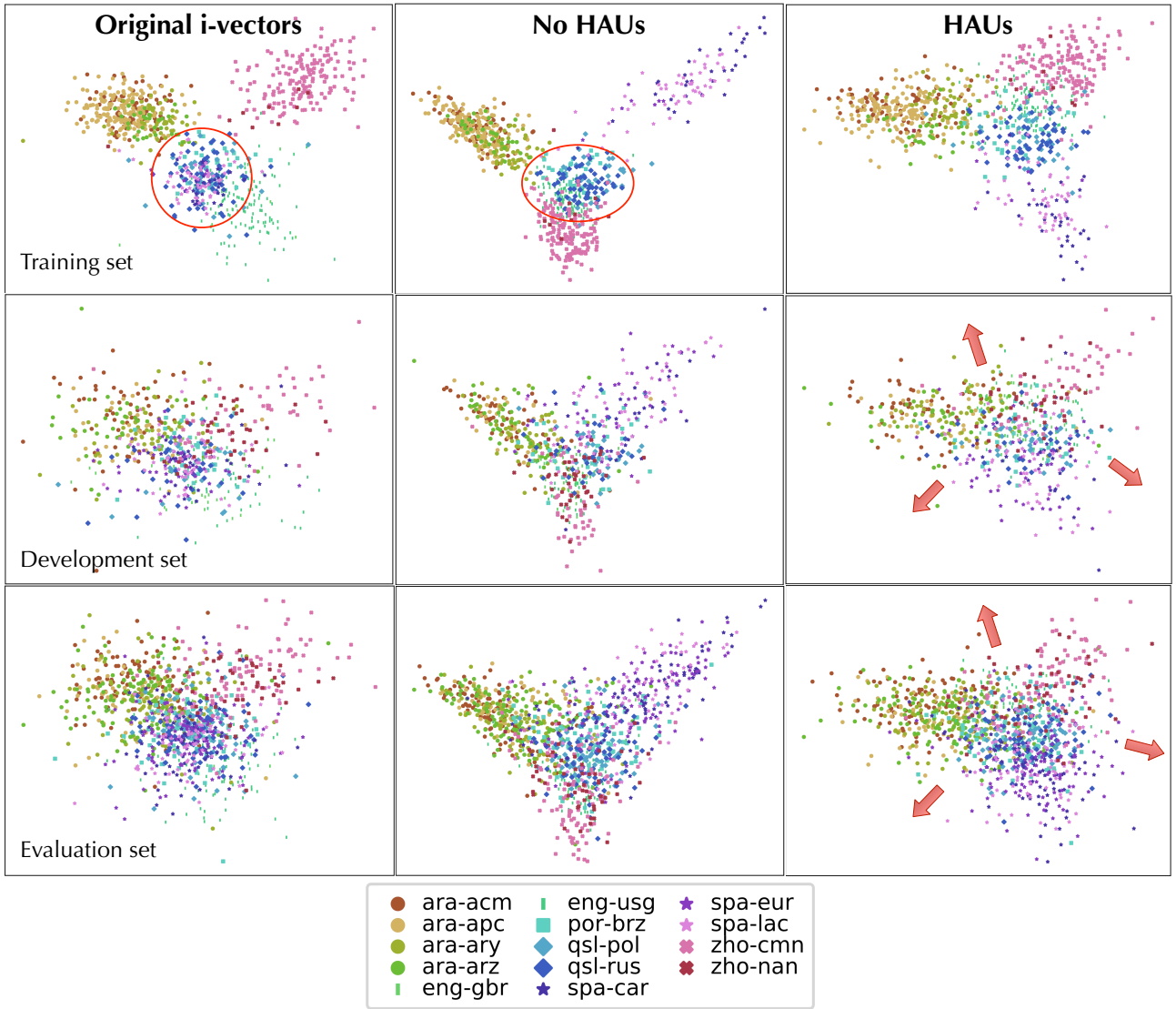


Figure 2: Comparison of learnt representation on training, validation and evaluation sets among the three learnt representations: original i-vectors, last hidden activation from the “No HAU” network, last hidden activation from the “HAUs” network.

area.

Table 4: The two corpora statistics, the numbers in the brackets indicate number of new speakers in Conversational corpus.

Dialects	Read speech		Conversational speech	
	#Spk	Hour	#Spk	Hour
Kautokeino	4	1.03	-	-
Karaskjoki	6	0.72	6(1)	1.5
Ivalo	6	0.72	7(1)	0.72
Utsjoki	5	1.07	6(1)	1.03
Inari	4	0.73	-	-
Total	25	3.26	19	4.28

3.2.2. Baseline systems

We use the state-of-the-art i-vector approach as described in Section 3.1.2. To compensate for the lack of sufficient development data, we first trained the UBM and the T-matrix using the

Finnish language *PERSO* corpus [27] corresponding to 30916 utterances in total. The backend is replaced by *heteroscedastic linear discriminant analysis* (HLDA) [20] and the cosine similarity is used subsequently.

We use the combination of CNN, LSTM and feedforward neural network as described in [16]. The network design is specifically fine-tuned for our task and re-scaled to match the size of the North Sami corpora. The fine-tuning strategy is gradually augmenting the network architectures by increasing the number of layers following by a number of hidden units [16]. The network consists of two convolutional layers with 32 and 64 features map of 3×3 kernel, then follows by a bidirectional LSTM of 256 units each. Finally, we include a projection layer of 512 units and an output layer.

3.2.3. Results and analysis

As discussed in [28], all the experiments are repeated three times to minimize the effect of random initialization on a small

dataset, the final reported numbers are the mean and the standard deviation of these experiments. The whole conversational corpus is left out for evaluation since we want to validate the generalization ability under different session and recording. We use “leave-one-speaker-out” (LOSO) strategy [20], i.e. one speaker from each dialect is randomly selected for validation process, and another speaker is selected for testing. The remaining set of speakers is used for training our classifier. As the scheme used in [8], the results are comparable to previous i-vector approach [8]. It is notable that each repetition is done with a totally different set of speakers in order to preserve the generalizability of the experiments.

Table 5 emphasizes the importance of adaptive objectives and HAUs in improving classifier performance in case of under-resourced languages. Moreover, our best network significantly outperforms the i-vector system.

Table 5: Results reported in term of $C_{\text{avg}} \times 100$

Methods	Read speech	Conv. speech
cross-entropy	11.66 \pm 1.66	13.56 \pm 0.99
BCE	10.40 \pm 1.03	10.53 \pm 2.43
BCE + HAUs	8.73 \pm 3.24	6.81 \pm 0.58
i-vector [8]	17.79	–

Even though the HAUs objective has been applied to a complicated architecture, it was able to encapsulate the effect of majority language in the network architectures using only supervised data. This opens up the potential of using end-to-end deep learning approach for under-resourced corpora.

4. Conclusions

In the present study, we attempted to model jointly, as an auxiliary task, the effect of language family in the audio encoding. We observed the potential benefit of modeling hierarchical structures implicitly in the learning algorithm, and presented a novel approach to integrate meta-information learning into deep classifier, which outperformed the common approaches to multi-task learning. We also demonstrated that by modifying the training objective that takes into account the prior information of multiple internal factors will help the learned model to generalize better. Moreover, we studied how the method’s applicability can be extended to boost end-to-end dialect recognition for under-resourced languages like North Sami. As a future work, we plan to investigate how to lessen the effect of supervision in our model, as the supervised model can too easily lack robustness against unforeseen dataset drift, as observed for example in the LRE15 evaluation set. Consequently, we will focus on using semi-supervised learning to explicitly model channel variation.

5. Acknowledgments

This research was partly funded by the Academy of Finland (grant #313970). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

6. References

- [1] Kong Aik Lee, Haizhou Li, Li Deng, Ville Hautamaki, and et al., “The 2015 nist language recognition evaluation: the shared view of i2r, fantastic4 and singams,” *Interspeech*, 2016.
- [2] Dario Amodei, Rishita Anubhai, and Eric Battenberg et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” *CoRR*, vol. abs/1512.02595, 2015.
- [3] H. Behravan, V. Hautamaki, S.M. Siniscalchi, T. Kinunen, and Chin-Hui Lee, “i-vector modeling of speech attributes for automatic foreign accent recognition,” *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 1, pp. 29–41, Jan 2016.
- [4] Fred Richardson, Douglas A. Reynolds, and Najim Dehak, “A unified deep neural network for speaker and language recognition,” *CoRR*, vol. abs/1504.00923, 2015.
- [5] NIST, “The 2015 nist language recognition evaluation plan,” 2015.
- [6] NIST, “Nist 2017 language recognition evaluation plan,” 2017.
- [7] Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez, “Language identification in short utterances using long short-term memory (lstm) recurrent neural networks,” *PLOS ONE*, vol. 11, no. 1, pp. 1–17, 01 2016.
- [8] Kristiina Jokinen, Trung Ngo Trong, and Ville Hautamäki, “Variation in spoken north sami language,” in *Interspeech 2016*, 2016, pp. 3299–3303.
- [9] Orhan Firat, KyungHyun Cho, and Yoshua Bengio, “Multi-way, multilingual neural machine translation with a shared attention mechanism,” *CoRR*, vol. abs/1601.01073, 2016.
- [10] Richard Caruana, “Multitask learning: A knowledge-based source of inductive bias,” in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 41–48, Morgan Kaufmann.
- [11] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook, “Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 845–850, Association for Computational Linguistics.
- [12] Sebastian Ruder, “An overview of multi-task learning in deep neural networks,” *CoRR*, vol. abs/1706.05098, 2017.
- [13] Dong Wang and Thomas Fang Zheng, “Transfer learning for speech and language processing,” *CoRR*, vol. abs/1511.06066, 2015.
- [14] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, 10 2017.
- [15] Frederic Morin and Yoshua Bengio, “Hierarchical probabilistic neural network language model,” in *AISTATS05*, 2005, pp. 246–252.
- [16] Trung Ngo Trong, Ville Hautamaki, and Kong Aik Lee, “Deep language: a comprehensive deep learning approach to end-to-end language recognition,” *Odyssey: the Speaker and Language Recognition Workshop*, 2016.

- [17] Alexandre Dalyac, Prof Murray Shanahan, and Jack Kelly, “Tackling class imbalance with deep convolutional neural networks,” *Thesis, Imperial College London*, 2014.
- [18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 05 2015.
- [19] Andrew O. Hatch, Sachin Kajarekar, and Andreas Stolcke, “Within-class covariance normalization for svm-based speaker recognition,” in *Proc. of ICSLP*, 2006, p. 14711474.
- [20] Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen, “Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish,” *Speech Communication*, vol. 66, pp. 118–129, 2015.
- [21] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*, Cambridge University Press, New York, NY, USA, 2000.
- [22] Matthew D. Zeiler, “ADADELTA: an adaptive learning rate method,” *CoRR*, vol. abs/1212.5701, 2012.
- [23] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” *ArXiv e-prints*, Nov. 2016.
- [24] Lutz Prechelt, *Neural Networks: Tricks of the Trade: Second Edition*, chapter Early Stopping — But When?, pp. 53–67, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [25] Kristiina Jokinen, “Open-domain interaction and online content in the sami language,” *Proceedings of the Language Resources and Evaluation Conference (LREC-2014)*, June 2014, Reykjavik, Iceland.
- [26] Kristiina Jokinen, Katri Hiovain, Niklas Laxström, Ilona Rauhala, and Graham Wilcock, *DigiSami and Digital Natives: Interaction Technology for the North Sami Language*, pp. 3–19, Springer Singapore, Singapore, 2017.
- [27] H. Kallio, “Perso databases for finnish speech synthesis,” *Institute of Behavioural Sciences, University of Helsinki*, 2013.
- [28] Trung Ngo Trong, Kristiina Jokinen, and Ville Hautamäki, “Enabling spoken dialogue systems for low-resourced languages: End-to-end dialect recognition for north sami,” in *Proceedings of the 9th International Workshop on Spoken Dialogue Systems Technology*, Singapore, 2018.