Enabling Spoken Dialogue Systems for Low-Resourced Languages – End-to-End Dialect Recognition for North Sami

Trung Ngo Trong[1], Kristiina Jokinen[2], Ville Hautamäki[1]

[1] University of Eastern Finland, Joensuu, Finland
firstname.lastname@uef.fi

[2] AI Research Center, AIST Tokyo Waterfront, Japan
firstname.lastname@aist.go.jp

Abstract In this paper, we tackle the challenge of identifying dialects using deep learning for under-resourced languages. Recent advances in spoken dialogue technology have been strongly influenced by the availability of big corpora, while our goal is to work on the spoken interactive application for the North Sami language, which is classified as one of the less-resourced languages spoken in Northern Europe. North Sami has various variations and dialects which are influenced by the majority languages of the areas in which it is spoken: Finnish and Norwegian. To provide reliable and accurate speech components for an interactive system, it is important to recognize the speakers with their Finnish or Norwegian accent. Conventional approaches compute universal statistical models which require a large amount of data to form reliable statistics, and thus they are vulnerable to small data where there is only a limited number of utterances and speakers available. In this paper we will discuss dialect and accent recognition in under-resourced context, and focus on training an attentive network for leveraging unlabeled data in a semi-supervised scenario for robust feature learning. Validation of our approach is done via two DigiSami datasets: conversational and read corpus.

1 Introduction

Recent advances in dialect and accent recognition have been strongly influenced by the availability of big corpora for popular languages. Dialogue technology applications for major languages are widely available, but for many languages this is not the case: there is no commercial interest in developing speech technology nor corpora or resources to enable further development in interactive applications. It is thus important to study how to create such technology for languages which are under-resourced in that the available data is not large or it is not in the digital for-

mat ready to be used for technical applications. The lack of digital "presence" can be threatening to such languages as the speakers are forced to compromise and trade off their native language for languages which provide better and wider communication with the world and the society. The recent interest in revitalizing such languages has initiated both general and scientific effort to collect and develop tools and applications for less-resourced languages (Crystal 2000; Besacier et al. 2014; Jokinen 2014). From the research and development point of view, the focus has been on enabling technology which allows creating applications and technology given low data resources and shortage of staff, bringing in questions of how to best address basic speech technology needs with minimum effort available.

Robot systems are getting popular as communicating devices, and there is a wide range of applications from small sociable robotic devices to advanced interactive applications such as WikiTalk (Wilcock and Jokinen 2014) and ERIKA (Kawahara et al. 2016) which enable natural language human-robot dialogues. Much research goes into development of speech technology, but applications in social robotics with the human-centered view as the core concept to support natural interaction capabilities have gradually been brought into the center of dialogue system development. The main hypothesis is that the more engaging the interaction is in terms of communicative competence, the better results are obtained by the system in terms of enjoyment and reliability. Concerning under-resourced languages, new technology can play a pivotal role in boosting revitalization of threatened languages, as the language users can see the language as a meaningful and useful part in the globalized and technologized world, cf. Ó Laoire (2008).

Multilingual aspects have been addressed by Laxström et al. (2016), who point out the need for software localization and internationalization to produce systems that can be used in different language and cultural contexts. Multilingualism relates to the need to cater for speakers with different dialects and speech accents so as to allow the spoken interaction to proceed smoothly. In real world applications, dialect and accent recognition contributes to the performance of speech processing systems, and the task has attracted increased attention in the speech community. Spoken dialogue systems are typical applications which can benefit from enhanced multilingualism: if they can recognize the user's language preferences, they can customize the interface to digital services accordingly (Dehak et al. 2011). The models can enhance the performance of processing tasks such as SR and ASR, and contribute positively to the performance of the whole system, and consequently, to the user experience and evaluation.

In our previous work, we studied North Sami spoken variation in Finland and in the Finnmark area in Norway (Jokinen et al. 2016). We chose North Sami as our target language, since it is the object of research in the DigiSami project (Jokinen et al. 2017), which concerns speech and language technology to support small Finno-Ugric language communities. North Sami is an official language in the six northernmost counties in Norway, and legally recognized in Finland and Sweden. The Sami speakers are at least bilingual and can also speak the majority language of the country they live in (Norwegian, Swedish or Finnish), while North Sami is used as a lingua franca among the Sami people (Jokinen et al. 2017). The speaker's

country of origin can be fairly easily distinguished based on their speaking manner. In Jokinen et al. (2016), we hypothesized that the variation in North Sami dialects is due to the majority language, rather than individual variation, i.e. that there is more variation among the speakers of North Sami who live in the different majority language locations in Norway and in Finland, than among the speakers who live in different location within the same majority language context.

In this paper we continue the work on language change, assimilation, and dialect variation based on the North Sami data. We address the challenge of identifying minority dialects in a restricted data context using deep learning. Our approach augments neural architecture to form a robust and consistent dialect representation from a small corpus, while keeping an end-to-end design to maximize the potential application to similar problems. Specifically, the experiments establish two crucial situations recognizing minority languages. First, the set of samples is available only for development and consists of a small number of samples from different sources which were gathered from distinct contexts and speakers. Second, a partial set of dialects is presented during the training process, and the algorithm performs semi-supervised learning to efficiently recognize new dialects during the test time. We validate our approach through systematic experiments on two DigiSami datasets: conversational and read corpus. Our experimental findings are corroborated by outperforming the recent state-of-the-art i-vector approach.

The paper is structured as follows. We discuss the DigiSami datasets and visualise their properties in Section 2. We describe the Deep Learning method focusing on its use for small corpus dialect recognition in Section 3. We present our experiments and results in Section 4, and conclude with future views in Section 5.

2 DigiSami datasets

Availability of large corpora in speech processing has been one of the major driving forces advancing speech technologies. This has allowed recent state-of-the-art systems to obtain impressive performance in recognizing spoken languages (Lee et al. 2016, Li et al. 2013, Amodei et al. 2015). As for under-resourced languages, research is carried out in several projects and initiatives which focus on data collection and speech technology development. Most of the development for those languages have been concentrated on two directions: bootstrapping the system using adaptation of pre-trained model (Thomas et al. 2013), and introducing closely related ``out-of-languages'' data (Besacier 2014). These approaches require additional corpus, which is problematic for under-resourced languages.

The DigiSami project (Jokinen et al. 2017) aims to study the effect of digitalisation on small endangered languages, and to support visibility and revitalisation of Finno-Ugric language communities by creating digital content as well as developing language and speech technology tools, resources, and applications that can be used for automatic speech and language processing. The project focuses on the North Sami language, the largest of the Sami languages with about 20000 speak-

ers, and explores various spoken language issues (speaker identification, multimodal conversation analysis, laughing), with the challenging goal of demonstrating viability of an interactive dialogue system in the North Sami language, SamiTalk, following the multilingual open-domain robot application WikiTalk (Wilcock et al. 2017). WikiTalk is an interactive robot application that enables users to find out more about subjects that interest them by discussing with the humanoid robot. They can navigate through the Wikipedia articles, ask for more information on interesting subjects, and get the robot to read the related Wikipedia article for them (Jokinen and Wilcock, 2012, 2013).

The project organized Sami language data collection and Wikipedia article writing through series of community events in the central Sami speaking areas. The participants took part in three different tasks: discussion and writing Wikipedia articles, reading aloud of existing Wikipedia texts, and taking part in a free conversation which was video recorded. Locations were selected to represent different North Sami dialects, and consisted of three villages in Finland: Utsjoki (Ohcejohka), Inari (Anár) and Ivalo (Avvil), and two villages in Norway: Kautokeino (Guovdageaid) and Karasjoki (Kárásjohka). See more of the DigiSami data and data collection in (Jokinen 2014; Jokinen & Wilcock 2014).

There were 28 participants, 10 men and 18 women with age range 16-65 years. They (or their parents in case of under-aged participants) gave explicit agreement to allow the data to be used for research purposes. All the participants were native speakers of North Sami, and almost all (26) reported using North Sami daily. All participants were bilingual and spoke either Finnish (Utsjoki, Ivalo, Inari), or Norwegian (Kautokeino and Karasjoki). Most participants had lived their life in the Sápmi area, although not in the same place. Ten participants had also lived in bigger cities in the southern part of the area for a short period of time.

The read speech part of the corpus contains speech samples from 28 participants who read Wikipedia articles written in North Sami. The conversational corpus has eight casual conversations with two or three participants, and the topics vary from everyday life (next vacation, driving school, cars) to translation between Sami and other languages and to technological tools that have been made to help writing North Sami more correctly. Annotation of the corpus was done with Praat and consists of 5 time-aligned tiers: a phonological/phonetic transcription, the words, the sentence in orthographic form, a Finnish translation, and remarks on things like dialectal variation. Conversations are also annotated with respect to topics and laughter (Hiovain and Jokinen 2016). The corpus has been made available for general use through the CSC website.

Overview of the datasets is given in Table 1, and Figure 1 visualizes the closeness of the dialects in the read and conversational corpora. There is a separation between read speech and conversational speech. We observe that conversational speech is more separated than read speech in the majority languages (Finnish and Norwegian), which can be explained by the formal mode of speaking in the read speech versus informality of the conversational speech. We also noticed occasional code-switching in conversational speech to majority language. The samples group into isolated regions, but dialects do not form clearly separated clusters. The

same dialect can form multiple clusters, which possibly represent the effect of unwanted variation (i.e. speaker variation, recording variation, etc).

Dialects	Read corpus		Conversational corpus	
	#Speakers	Duration (hour)	#Speakers	Duration (hour)
Kautokeino	4	1.03	-	-
Karasjoki	6	0.72	6	1.5
Ivalo	6	0.72	7	0.72
Utsjoki	5	1.07	6	1.03
Inari	4	0.73	-	-
Total:	**25**	**3.26**	**19**	**4.28**

Table 1 DigiSami data overview.

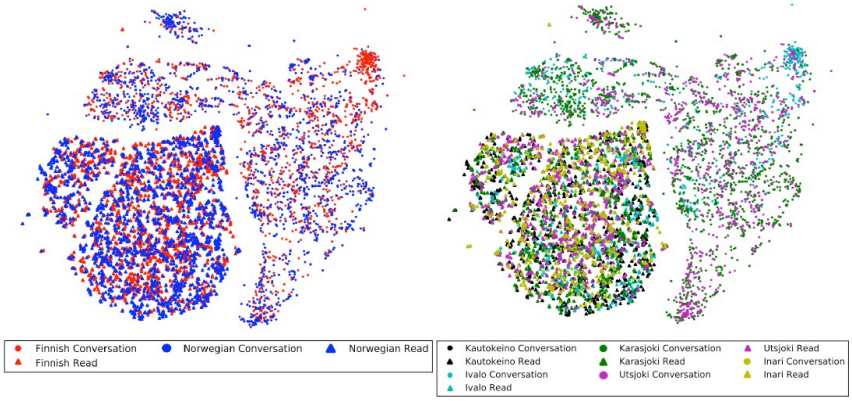

Figure 1 t-SNE visualization of MFCCs features with 10 frames for left and right context for the conversational corpus (dot) and read corpus (triangle). Left figure highlights the majority languages, Right one illustrates five different North Sami dialects.

3 Deep learning for small corpus dialect recognition

We tackle the problem discriminative learning for a small corpus in dialect recognition by addressing certain challenges. First, the complexity of speech signal is emphasized in the small dataset, with a limited number of utterances per speaker. Conventional approaches to dialect recognition compute universal background models from closely related languages (Lee et al 2016; Amodei et al. 2015; Richardson et al. 2015), and as a result, the system requires big data to compute reliable statistics. For minority languages, collecting additional data is difficult, and as the language is typically distinguishable from the prominent language groups, introducing external corpus might lead to an unpredictable bias (Thomas et al. 2013). Moreover, separating feature learning and discriminative learning can lead to an unoptimized representation for the classification objective. We thus simplified our approach by constructing an end-to-end trained deep neural network that takes into account the channel variation to learn robust dialect representation.

Second, deep networks confront two crucial issues as non-linear models. Theoretically, neural networks can approximate Bayesian posterior probabilities under the assumptions which concern accurate estimation: the number of parameters is enough, there are enough training examples, and the priori class distribution of training set must be representative for the whole data set (Zhang and Quin 2013; Morin and Bengio 2005; Kirkpatrick et al. 2017). In recent years, much effort has been directed to advance neural architectures and optimization algorithms for deep learning (Ngo Trong et al. 2016; Dalyac et al. 2014; Zhang et al. 2016; Prechelt 2012). However, the issue concerning small, imbalanced datasets remains open. Inspired by the new perspective in understanding generalizability of deep network (Sainath et al. 2014), we employ the implicit regularization techniques to directly combat overfitting within the network architectures.

3.1 Supervised attention for language identification

We improve the language identification accuracy by forcing the model attended to speech segments within the utterances. Since the speech was recorded in various conditions, from conversations with ambient noise to formal reading sessions. There are differences among the dialect distribution of training and evaluation data, and the limited amount of training requires that the training process extracts a more precise representation. On the other hand, conventional attention-based approaches require large datasets since the attended weights are automatically learnt together with the main task objective (Xu et al. 2015; Bahdanau et al. 2015). We introduce a supervised attention algorithm for the speech processing task as inspired from (Mi et al. 2016). We compute the distance between the machine attentions and the "true" alignments of speech segments, and integrate the cost to the LID objective. The energy-based Voice Activities Detection (VAD) is used to generate the labels to supervise the attention network. Also, soft attention mechanism (Xu et al. 2015) is implemented to handle uncertainty of the VAD labels.

3.2 Semi-supervised end-to-end learning

Fig. 2 (next page) shows the proposed architecture for semi-supervised end-to-end (SSEE) learning. The design is composed of three deep learning architectures: convolutional neural network (CNN), long short-term memory network (LSTM), and fully connected network (FNN) (LeCun, 2015). The semi-supervised function is introduced by the convolutional decoder which learns to reconstruct the original signal. We do not apply decoding process after LSTM, since the recurrent neural network learns temporal patterns, and decoding of temporal signals involves alignment of long sequences which requires additional constraints and weakens our main objective to learn a robust features representation. Moreover, the weights of the decoder and encoder are tied, and random Gaussian noise is presented in the encoder during the training process. The learning process is a joint optimization of discriminative objective and reconstruction cost balanced by the hyperparameter α

$$\alpha * \sum_{i}^{n} log(p_\theta(y_i|X_i, z_i) + (1-\alpha)E[||X_i - \hat{X}_i||_2]. \quad (1)$$

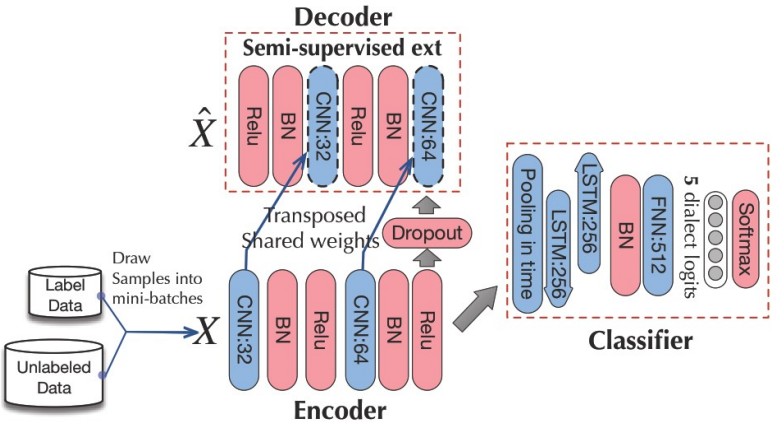

Figure 2 Architecture of a semi-supervised end-to-end dialect recognition system with the different algorithms. A fully-supervised network can be achieved by removing the decoder.

By including the second term, and the weight of encoder and decoder are tied, optimizing the objective also denoises the corrupted encoder. In order to stabilize the optimization process and enforce the network learning robust statistics from training set, we introduce batch normalization after each layer using Eq. 2 where batch normalization $\mathrm{BN}_{\gamma,\beta}$ is given by Eq. 3. As a result, the network seeks for the optimal weights and layers statistics that preserve the most representative features within its convolutional structures.

$$f_W(\mathrm{BN}_{\gamma,\beta}(X,\epsilon)) = X \qquad (2)$$

3.3 Compensate channel variances using implicit regularization

Regularization is used to prevent the learning algorithm from overfitting the training and thus boost the model's generalizability. The two main strategies for regularizing a neural network are explicit regularization and implicit regularization (Zhang et al. 2016). Explicit regularization applies prior to constraint network parameters (e.g. norm constraint, dropout) and concentrates on penalizing the algorithm from overfitting; the resulted model is neither necessary nor sufficiently generalized. Implicit regularization considers the mismatch between the training set and the population, and integrates its policies into the optimization process. Early stopping and batch normalization (BN) have shown to be effective approaches to implicit regularization. We use generalization loss (GL) as early stopping criterion (Prechelt 2012) and decrease learning rate by 2 whenever the network drops its validating score. We also modify BN to include internal noise as suggested in (Radford et al. 2015) and shown in formula (3)

$$y = f\left(\frac{(X - E[h])}{\sqrt{\mathrm{Var}(X)}}\gamma + \epsilon + \beta\right), \qquad (3)$$

where ε is a residual term that explaines the differences between training and evaluation data. It is represented as Gaussian noise, and added to the normalized input before applying activation to force BN to learn a more robust normalized

space. Conversely, the approach in Eq. 4 creates instable statistics and decelerates convergent process.

$$y = f(\mathrm{BN}_{\gamma,\beta}(X)) + \epsilon \qquad (4)$$

3.4 Bayesian cross-entropy objective

The modified cross-entropy (Dalyac et al. 2014) takes into account prior distribution of training set, and scales the loss value appropriately for each class

$$L(\theta|(X,y)) = -\frac{1}{Kn}\sum_{i=1}^{n} y_i * \frac{\log(f(x_i,\theta))}{p(y_i)}, \qquad (5)$$

n is the number of training examples, K is the number of classes, and $p(y_i)$ is the probability of class y_i given our training set. This objective heavily relies on the assumption that the training set encapsulates the same distribution as the population. Since the assumption is unlikely to be sound for small datasets, we use mini batch statistics to aggregate the prior probability of each class, i.e.

$$p(y_c) = \sum_i^n \mathbb{1}(y_i = c)/n$$

This approach has been proved to stabilize the gradients and lead to better results in our experiments. We also found out that softmax activation outperforms other activation functions (i.e. sigmoid, rectifier, tanh), and is more stable for learning imbalanced data due to its normalization term, the gradients are equally distributed to anti-model neurons. Consequently, the deep network becomes a probabilistic inference model, since the parameters define a probability distribution of discrete random variable for each class conditional on the training data.

4 Experiments in dialect recognition

All the experiments were repeated three times to minimize the effect of random initialisation, and the final reported numbers are the mean and the standard deviation of the experiments. All the audio files were down-sampled to 16 kHz and partitioned into 30 seconds chunks. Our experiments have showed that log mel-filter banks features with Δ and Δ Δ are more suitable for deep network, and the same observation has been found in (Ossama et al. 2014, Trong et al. 2016). Using RNN enables us to leverage longer temporal windows, we segment each utterance into chunks of 200 consecutive frames, and each chunk is shifted forward 100 frames to form the next sample. One may finds augmenting the data by decreasing the shift distance, however, our experiments had provided no improvement with smaller distance. As suggested in (LeCun et al. 1998), we normalize our speech frames using global mean and variance calculated from given training set.

4.1 Evaluation metrics

Results are reported in terms of average detection cost (C_{avg}) which is the mean of all the binary detection cost (C_{DET}) for each language. C_{DET} is defined as in (Li et al. 2013):

$$\begin{aligned} C_{\mathrm{DET}} &= C_{\mathrm{miss}} P_{\mathrm{tar}} P_{\mathrm{miss}}(L_a) \\ &+ C_{\mathrm{fa}}(1-P_{\mathrm{tar}}) \frac{1}{J-1} \sum_{k \neq j} P_{\mathrm{fa}}(L_j, L_k) \end{aligned} \quad (6)$$

where P_{miss} denotes the miss probability (or false rejection rate), i.e., a test segment of dialect L_i is rejected as being in that dialect. $P_{fa}(L_i; L_k)$ is the probability of a test segment of dialect L_k accepted as being of dialect L_i. The costs, C_{miss} and C_{fa} are both set to 1 and P_{tar}, the prior probability of a target accent, is set to 0.5 as in (Li et al. 2013).

4.2 Baseline system

We impose the state-of-the-art i-vector approach to LID as our baseline (Lee et al. 2015; Richardson et al. 2015), the system has been implemented for our task in (Jokinen et al. 2016). An utterance is represented using the fixed length and low-dimensional latent variable vector in the total variability space (Dehak et al. 2011). This is commonly called an i-vector, and it contains the variability in the utterance, such as dialect, speaker and the recording session. The Gaussian mixture model (GMM) supervector, M, of an utterance is represented as

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \quad (7)$$

where m is the utterance independent component (the universal background model or UBM supervector), T is a rectangular low rank matrix and w is an independent random vector of distribution N(0; I). T represents the captured variabilities in the supervector space. It is estimated by the expectation maximization (EM) algorithm similar to estimating the V matrix in joint factor analysis (JFA) (Matrouf et al. 2007), with the exception that every training utterance of a given model is treated as belonging to a different class. The extracted i-vector is then the mean of the posterior distribution of w.

As the extracted i-vectors contain both intra- and inter-dialect variability, we use heteroscedastic linear discriminant analysis (HLDA) to project i-vectors onto a space where inter-dialect variability is maximized and intra-dialect variability is minimized. In standard HLDA technique, the vectors of size n are projected into subspace $p < n$, using HLDA matrix

$$\mathbf{A} \in \Re^{n \times n}$$

Within-class covariance normalization (WCCN) is used to compensate unwanted intra-class variations in the variability space (Behravan et al. 2015). Given two i-vectors w_{test} and $w_{dtarget}$ for dialect d, cosine similarity score t is computed:

$$t = \frac{\hat{\mathbf{w}}_{\mathrm{test}}^{T} \hat{\mathbf{w}}_{\mathrm{target}}^{d}}{\|\hat{\mathbf{w}}_{\mathrm{test}}\| \left\|\hat{\mathbf{w}}_{\mathrm{target}}^{d}\right\|} \quad (8)$$

where

$$\hat{\mathbf{w}}_{\mathrm{test}} = \mathbf{A}^{T} \mathbf{w}_{\mathrm{test}} \quad (9)$$

Further, w^{d}_{target} is the average i-vector over all the training utterances in dialect d. This score a calculated for all target languages, and the dialect is identified by the highest degree of similarity. Only the dialect labels are involved in computing HLDA, hence, the system doesn't know which utterances belong to Norwegian or Finnish, and provides unbiased results concerned the effect of majority languages.

4.3 End-to-end deep learning systems

Following (Trong et al. 2016, Sainath et al. 2015), we design our network using multiple neural architectures which are complementary in their modelling capabilities to capture different patterns. Figure 2 presented the architecture and the different algorithms used.

While FNN using multiple processing layers is able to extract hierarchical representations that benefit the discriminative objective, CNN has ability to extract local invariant features in both time and frequency domain (Ganapathy et al. 2014). RNN combines the input vector x_t (i.e. t-th frames of utterances) with their internal state vector to exhibit dynamic temporal pattern in signal. As sequence-training is critical for speech processing, conventional FNN approaches have been proven inefficient in both language and speaker identification task (Lopez-Moreno 2014, Ganapathy et al. 2014). Our observation shows that DigiSami datasets contains long conversation with continual silence between each talk, hence, the frames-level features extracted by FNN can introduce extra biases and noises to the network. As a result, our algorithm focuses on adapting CNN and RNN architectures to address the difference between distribution of training and evaluation data. We further compare our networks to the approaches in (Lopez-Moreno et al. 2014, Gonzalez-Dominguez et al. 2014, and (Ganapathy et al. 2014), where the networks' hyperparameters (i.e. number of layers, number of hidden units, activation function, and parameters initialization) are fine-tuned for our task. The designs of the networks are shown in Table 2.

Network	Design	# of parameters
FNN (1)	FNN(2560-2560-1024)	2.8×10^6
CNN (2)	CNN(32-64-128-256)	2.4×10^5
LSTM (3)	RNN(512)	6.1×10^6
Our system	CNN(32-64);RNN(256-256);FNN(512)	3.2×10^6

Table 2 Different end-to-end neural architectures. (1) = Lopez-Moreno et al. (2014), (2) = Ganapathy et al. (2014), (3) = Gonzalez-Dominguez et al. (2014).

4.4 Supervised language identification

The algorithms are developed on the read speech corpus using a restricted closed training set, and their performance is verified on both read and conversational speech corpus. For training, we randomly split the read speech corpus into three datasets: training set (50% of the corpus), validation set and test set (25% each). The segmentation process ignores speaker information, and so the three sets contain three disjoint sets of utterances from the shared speaker space. Validation set is used for early-stopping. The test set from the read speech corpus is used for test-

ing, together with the conversational speech corpus which is not used for training. Table 3 shows the results of the four different architectures. All the systems seem to generaliz well, however, they are overfitting to individual speakers which is indicated by poor performance on conversational data.

However, the result in Table 3 show the effect of channel variation as the data splitting used a shared speaker set. We thus split the development data so that one speaker from each dialect is randomly selected for validation and another speaker for testing, cf. (Jokinen et al. 2016, Behravan et al. 2016). The remaining speakers are used for training the classifier. The results of the LOSO (leave-one-speaker-out) method are shown in Table 4, and they are comparable to our previous i-vector approach (Jokinen et al. 2016). Table 4 emphasizes the importance of implicit regularization and the multiple-architecture design for end-to-end learning. Moreover, the best network outperforms i-vector system in both datasets.

	C_{avg} x 100	
Networks	Read speech (test set)	Conversational speech
FNN (1)	2.56 +/- 0.25	22.60 +/- 2.42
CNN (2)	0	17.76 +/- 0.84
LSTM (3)	0	15.60 +/- 1.1
Our system	0	21.05 +/- 0.25

Table 3 Performance of the different network designs on a shared speaker set. (1) = Lopez-Moreno et al. (2014), (2) = Ganapathy et al. (2014), (3) = Gonzalez-Dominguez et al. (2014).

	C_{avg} x 100	
Networks	Read speech (test set)	Conversational speech
FNN (1)	32.30 +/- 1.60	24.67 +/- 2.36
CNN (2)	26.06 +/- 1.97	23.82 +/- 2.52
LSTM (3)	25.77 +/- 2.69	21.03 +/- 1.85
Our (Eq. 3)	14.68 +/- 0.42	19.78 +/- 3.48
Our (Eq. 4)	18.98 +/- 1.22	22.08 +/- 2.32
Our (crossentropy)	18.49 +/- 1.75	19.26 +/- 1.34
i-vector (4)	17.79	-

Table 4 Performance of different network designs using LOSO. . (1) = Lopez-Moreno et al. (2014), (2) = Ganapathy et al. (2014), (3) = Gonzalez-Dominguez et al. (2014

4.5 Semi-supervised scenario

We also compared the performance of a semi-supervised and full-supervised system. We used the same configuration as described in the LOSO experiment except that we removed the label from the validation set and feed it into the unsupervised system as unsupervised samples. Both the labelled and unlabelled data are shuffled and mixed into mini-batches for training.

Table 6 shows the results on both data sets and emphasizes the role of α in balancing the supervised and unsupervised objectives for the final performance. It should be noted that the value of α also variates depending on the ratio between the amount of supervised and unsupervised data available during training process.

We chose the optimized α = 0.5 for our SSEE system. The results in Table 7 indicate that semi-supervised system outperforms the supervised one in both dataset. As a result, we conclude that SSEE has learnt an internal structure within the unsupervised samples to support the discriminative task of dialect recognition.

α	Read speech (test set)	Conversational speech
0.1	24.68 +/- 0.22	22.42 +/- 1.91
0.2	22.12 +/- 3.26	18.83 +/- 2.13
0.5	**19.78 +/- 3.48**	**16.52 +/- 1.33**
0.8	29.51 +/- 1.21	25.12 +/- 1.84

Table 5 Semi-supervised learning with different α values.

	Fully supervised	Semi-supervised
Read speech (test set)	14.68 +/- 0.42	12.42 +/- 1.79
Conversational speech	19.78 +/- 3.48	16.52 +/- 1.33

Table 6 Performance of different network designs on LOSO configurations.

5 Discussion and conclusion

The goal of the study is to enable automatic spoken interaction in the less-resourced North Sami language. We have focused on the dialect recognition task, which is one of the main issues in speech technology in general. In this paper we presented the first profound study concerning end-to-end learning on a small corpus for dialect recognition. Our results indicate the potential of end-to-end deep learning approach, and also validate the possibility of applying semi-supervised learning for auditory signal to improve the performance in restricted data context.

The results also support localization of speech applications to endangered languages. Such applications can be beneficial to these languages in multiple ways. They can provide motivation to use the language and have a favourable effect on the prestige of the language. For instance, there exists a growing number of Wikipedia articles in North Sami, and the SamiTalk application, based on the existing WikiTalk technology (Jokinen and Wilcock 2012), allows the user to use North Sami Wikipedia by conducting a conversation with a humanoid robot in North Sami. The speech components need to be integrated and tested on the robot software. They are being developed separately: the DigiSami project worked on speech recognition (Leinonen 2015), and recently a commercial company has started a project on North Sami speech recognition. The dialect recognition component described in this paper is one of the enabling technologies that can be used in the development of interactive applications for North Sami.

Acknowledgements. The paper is partially based on results obtained from the Academy of Finland project *Fenno-Ugric Digital Citizens* and the *Future AI and Robot Technology Research and Development* project commissioned by the New Energy and Industrial Technology Development Organization (NEDO) in Japan.

References

Amodei, D., Anubhai, R., Battenberg E. et al., (2015). Deep speech 2: End-to-end speech recognition in English and Mandarin. CoRR, vol. abs/1512.02595.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y. (2015). End-to-end attention-based large vocabulary speech recognition. CoRR, vol. abs/1508.04395.

Behravan, H., Hautamäki, V., Siniscalchi, S.M., Kinnunen, T. and Lee, C-H. (2016). I-vector modeling of speech attributes for automatic foreign accent recognition. Audio, Speech, and Language Processing, IEEE/ACM Transactions, vol. 24, no. 1, pp. 29–41.

Besacier, L., Barnard, E., Karpov, A. and Schultz T. (2014). Automatic speech recognition for under-resourced languages: a survey. Speech Communication, 56, pp. 85–100.

Crystal, D. (2000). English as a Global Language, Cambridge.

Dalyac, A., Shanahan, M., and Kelly, J. (2014). Tackling class imbalance with deep convolutional neural networks. Thesis, Imperial College London.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788–798.

Ganapathy, S., Han, K., Thomas, S. and et al. (2014). Robust language identification using convolutional neural network features. Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.

Gonzalez-Dominguez, J., Lopez-Moreno, I. and Sak, H. (2014). Automatic language identification using long short-term memory recurrent neural networks. Interspeech.

Hiovain, K., Jokinen, K. (2016). Acoustic Features of Different Types of Laughter in North Sami Conversational Speech. Proceedings of the LREC Workshop Just talking – casual talk among humans and machines, Portorož, Slovenia.

Jokinen, K. (2009). Constructive Dialogue Modelling – Speech Interaction with Rational Agents. John Wiley & Sons, Chichester, UK.

Jokinen, K. (2014). Open-domain interaction and online content in the Sami language. Proceedings of the Language Resources and Evaluation Conference (LREC 2014).

Jokinen, K., Trong, T. N., Hautamäki, V. (2016). Variation in Spoken North Sami Language. Interspeech-2016, pp. 3299–3303.

Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I., Wilcock, G. (2017). DigiSami and Digital Natives: Interaction Technology for the North Sami Language. In: Jokinen, K. and Wilcock, G. (eds.) Dialogues with Social Robots. Springer. pp. 3-19.

Jokinen. K., Wilcock, G. (2013). Multimodal open-domain conversations with the Nao robot. In: Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialogue Systems into Practice, pages 213–224. Springer.

Jokinen, K. and Wilcock, G. (2014). Community-Based Resource Building and Data Collection. Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14), St Petersburg, Russia. pp. 201-206.

Kirkpatrick, K., Pascanu, R., Rabinowitz, N.C. and et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America, vol. 114 13, pp. 3521–3526.

Lee, K.A., Li, H., Deng, L., Hautamäki, V. and et al. (2016). The 2015 NIST language recognition evaluation: the shared view of i2r, fantastic4 and singams. Interspeech.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature, 521(7553): 436–444.

LeCun, Y., Bottou, L., Orr, G.B. and Müller, K. R. (1998). Efficient Back-Prop. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–50.

Leinonen, J. (2015). Automatic speech recognition for human-robot interaction using an under-resourced language. Master's thesis, Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo.

Li, H., Ma, B. and Lee, K.A. (2013). Spoken language recognition: From fundamentals to practice. Proceedings of the IEEE, vol. 101, no. 5, pp. 1136–1159.

Lopez-Moreno, I., Gonzalez-Dominguez, J. and Plchot, O. (2014). Automatic language identification using deep neural networks. ICASSP.

Matrouf, D. Scheffer, N. Fauve, B. G. B. and Bonastre, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. Interspeech, pp. 1242–1245.

Mi, H., Wang, Z. and Ittycheriah, A. (2016). Supervised attentions for neural machine translation, CoRR, vol. abs/1608.00112.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. AIS-TATS05, pp. 246–252.

Ó Laoire, Muiris (2008). Indigenous Language Revitalization and Globalization. Te Kaharoa, Vol. 1

Ossama Abdel-Hamid, Abdel-rahman Mohamed (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, pp. 1533–1545.

Prechelt, L. (2012). Neural Networks: Tricks of the Trade. Second Edition, chapter "Early Stopping — But When?" pp. 53–67, Springer, Berlin, Heidelberg.

Radford, A. Metz, L. and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, vol. abs/1511.06434.

Richardson, F., Reynolds, D.A., Dehak, N. (2015). A unified deep neural network for speaker and language recognition. CoRR, vol. abs/1504.00923.

Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A., Dahl, G. and Ramabhadran, B. (2014). Deep Convolutional Neural Networks for Large-scale Speech Tasks. Neural Networks, pp. 1–10.

Sainath, T., Vinyals, O., Senior, A. and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. ICASSP, pp. 4580–4584.

Thomas, S., Seltzer, M. L., Church, K. and Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6704–6708.

Trong, T.N., Hiovain, K., Jokinen, K. (2016). Laughing and co-construction of common ground in human conversations. The 4th European and 7th Nordic Symposium on Multimodal Communication, Copenhagen, Denmark.

Trong, T.N., Hautamäki V., Lee, K.A. (2016). Deep Language: a comprehensive deep learning approach to end-to-end language recognition. Speaker Odyssey, Bilbao, Spain.

Wilcock, G., Jokinen, K. (2014). Advances in Wikipedia-based Interaction with Robots. Proceedings of the ICMI Workshop on Multi-modal, Multi-Party, Real-World Human-Robot Interaction, pp. 13-18.

Wilcock, G., Jokinen, K. (2015). Multilingual WikiTalk: Wikipedia-based talking robots that switch languages. Proceedings of the SIGDial Conference, pp. 162-164.

Wilcock, G. Laxström, N., Leinonen, J., Smit, P., Kurimo, M., Jokinen, K. (2016). Towards SamiTalk: A Sami-Speaking Robot Linked to Sami Wikipedia. In: Jokinen, K., Wilcock, G. (eds.) Dialogues with Social Robots. Springer, pp. 343-351

Wilcock, G. and Jokinen, K. (2013). WikiTalk human-robot interactions. Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI), pp. 73–74.

Xu, K., Ba, J., Kiros, R., Cho, K. et al. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Proceedings of the 32nd International Conference on Machine Learning, pp. 2048–2057.

Zhang, S. and Y. Qin (2013). Semi-supervised accent detection and modelling. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7175–7179.

Zhang, Z., Bengio, S., Hardt, M. Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. ArXiv e-prints, Nov. 2016.