Kurdish spoken dialect recognition using x-vector speaker embeddings

Arash Amani, Mohammad Mohammadamini, Hadi Veisi

Kurdish spoken dialect recognition using x-vector speaker embeddings

Arash Amani[1], Mohammad Mohammadamini[2], and Hadi Veisi[3]

[1] Asosoft Research group, Iran
amani.ara@gmail.com
[2] Avignon University LIA (Laboratoire Informatique d'Avignon), Avignon, France
mohammad.mohammadamini@univ-avignon.fr
[3] Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran
h.veisi@ut.ac.ir

Abstract. This paper presents a dialect recognition system for the Kurdish language using speaker embeddings. Two main goals are followed in this research: first, we investigate the availability of dialect information in speaker embeddings, then this information is used for spoken dialect recognition in the Kurdish language. Second, we introduce a public dataset for Kurdish spoken dialect recognition named Zar. The Zar dataset comprises 16,385 utterances in 49h-36min for five dialects of the Kurdish language (Northern Kurdish, Central Kurdish, Southern Kurdish, Hawrami, and Zazaki). The dialect recognition is done with x-vector speaker embedding which is trained for speaker recognition using Voxceleb1 and Voxceleb2 datasets. After that, the extracted x-vectors are used to train support vector machine (SVM) and decision tree classifiers for dialect recognition. The results are compared with an i-vector system that is trained specifically for Kurdish spoken dialect recognition. In both systems (i-vector and x-vector), the SVM classifier with 86% of precision results in better performance. Our results show that the information preserved in the speaker embeddings can be used for automatic dialect recognition.

Keywords: Speaker embeddings · x-vector · Kurdish language · dialect recognition · Zar dataset.

1 Introduction

Spoken dialect/language recognition is the automatic identification of a dialect/ language from speech utterances. Spoken dialect/language recognition can be used as a preprocessing step for other speech technologies such as speech dictation, speech to speech translation, speech assistants, etc [1]. In general, dialect recognition is more difficult than language recognition because the dialects of specific language are almost similar [2].

Earlier generations of language recognition are based on speaker modeling frameworks such as Gaussian Mixture Model (GMM), [4], FA [5] and i-vector [3] systems. Recently, the x-vector speaker modeling is adopted for language

recognition [6] [7].The x-vector is a deep learning speaker embedding system that is already proposed for speaker recognition. In the speaker recognition pipeline the Deep Neural Network is trained to classify the speakers and a compact fixed-sized representation for each utterance extracted from a hidden layer [8]. The language recognition recipe of the system is trained with language labels [7]. The original version of x-vectors (i.e. trained for speaker recognition) has shown that besides the speaker's information, it also encodes different characteristics (e.g. gender [9],) and environment variability (room, microphone [9], type of noise [10]).

This paper follows two goals: Firstly, we investigate the possibility of doing dialect recognition with an embedding system that is already trained for speaker recognition. In doing so, we use the x-vector system that classifies speakers. The trained system is used to extract x-vector representation for our dataset that the dialect recognition is done on it. In [6] [7] the language/dialect recognition is done with an embedding network that is specifically trained for language recognition but in our work the x-vector network trained for speaker recognition is used. Using speaker embeddings for dialect recognition had importance for several reasons. Firstly, the speaker recognition data and recipes are more accessible. Secondly, the dialect feature could be interesting in forensic speaker recognition or privacy preserving voice applications.

The second goal of this research is to present the first dataset for dialect recognition in the Kurdish language. Kurdish is an Indo-European language which is spoken by more than 30 million people in Kurdistan (a region between Turkey, Iran, Iraq, Syria) and by Kurdish diaspora [11]. In a broader accepted taxonomy, the Kurdish dialects are categorized in five main branches: Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish (Laki and Kalhori), Hawrami (Gorani) and Zazaki [12]. For each dialect there are several sub-dialects. For example, the Central Kurdish branch has Ardalani, Mukriyani, Sorani and Jafi sub-dialects [13]. In this work, we have focused on the main branches and we didn't consider the sub-dialects of each dialect. Spoken dialect recognition is an important task for Kurdish because it can be used as a preprocessing system that can be integrated with automatic speech recognition systems to have specific acoustic models for each dialect.

Kurdish language is not studied broadly in the domain of speech processing. There are just few researches focused on speech processing for this language. Preparing speech resources for this language is crucial to foster research on this language. Asosoft speech corpus is the only speech corpus for the Kurdish language [13]. This corpus is usable for automatic speech recognition and text to speech tasks. The Asosoft speech corpus is in Central Kurdish and it can not be used for dialect recognition. Already in [14], a small data set with 100 utterances is used for speaker recognition in Kurdish language. It deserved to be mentioned that there are some attempts on Kurdish text dialect recognition [15] [16] but to the best of our knowledge the current paper is the first attempt of spoken dialect recognition in the Kurdish. In our work we present a dataset with 16,385 utterances in 49h36mi hours for five main dialects of the Kurdish

language. The covered dialects are Northern Kurdish, Central Kurdish, Southern Kurdish, Hwarami and Zazaki.

The next parts of this paper are organized as: in Section 2. the data collection and data prepossessing steps are presented. In Section 3 the methodology is described. The experimental setup and results are presented in Section 4 and finally, in Section 5 the obtained results are discussed.

2 Data collection

In this section a speech corpus for Kurdish dialect recognition is introduced. The corpus named *Zar* which means dialect in Kurdish. In collecting and designing the speech corpus we considered several points. First of all, we tried to cover all Kurdish dialects. The available Kurdish corpora normally include Northern and Central Kurdish that covers the most population of Kurdish speakers and are used more in both written and spoken form. We covered all Kurdish dialects, even those such as Hawrami that is a zero-resourced dialect and an endangered dialect.

The data is collected mainly from Kurdish public TV and radio websites. The major part of the data is scrapped from the web, even though a smaller part downloaded from the Telegram Channels manually. Table 1 shows the source and original type of scrapped data. The files are a combination of news, dialogue and music in different languages dialects of Kurdish language that we have to review to remove other sounds that include non-proper Kurdish dialects and non Kurdish parts. Each segment may consist of more than one speaker and corrupted by noise. The speakers are male and female. The speaker can be anywhere in studio, in nature, on the street and speak through smart phone or telephone. The data collection and processing is done in the following steps:

1. The collected files were in different audio and video formats. The data is converted into mono channel wav format.
2. The files shorter than 8 seconds were merged into bigger ones to have enough information in dialect recognition.
3. These files were reviewed manually. Music, other languages or other dialects have been removed from each file if it exists.

The final version of the Zar dataset included 16,385 files results in 49h36m in five dialects of the Kurdish language. We tried to have a balanced data for each dialect. The minimum length of speech files is 8 second and the maximum size is 39 seconds, The average length of the files is 10.9 seconds. The general specification of the dataset is shown in Table 2. In Fig. 1 the distribution of file duration for all dialects is shown. The Zar dataset is publicly available in the Github repository [4].

Table 1. Source of Data

Dialect	Source	Type
Hawrami Kurdish	www.sterktv.net	Video
	https://t.me/hawramanhaneberchem	Voice
Zazaki Kurdish	http://sterktv.net/	Video
	https://globalrecordings.net/en/language/12048	Video
Northern Kurdish	https://www.dengeamerika.com/	Audio
	http://www.denge-welat.org/	Video
Central Kurdish	https://www.dengiamerika.com/	Audio
Southern Kurdish	https://prim.dideo.ir/	Video
	https://t.me/radioro$_k$*urdestan*	Audio
	https://t.me/shervaadab$_r$*azi*	Audio

Table 2. General specifications of the Zar dialect recognition dataset.

Dialect	Number	Min Len (Sec)	Max Len (Sec)	Avg Len (Sec)	Total (duration)
Hawrami Kurdish	1876	8	36	11.2	5:49:42
Zazaki Kurdish	3839	8	39	10.4	11:08:23
Northern Kurdish	3603	8	37	11.0	11:00:57
Central Kurdish	3386	8	37	11.1	10:24:08
Southern Kurdish	3681	8	38	11.0	11:13:22
Total	16,385	8	39	10.9	49:36:32

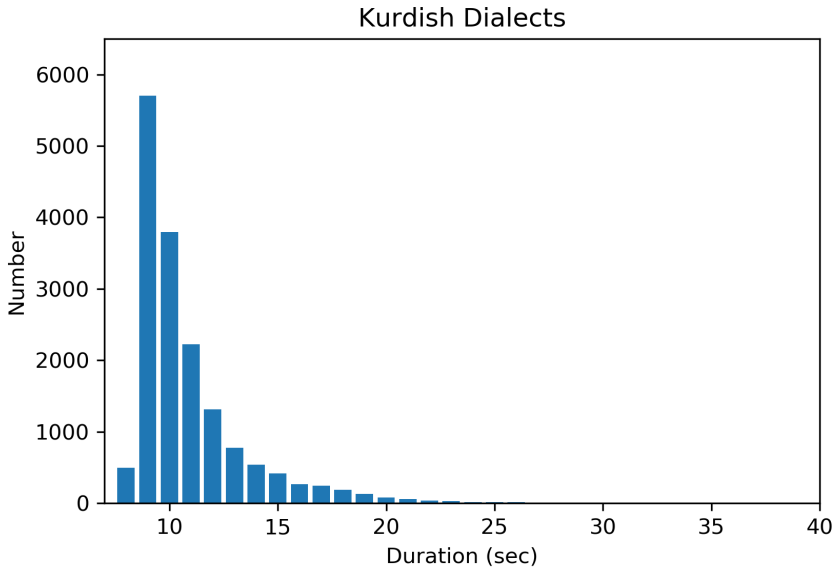

Fig. 1. The distribution of file duration in Zar corpus.

3 Proposed Method

In this section the proposal method is described. The previous research shows that x-vector embeddings preserve different characteristics of a speech file. Until now, the researchers have shown that besides the speaker specific information, the x-vectors holds both types of speaker and environment information. Among the speaker characteristics it was shown that x-vectors holds gender [9], accent, emotion [17] In regard to environment information it is shown that x-vectors can be classified in terms of the recording environment, recording device, type of noise etc [18]. This behaviour of x-vectors gives us the power of applying this representation to other speech processing applications. In this paper we tried to explore the availability of dialect information in the x-vectors. The experiments are done in this steps:

Step 1. Feature extraction: The x-vector systems was trained on MFCC features with 25 frame length.

Step 2. x-vector: In this work the TDNN is used for x-vector extraction [8]. The x-vector network was trained with the clean and augmented version of Voxceleb2. The Musan corpus was used for data augmentation [19]. The x-vector has 512 dimension.

Step 3. Dialect recognition: To do dialect recognition, we trained Support Vector Machine and Decision Tree models on x-vector speaker embeddings. 5-fold cross validation method is used to train and test models. The criterion used on DT is *entropy*. The code is available in the Github repository. Results are discussed Section 4.

The configuration of the x-vector dialect recognition is presented in Fig. 2.

4 Results and Discussion

In this section the obtained results are discussed. As it was mentioned in Section 3, the experiments are done in three steps. Firstly, the MFCC features are extracted. Then, two systems (i.e. x-vector, and i-vector) are used to achieve a fixed length representation for each utterance. Finally, two classifiers are trained for dialect recognition. In our experiments we used decision tree and support vector machine to classify the dialects. The results are presented in table 3.

[4] https://github.com/ArashAmani/Kurdish-Dialect-Recognition

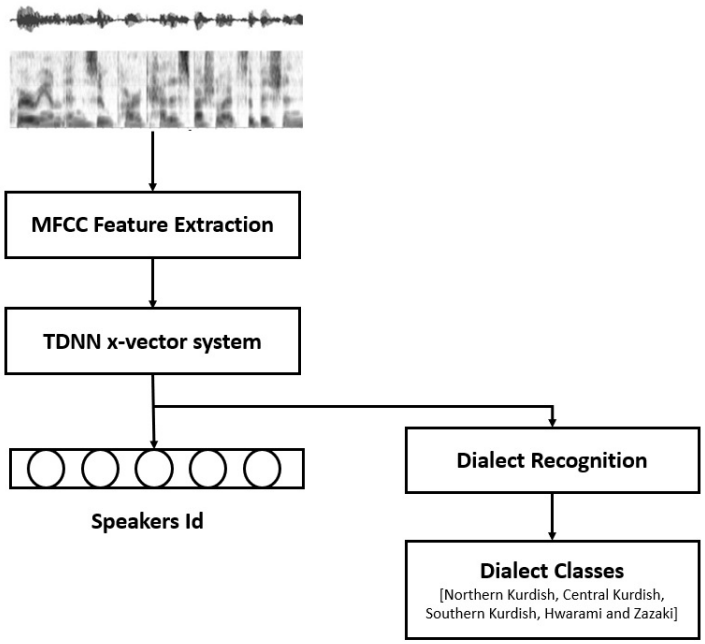

Fig. 2. The configuration of the dialect recognition system using x-vector speaker embeddings.

Table 3. Kurdish dialect recognition with x-vector speaker embeddings

System	**x-vector**			**i-vector**		
Metric	Precision	Recall	F-measure	Precision	Recall	F-measure
DT	0.65	0.64	0.64	0.67	0.66	0.66
SVM	0.87	0.86	0.86	0.87	0.83	0.83

From the results in Table 3 we can see that x-vector speaker embedding, beside capturing the speaker features, preserve language/dialect information. The results show that however the x-vector is trained basically for speaker recognition system but its performance is better or same as an i-vector system that is trained specifically for dialect recognition. This property of x-vector speaker embedding show that it is important to do more exploration in the characteristics preserved in x-vector speaker embeddings.

Another potential reason behind this ability of the x-vector system comes from the channel and environment of the recorded data. Because in our dataset the recording environment for a specific dialect almost are same and it is shown already that x-vectors are sensitive to the recording environment and it is possible to classify the vectors coming from different acoustic environments.

5 Conclusion

In this paper we introduced the first spoken dialect recognition corpus for the Kurdish language. The corpus comprises five dialects of the Kurdish language in 16385 files resulting 49h36m speech data. Also, we explored the availability of dialect information in x-vector speaker embedding. We observed that x-vector speaker embedding beside the speaker characteristics holds the dialect information. It means that having the speaker embedding belonging to a specific person it is possible to find the language/dialect of that speaker. This feature can be interesting to develop privacy preserving speech applications. Since, x-vectors are sensitive to diffident environment variabilities such as recording device, channel, noise, reverberation etc, developing more strict protocols for probing dialect information in x-vector systems is important. Also we think that domain adaptation at the speaker modeling network or transformations on the the extracted embeddings make the dialect recognition systems based on speaker embeddings more efficient.

References

1. H. Li, B. Ma and K. A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," in Proceedings of the IEEE, vol. 101, no. 5, pp. 1136-1159, May 2013, doi: 10.1109/JPROC.2012.2237151.
2. F. Biadsy, H. Soltauy, L. Manguy,J. Navratily, and J. Hirschberg,"Discriminative phonotactics for dialectrecognition using context-dependentphone classifiers," in-Proc. IEEE Odyssey:Speaker and Language RecognitionWorkshop, Brno, Czech Republic, 2010,pp. 263–270.
3. Wei Wang, Wenjie Song, Chen Chen, Zhaoxin Zhang, Yi Xin, I-vector features and deep neural network modeling for language recognition, Procedia Computer Science, Volume 147, 2019, Pages 36-43, ISSN 1877-0509,
4. Pedro A. Torres-Carrasquillo, Terry P. Gleason and Douglas A. Reynolds, Dialect identification using Gaussian Mixture Models 2004
5. Y. Lei and J. H. L. Hansen, "Factor analysis-based information integration for Arabic dialect identification," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 4337-4340, doi:10.1109/ICASSP.2009.4960589.
6. Abualsoud Hanani, Rabee Naser, Spoken Arabic dialect recognition using X-vectors, Natural Language Engineering, Cambridge university press. 2020
7. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S. (2018) Spoken Language Recognition using X-vectors. Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, 105-111, DOI: 10.21437/Odyssey.2018-15.
8. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.
9. D. Raj, D. Snyder, D. Povey and S. Khudanpur, "Probing the Information Encoded in X-Vectors," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 726-733, doi: 10.1109/ASRU46091.2019.9003979.

10. Mohammad Mohammadamini, Driss Matrouf, Jean-Francois Bonastre, Romain Serizel, Sandipana Dowerah, Denis Jouvet, Compensate multiple distortions for speaker recognition systems, EUSIPCO 2021
11. Hadi Veisi, Mohammad MohammadAmini, Hawre Hosseini, Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus, Digital Scholarship in the Humanities, Volume 35, Issue 1, April 2020, Pages 176–193, https://doi.org/10.1093/llc/fqy074
12. Malmasi, Shervin, Subdialectal Differences in Sorani Kurdish, Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,Osaka, Japan, 2016
13. Hadi Veisi, Hawre Hosseini, Mohammad Mohammadamini (LIA), Wirya Fathy, Aso Mahmudi, Jira: a Kurdish Speech Recognition System Designing and Building Speech Corpus and Pronunciation Lexicon, https://arxiv.org/abs/2102.07412v1, 2021
14. Zrar Khalid Abdul, Kurdish speaker identification based on one dimensional convolutional neural network, Vol. 7, No. 4 (Special Issue), pp. 566-572 Computational Methods for Differential Equations, 2019.
15. Hossein Hassani and Oussama H. Hamid, Using Artificial Neural Networks in Dialect Identification in Less-resourced Languages - The Case of Kurdish Dialects Identification
16. Hossein Hassani, Dzejla Medjedovic, Automatic Kurdish Dialects Identification, Conference: Fifth International Conference on Natural language Processing (NLP - 2016)At: Sydney, Australia
17. Raghavendra Pappagari, Tianzi Wang, JesÃºs Villalba, Nanxin Chen, Najim Dehak , X-Vectors Meet Emotions: A Study On Dependencies Between Emotion And Speaker Recognition, ICASSP 2020.
18. Nandwana, M.K., Lomnitz, M., Richey, C., McLaren, M., Castan, D., Ferrer, L., Lawson, A. (2020) The VOiCES from a Distance Challenge 2019: Analysis of Speaker Verification Results and Remaining Challenges. Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 165-170, DOI: 10.21437/Odyssey.2020-24.
19. David Snyder and Guoguo Chen and Daniel Povey, MUSAN A Music, Speech, and Noise Corpus, 2015, arXiv:1510.08484v1