Automatic Language Identification in Short Utterances

Diptanu Sarkar
Rochester Institute of Technology
ds9297@rit.edu

December 6, 2019

Abstract

Language Identification (LID) in Natural Language Processing (NLP) is the process of identifying the spoken language in speech utterances. In the last decade, the interest and functional application of speech processing systems have grown exponentially. The proliferated use of hands-free voice-operated devices, speech-to-speech translation systems requires low latency, reliable automatic speech identification systems. This article examines three different models to recognize languages automatically in speech. The first model uses Dynamic Hidden Markov Networks (DHMNet) for LID in utterances. Another model utilizes Deep Neural Network (DNN), and the third uses the recently developed Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN). Finally, comparing three different models, it is shown that a fusion of LSTM RNN and DNN model gives better results than the state-of-the-art models when applied to short utterances.

1 Introduction

Recent technological trends towards speech-to-speech translation systems and hands-free voice-operated devices have increased globally. Automatic speech recognition (ASR) systems performance heavily relies on automatic language identification (LID). Lexical similarity, multiple dialects of the language, various accent makes LID particularly challenging. For this reason, there's a need for a reliable, real-time automatic language detection system with remarkably low latency.

Previously, the research on LID mainly focused on the use of the phonetic content of the speech signal to identify multiple languages. Gaussian mixture model (GMM) and i-vector based model are also proposed utilizing the acoustical features of an utterance. However, one of the major drawbacks of these models is latency. The quality of output degrades promptly as the utterance duration decreases.

In [11], the author proposed a model based on Dynamic Hidden Markov Network (DHMNet) [10], which is a never-ending learning model. Lopez-Moreno et al. [9] proposes two different models using Deep Neural Networks (DNN). In [4], the authors introduce Long Short-Term Memory (LSTM) based model for LID.

Section 2 presents previous work in LID and a summary of the models used. Section 3 explains the architecture of the models, Section 4 summarizes the experiments and results. Section 5 contains some merits and demerits of the proposed models. Section 6 discusses the learning outcome of the term paper. Finally, Section 7 summarizes the conclusion and direction for future work.

2 Background

Previous researches on LID mainly focused on the use of the phonetic content of the speech signal to identify multiple languages. In a popular technique called Parallel Phone Recognition and Language Modeling (PPRLM) [14], the input utterance converted into phone sequences by recognizers and the probability of the phone sequence in a language is calculated by the language n-grams. However, the disadvantage is the expense or unavailability of excellent phone labels available during the training. Another Gaussian mixture model (GMM) based approach leverages the acoustical differences between languages. Yet, the performance improvement is not significant enough.

Later, an i-vector approach is proposed in [8]. An i-vector is represented as a dimensionally reduced vector of a whole speech. Even though i-vector models are highly successful in acoustic feature extractions and speaker affirmation tasks. In LID i-vector model's accuracy degrades for shorter utterances and also computation heavy in real-time application.

2.1 Dynamic Hidden Markov Network

In [11], the author proposed a model based on Dynamic Hidden Markov Network (DHMNet) [10], which is a never-ending learning model with the ability to represent the utterance manifold embedded in feature space. The number of nodes and structure of the network is automatically determined and can change depending on the data distribution [11]. For the language identification task, the DHMNet trained on labeled feature vectors of the data sequence. A discrete label of the probability distribution assigned to each DHMNet state and transition. During the test, the language labels probabilities calculated over the best state sequence. The highest probability score account for the label of the identified language.

2.2 Deep Neural Network

Lopez-Moreno et al. [9] proposes two different Deep Neural Networks (DNN) based models, motivated by the success of DNN in machine learning applications - object recognition, acoustic modeling. The first model is DNN based LID classifier which inputs acoustic features and outputs the calculated probabilities of different languages. The second model is a hybrid model using the DNN and i-vector system, which is taking the best from both the approaches. The bottleneck features of the utterances extracted by the DNN model, later used as an input to the i-vector system, for better performance. The model does not require training on transcribed audios, which are usually harder to get. The LID evaluation is consistent with the DNN optimization.

2.3 Long Short-Term Memory Recurrent Neural Network

In [4], the author proposed Long short-term memory model to identify the language of the utterances. LSTM can store information from previous inputs through long-duration, which makes them more suitable from DNNs while modeling sequential data. The memory blocks of the LSTM contains memory cells with self-connections to store the temporal state of the network. The multiplicative units are to control the in-flow and out-flow to rest of the network [4]. LSTM uses the temporal differences in the acoustic features of languages and learns complex long-range features to classify the utterance.

3 Architecture

3.1 Dynamic Hidden Markov Network

The DHMNet used in the language identification task consists of hidden Markov states with self-connection and lateral connection among neighboring states. Single Multivariate Gaussian function with a fixed diagonal covariance matrix models the input linguistic features and which are represented by each state. Paths in the network represent the learned acoustic patterns. For unseen patterns, new states and transitions added to the network. Noise will also create these new states. However, spurious states would be removed from the network as they maybe never visited again, and eventually considered as dead and removed from the network.

In the DHMNets, different states represent different features. The model utilizes the competitive Hebbian rule, so the state network is a topology representing network. The competitive Hebbian rule - for each input vector, connects to adjacent nodes by the edges. When the network changes dynamically, neighborhood relations also change. For input utterance represented by feature vectors, the neighbor and the path preserving properties of the network stores the best state sequences.

The DHMNet is capable of unsupervised learning. However, abstract knowledge about words and languages unobtainable from acoustic signals and can only

be achieved by supervised learning through labeled data. During the learning, labeled data ('lang-n' for nth language and 'sil' for silence) is used to incrementally updated the best path in the network for acoustic features.

[width=3.2in]HMM.jpg

Figure 1: Semantic structure of Dynamic Hidden Markov network.

3.2 Deep Neural Network

Recently, DNNs has achieved higher accuracy in speech recognition tasks over classical GMMs. The DNNs are a more complex multilevel distributed structure that makes DNNs a more compact model than GMMs. The DNN models can also utilize a large amount of data to make a better and robust model.

3.2.1 The DNN Architecture

The DNN model proposed in [9] is a fully-connected neural network with four hidden layers and rectified linear units (ReLU) activation function. Cross-entropy function is used to train the model during backpropagation. Softmax function in the output layer is used with the same number of dimensions as the number of target languages. The DistBelief framework's [2] asynchronous stochastic gradient descent is used to train the model with a 0.001 learning rate and 200 samples mini-batch size. The architecture works at frame-level, we can decide the language of the utterance in each new frame, which is the main advantage of DNNs over GMMs. In the final layer, the score for the language is calculated by multiplying the output probabilities obtained in each frame.

[width=3.4in]CNN.jpg

Figure 2: Architecture of the Deep Neural Network. Generated using [12].

3.2.2 A hybrid architecture: DNN and i-vector

The motivation behind a hybrid architecture is to learn better feature representation from the discriminative ability of the DNN model and also to utilize the generative modeling in the i-vector system.

In this approach, the bottleneck features [6] are extracted from the DNN trained for the language Identification task. The bottleneck features are low-dimensional input features of the utterance with non-linear transformation. The last layer of the DNN based architecture replaced by a 40-dimensional bottleneck layer and used as input for the i-vector system.

3.3 Long Short-Term Memory Recurrent Neural Network

LSTM can store information from previous inputs through long-duration [3], which makes them more suitable from DNNs while modeling sequential data. The internal layer contains memory blocks and cells with self-connections to save different states. The multiplicative units control the inflow and outflow of the information.

The proposed LSTM architecture contains 512 memory cells. The input to the network is 39-dimensional perceptual linear predictive (PLP) [7] features retrieved from acoustic frames. PLP features are a low dimensional representation of speech, which considers loudness and intensity of the human voice. The model trained within a distributed training framework using asynchronous stochastic gradient descent (ASGD) and the truncated backpropagation through time (BPTT) learning algorithm [1] with exponentially decaying learning rate 1e-04. During training, the inputs split into chunks of 2.5 second to 3 second utterances for better randomization of the gradient. The model computes a mapping from the input sequences of features to the output sequences using the softmax activation function. The final score is the average log values of softmax output in each frame for the target language.

4 Experiment and Analysis

4.1 Dynamic Hidden Markov Network

ATR multilingual travel domain speech database [13] is used to experiment with the DHMNet model. It consists of studio recordings form many speakers in 3 different languages - English(en), Japanese(jp), and Chinese(ch). All data divided into speech and silence (sil) regions. For training the DHMNet, 1000 utterances of length 2 seconds per language from both males and females used. For testing is conducted on different 200 utterances of both males and females per language. Two separate tests were performed - Test 1 and Test 2. Test 1 consists of the dataset with an average speech length of 2.6 seconds. For Test 2, utterances that are longer than 5 seconds are used.

The result from Test 1, shows that DHMNets performance gets better with the increased number of states and achieves about 85% of accuracy. However, to use the system in real-time, the LID should not wait for the completion of the utterance. Hence for Test 2, forced LID decisions are processed after 1, 2, 3, 4, and 5 seconds. The result shows that the longer the speech, the better the result. For 5 seconds utterances, the model has achieved an 89% identification rate, and for 3 seconds the accuracy is 87%, which is still acceptable for real-time operations.

4.2 Deep Neural Network

The NIST Language Recognition Evaluation Dataset 2009 (LRE'09) [5] is used to evaluate both the language identification system. The dataset comprises

of two different audio sources: Conversational Telephone Speech (CTS) and Voice of America(VOA) news. For the experiment, two different evaluation sets from LRE'09 - LRE09_FULL and LRE09_BDS used. LRE09_FULL represents the whole LRE' 09 datasets with 23 languages, and original train and test files. The LRE09_FULL dataset is used mainly to compare the result with the current state-of-the-art LID. LRE09_BDS is a balanced dataset of 8 different languages, which facilitates new experiments in a controlled setting.

Results using both the dataset shows that DNN based model outperforms the current i-vector system in shorter utterances (¡ 10 seconds). However, the i-vector model performs slightly better in case of longer duration utterances. Nevertheless, the bottleneck model using DNN and i-vector is more robust than the i-vector and DNN based model. The bottleneck model with shorter utterances (3 seconds) records 96% accuracy using the LRE09_BDS dataset and 89% with the LRE09_FULL dataset.

4.3 Long Short-Term Memory Recurrent Neural Networks

For the evaluation subset of 8 languages containing short utterances (3 seconds) from the official NIST Language Recognition Evaluation Dataset 2009 (LRE'09) [5] is used. The LSTM model then is compared with DNN with varying numbers of hidden layers. Interesting comparisons show that the DNN model with 4 hidden layers outperforms the models with 2 hidden layers and 8 hidden layers.

The LSTM model surpasses the DNN model (with 4 hidden layers) performance with 20 times lower number of parameters (1M vs 20M). Furthermore, fusions of different models are also examined. Interestingly, the combination of the LSTM and the DNN system performs notably better than the i-vector system. The fusion model also achieves better results than the original LSTM model.

5 Discussion

The DHMNet based discriminative model [11] is the first state-of-the-art to achieve high accuracy in shorter utterances (3 seconds). Even though the model is less computationally expensive, the performance is not extensively utilizable in a real-life setting due to the limitation of the number of languages the model recognizes. Also, the performance of the model deteriorates when identifying two or more lexical similar languages.

The models proposed by Lopez-Moreno et al. [9] outperforms the previous state-of-the-art's, achieving very high accuracy in shorter utterances. However, DNNs need very high computation (20 M parameters) and comparatively larger datasets to train. In the case of longer utterances (¿ 10 seconds), i-vector would be an ideal choice instead of DNN based models due to lower computation cost and simplicity of the models.

Gonzalez-Dominguez et al. [4] proposed LSTM based model surpasses the previous DNN models with 20 times lower training parameters. The model

also generalizes the unseen languages robustly. However, the model requires a moderately balanced dataset to perform well. The authors compare a different combination of models, to benefit from both discriminative and generative properties of different models. However, the architectures of those models are not discussed.

6 Learning Outcome

Multiple state-of-the-art models used for automatic language identification task is learned. Also, the trade-offs between different models are reviewed. For example - an i-vector model can be utilized with less computation when the number of languages to recognize is less, and the utterances are moderately short (¿ 5 seconds). If a large amount of data is available, even shorter utterances can be recognized using a combination of DNN and i-vector system. Also, in some cases, instead of using only the generative or discriminative model, a combination of both can be useful and more effective.

7 Conclusion and Future Work

This article provides an extensive comparison of state-of-the-art automatic language identification models in utterances. The LSTM based model achieved the highest accuracy of 96.6% in short utterances (3 seconds) with 8 different languages. The model can be used more reliably in real-time.

Future work can be focused on applying the model to more lexical similar languages and also experiment with different models. One interesting area of research would be to work with an unbalanced dataset as in the real-world getting speech utterances in all languages is challenging.

References

[1] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

[2] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.

[3] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, 3(Aug):115–143, 2002.

[4] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno. Automatic language identification using long short-term