SAYS WHO? DEEP LEARNING MODELS FOR JOINT SPEECH RECOGNITION, SEGMENTATION AND DIARIZATION

Amitrajit Sarkar, Surajit Dasgupta, Sudip Kumar Naskar, Sivaji Bandyopadhyay

Jadavpur University, India

ABSTRACT

The field of speech recognition has seen tremendous advances in the recent past owing to the development of powerful deep learning architectures. However, the closely related fields of speech segmentation and diarization are still primarily dominated by sophisticated variants of hierarchical clustering algorithms. We propose a powerful adaptation of the state-of-the-art Speech Recognition models for these tasks and demonstrate the effectiveness of our techniques on standard datasets. Our architectures are a combination of Bidirectional Long Short Term Memory (LSTM) Networks, Convolutional Networks, and Fully Connected Networks, trained by Gradient Descent to minimize the Cross Entropy and the Connectionist Temporal Classification (CTC) losses. We adapt the Libri Speech corpus for the task of segmentation and diarization. We obtained comparable results with respect to state-of-the-art in both tasks.

Index Terms: Speech recognition, speech diarization, speech segmentation, deep learning, neural networks.

1. INTRODUCTION

Automatic speech recognition (ASR) deals with the transcription of speech signals into text, while speaker diarization partitions an audio stream into contiguous segments based on the identity of the speaker. The former answers the question, "what was said?", while the latter answers the question, "who spoke when?". In conjunction, they can be used effectively to transcribe conversations, a task especially important for broadcasts and meetings [1, 2]. To the best of our knowledge, there has not been any effort to jointly model these two tasks. Moreover, the state-of-the-art in both rely on very different methods. In this paper, we propose a joint model in the deep learning paradigm. Current state-of-the-art speech diarization and speaker identification systems rely on long pipelines, with several phases employing complex, hand engineered processing stages or features. Deep learning has accelerated the field of ASR to yield highly accurate end-to-end systems. The objective of the proposed work is to leverage the recent developments in ASR to propose end-to-end solution for speech diarization and segmentation.

We proposed a deep architecture based on Deep Speech 2 [3]. Our system utilizes a combination of convolutional, bidirectional LSTM and fully connected layers. We minimize the cross entropy and CTC losses using gradient descent. We performed variety of experiments on the proposed model. For evaluation, we used an adaptation of the Libri Speech corpus[1] [4] for the diarization and segmentation tasks.

2. RELATED WORK

It has been shown that recurrent neural networks (RNN) perform well in end-to-end speech recognition[5]. These models broadly exploit two different paradigms, used to map variable length audio sequences directly to variable length transcriptions. The RNN encoder-decoder paradigm uses an encoder RNN to map the input to a fixed length vector. A decoder network then expands the fixed length vector into a sequence of output predictions [6]. Adding an attentional mechanism to the decoder greatly improves the performance of the system, particularly with long inputs or outputs [7].

The other commonly used technique for mapping variable length audio input to variable length output is the CTC loss function [8] coupled with an RNN to model temporal information. The CTC-RNN model and its derivatives perform well in end-to-end ASR [9, 3, 10].

[1] http://www.openslr.org/12/

The general approach to speaker diarization is a two-step process of segmentation and clustering [11, 12]. Metric based segmentation is the most popular, relying on metrics such as the Bayesian Information Criterion, Generalized Likelihood Ratio, etc. Clustering may be carried out top down, or bottom up, the latter being slightly more popular.

The bottom-up approach estimates a number of clusters or models and aims at successively merging and reducing them until only one remains for each speaker. Clusters are usually modeled with Gaussian Mixture Models (GMMs). Upon merging, a single new GMM is trained on the data that was previously assigned to the two individual clusters. The top-down approach initially models the entire audio stream with a single speaker model. It then proceeds to successively add new models to it, until all the speakers are considered to be accounted for.

3. MODEL

Our model can be semantically decomposed into three parts which are discussed in the following subsections. However, it is to be noted that the entire deep learning architecture is trained and deployed jointly and the breakup is only for the ease of discussion.

3.1. Recognition Module

The ASR module is adapted from Baidu's DeepSpeech 2 [3] architecture. This module consists of a combination of convolutional, recurrent and fully connected layers.

The hidden representation at a layer l is represented by the vector h^l. We use h^0 to represent the input x. The lower layers consist of one or more convolutions over the time dimension of the input. For a context window of size c, the i^{th} activation at time-step t of the convolutional layer is given by Equation 1, where $\odot$ denotes the element-wise product of the i-th filter and the context window of the previous layers' activations, and f denotes a unary non-linear function.

$$h^l_{t,i} = f(w^l_i \odot h^{l-1}_{t-c:t+c}) \tag{1}$$

As our non-linearity, we used the clipped rectified-linear unit (ReLU) function $\sigma(x) = min(max(x, 0), 20)$.

One or more bidirectional recurrent layers follow the convolutional layers. The forward ($\overrightarrow{h}^l_t$) and backward ($\overleftarrow{h}^l_t$) recurrent layer activations are computed as in Equation 2.

$$\begin{aligned} \overrightarrow{h}^l_t &= g(h^{l-1}_t, \overrightarrow{h}^l_{t-1}) \\ \overleftarrow{h}^l_t &= g(h^{l-1}_t, \overleftarrow{h}^l_{t+1}) \end{aligned} \tag{2}$$

The output activations for the layer are formed from the sum of the two sets of activations as $h^l = \overrightarrow{h}^l + \overleftarrow{h}^l$. The function g is the standard recurrent operation, given as $\overrightarrow{h}^l_t = f(W^l h^{l-1}_t + \overrightarrow{U}^l \overrightarrow{h}^l_{t-1} + b^l)$, where W^l is the input-hidden weight matrix, $\overrightarrow{U}^l$ is the recurrent weight matrix and b^l is a bias term. The bidirectional recurrent layers are followed by fully connected layers, as in Equation 3.

$$h^l_t = f(W^l h^{l-1}_t + b^l) \tag{3}$$

The output layer L performs a softmax operation, computing the probability distribution over characters, as given by Equation 4.

$$p(l_t = k|x) = \frac{\exp\left(w^L_k \cdot h^{L-1}_t\right)}{\sum_i \exp\left(w^L_i \cdot h^{L-1}_t\right)} \tag{4}$$

We also introduced a special end-of-segment character. Our system uses this character to indicate a change in speaker. This is externally synchronized with the outputs of the segmentation and diarization modules. The recognition module is trained using the CTC [8] loss function.

3.2. Segmentation Module

The segmentation module uses the features learnt by the outputs of the recurrent layers of the recognition module, to estimate the probability of an end-of-segment marker. This is essentially a binary classification task. We use multiple fully connected layers for this task, as in Equation 5.

$$h^l_t = f(W^l h^{l-1}_t + b^l) \tag{5}$$

In the intermediate layers we use the ReLU for our non-linearity f. The last of these feedforward layers is fed through a sigmoid function, to allow restrict the output range to $(0, 1)$ as required for predicting probabilities. Finally, a threshold is used to determine whether an end-of-segment marker is predicted or not. This threshold is a hyperparameter tuned on the development set.

The module is trained to minimize the cross entropy of prediction, as in Equation 6, where l_t and h_t represent the labels and the predictions, respectively. The labels $l_t = 1$ at segment boundaries, and 0 otherwise.

$$\sum_t l_t \log(h_t) + (1 - l_t) \log(1 - h_t) \tag{6}$$

3.3. Diarization Module

The diarization module accepts the inputs from both the recognition as well as the segmentation modules to perform diarization. The features learnt by the recurrent layers of the recognition module are selected for further processing by the outputs of the segmentation layer. It is to be noted that there is a one-to-one correspondence between the frames in the output of the segmentation module, the features of the recognition module and the frames in the original audio signal.

Once we have a set of features corresponding to the predicted end-of-segment points, h_i, we construct a similarity matrix as shown in Equation 7.

$$\begin{pmatrix} h_{11} & \dots & h_{1n} \\ \dots & \dots & \dots \\ h_{n1} & \dots & h_{nn} \end{pmatrix} = \frac{1}{2}(1+\begin{pmatrix} \frac{h_1^T}{|h_1^T|} \\ \dots \\ \frac{h_n^T}{|h_n^T|} \end{pmatrix} \cdot \begin{pmatrix} \frac{h_1}{|h_1|} & \dots & \frac{h_n}{|h_n|} \end{pmatrix}) \tag{7}$$

Each h_{ij} predicts the probability of the i^{th} and j^{th} segments belonging to the same speaker. This module is also trained to minimize the cross entropy of prediction, as in Equation 8, where l_{ij} are the labels and h_{ij} are the predictions. The labels $l_{ij} = 1$ if segments i and j belong to the same speaker, and 0 otherwise.

$$\sum_{i,j} l_{ij} \log(h_{ij}) + (1 - l_{ij}) \log(1 - h_{ij}) \tag{8}$$

3.4. Joint Model

The three modules described above is trained and used jointly as an end-to-end system. The structure of the joint model is outlined in Figure 1.

In the figure, it is to be noted that common features are obtained from the combination of convolutional and bidirectional recurrent layers. These are used by the three modules or heads for recognition, segmentation and diarization. The mathematics underlying these modules has been described in the previous subsections.

4. DATASET

We realized in the early stages of our work that datasets containing sufficient information at the required level of granularity for the joint task of segmentation, diarization and recognition were unavailable. Hence we decided to adapt the widely popular open source Libri Speech[2] dataset [4] for ASR to our task and created synthetic training data[3]. The construction of the new dataset is outlined below.

We gathered the set of distinct speakers, and collected all audio samples along with their corresponding transcriptions. n-segment speech samples were generated by computing n random derangements of the set of speakers. We randomly selected and combined samples of the corresponding speakers in the derangements, keeping track of the segmentation points and build the diarization similarity matrix. This matrix represents the similarity between two snippets from the point of view of performing diarization. The higher the similarity, the greater the chances of it being the same speaker. It is to be noted that this matrix reduces to the identity matrix if the constructed sample was composed of unique speaker segments. In any case, the matrix is always symmetric. The process for the generation of the dataset is outlined in figure 2. The idea is illustrated on a database of 5 speakers, for 4 speech segments. Since the 1st and the 3rd speech segments are from the same speaker, the similarity matrix for the 1st generated clip is $a_{i,j} = 1 \forall (i,j) \in \{(1,1),(1,3),(2,2),(3,1),(3,3),(4,4)\}$ and $a_{i,j} = 0$ otherwise.

5. EXPERIMENTS

5.1. Setup

We carried out our experiments and report results only on the two-speaker subset of our dataset. We implemented our model[4] in the TensorFlow[5] framework. We first trained the feature extractor and the recognition head. We fine-tuned a freely available pretrained model[6] for the task.

While training the segmentation and diarization heads, we kept the weights of the feature extractor layers fixed. The remaining network parameters were initialized using Xavier initialization [13]. We trained for 16 hours on a Quadro M6000 to obtain the results (cf. Section 5.2), using the Adam Optimizer. We trained our model using various learning rates, differing uniformly on a logarithmic scale, for a fixed number of initial epochs. Following this we continued to train the most

[2] http://www.openslr.org/12

[3] https://github.com/aaiijmrtt/SAYSWHO

[4] https://github.com/aaiijmrtt/SAYSWHO

[5] https://www.tensorflow.org/

[6] https://github.com/SeanNaren/deepspeech.torch

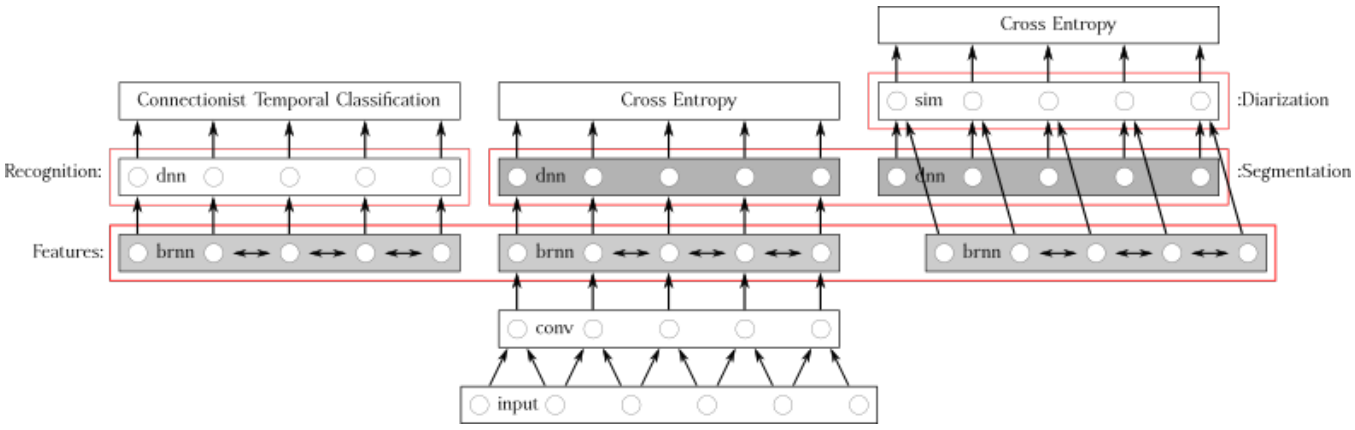

Fig. 1. Schematic diagram of the proposed model.

Speaker IDs		Audio Clips
1	→	1.1 ... 1.11
2	→	2.1 ... 2.7
3	→	3.1 ... 3.5
4	→	4.1 ... 4.13
5	→	5.1 ... 5.17

Derangements				
2	5	1	3	4
3	1	4	5	2
2	4	5	1	3
4	3	5	2	1

→

Generated Clips			
2.7	3.4	2.3	4.10
5.12	1.6	4.3	3.2
1.5	4.13	5.6	5.8
3.5	5.15	1.2	2.6
4.11	2.5	3.1	1.9

Fig. 2. Dataset generation process.

promising models for an extended number of epochs on the training set. Finally, we fine-tuned the models by manually monitoring the learning rates for the final few epochs. The threshold for segmentation is also treated as a model hyperparameter.

5.2. Results

To the best of our knowledge, no evaluation metric has yet been proposed for the joint task attempted in this work. The subtasks themselves have well established evaluation measures, however. Therefore, we separately evaluated the various subtasks.

For evaluation, we used word error rate (WER) and diarization error rate (DER), the most common evaluation metrics for ASR and speech diarization, respectively. On the ASR subtask, we obtained a WER of 12 on the Libri Speech corpus and 8.25 on the AN4 corpus. This is comparable to the WER of 7.89 reported by Deep Speech [9] and 5.33 reported by Deep Speech 2 [3] on the Libri Speech corpus. The state-of-the-art DER performance varies between 3 to 20 [14]. Our system achieved a DER of 2.51.

We also computed precision, recall and f1 scores for the segmentation task. Since we theoretically treated the subtask as a binary classification problem over the speech frames, we highlight the high precision, recall and f1 scores of 0.82, 0.87 and 0.84, respectively, on the test set as indicative of the success of our system. We also noted that the subtask is deceptively challenging, since end-of-segment markers are sparse, $\leq 1\%$ for most samples. To tackle this challenge effectively, we bias the cross-entropy calculations, penalizing the model by a factor of 100 for missing an end-of-segment marker. Furthermore, we allow the model to predict an end-of-segment marker within a small window of 2 frames in either direction of time to account for small variations.

6. CONCLUSION

We presented a novel deep learning model which jointly performs the tasks of speech identification, segmentation and diarization. We trained our model on an adaptation of the open source Libri Speech dataset, developed to include segmentation and diarization labels. Our results are comparable to the state-of-the-art systems in each of the tasks.

7. REFERENCES

[1] Xavier Anguera Miró and Jean-François Bonastre, "Fast speaker diarization based on binary keys," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 2011, pp. 4428–4431.

[2] Xavier Anguera, Chuck Wooters, and José M. Pardo, "Robust speaker diarization for meetings: ICSI rt06s evaluation system," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[3] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," *CoRR*, vol. abs/1512.02595, 2015.

[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[5] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 1764–1772.

[6] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[7] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, "Listen, attend and spell," *CoRR*, vol. abs/1508.01211, 2015.

[8] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 2006, pp. 369–376.

[9] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[10] Tara N. Sainath, Oriol Vinyals, Andrew W. Senior, and Hasim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4580–4584.

[11] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Lluc Gauvain, "Improving speaker diarization," in *IN PROC. FALL 2004 RICH TRANSCRIPTION WORKSHOP (RT-04*, 2004.

[12] Margarita Kotti, Vassiliki Moschou, and Constantine Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.

[13] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, 2010, pp. 249–256.

[14] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb 2012.