

# CS753 : Automatic Speech Recognition

## Assignment 3

Arka Sadhu  
140070011

### Abstract

The problem statement considered here is that given an audio recording from a radio station we want to isolate the parts of the recordings that correspond to human speech. This will be useful for further downstream tasks like sentiment classification, audio summarization etc. The dataset we have includes audio clips annotated with intervals that contain human speech. Here we present two methods for this task. First is the use of Conditional Random Fields (CRF) which uses a discriminative model as opposed to the generative model used by Hidden Markov Models (HMM) and also circumvents bias problems inherent in other discriminative models like Maximum Entropy Markov Models (MEMMs). Second is the use of Segmental Recurrent Neural Networks which defines a joint probability distribution over segmentations of the input and labellings of the sequence.

### Problem Statement

We define the observation sequence as observation vectors upto time  $T$  (which is assumed to be fixed) as  $O = [o_1, o_2, \dots, o_T]$ . We also assume that we already know the mapping from the observation vector index to the actual speech duration which is to say that we know some observation vector  $o_i$  would correspond to the duration  $t_i : t_j$ . Therefore our only task is to label the observation sequence vectors. The label space we consider here is  $l = \{l_h, l_n\}$  with  $l_h$  corresponding to human speech and  $l_n$  corresponding to anything other than human speech which we are not bothered with. So our task reduces to predicting the label sequence  $Y = [y_1, y_2, \dots, y_T]$  such that each  $y_i \in l$ . Another formulation we introduce is to explicitly group together consecutive same labels. In this case we would like to predict the sequence  $Z = [z_1, z_2, \dots, z_p]$  and  $Y = [y_1, y_2, \dots, y_p]$  such that  $z_i \in \mathbb{Z}_+$  and  $y_i \in l$  where  $z_i$  represents the duration of a segment and  $y_i$  is the label corresponding to the segment. The first formulation is used in the conditional random fields and the second formulation is used in the segmental recurrent neural networks. The training data which is available to us is audio clips with intervals of annotated human speech. Thus for training we can get the data for either formulation.

### Methodology

Here we describe both methods Conditional Random Fields (CRF) (Lafferty, McCallum, and Pereira 2001) and Segmental Recurrent Neural Networks (sRNN) (Kong, Dyer, and Smith 2015) in more detail.

#### Conditional Random Field (CRF)

Conditional Random field is a discriminative framework which is able to model the sequence of data better in the sense of relaxing strong independence conditions in Hidden Markov Models (HMM) and overcoming the label bias problem inherent in Maximum Entropy Markov Models (MEMMs).

The independence condition in HMM is required because we are modeling the joint probability distribution  $P(Y, X)$ . Here  $Y$  is the label sequence and  $X$  is the input sequence. It is necessary to make the independence assumption for the probability distribution to be tractable. CRF overcomes this by instead modelling  $P(Y|X)$  which is the probability distribution of  $Y$  conditioned over  $X$  and not explicitly modeling the distribution  $P(X)$ . CRF gets away by not modeling  $P(X)$  because at test time the observations are anyway fixed. This also allows for context dependent information to be captured because the label sequence  $Y$  is no longer needed to be decided only by the current observation.

The paper also compares CRF to other discriminative framework like MEMMs and claims like models like MEMMs are inherently biased towards states which have smaller number of output branches. The reason for this is that when it is at a particular state the competition of the next state is only from the current state. In other words we are only comparing transitions to the next state from the current state instead of looking at every other transition from any state to any other state (including self loops as well). To avoid this either the finite state machine can be determined but this can lead to explosion of number of states or to start with fully connected model but that doesn't allow us to use our domain knowledge or priors. This problem is circumvented in CRF because the individual transitions do not explicitly choose the transition rather only amplify or dampen the probability mass they receive.

**Actual Model** Let  $X$  be the random variable over data sequence and  $Y$  be the random variable over label sequence.

Even though  $X$  and  $Y$  are jointly distributed we will try to get the probability distribution  $p(Y|X)$  and we choose not to model explicitly the distribution  $p(X)$  as it is already fixed during runtime.

We consider a graph  $G = (V, E)$  characterized by vertex set  $V$  and edge set  $E$  and let  $Y = (Y_v)_{v \in V}$  be indexed by the vertices of the vertex set. Then we define  $(X, Y)$  to be a conditional random field if the random variables  $Y_v$  when conditioned on  $X$  obey the Markov property with respect to the graph:  $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \in nbd(v))$ . Here  $nbd(v)$  is the set of vertices one-hop distance away from the vertex  $v$  or in the 1-neighbourhood of  $v$ .

## **Segmental Recurrent Neural Networks (sRNN)**

### **Task**

Here we describe how the two models explained can be used for the speech extraction task described in the problem statement.

## **Conditional Random Field (CRF)**

## **Segmental Recurrent Neural Networks (sRNN)**

### **Benefits and Shortcomings**

Here both Benefits and Shortcomings of the proposed models are given.

## **Conditional Random Field (CRF)**

## **Segmental Recurrent Neural Networks (sRNN)**

### **References**

- Kong, L.; Dyer, C.; and Smith, N. A. 2015. Segmental recurrent neural networks. *CoRR* abs/1511.06018.
- Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.