Encoding of speaker identity in a Neural Network model of Visually Grounded Speech perception

Master's thesis
Communication and Information Sciences
Specialization track Data Science: Business and Governance
Tilburg University - School of Humanities

Mark van der Laan
m.l.vdrlaan@tilburguniversity.edu
ANR: 633762
SNR: 2011906

Supervisor: dr. G. Chrupała
Second reader: dr. D. Hendrickson

July 15, 2018

Abstract

This thesis presents research on how the unique characteristics of a voice are encoded in a Recurrent Neural Network (RNN) trained on Visually Grounded Speech signals. Multiple experiments were executed to determine to what extent speaker identity is encoded. These experiments were executed against raw Mel Frequency Cepstral Coefficient (MFCC) vectors, a convolutional layer and the recurrent layers of the RNN. This thesis also describes the process of annotating perceived gender of the speakers in two audio caption corpora. The outcomes of this procedure were used to examine how well the trained RNN could distinguish male and female speakers in each layer.

The most important takeaway from these experiments is that in general gender and speaker encoding are most prevalent in the first few layers of the RNN. This finding aligns with the research that has been done on encoding of phonemes in this type of model: 'form-related aspects' are most prevalent in the first few layers while semantics are better encoded in the deeper layers.

An experiment comparing the performance between male and female speakers revealed that differences depending on the dataset occur, which were most notable in the classification on the MFCC vectors and the convolutional layer. It is not entirely certain why these differences occur but an important reason could be that the amount of male and female speakers differ per dataset.

Preface

This thesis is written to fulfill the Master program of Data Science: Business and Governance, which is a specialization track within the Master program of Communication and Information Sciences.

I would like to thank my supervisor dr. Grzegorz Chrupała for all the work he has done to make the execution of this thesis possible and the valuable feedback that he provided me with. With help of dr. Chrupała the Neural Network presented in chapter 3 was trained so that I could use the output to execute the experiments in this thesis. Additionally, I also would like to thank Alessandro Hazenberg for helping me to label the data and my family for their support.

Yours sincerely,

Mark van der Laan
Tilburg July 15, 2018

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The scientific field of computational linguistics is concerned with finding methods to systematically acquire important characteristics of natural language (Wikipedia, 2018). A subfield of computational linguistics is speaker recognition which can be defined as: *'extract, characterize and recognize the information in the speech signal conveying speaker identity'* (Reynolds, 2002). Within the field of speaker recognition multiple facets are researched like speaker identification (i.e. identify a speaker from a known set of speakers), speaker verification (i.e. reject the identity of the speaker based on the provided signal) and speaker diarisation (i.e. identify speakers in speech signals with multiple speakers) (Reynolds, 2002; Tranter and Reynolds, 2006).

This thesis mainly focuses on speaker identification. More specifically, speaker identification is researched within the context of Recurrent Neural Networks trained on Visually Grounded Speech signals. This type of Neural Network is trained on visual (i.e. images) and auditory (i.e. spoken captions) input to learn syntax and semantics from continuous speech. Contrary to older methods, like using text transcriptions, this method requires less supervision (Harwath and Glass, 2015). From a societal perspective, this method is relevant because it also allows for analysis on rarer languages which don't have large amounts of text transcriptions available. Moreover, some spoken dialects do not have a specific writing system so methods based on text transcriptions cannot encode these dialects.

Within the context of RNNs based on Visually Grounded Speech signals how speaker identification is encoded is currently not known. Therefore, the research question for thesis is:

'From which part of the Recurrent Neural Network can speaker identity be decoded with the best accuracy?'

Previous work looked into this type of Neural Network to see how phonemes are encoded. Researchers (Alishahi et al., 2017) found that phoneme encoding follows a certain pattern: 'form-related aspects' are more prevalent in the first few layers of the Neural Network while semantic information is better encoded in the deeper layers. This finding aligns with the general understanding of deep Neural Networks, which presumes that deeper layers can learn more complex patterns (LeCun et al., 2015). The hypothesis for this research question is that speaker identification also follows this pattern. This implies that the first layers better encode speakers due to the fact that speaker identification is a form-related aspect. To further analyze how speaker identity is encoded in the RNN, I also executed two experiments regarding the encoding of perceived gender [1]. During the execution of these two experiments I also examined whether the encoding of perceived gender suffers from any bias (e.g. females consistently score better than males).

The most important finding of this thesis is that speaker identification (i.e. encoding of speakers and perceived gender) is most notable in the first recurrent layers. Moreover, the experiment regarding performance differences between male and female speakers indicates that there are

[1] The annotation of the perceived gender of the speakers can be found in the data/annotation directory in https://github.com/markvdlaan93/vgs-speaker-identification.

differences but they aren’t consistent and most likely caused by the representation of the data.

Chapter 2

Background

In this chapter I will provide background information and previously established work that is relevant to this thesis. In section 2.1, I will provide an overview of what already has been done on the automatic identification of speakers. In section 2.2 previous work on Visually Grounded Speech is provided, while section 2.3 gives an overview of literature that considers gender bias in Machine Learning.

2.1 Speaker identification

The first attempts of speech recognition date back to at least the seventies of the 1970s (Huang et al., 2014). The urge for speaker recognition systems came from different disciplines like law enforcement (e.g. monitoring prison calls), access control (e.g. giving access to a system based on voice verification) and technology firms (e.g. personalizing web content) (Reynolds, 2002). The performance of first generation speech recognition systems were hindered by constraints that occurred during that era like limited computational power, lack of understanding how to cancel out background noise and lack of advanced models like Neural Networks or Support Vector Machines.

In recent years, the scientific community have put a lot of effort into enhancing the performance of speech recognition systems. Before 2010, solely using Hidden Markov Models (HMM) was a popular method to process speech signals that already used MFCC vectors as input. However, a HMM is limited in terms of expressive power which implies that it is difficult to model complex signals with this particular type of model (Huang et al., 2014). Nowadays, the use of different types of Neural Networks is popular in high performance systems. Although Neural Networks do need a great amount of data to be effective, in general they can model complex signals with high performance. A specialized form of a Neural Network, called a Recurrent Neural Network (RNN), is suitable for modeling sequential data. This is of great importance for speaker recognition. Solely using recurrent architectures like Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU) have proven to be more effective than previously established methods, like combining a RNN with a HMM (Graves et al., 2013). The model used to analyze the encoding of speakers in this thesis uses a variant of the GRU architecture.

2.2 Visually Grounded Speech

Models with speech and accompanying text have been a popular method in Machine Learning to learn from speech signals (Harwath and Glass, 2015). Researchers presented a model (Harwath and Glass, 2015) that substituted text with relevant images. This methodology requires less supervision than using text transcriptions. However, it is harder to analyze and extract linguistic characteristics from speech signals, compared to using text transcriptions, because the same word pronounced by different speakers can result in differences in the analysis. Therefore, the authors decided to only train the model on words instead of whole sentences in order to reduce the complexity. The conclusion of this research is that a model focused on learning words in speech signals and accompanying images can learn linguistic constructs. An interesting factor is that this way of learning resembles the way humans learn more closely, because humans are enabled to learn by just grounding stimuli in their sensory environment.

In (Harwath et al., 2016) the authors also developed a model based on Visually Grounded Speech signals and extended this idea. Contrary to the previous study (Harwath and Glass, 2015), this research focuses on continuous speech signals instead of just words. Similar to previous research (Harwath and Glass, 2015), this study grounds images and speech signals by mapping the modalities in the same semantic space, where a pair that describes the scene has a higher similarity score than a pair which does not. This research revealed that this type of model can learn from continuous speech.

Previous studies (Harwath and Glass, 2015; Harwath et al., 2016) used a convolutional architecture. In (Chrupala et al., 2017) they also used images and continuous speech directly, but instead of a convolutional architecture, a recurrent multi-layer architecture was used. This recurrent multi-layer architecture performs better than the previously proposed convolutional architecture. Research also revealed how the different layers encode linguistic information such as semantics. They found that 'form-related aspects' (e.g. how different speakers pronounce the same word) and semantics (e.g. identifying which words are important in a sentence) are better encoded in different parts of the Neural Network.

This recurrent multi-layer architecture was used to analyze how phonemes are encoded (Alishahi et al., 2017). The model was trained on synthetic speech from the Microsoft Common Objects in Context (MS COCO) dataset. The results of the executed experiments, like phoneme decoding and phoneme discrimination, aligned with what has been found in (Chrupala et al., 2017): semantics are more prevalent in the deeper layers while form-related aspects are more prevalent in the first few layers.

2.3 Gender bias in Machine Learning

Multiple characteristics of the human voice influence how models interpret and classify speech signals. Characteristics like gender, accent and age are important to the performance of speaker identification models (Abdulla et al., 2001). If these characteristics are insufficiently addressed certain speakers could have higher error rates than others.

Researchers (Tatman, 2017) examined how accent and gender influence the results in the Automatic Speech Recognition (ASR) system of Youtube. The most important finding is that females and people from Scotland consistently obtain lower rates of accuracy than males and other accents. Conclusions are based on a so-called 'accent challenge' with a low amount of participants (N=80). This experiment required the participants to pronounce a list of words which can clearly discriminate between different accents within a language.

Research on gender bias in speaker identification is limited and it is therefore not possible to give definitive answers on why differences occur (Abdulla et al., 2001; Tatman, 2017). One possible reason is that there is an imbalance in the data. For example, there is a possibility that the model of the Youtube ASR is trained on more male than female feature vectors. Another possible reason for bias is differences in pitch between males and females (Latinus and Taylor, 2012). Pitch differences occur due to biological reasons (e.g. length of the vocal tract) and the style of the speaker (Meena et al., 2013). The difference in pitch is a factor that may contribute to performance differences, however, in the analysis of the ASR of Youtube (Tatman, 2017) no significant effect was found.

Inspired by this study, an experiment in this thesis is presented which examines whether any performance differences between male and female speakers occur. I have specifically chosen gender because this is the only mentioned characteristic that is available for analysis. Caution must therefore be taken to interpret results, because confounding variables like accent and age may influence the results.

Results in this thesis cannot directly be compared to (Tatman, 2017) because bias is examined in a restrained environment based on words instead of continuous speech fragments. Continuous speech fragments are substantially more subject to background noise. Moreover, the experiments in this thesis differ from (Tatman, 2017) because the accuracy scores are obtained through a system which focuses on speech recognition.

Chapter 3

Methods and datasets

This chapter describes the specifics of the dataset and how the research is going to be validated. The key characteristics of the dataset are explained in section 3.1. In section 3.2, I describe the procedure that is used to label the perceived gender of each speaker in each dataset [1]. Finally, remarks on the distribution of the speakers in the dataset are specified in section 3.3 and section 3.4 explains the characteristics of the models that were used.

3.1 Data

Contrary to the dataset used in the experiments on phoneme encoding (Alishahi et al., 2017), the dataset for this thesis consists of human spoken utterances instead of synthetic speech. I started with the Flickr8K audio caption dataset (Harwath and Glass, 2015; Hodosh et al., 2013) which consists of 8,000 images with five spoken captions each generated by more than 180 participants in Amazon's Mechanical Turk. I also use the Places dataset in this thesis (Harwath et al., 2016; Harwath and Glass, 2017). Like the Flickr8K dataset, the Places dataset consists of human spoken utterances collected through Amazon's Mechanical Turk but this dataset is much larger with more than 200,000 utterances and 1,400 participants and is contextually richer than the Flickr8K dataset (Harwath and Glass, 2017). This becomes clear when the audio recordings of the Places validation set are played: the audio records are longer, contain more words and the pauses between words are longer.

3.2 Labeling of gender

To annotate the perceived gender of the speakers, I listened for each speaker to an audio recording provided in the validation dataset of Flickr8K and Places. For each speaker, to make sure no mistakes are made, the audio recording are also labelled by a second listener. After the second round, the results of the first two rounds were compared and I checked whether any differences between these two rounds occurred. For the audio recordings of the speakers which didn't match in the first two rounds, another audio recording was selected to determine whether the speaker was male or female.

3.3 Distribution

To determine whether the classes are imbalanced, I decided to plot the distribution of the speakers presented in Figure 3.1. Both datasets have in common that distribution is skewed towards a few speakers which recorded most of the audio captions in the dataset. Caution must therefore be taken when evaluation metrics like accuracy are interpreted (accuracy and F1-scores are both

[1] All the code that have been used to execute the experiments can be found in https://github.com/markvdlaan93/vgs-speaker-identification.

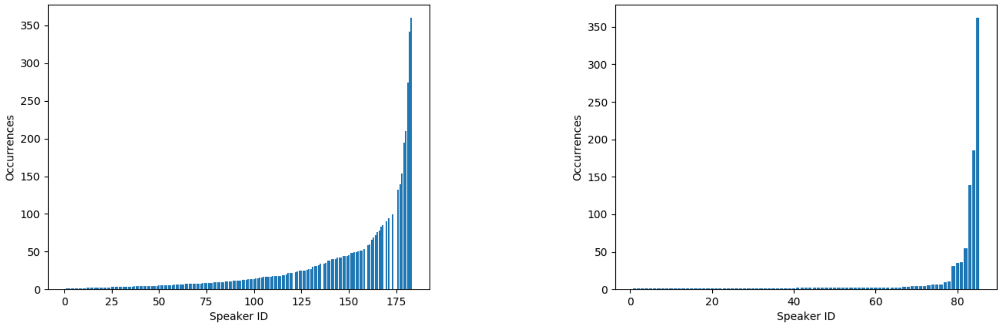

Figure 3.1: Amount of feature vectors per speaker for the Flickr8K dataset (left) and the Places dataset (right)

presented for each experiment). The majority speaker in the Flickr8K dataset accounts for 360 entries (7.2% of the total amount of feature vectors) while the majority speaker of the Places dataset accounts for 362 entries (36.2% of the dataset). These numbers will be used in the speaker identification experiments as a baseline. For gender classification, which has two classes, the baseline will be based on the percentage of the most common group. For the Flickr8K dataset this is 53.95% and for the Places dataset 50%.

The results of section 3.2 are presented in Table 3.1. For the Places dataset it was necessary to remove one speaker with a single entry because for some reason the audio record was not available in the package of the dataset (Harwath et al., 2016) although the supplemented instructions said otherwise.

	Flickr8K	**Places**
Male	2,697	550
Female	2,303	449
Total	5,000	999

	Flickr8K	**Places**
Male	79	43
Female	104	41
Total	183	83

Table 3.1: Amount of male and female speakers entries (left) and total amount of male and female speakers in the validation sets (right).

3.4 Models

This section provides an overview of the models that I have used in order to execute the experiments and see whether speaker identity and perceived gender are encoded in the Neural Network. In subsection 3.4.1 I will give a quick overview of the model that is used to train the data while subsection 3.4.2 presents the models that were used to classify each layer of the validation set.

3.4.1 Recurrent Highway Network

The model used to train the data is called a Recurrent Highway Network (RHN) (Zilly et al., 2017) which is similar to a model based on the GRU architecture (Chrupala et al., 2017). Full explanation of the model is provided in (Chrupala et al., 2017), I will only give a short overview here.

For a model based on Visually Grounded Speech, one requires both an image and a speech encoder. Together, the encoders need to make sure that spoken captions and images that are related to each other do have a closer distance measure than unrelated captions and images. The distance measure in this model is the cosine similarity calculated by the following loss function:

$$\sum_{u,i}\left(\sum_{u'} max[0,\alpha + d(u,i) - d(u',i)] + \right. \tag{3.1}$$
$$\left. \sum_{u'} max[0,\alpha + d(u,i) - d(u,i')]\right)$$

In this formula, u stands for utterance and i for image and u and i together is an image-caption pair which describe the same scene. U' and i' are the set of utterances and images that not describe the same scene.

The image encoder used to process the images is a pre-trained VGG-16 model (Simonyan and Zisserman, 2014) which is fed with vector representations of the image. The utterance encoder has the following structure:

$$enc_u(\mathbf{u}) = unit(Attn(RHN_{k,L}(Conv_{s,d,z}(\mathbf{u})) \tag{3.2}$$

In this model input u consists of MFCC vectors and the RHN with k recurrent layers and L depth uses the output of a convolutional layer with size s, d dimensions and stride z. The model for both datasets is initialized with feeding the MFCC vectors (37 dimensions for the Flickr8K dataset and 13 dimensions for the Places dataset) to the convolutional layer of 64 hidden units. The Recurrent Highway Network consists of four layers with each 1,024 dimensions while the attention layer is a Multilayer Perceptron (MLP) which calculates the average of the activations. The feature vectors after the attention layer have 1,024 dimensions (throughout the thesis I will refer to this output as the embedding layer).

3.4.2 Linear models

For each layer in the trained Neural Network (i.e. convolutional layer, recurrent layers and embedding layer) I have at my disposal the normalized mean activation values for each feature vector in the dataset. The Flickr8K dataset has a validation set of 5,000 feature vectors and the Places validation set 1,000. The main task is to feed each layer of mean activations values (i.e. matrices of 5,000 by 1,024 for Flickr8K and matrices of 1,000 by 1,024 for the Places dataset) into a linear classifier and supplement it with the labels I also have at my disposal. This implies that the experiments in this thesis differ from training of the model itself because I used strongly supervised learning as opposed to the weak supervision of Visually Grounded Speech signals.

The Stochastic Gradient Descent (SGD) classifier of the Python Scikit-Learn package was used in conjunction with Grid Search and Cross Validation. In Table 3.2 an overview is given of the parameters that I used to tune each layer. For the loss function I tried 'log' which gives Logistic Regression and 'Hinge' which gives a linear Support Vector Machine. Because of limited computational resources I decided to minimize Cross Validation only to three folds for each layer within each model.

	Values
Learning rate	0.01, 0.001, 0.0001 and 0.00001
Loss function	Log and Hinge
Penalty	L_1, L_2 and ElasticNet

Table 3.2: Values used in Grid Search in conjunction with K-fold Cross Validation.

Chapter 4

Experiments

In this chapter, I will present a number of experiments to demonstrated to what extent speakers are encoded in the proposed RNN. In section 4.1, I explain every aspect that is relevant to executing the experiments, followed by the results of the experiments in section 4.2.

4.1 Experimental setup

In subsection 3.4.2, I have explained that each layer is trained with 3-fold Cross Validation in conjunction with Grid Search. To verify whether the model generalized sufficiently, I split the data into a training and test set. The training set was internally split by Scikit-learn into stratified partitions of equal length. This means that this method accounts for class imbalances presented in section 3.3. After training with Cross Validation and Grid Search the model was trained again but this time on the entire dataset. This methodology is repeated for each layer within each dataset and the code for running the experiments is partially taken from the Scikit-learn documentation (Scikit-learn, 2017a). To examine whether the presented Neural Network suffers from performance differences between male and female speakers, the accuracy and F1-scores per gender are derived from the experiment on gender encoding.

4.1.1 Splitting the data

To account for the small size of the Places dataset (i.e. each layer consists of more dimensions than feature vectors), I decided to split the training and test sets in 60% / 40% proportions. Because the amount of feature vectors is substantially larger for the Flickr8K dataset, I decided to split the experiments on the Flickr8K into 67% for the training set and the remaining part for the test set (i.e. 33%).

4.1.2 Stratification

By default, the 'train and test split'-function of Scikit-learn is not stratified (Scikit-learn, 2017c). Due to time constraints I decided to not stratify the classification of speaker identity on both the Flickr8K and Places datasets. The 'train and test split'-function can only be used in stratification mode, when each class in the dataset has at least two members. For gender classification this is not a problem because each class (i.e. male and female) has at least two members, so no extra pre-processing is required. For speaker identification, this would require me to do more pre-processing which I ultimately did not have time for. However, I checked whether remaining single entries were considered in the calculation of the evaluation metrics like the accuracy and F1-score. Internally Scikit-learn ignores these observations by giving them a zero as F1-score and not considering the single entry speakers when averaging the evaluation metric over the classes.

4.1.3 Evaluation metrics

The 'Grid Search Cross Validation'-function of Scikit-learn only has the possibility to provide a single scoring function. The parameter 'f1_weighted' is used which calculates the F1-score per class and averages the results based on the amount of instances in a certain class (Scikit-learn, 2017b). In section 3.3, I explained that I will provide both accuracy and F1-scores of each experiment. However, since the scores on the training folds will be F1-scores, I take this metric as a guideline for interpreting the results. If there is any notable difference between the F1-score and accuracy for a certain layer, I will mention that and refer to Appendix A.

4.2 Results

For each proposed experiment I will provide the results in this section. In subsection 4.2.1, I will start with explaining observations for speaker identification on the Flickr8K data while subsection 4.2.3 focuses on speaker identification in the Places dataset. In subsection 4.2.2 and subsection 4.2.4, I will elaborate on gender classification and subsection 4.2.5 will provide results for the research regarding gender bias.

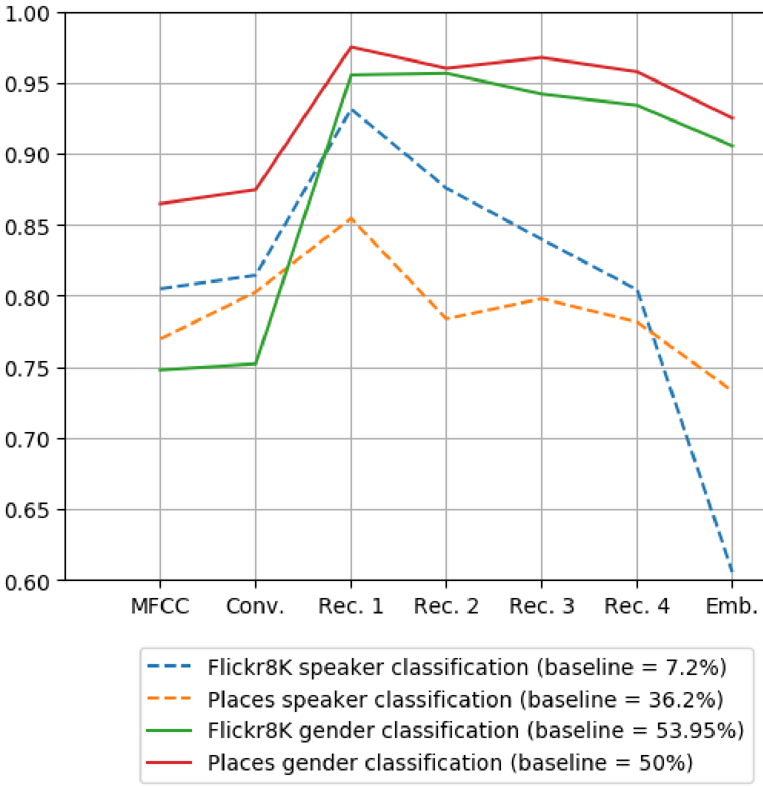

Figure 4.1: F1-scores for each experiment.

4.2.1 Speaker identification in Flickr8K dataset

In Table 4.1 an overview is given of the optimal parameters for each layer. For each layer the optimal loss function, penalty and learning rate are presented. The column 'Average training F1-score' specifies the average F1-score over each fold for the optimal parameters. The standard deviation is an important indicator to provide insight into how consistent the predictions are across the different folds. For speaker identification in the Flickr8K dataset Table 4.1 shows that the standard deviation is small (not greater than 0.05). This overview also shows that for each layer the regularization parameter L_2 returned the best results.

In Figure 4.1 the performance of each experiment is presented. For speaker identification in Flickr8K dataset the classification on the raw MFCC vectors and the convolutional layer consists of approximately the same results which is followed by an increase in the first recurrent layer. After the first recurrent layer, the performance descends steadily till a point where the fourth recurrent layer and embedding layer are performing even worse than the raw MFCC vectors and the convolutional layer.

	Loss function	Penalty	Learning rate	Average training F1-score	Std.
MFCC	Log	L_2	0.001	0.800	0.022
Convolutional	Log	L_2	0.001	0.803	0.030
Recurrent 1	Hinge	L_2	0.001	0.925	0.026
Recurrent 2	Hinge	L_2	0.01	0.871	0.038
Recurrent 3	Hinge	L_2	0.01	0.816	0.039
Recurrent 4	Hinge	L_2	0.01	0.773	0.049
Embedding	Hinge	L_2	0.0001	0.580	0.031

Table 4.1: The optimal parameters for each layer for speaker identification on the Flickr8K dataset.

4.2.2 Gender identification in the Flickr8K dataset

Similar to the results presented in subsection 4.2.1, the folds in the training dataset don't have much variance. The pattern consists of relatively low scores for the MFCC and convolutional layer followed by an increase (see Figure 4.1). Contrary to the results in subsection 4.2.1, the second recurrent layer scores approximately the same as the first layer. This pattern is, however, not visible in the average F1-scores of the training dataset (see Table 4.2). After the second layer the performance decreases but the drop in the embedding layer isn't of the same magnitude as that of speaker identification in Flickr8K.

	Loss function	Penalty	Learning rate	Average training F1-score	Std.
MFCC	Hinge	L_2	0.01	0.757	0.027
Convolutional	Hinge	L_1	0.01	0.765	0.024
Recurrent 1	Log	ElasticNet	0.001	0.949	0.010
Recurrent 2	Hinge	L_2	0.01	0.938	0.015
Recurrent 3	Hinge	L_2	0.01	0.929	0.003
Recurrent 4	Log	L_2	0.01	0.919	0.014
Embedding	Log	ElasticNet	0.0001	0.896	0.006

Table 4.2: The optimal parameters for each layer for gender identification on the Flickr8K dataset.

4.2.3 Speaker identification identification in Places dataset

In Table 4.3 an overview is given for the optimal parameters for each layer. Contrary to the other results the variance for speaker identification on the Places dataset is much higher (around +/- 0.2). Similar to the other experiments the pattern starts with relatively low scores for MFCC and convolutional layer followed by a sharp increase. After the sharp increase, the performance drops slightly in the second recurrent layer followed by a small increase in third layer. The fourth recurrent and embedding layer encode speaker identification worse than the other layers.

	Loss function	Penalty	Learning rate	Average training F1-score	Std.
MFCC	Log	ElasticNet	0.001	0.779	0.220
Convolutional	Log	ElasticNet	0.001	0.812	0.189
Recurrent 1	Hinge	L_2	0.01	0.866	0.197
Recurrent 2	Hinge	ElasticNet	0.01	0.778	0.276
Recurrent 3	Hinge	L_2	0.01	0.774	0.273
Recurrent 4	Hinge	ElasticNet	0.01	0.775	0.280
Embedding	Hinge	L_2	0.0001	0.695	0.264

Table 4.3: The optimal parameters for each layer for speaker identification on the Places dataset.

4.2.4 Gender identification in the Places dataset

Consistent with encoding of male and female speakers in the flickr8K dataset, the results in Table 4.4 show that the variance in the training folds is low. Figure 4.1 shows that the largest performance score is reached in the first recurrent layer. After the third recurrent layer, similar to aforementioned experiments, the performance drops.

	Loss function	Penalty	Learning rate	Average training F1-score	Std.
MFCC	Log	L_1	0.01	0.897	0.009
Convolutional	Hinge	ElasticNet	0.001	0.897	0.014
Recurrent 1	Log	L_2	0.001	0.974	0.000
Recurrent 2	Log	L_2	0.001	0.965	0.018
Recurrent 3	Log	ElasticNet	0.001	0.965	0.018
Recurrent 4	Log	ElasticNet	0.001	0.966	0.006
Embedding	Hinge	L_1	0.001	0.951	0.951

Table 4.4: The optimal parameters for each layer for gender identification on the Places dataset.

4.2.5 Gender bias

The results for both the Flickr8K and Places dataset are presented in Figure 4.2. For both datasets and both genders the performance on the raw MFCC vectors and convolutional layer are approximately equal but low compared to the other layers. After the convolutional and MFCC layers, for each dataset and for each gender a large increase is observed. Differences between gender are most notable in the MFCC and convolutional layer. After the MFCC and convolutional layer, both genders in both datasets are much closer to each other.

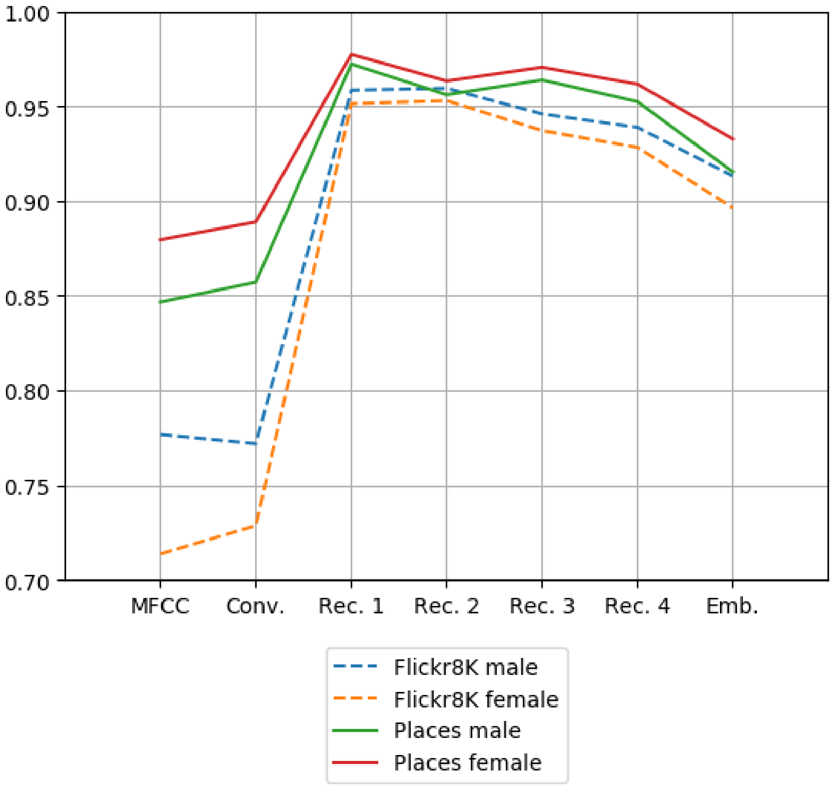

Figure 4.2: F1-scores for speaker identification split by perceived gender in both the Places and Flickr8K dataset.

Chapter 5

Discussion

In this chapter the findings from chapter 4 are discussed. In section 5.1, I will discuss the results of the experiment on speaker identification on the Flickr8K dataset. Next in section 5.2, I present results on gender identification in the Flickr8K dataset followed by the analysis of both speaker and gender identification in the Places dataset in section 5.3 and section 5.4. This section ends with an analysis on how performance differences between male and female should be interpreted in section 5.5.

5.1 Speaker identification in Flickr8K dataset

The results in subsection 4.2.1 reveals that the standard deviation between the average F1-scores of the different folds is low. Based on the experiments of phoneme encoding (Alishahi et al., 2017), I assumed that speaker identification would follow the same trend. From all the experiments that are executed in this thesis, speaker identification in Flickr8K dataset follows this pattern most consistently.

In subsection 4.1.2, I mentioned that, I didn't use the stratification option in the procedure of splitting between the test and train sets. This may cause poor generalizability because the speakers aren't equally divided over the train and test set. If so, the average F1-score over the folds and the F1-score over the test set should have differed significantly which is not the case. Another finding is that the classification on the convolutional layer and MFCC vector follow the same pattern as the experiments presented in phoneme encoding (Alishahi et al., 2017) and each layer scores far above the mentioned baseline of 7.2%.

5.2 Gender identification in Flickr8K dataset

In contrast with the other experiments, the encoding of perceived gender in the Flickr8K dataset is most notable in the first and second recurrent layer of the model. Although the second recurrent layer scores higher than the first layer, the difference between the two layers is negligible (only 0.0012 difference). Consistent with results in other experiments, the performance drops after the first two recurrent layers and the model scores better than the baseline of 53.95%. Comparing the performance in the training (see Table 4.2) and test set (see Figure 4.1 and Appendix B) reveals that confirmation of the hypothesis is more evident in the training dataset. Why this difference occurs is not certain because for the experiment on encoding of gender the train and test sets are stratified. Moreover, the trained model generalizes adequately because for every layer the score on the test set doesn't differ much compared to the training set (see Appendix B).

5.3 Speaker identification in Places dataset

Contrary to the results presented in section 5.1, the presented results in subsection 4.2.4 suggest that classification on the Places dataset does suffer from more variance. This is probably caused by the lack of stratification and the limited amount of data available, by which I mean that each fold doesn't have a wide range of examples available for a larger portion of the speakers.

Similar to the results showed in section 5.1, the first recurrent layer gives the best results. However, the trend after the first recurrent layer is less similar to speaker identification on the Flickr8K dataset but a downward trend is visible. Especially if only the average F1-score of the folds are considered in Table 4.3 the results confirm the hypothesis. Looking at the difference between training and test F1-score for each layer, I don't observe much difference between the scores. In fact, the trained model for each layer generalizes well and in most cases only differs +/- 0.01. Consistent with the average F1-scores of the training data, the accuracy scores in Figure A.1 also show an exclusively descending line starting from first recurrent layer. Comparing the performance to the baseline of 36.2%, each layer scores better than this threshold.

5.4 Gender identification in the Places dataset

Compared to the experiment on encoding of perceived gender on the Flickr8K dataset, encoding of gender is more notable in every layer in the Places dataset. Why gender is encoded better in the Places dataset is not entirely certain. One could argue that the Places dataset is contextually richer which leads to better performance (i.e. the model has more information to its disposal). However, this is not entirely consistent with the experiments on speaker identification in both datasets. Except in the embedding layer, the encoding of speakers is more notable in all layers in the Flickr8K dataset compared to the Places dataset. Another reason could be that the Places dataset has less speakers (83 speakers in the Places dataset versus 183 speakers in the Flickr8K dataset) which leads to less variability in the input signals.

Similar to the other experiments, generalizability of the experiment on encoding of males and females in the Places dataset is adequate and the model performs better than the baseline (50%). For some layers the performance is slightly lower in the test set (e.g. the fourth recurrent layer has a F1-score of 0.966 compared to 0.9575 in the test set) than in the training set.

Comparing the performance of gender and speaker identification in both the Flickr8K and Places dataset, the most notable difference is that the experiments on speaker identification do descend more rapidly through the recurrent layers.

5.5 Gender bias

Inspired by previously executed experiments (Tatman, 2017), I decided to research whether gender bias occurs in Neural Networks based on Visually Grounded Speech signals. The results for the convolutional and MFCC layer of both datasets have conflicting scores. In the Flickr8K dataset the first layers score consistently better for male speakers than for female speakers, while for the places dataset it is the other way around. Analyzing all layers, this pattern continues to occur although the differences in the recurrent and embedding layers are negligible. Comparing the F1-scores to the accuracy scores in Figure A.2, most of the layers are consistent except the second recurrent layer for females in the Places dataset.

It is hard to say why theses differences occur. Although research cannot be directly compared, in (Tatman, 2017) it became clear that unbalanced classes may be an issue which could cause differences in performance. The distribution of Places and Flickr8K datasets are given in section 3.3. Analyzing the differences in gender between the datasets reveals that Flickr8K consists of more males and Places consists of more females. This can possibly clarify why gender scores for the different datasets give different results: there is bias in the data. Although Table 3.1 indicates that the amount of feature vectors does not differ much between male and female speakers, the algorithm gets a more diverse set of male speakers in the Flickr8K dataset and female speakers in the Places dataset. Even though the data is stratified, it does not account for this issue because stratification only makes sure that the split groups consist of a proportional amount of male and female examples (Dwyer, 2015). Techniques like upsampling and undersampling can be potentially used to solve this problem (Wikipedia, 2017).

Another reason for differences between male and female scores could possibly be contributed to pitch differences (see section 2.3). If so, then the scores should differ in favour of one class in each dataset. This is, however, not the case thus it is not likely that pitch differences causes different performance scores for male and female speakers.

Chapter 6

Future work and conclusion

In this chapter, I will give a final conclusion based on the results presented in chapter 4 and chapter 5. I will also specify future work in section 6.2 which is necessary to further examine the extent to which speaker and gender encoding are processed by RNNs based on Visually Grounded Speech.

6.1 Conclusion

In this thesis, I presented novel research on the encoding of perceived gender and speaker identity in Recurrent Neural Networks based on Visually Grounded Speech signals. Based on the results in chapter 4 and chapter 5, I conclude that in general the hypothesis is confirmed. Especially taken the F1-scores of the training set into consideration, speaker and gender encoding are most prevalent in the first few layers of the RNN.

For encoding of male and female speakers (see chapter 1 for location of annotation files) the best results were achieved in the Places dataset in all layers. Contrary to gender encoding, speaker encoding follows a different pattern, namely one in which encoding of speakers achieves the best results in the Flickr8K dataset. Comparing encoding of gender and speaker identity, the most notable finding is that the prevalence of speaker encoding descends more quickly through the recurrent layers than encoding of perceived gender.

Research on gender bias revealed that although there are differences for gender classification in both datasets, this is most likely caused by the representation of the data. It is not likely that pitch contributed to these performance differences because in that case the bias should be consistently in one direction (i.e. males do get higher classification scores or the other way around).

6.2 Future work

I mentioned in subsection 4.1.2 that I didn't have time to stratify the train and test split in the experiments for speaker identification. Unstratified train and test sets could potentially explain why the pattern is less clear in the test set. In future work it is therefore important to research how this limitation affects the performance. Besides stratifying the data, the use of a Gradient Reversal layer would also be interesting to see how it affects the performance scores of each layer in each dataset. By using a Gradient Reversal layer (Ganin et al., 2016) the representation of the speaker can be made more invariant which could lead to better performance.

Bibliography

Abdulla, W. H., Kasabov, N. K., and Zealand, D.-N. (2001). Improving speech recognition performance through gender separation. *changes*, 9:10.

Alishahi, A., Barking, M., and Chrupala, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. In *CoNLL*, pages 368–378. Association for Computational Linguistics.

Chrupala, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *ACL (1)*, pages 613–622. Association for Computational Linguistics.

Dwyer, A. (2015). Managing unbalanced data for building machine learning models. http://www.simafore.com/blog/handling-unbalanced-data-machine-learning-models.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE.

Harwath, D. and Glass, J. (2017). Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–517. Association for Computational Linguistics.

Harwath, D., Torralba, A., and Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

Harwath, D. F. and Glass, J. R. (2015). Deep multimodal semantic embeddings for speech and images. In *ASRU*, pages 237–244. IEEE.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103.

Latinus, M. and Taylor, M. J. (2012). Discriminating male and female voices: differentiating pitch and gender. *Brain topography*, 25(2):194–204.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Meena, K., Subramaniam, K. R., and Gomathy, M. (2013). Gender classification in speech recognition using fuzzy logic and neural network. *Int. Arab J. Inf. Technol.*, 10(5):477–485.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on*, volume 4, pages IV–4072. IEEE.

Scikit-learn (2017a). Parameter estimation using grid search with cross-validation. http://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html.

Scikit-learn (2017b). Sklearn metrics f1 score. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

Scikit-learn (2017c). Sklearn model selection train test split. http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Tatman, R. (2017). Gender and dialect bias in youtube's automatic captions. In *EthNLP@EACL*.

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

Wikipedia (2017). Oversampling and undersampling in data analysis - Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis.

Wikipedia (2018). Computational linguistics - Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Computational_linguistics.

Zilly, J. G., Srivastava, R. K., Koutník, J., and Schmidhuber, J. (2017). Recurrent highway networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4189–4198, International Convention Centre, Sydney, Australia. PMLR.

Appendix A

Accuracy scores for experiments

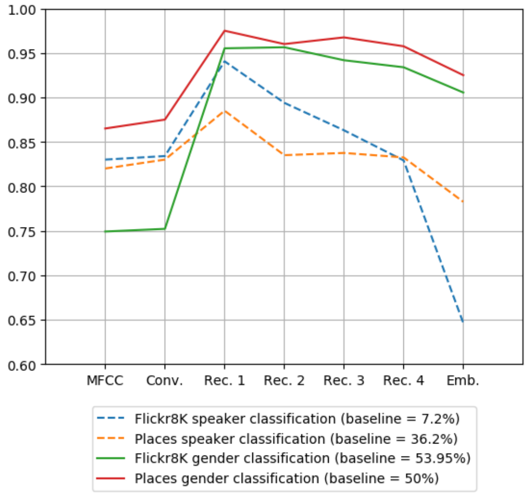

Figure A.1: Accuracy scores for each experiment.

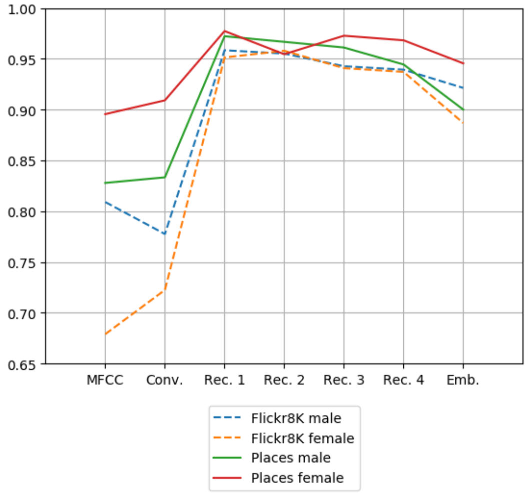

Figure A.2: Accuracy scores for gender in both the Places and Flickr8K dataset.

Appendix B

Difference between training and test scores

	Training F1-score	Test F1-score	Difference
MFCC	0.800	0.8049	0.0113
Convolutional	0.803	0.8143	0.0113
Recurrent 1	0.925	0.9313	0.0063
Recurrent 2	0.871	0.8756	0.0046
Recurrent 3	0.816	0.8396	0.0236
Recurrent 4	0.773	0.8042	0.0312
Embedding	0.580	0.6055	0.0255

Table B.1: Exact difference (test score minus training score is the difference) in F1-score between prediction on the training and test set for the Flickr8K speaker identification experiment.

	Training F1-score	Test F1-score	Difference
MFCC	0.757	0.7477	-0.0093
Convolutional	0.765	0.7520	-0.0130
Recurrent 1	0.949	0.9552	0.0062
Recurrent 2	0.938	0.9564	0.0184
Recurrent 3	0.929	0.9418	0.0128
Recurrent 4	0.919	0.9339	0.0149
Embedding	0.896	0.9054	0.0094

Table B.2: Exact difference in F1-score between prediction on the training and test set for the Flickr8K gender identification experiment.

	Training F1-score	Test F1-score	Difference
MFCC	0.779	0.7695	-0.0095
Convolutional	0.812	0.8026	-0.0094
Recurrent 1	0.866	0.8544	-0.0116
Recurrent 2	0.778	0.7836	0.0056
Recurrent 3	0.774	0.7979	0.0239
Recurrent 4	0.775	0.7814	0.0064
Embedding	0.695	0.7329	0.0379

Table B.3: Exact difference in F1-score between prediction on the training and test set for the Places speaker identification experiment.

	Training F1-score	**Test F1-score**	**Difference**
MFCC	0.897	0.8647	-0.0323
Convolutional	0.897	0.8746	-0.0224
Recurrent 1	0.974	0.9750	0.001
Recurrent 2	0.965	0.9600	-0.0065
Recurrent 3	0.965	0.9675	0.0025
Recurrent 4	0.966	0.9575	-0.0085
Embedding	0.951	0.9249	-0.0261

Table B.4: Exact difference in F1-score between prediction on the training and test set for the Places gender identification experiment.