Sergei Popkov,
Machine Learning for Speech [UEF Summer School 2019], learning diary questions.

Monday.1) Written language consists of sequences of characters, words, sentences. What are the basic “building blocks” of spoken language, and how can they be quantitatively characterized / represented?

Basic building blocks are as follows:

1) segments – minimal fragments of speech, like vowels and consonants, may be quantitatively represented as a minimal units of speech sounds (phones), lingustic parts (phonemes), including allophones – different phonetic realizations of a phoneme;

2) syllables – minimal fragments of words that can be comfortably pronounced (produced) independently (in isolation) from each other. Intuitively represented as a count of vowels in a word;

3) words – minimal units of a language that are able to keep individual meaning and, therefore, are recognizable only by the people who know the corresponding language. The words can be counted and divided according to their boundaries;

4) phrases – language complete expressions consisted of the words. Can be detected and represented in speech based on the individual rhythmic properties as well as the intonation of a speaker;

5) suprasegmentals – various features of the speech, applied to another block of speech (such as tone, accent, stress, tempo, rhythm, etc). Usually suprasegmentals do not have individual countable characteristics and are viewed as a whole with another building block they depend on, although the countable measurement methods can be applied in some sense or given perspective (for example, we can notice that the word can’t have more than one stress, this observation allows us to count the quantity of stresses in the given phrase or statement).

Monday.2) Explain the principal difference of maximum likelihood parameter estimation and Bayesian parameter estimation. Under which circumstances one uses each approach?

Maximum likelihood gives point estimates of model parameters and treats the parameters as a constants (assuming their derivatives are equal to zero) to obtain the values, while Bayesian inference treats the parameters as random variables, and use prior parameter values (represented by distribution) to obtain distributions. Maximum likelihood provides a consistent approach to estimate the parameters (most useful when we can provide a relatable likelihood function for the distribution explored in the given task), but Bayesian estimation is more scalable and universal approach (most useful for the big data or stochastic environments).

Tuesday.1) In lectures, we discussed both Gaussians and Gaussian mixture models. What is the principal difference of the two? What kind of data can be modeled with them? Give at least one (speech) example in each case.

Gaussian mixture model is an unsupervised learning approach, which uses a given number of components – Gaussians – to generate the clusters determining the classes for the examined data. Gaussians are basically the functions that follow the rules of Gaussian distribution. A single Gaussian can be seen as a GMM with parameter C=1, where C is a quantity of the clusters. A Gaussian may be used to recognize vowels based on their spectrogram, whereas a proper GMM with C>1 can be applied to solve more sophisticated tasks, such as a speaker verification system built to estimate the likelihood that the input voice indeed belongs to the certain corresponding identity, not an impostor or another person.

Tuesday.2) Describe differences between CNN, fully connected (Dense) and RNNs.

A dense layer of the neural network is a default (most simple) approach to build a feed-forward artificial neural network. Unlike convolutional neural networks (CNNs), the dense layer is fully connected and simply transmit the signal from one perceptron to another using the given activation function. CNN, on the other hand, prefers local connectivity over global one, that is, uses the subset of weights to perform a convolution (applying the filter) where nearby inputs are locally connected to the nearby outputs, overlapping (or grouping) the corresponding neurons to gather the features – more detailed and specified representation of the input data. Local connectivity helps CNNs to learn faster than the dense layers, and pooling layers (implemented to reduce the dimensionality) help to simplify the data, too. Recurrent Neural Networks (RNNs) do not follow feed-forwarding principle, as their outputs can fed back to the inputs of a previous layer. Such design provides them with a memory that usual feed-forward networks can't have. This property makes the RNN quite useful for the sequential (temporal, time-related) data processing (such as text (sequence) translation, speech analysis (sequence classification), chatbot implementation, text prediction (sequence generation) and so on).

Wednesday.1) Why you would not typically use RNN in practice and what model you would use?

Training RNN is much more difficult than an ordinary feed-forward network. The gradient is rather unstable (explodes - goes way too big, or vanishes - becomes insignificantly small), although there's a trick to prevent the gradient explosion called the gradient clipping (setting a threshold to cut the length of the gradient vector without chainging its direction). To prevent the gradient vanishing, though, it is recommended to use another model called LSTM (Long Short-Term Memory) neural network. The hidden state activation of such modified RNN responds to the closest activations of the neighbors ("short-term" memory), but the network weights are changed along with the long sequental computations, allowing to remember the input of the cell for indefinite time ("long-term" memory). This new design of the traditional RNN allowed an activation state to be used in such a way, somewhat similar to the usual network weights, that provides the means for the preservation of the information over long distance, thus solving the problem of the input values sensitivity (which decays over time in traditional RNN models) and gradient vanishing.

Wednesday.2) What is equal error rate (EER)? For what purposes is it used in the field of speech processing?

EER is one of the metrics that summarize the system performance. This value shows the point on the DET curve where the false rejection rate is equal to the false acceptance rate. This value is used for the speaker verification problem solution, representing the most balanced trade-off between security and convenience of the application.

Wednesday.3) What are i-vector and x-vector? For what purposes are they used? What is the main difference of these two techniques?

Identity vector (i-vector) is used to represent the variable-length speech utterances in a form with fixed length and low dimensionality. The identity vector contains the person voice characteristics and channel factor (it has, basically, two major components for speaker and channel subspace, respectively). x-vector is an implementation of the i-vector task solved with a supervised model based on a deep neural network. The main difference is that joint factor analysis based i-vector approach follows a generative modeling paradigm, whereas neural networks based x-vector approach is discriminatively trained.

Wednesday.4) Navigate to YouTube (or any other Internet resource) and find at least three different examples of DeepFake videos. How convincing do they look (and sound) to you? Would you expect your selected examples to be detectable as fake videos using a machine learning approach?

At first glance, I was rather confused, as the very experience of seeing the deepfake leads to pretty troublesome thoughts, like, who can I believe these days... Some examples (like the app here https://www.youtube.com/watch?v=9OIFVm0dPLw at 1:30) are mostly recognizable as fake ones. While other, more complicated (like examples at this TED conference speech https://www.youtube.com/watch?v=o2DDU4g0PRo) are not seem to be easily distinguishable even by human. This example, https://www.youtube.com/watch?v=2daN4eRTs4A , if it would not be discarded for the fact that the face morphing is impossible, is one of the hardest to recognize, in my opinion (that is, if the frames would be presented as a bunch of non-sequental clips showing the final stages of the morphing independently each time).

Friday.1) What acoustic changes we observe in voices of boys and girls when they grow up?

Due to anatomic changes related to the growth of the child, its voice drastically changes over time, though in the different ways in regards of the gender. For example, the male fundamental frequency is keeping fading (falling down) over time since the puberty, till the adulthood. The formant frequencies of male speakers decrease faster and reach much lower values than those of female speakers. Female fundamental frequency is gradually and linearly decreasing over time of the girl growth, but never fall as rapidly as the fundamental frequency of the male child. In general, gender differences start slowly to appear at 6-10 years, take a sudden turn approximately at 11 years, become obvious by age 12, and can be described as easily noticeable around age 15.

Friday.2) Explain why considering the group level dependencies of the data in models is important in speaker recognition.

The group-level dependencies can cause their own unwanted influence on the data, as the model may learn these unwanted group dependencies (not related to the individual characteristics of each provided sample within group) during training, making training process incorrect and problematic. To avoid this, the mixed-effect models are applied to help making correct group-level predictions. The primary idea behind the mixed models is that they incorporate fixed and random effects, where a “fixed” effect is a constant parameter, and a “random” effect is a random variable parameter.

Friday.3) Explain the main difference between supervised, unsupervised, and semi-supervised machine learning.

The supervised learning uses labeled training examples that are considered trusted, i. e. data that is supposed to be true and verified. A supervised learning model tries to find the patterns in the given labeled data to correctly predict such labels for new examples unexamined before. The unsupervised learning works with somewhat “raw”, unlabeled data, there’re no hints about its relevance to any class. The model must find the relevant structure of the data that correlates the given sample to the certain cluster. The semi-supervised machine learning model uses both approaches: it tries to gather information from the given labeled data (usually, quite not much, because the labeled data generally is not cheap to produce), while investigating the rest of the data provided – that is, unlabeled data – to create the structure of clusters in the same way as unsupervised approach does.

Friday.4) In the lectures we discussed one of the large and widely adopted datasets known as VoxCeleb. Explain what this database is, and for what purposes it is used. Explain in brief the collection process of this database. What kinds of machine learning techniques were used?

The VoxCeleb project is the huge database (more than 2000 hours of audio and video data in total) of short clips (not shorter than 3 seconds long, though) containing human speech extracted from the interviews. The database uses openly available YouTube videos as sources of the collected data. The whole process involves obtaining data and performing speaker verification (confirming the identity of the speaker) using CNN-based facial recognition. The database can be applied in various fields, for research and commercial purposes. It was mostly used (as of now) in such applications as the speaker identification, the speech separation, the emotion recognition, the face generation and many others.