Good morning ladies and gentleman, welcome to the presentation of my master thesis, on which i worked together with MR.Lukas Mauch for the past half year. In the coming 30min i'll give a topic about "Deep neural network for speech emotion recognition".

As we know most of the current work of machine learning field has dedicated to computer vision and natural language processing and in NLP the linguistic information is recognized whereas the paralinguistic is discarded. But a more natural human-machine interaction requires also paralinguistic information where the personality of the speaker can also be recognized, such as age, gender and emotion.
To deal with emotion recognition we should firstly recognize the speech emotion data in the view of pattern recognition, where the emotion data is considered as ....
The state of the art in emotion recognition mostly based on GMM-HMM, where GMM tries to model the data distribution of speech data, however the work requires a lot of hand-crafted features and the number of component of GMM is generally restricted due to computation cost.
In order to model the speech emotion sufficiently, we exploit a technique called deep learning, which is a quite new field of machine learning and is currently widely disscussed.
With deep architecture we can extract complex structure and building internal representations via unsupversied learning, those ideas have seen success in vision and audio processing ....

The content of this talk can be divided into following parts. Firstly the foundation of emotion recognition is shortly introduced which followed by second chapter to talk about CRBM in detail. And, in the third chapter i going to give a introduction to deep neural network regarding its basic structure and functions and how they are trained. Afterwards the Long short term memory for sequential modelling is discussed and then the result of this work is showed. Finally i am going to draw a short conclusion and give a outlook of the future research.

MFCC is one of the most commonly used features in speech emotion recognition.
THe ...spectrum is calcuated ... and tranformed with the following equation to mel-scale in order to ....

The framework of emotion recognition is illustrated in the figure ...
Then those are the foundation of this work and in the next we are going to talk about CRBM

To extract the emotion representations from the pre-processed MFCC features we build a temporal model named CRBM
In order to understand how this works we firstly take a look at the basic concept called RBM, which...P of a set of training data x, and probability distribution has some parameters to be learned during the training.
RBM is trained in ...
RBM is

A RBM defines a basic structure showed with this figure. It has one visible layer of binary units denoted as vector x for receiving input data and one hidden layer of binary units denoted as vector h. for representations. each unit in visible layer is connected to each unit in hidden layer and the inter-layer connection is specified with the weight matrix W. b and c denotes the visible and hidden bias vector repectively.
Notice that the restriction of this model is that there is no intraconnection within each layer that is also the reason why this is called boltzmann machine.

There are several important definitions for understanding RBM.
The first concept is Energy FUnction called E subscript theta, theta denotes the parameter set W,b and c. The energy function defines the structure of the RBM, in other words the is related to how the visible and hidden layer connected with each other.
Then we have a joint distribution P of x,h... , and Z is called partition funtion in order to normaliz e the righ-hand side term sothat it should be a sufficient probability distribution.
from the definition of the probability distribution we can see that if a higher probability is desired, the energy function should be small. if you are familiar with thermodynamic, you may have heard about Boltzmann distribution which tells us about the state of the system depending on its energy. With lower energy the system is by definition more stable and this idea is borrowed to build RBM.
Another concept Free Energy is defined and will be used in training RBM, we are going to cover this part later.

The inference of RBM is quite straightforward it requires only the mathematical calculation where the marginal distribution is calcualted and using Bayes' theorem we can then get the conditional distribution. Then it allows us to calcualte the probability of one particular taking value one given the corresponding input or hidden vector.

The RBM is a static model. However the emotion in speech varies over both short and long time period and consequently we need a temporal model to capture those short and long term variation.
Therefore before we come to the training of the energy-based model, i am going to introduce an extention of RBM named CRBM
The matrix A and B are weight parameters of history visible units to current vis, and history visible units to current hidden units.
Now the visible units are
same as RBM the energy function defines how the units connected with each other, where tilde b is defined as the original bias of visible layer plus the weight matrix A times history visible units, x subscript smaller than t, and the number of history input is specified by N, as shown in the figure intop. similarly case for the hidden bias tilde c. now the parameter set consists of ....and the free energy is also changed.

Now let's see how can the energy-based model be trained with MLE. Firstly i am going to introduce the KL-div, which defines the difference between two probability distributions by calculating the follow equation, here the integral is used when we have a continuous distribution and for the case of crbm where the distribution is discrete, the integral is substitued by the sum over x. As previous discussed, we want to use a model to approximate the true data distribution, here the Q...P is .... and the brackets denotes the ...
the first term on the right hand side is the distribution of the data, it can be treated as a constant value during the optimization so the objective remains only the second term

With the definition of free energy we can now rewrite the likelihood distribution its partial derivative with respect to the parameter set.
if we average the derivative over the data , the objective function —(how become expectation over ...???
but the problem is that the calculation of the second term on the righthand side requires all configuration of the model which makes it impossible to be done directly, so we need to do sampling sufficiently to approximate the model distribution.

the technique used for sampling is Gibbs sampling where it performs a markov chain starting from the visible layer at time 0 and the hidden state at time 0 can be calculated with conditional probability, then the visible at time 1 is again calculated with conditional probability.. so this is a full step of Gibbs sampling.
The gibbs step repeats up to k full steps, where k is a pre-defined variable.

With k=0, the probability represente the distribution of input data which is independent of parameter set theta.
if Gibbs-steps is performed with big enough k steps e.g. up to infinite then the markov chain is guarantted to converge to ....: [click] so the model distribution can be approximated with the sampled distribution and the objective function can be rewritten as:

But in practise we cannot run the steps waiting until the chain converges and instead of optimizing the divergence between 0 step and infinite step, another concept is introduced by Hinton named Contrastive Divergence
where the difference of KL between P0 and Pinfinity and KL bet . . P1 and Pinfinity is minimized by calculating the following equation. It has already been known from many experiment that by ruuning one full gibbs step we can already get good approxmation of model distribution
From the literature by Hinton, we know that the third term of righthand side is irrelavant in optimization, so we can omit it during the training and simpify the CD.
Then we have the updating rule for the parameter set.

So in the next we are going to see the Deep Neural Network. The origin of artificial neural network trace back to the 60 and 70 in last century. To immitate how the human brain works, the scientists have used the aritificial neurons and connections to build up the structure of neural network which is illustrated with the figure on the right. The structure shows a single hidden layer feedforward neural network. The data are fed forward from input layer on the bottom and through a hidden layer then to the output layer on the top.
the preactivation a of x of the hidden layer can be calculated by multiplying the input vector with weight matrix W superscript 1, adding the bias vector to the product.
mapping the preact a by a non-linear activation function f we can calcualte the activation of the hidden layer.
similarly the activation of the output layer hat y is calcualted based on the hidden layer. The output layer can be prediction or classification or fed as input of further network, depending on different activation function.
If the number of hidden units is more than 1, then we'll have a multi-layer neural network or DNN.

Training the neural network is generally done via supervised learning to minimize the empirical loss.