1 Psychoaccoustics

- Experiments on the tempo sensitivity on humans have shown that the ability to notice tempo changes is proportional to the tempo, with the JND (just noticable difference) beeing around 2-5% [1].

2 Beat Tracking System

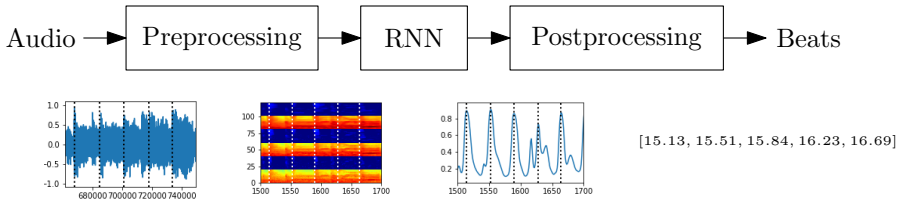

Figure 1: Beat tracking system signal flow

3 Preprocessing

The first step in the approach consists of preprocessing the original data. Data preprocessing refers to all transformations on the raw data before the resulting training set is fed to the machine learning algorithm. It includes different methods such as normalization, transformation and feature extraction.

The data set contains raw pulse code modulated (PCM) audio signals stored as WAV files. For the sake of consistency and also to reduce computational complexity the audio signal is resampled at a sampling rate $f_s = 44.1\,\text{kHz}$ and converted to a monaural signal by averaging both stereo channels.

- Slice audio $x(t)$ into frames $x_n(t)$, $n = 1, 2, \ldots, N$, where N is the number of frames
- Compute complex spectrogram $X(n, k)$ with FFT
- Complex spectrogram is converted to the power spectrogram $S(n, k) = |X(n, k)|^2$
- Mel-spectrogram $M(n, m) = \log\left(S(n, k) \cdot F(m, k)^T + 1.0\right)$

4 Machine learning model

Binary Classification problem:

- Beat (class 1)
- No beat (class 0)

4.1 Data representation

General: After preprocessing the audio we obtain the observation set $O = \left\{\mathbf{x}^{(\alpha)}, \mathbf{y}_T^{(\alpha)}\right\}_{\alpha=1}^{p}$, with p samples in total. As elements the set cointains:

- feature vector: $\mathbf{x} \in \mathbb{R}^{(\text{sequence length}, \text{input size})}$
- true label: $\mathbf{y}_T \in \{0, 1, \ldots, (\text{number of classes} - 1)\}^{(\text{sequence length})}$

Model Böck:

- $p = 698$
- $\mathbf{x} \in \mathbb{R}^{(3015,\, 120)}$
- $\mathbf{y}_T \in \{0, 1\}^{(3015)}$

4.2 Model class

As the model class we choose a recurrent neural network as seen in Fig.

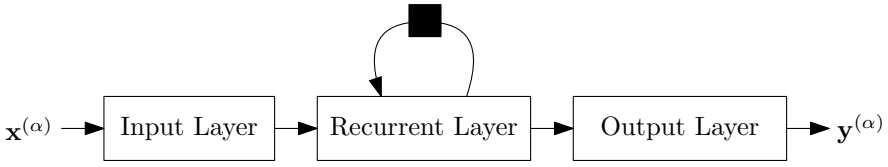

Figure 2:

Model Böck:

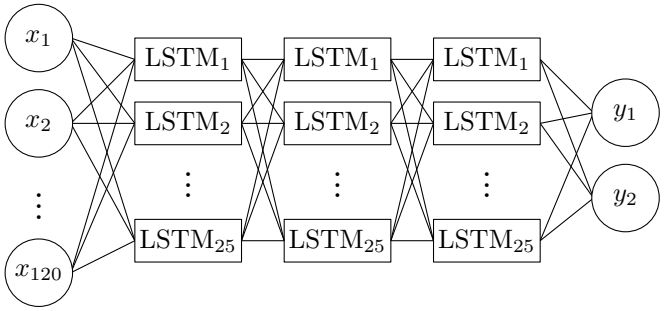

Figure 3:

LogSoftmax: A LogSoftmax normalization is added the last layer of the neural network to obtain the log-probabilities for the differnent classes.

$$\text{LogSoftmax}(x_i) = \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right) \tag{1}$$

4.3 Performance measure

Loss Function Cross entropy is defined for two probability distributions p and q as

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x). \tag{2}$$

In machine learning cross entropy can be used as a loss funciton to measures the performance of a classification model. The probability p_i is the true label (binary indicator 0 or 1), where as the distribution q_i is the predicted value of the current model.

4.4 Optimization

4.5 Valididation

Test set method Split the observarions into to disjunct subsets

$$\text{observations}\begin{cases} \text{training data } \left\{\left(\mathbf{x}^{(\alpha)}, \mathbf{y}_T^{(\alpha)}\right)\right\},\ \alpha\in\{1,\ldots,p\} \\ \rightarrow E^T \text{ selects model parameters} \\ \\ \text{test data } \left\{\left(\mathbf{x}^{(\beta)}, \mathbf{y}_T^{(\beta)}\right)\right\},\ \beta\in\{1,\ldots,q\} \\ \rightarrow \widehat{E}^G \text{ estimates generalization error} \end{cases}$$

n-fold cross-validation

5 Postprocessing

As a postprocessing step we make use of a dynamic Bayesian network (DBN) which jointly infers tempo an phase of the beat.

Hidden variables:

- ω - the tempo
- ϕ - the position inside the beat period

In order to infer the hidden variables from an audio signal, we specify three entities:

- **transition model**: describes the transitions between the hidden variables
- **observation model**: takes the beat activations from the neural network
- **initial distribution**: encodes prior knowledge about the hidden variables

5.1 Transition model

- The number of observations M per beat at tempo T in BPM is defined as

$$M(T) = \frac{60}{T} f_r \quad \left[\frac{\text{frames}}{\text{beat}}\right] \tag{3}$$

- The beat position state is dependent on the tempo by using exactly one state per audio frame.
- $\Phi \in \{1, 2, \ldots, M(T)\}$ denotes the position inside the beat period (pib)
- The tempo space corresponds to integer valued beat positions in the interval $[M(T_{\max}), M(T_{\min})]$, with

$$N_{\max} = M(T_{\min}) - M(T_{\max}) + 1 \tag{4}$$

  different tempo states.
- The tempo in beat positions per time frame $\dot{\Phi} \in \{M(T_{\max}), M(T_{\max}) + 1, \ldots, M(T_{\min})\}$
- For example:

$$\begin{aligned} T_{\min} = 56\,\text{BPM} &\rightarrow M(T_{\min}) = 107 \\ T_{\max} = 215\,\text{BPM} &\rightarrow M(T_{\max}) = 28 \end{aligned}$$

$$\begin{aligned} \Phi &\in \{1, 2, \ldots, 107\} \\ \dot{\Phi} &\in \{28, 29, \ldots, 107\} \end{aligned}$$

- Transitions to each tempo is allowed but only at beat times

$$\omega_k = \begin{cases} \omega_{k-1}, & P(\omega_k|\omega_{k-1}) = 1 - p_\omega \\ \omega_{k-1} + 1, & P(\omega_k|\omega_{k-1}) = \frac{p_\omega}{2} \\ \omega_{k-1} - 1, & P(\omega_k|\omega_{k-1}) = \frac{p_\omega}{2} \end{cases} \tag{5}$$

- The probability of tempo change is heuristically set to $p_\omega = 0.002$

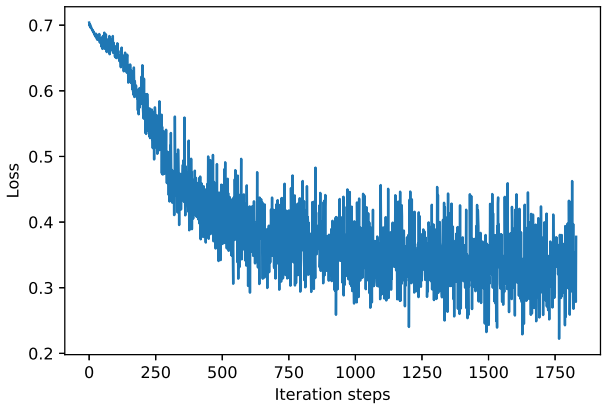

Figure 4: Loss function for 30 epechs with batch size 30 and learning rate 10^{-4}

5.2 Observation model

6 Evaluation

6.1 Network training

References

[1] Carolyn Drake and Marie-Claire Botte. Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics*, 54(3):277–286, 1993.