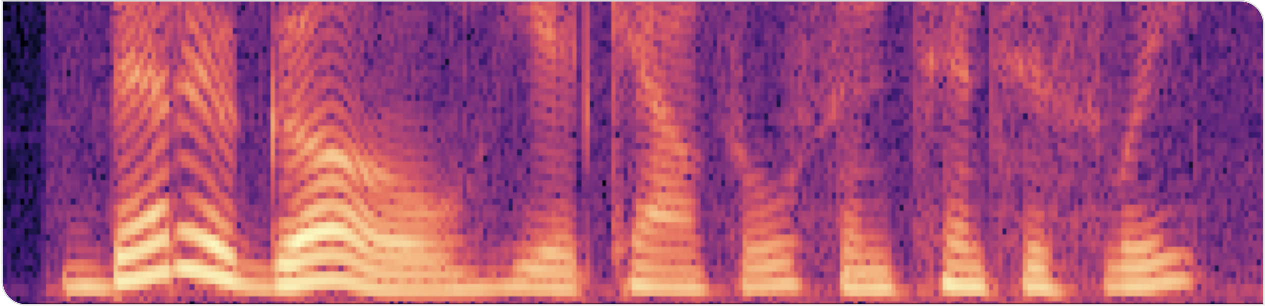
Karlsruhe Institute of Technology

Cross-lingual, Language-independent Phoneme Alignment

Bachelor Thesis Presentation

Niklas Bühler | 13. October 2021



Overview

Motivation

Language Death

- Roughly 40% of ≈7,000 languages are endangered[1]
- Documentary linguistics community needs the aid of automatic processing[2]

Goal

- Improve necessary technology to efficiently document *new* languages, especially their pronunciation

⇒ Phoneme Alignment of under-resourced languages

[1] Eberhard, Simons, and Fenning 2021.
[2] Woodbury 2003.

Research Question

Phoneme Alignment

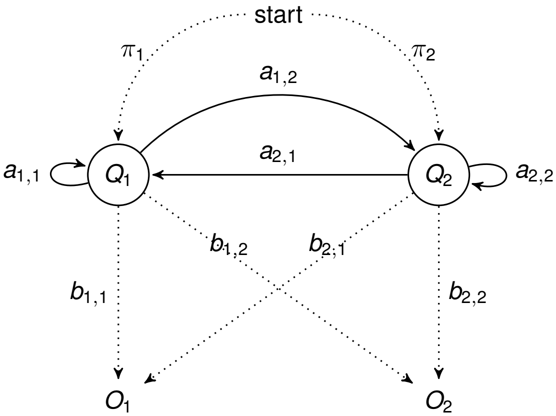
- Time-alignment of phonetic transcript and respective audio recording
- Standard method: Viterbi algorithm on hybrid HMM/GMM system[3]
- Alternative: HMM/ANN system[4]
  - Combines time-alignment capability of HMMs and discrimination-based learning of ANNs

Experiments

- Compare monolingual and multilingual approaches; as well as different neural network architectures
- Focus on under-resourced languages ⇒ cross-lingual methods

[3] Rabiner and Juang 1986.
[4] Franzini, K.-F. Lee, and Waibel 1990.

Hidden Markov Models



The Decoding Problem

Given an HMM λ and a possible observation sequence $o = o_1 o_2 \ldots o_T$, what is

$$q^* := \operatorname*{argmax}_{q \in Q^T} P(q, o \mid \lambda),$$

the most probable sequence of states the HMM might have attained while outputting o.

The Viterbi Algorithm

Data: HMM $\lambda = (S, V, \pi, A, B)$, output sequence $o = o_1 \dots o_T$

Result: Probability P^* of most probable state sequence $q^* = q_1^* \dots q_T^*$

for $1 \leq i \leq N$ do
- $\delta_1(i) = \pi_i b_i(o_1)$ — *// initialize the probabilities for all states in $t = 1$*

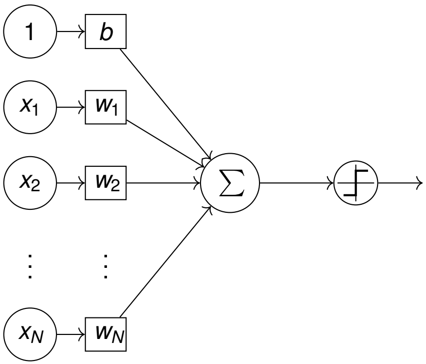
for $2 \leq t \leq T$ do — *// for all time steps*
- for $1 \leq j \leq N$ do — *// for all next states*
  - $\delta_t(j) = \max_{1 \leq i \leq N}[\delta_{t-1}(i) a_{ij}] b_i(o_t)$ — *// calculate each states probability iteratively*
  - $\Psi_t(j) = \operatorname{argmax}_{1 \leq i \leq N}[\delta_{t-1}(i) a_{ij}]$ — *// remember the most probable previous state*

$P^* = \max_{1 \leq i \leq N}[\delta_T(i)]$ — *// total probability of the most probable state sequence*

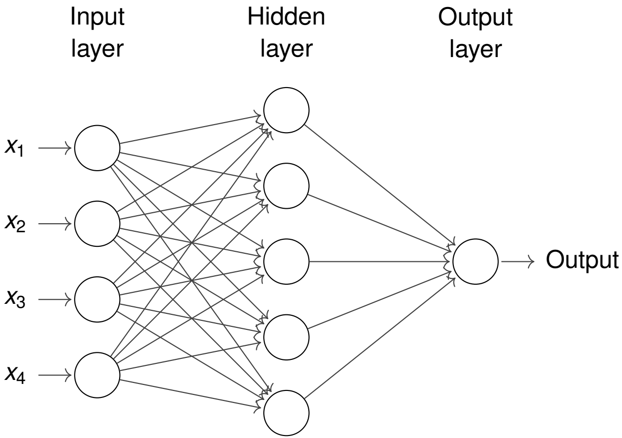
$q_T^* = \operatorname{argmax}_{1 \leq i \leq N}[\delta_T(i)]$ — *// most probable state in the last time step*

for $T - 1 \geq t \geq 1$ do
- $q_t^* = \Psi_{t+1}(q_{t+1}^*)$ — *// build the most probable state sequence*

The Perceptron

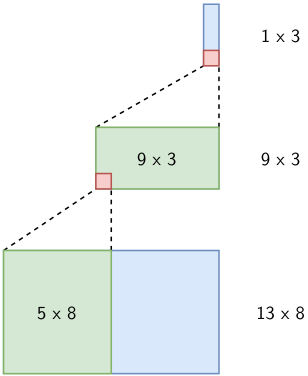


Feedforward Neural Network

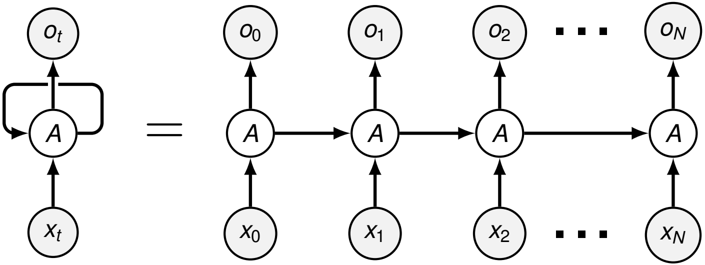


Time Delay Neural Networks

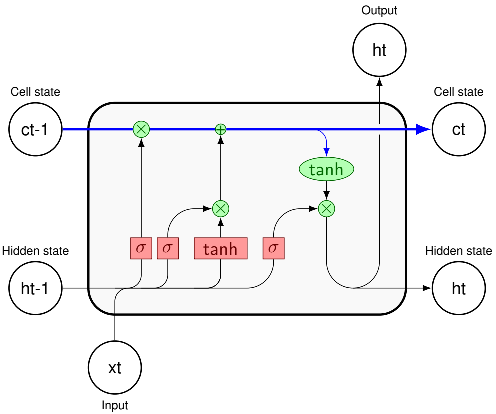
- Receives sequence of frames as input
- Connections between layers are shift invariant

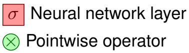


Recurrent Neural Networks

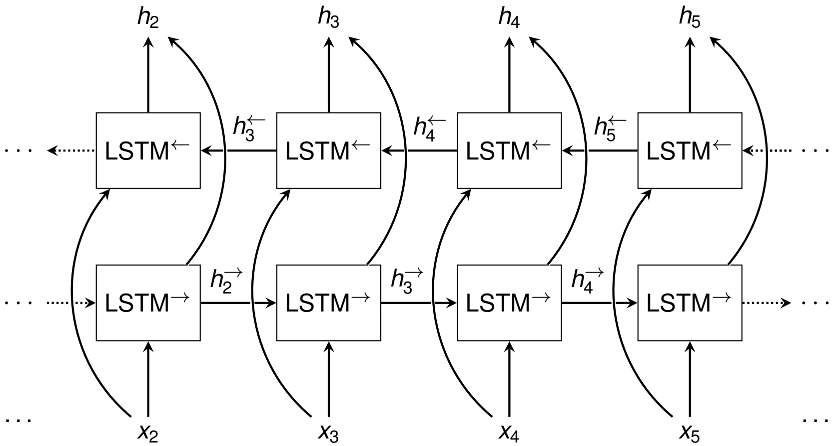


Long Short-Term Memory

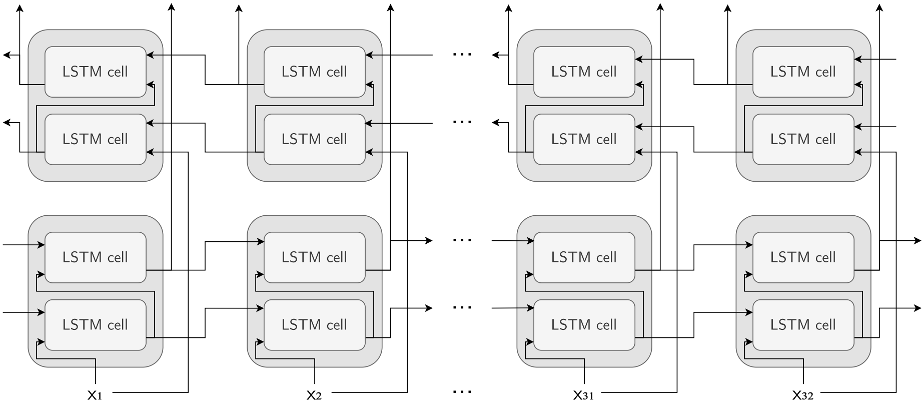




Bidirectional LSTM



Stacked (Bi-) LSTM



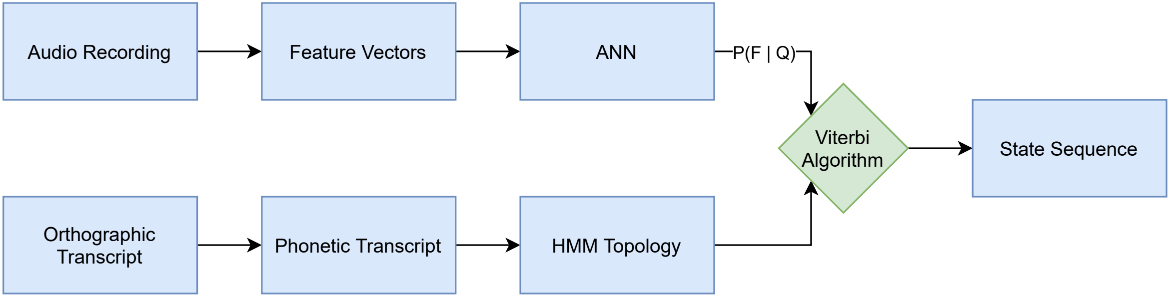
Related Work

- Graves and Schmidhuber 2005: *Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures*
  - BiLSTMs performed significantly better than unidirectional LSTMs
  - BiLSTMs also much faster to train and more accurate than standard RNNs or feedforward nets
- Franke et al. 2016: *Phoneme Boundary Detection using Deep Bidirectional LSTMs*
  - Promising results in phoneme boundary detection using BiLSTMs
  - Also regarding cross-lingual tasks

Related Work

- X. Li et al. 2020: *Universal Phone Recognition with a Multilingual Allophone System*
  - Supplementing language-independent phone distributions with language-dependent phoneme distributions
  - Improve performance by 2% phoneme error rate absolute
  - Improve phoneme recognition accuracy by 17% for unseen languages
- Müller, Stüker, and Waibel 2018: *Multilingual Adaptation of RNN based ASR systems* and Müller 2018: *Multilingual Modulation by Neural Language Codes*
  - Language adaptation techniques: Modulating the hidden layers of utilized RNNs using Language Feature Vectors
    - Extracted from bottleneck layer in language identification network
  - Decreased error rates in multilingual phoneme / grapheme recognition tasks
  - Extended by Multiplicative Language Codes and Adaptive Neural Language Codes

Hybrid HMM/ANN System



Bootstrapping a Multilingual Acoustic Model

- ANN does phoneme classification and provides a probability distribution over all (sub-)phonemes for every frame
- The evaluation output of the ANNs acts as acoustic model in the hybrid HMM/ANN system
- Bootstrap a multilingual model from a monolingual one:
  1. Map pronunciation dictionaries
  2. Roughly align the multilingual data set in a first iteration
  3. Create a first multilingual acoustic model
  - ⟲ Iterate steps 2 and 3 using the new acoustic model!

Toolkits, Libraries and Data sets

- Janus Speech Recognition Toolkit (Finke et al. 1997)
- PyTorch (Paszke et al. 2019)
- Common Voice (Ardila et al. 2019)
  - Data from the languages *en, de, ru, fr, es, sv*.

Training Data Set

- Build training data set from known languages *es, fr, ru, sv, de*
- 32,000 utterances per language $\Rightarrow$ 160,000 utterances $\approx$ 207 hours of speech recordings

Evaluation Data Set

- Build evaluation data set from target language: *en*
- 32,000 utterances $\approx$ 50 hours of speech recordings

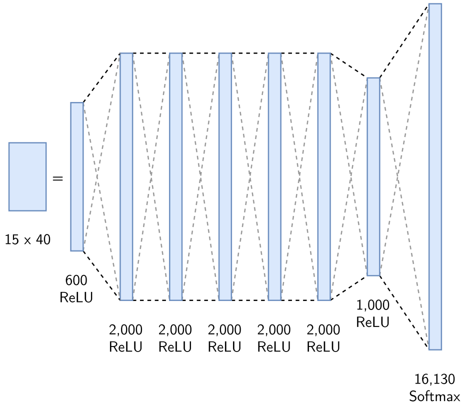
Experiments

- Networks are trained and utilized for phoneme classification, providing HMMs with emission probabilities
- Input: Preprocessed audio frames
- Output: Phoneme label in one-hot encoding
  - Softmax activation in the last layer
  - Cross-entropy loss function
- ReLU activation in hidden layers
- Minibatch size of 1024
- Split of 90/10 into training and validation set; data was shuffled during training
- Training for 8 epochs
- Pretrained network states in the second iteration of bootstrapping

Monolingual Feedforward Neural Network

Architecture

- Input: Context of 15 feature vectors of dimension 40 each
- Output: Probability distribution over 16,130 subphonemes



Monolingual Feedforward Neural Network

Epoch	Learning Rate η	Validation Accuracy
1	0.08	38.5%
2	0.08	41.4%
3	0.08	42.4%
4	0.08	42.7%
5	0.04	44.3%
6	0.02	45.5%
7	0.01	46.4%
8	0.005	47.0%
9	0.0025	47.5%
10	0.00125	47.7%
11	0.000625	47.9%
12	0.000313	48.1%
13	0.000156	48.1%

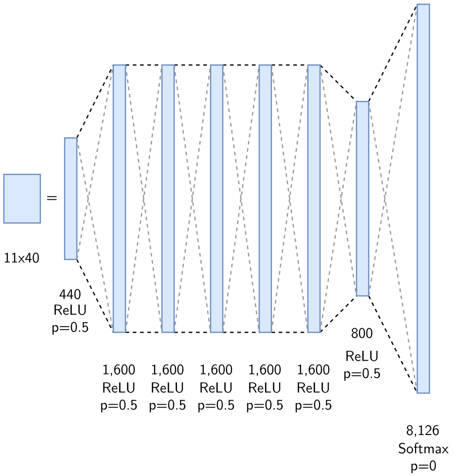
Training

- Stochastic Gradient Descent
  - Learning rate progression of $\eta = 0.08$ for four epochs, then halving it
- 13 Training epochs
- Final validation accuracy of 48.1%

Multilingual Feedforward Neural Network

Architecture

- Input: Context of 11 feature vectors of dimension 40 each
- Output: Probability distribution over 8,126 subphonemes
- Dropout with probability $p = 0.5$



Multilingual Feedforward Neural Network

Training

- Adam optimizer with initial learning rate of $\eta = 10^{-4}$
- 8 Training epochs in both iterations
- Final validation accuracy of 51.1% in the first iteration, and 48.4% in the second one

Epoch	Iteration 1	Iteration 2
1	45.8%	42.3%
2	49.6%	47.8%
3	50.6%	48.2%
4	51.1%	48.4%
5	51.1%	48.4%
6	51.2%	48.4%
7	51.2%	48.4%
8	51.1%	48.4%

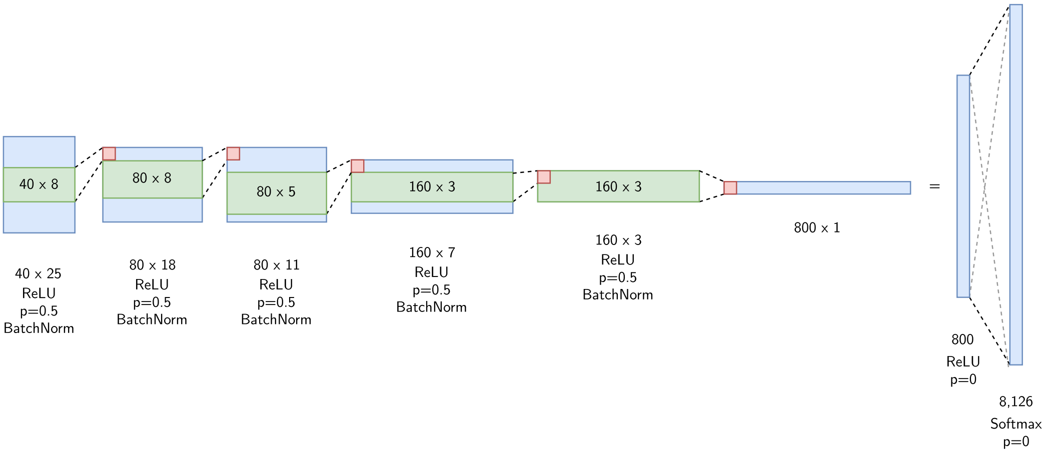
Table: Validation accuracies of the multilingual feedforward neural network, across both iterations of the bootstrapping process.

Multilingual Time Delay Neural Network

Architecture

- Input: Context of 25 feature vectors of dimension 40 each
  - Not stacked, but convolved with sliding filters with stride and dilation of 1
- Output: Probability distribution over 8,126 subphonemes
- Dropout with probability $p = 0.5$
- Each time delay layer also applies batch normalization

Multilingual Time Delay Neural Network



Multilingual Time Delay Neural Network

Training

- Adam optimizer with initial learning rate of $\eta = 10^{-3}$
- 8 Training epochs in the first iteration, 6 in the second one
- Final validation accuracy of 53.4% in the first iteration, and 47.6% in the second one

Epoch	Iteration 1	Iteration 2
1	46.7%	40.7%
2	52.1%	46.0%
3	52.7%	46.8%
4	53.1%	47.2%
5	53.3%	47.2%
6	53.3%	47.6%
7	53.4%	–
8	53.4%	–

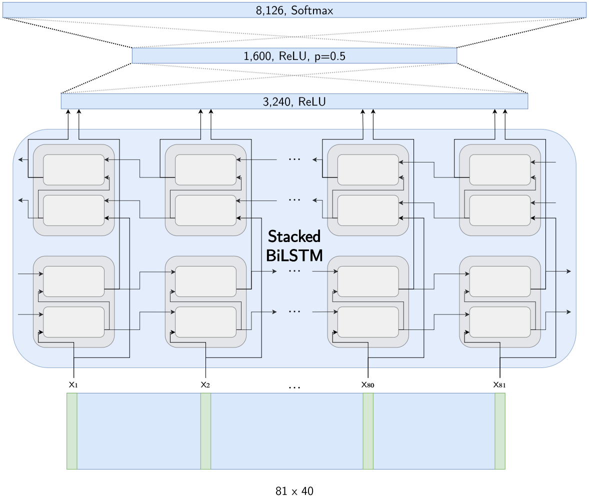
Table: Validation accuracies of the multilingual time delay neural network, across both iterations of the bootstrapping process.

Multilingual Stacked Bidirectional Long Short-Term Memory

Architecture

- Input: Context of 81 feature vectors of dimension 40 each
  - Neither stacked, nor convolved, but provided as a sequence over time in both time dimensions
- Two layers of BiLSTMs
- Hidden representations of size 20
- Stacked BiLSTMs have output dimensions $2 \times 81 \times 20 = 3,240$
- This output is concatenated and passed through a ReLU activation function into a new layer of size 1,600, again with ReLU and dropout with probability $p = 0.5$
- Output: Probability distribution over 8,126 subphonemes, via softmax

Multilingual Stacked Bidirectional Long Short-Term Memory



Multilingual Stacked Bidirectional Long Short-Term Memory

Training

- Adam optimizer with initial learning rate of $\eta = 10^{-4}$
- 8 Training epochs in both iterations
- Final validation accuracy of 53.0% in the first iteration, and 47.8% in the second one

Epoch	Iteration 1	Iteration 2
1	49.7%	44.6%
2	51.1%	46.0%
3	51.8%	46.6%
4	52.2%	46.9%
5	52.5%	47.3%
6	52.7%	47.5%
7	52.9%	47.5%
8	53.0%	47.8%

Table: Validation accuracies of the multilingual stacked BiLSTM neural network, across both iterations of the bootstrapping process.

Scoring Methods

Mean Squared Error Score

- Errors are given as deviations of predicted phoneme boundaries
- Letting Y_i be the point in time of transitioning from phoneme $i-1$ to phoneme i in the ground truth alignment and $\hat{Y}_i$ predicted point in time:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Scoring Methods

Box Score

- Inspired by Franke et al. 2016
- Counts the errors in the predicted phoneme boundaries, normalized by the total amount of phonemes in the alignment
- Error is defined as binary indicator, if prediction was correct
- Error tolerance of 20 milliseconds in both directions is granted

Overlap Score

- Phoneme overlap
- Defined as the total time of matching phonemes, divided by the total temporal length of the alignment

Results: Cross-lingual Phoneme Classification Accuracies

Experiment	Iteration 1	Iteration 2
Multilingual FFNN	39.5%	**36.8%**
Multilingual TDNN	39.6%	33.5%
Multilingual Stacked BiLSTM	**41.1%**	30.2%

Table: Comparison of the cross-lingual phoneme classification accuracies on the data set of the target language (English).

Results: Monolingual Feedforward Neural Network

- Serves as a baseline for the multilingual experiments

Results

- MSE: 0.1161
- Box: 41.03%
- Overlap: 69.86%

Results: Multilingual Feedforward Neural Network

- Results heavily depend on the applied scoring method
- MSE score slightly better than baseline
- Second iteration worse than first, probably linked to dropped validation accuracies in second iteration

First Iteration

- MSE: 0.1073
- Box: 7.09%
- Overlap: 41.61%

Second Iteration

- MSE: 0.1489
- Box: 2.29%
- Overlap: 18.32%

Results: Multilingual Time Delay Neural Network

- With MSE scoring, the TDNN system performed slightly worse than the multilingual feedforward system
- This is although the TDNN had a higher validation accuracy during training
- Also performed worse than the feedforward system with the other scoring methods
- Like with the feedforward system, the second iteration had worse results

First Iteration

- MSE: 0.1452
- Box: 1.46%
- Overlap: 11.33%

Second Iteration

- MSE: 0.1616
- Box: 0.44%
- Overlap: 1.81%

Results: Multilingual Stacked Bidirectional Long Short-Term Memory

- The BiLSTM system performed the worst out of all tested architectures, across all scoring methods
- This is despite it having the highest cross-lingual phoneme classification accuracy across the multilingual networks (in the first iteration)
- Again, the second iteration showed decreased performance

First Iteration

- MSE: 0.3180
- Box: 0.11%
- Overlap: 0.21%

Second Iteration

- MSE: 0.7250
- Box: 0.09%
- Overlap: 0.13%

Summary

- Goal: Apply cross-lingual, multilingual methods on phoneme alignment
- Built, trained and utilized three different ANN architectures in a hybrid HMM/ANN system to align multilingual data
- Iterated to bootstrap a multilingual acoustic model
- Utilized the resulting systems to cross-lingually align data from previously unseen target language
- Scored and compared the different experiments

Interpretation of Results

- All multilingual networks had higher phoneme classification accuracies than the monolingual system, at least in the first iteration
- However, in general, the multilingual systems did not outperform the monolingual system
  - Reason for missing transfer of improved results could be an imprecise cross-lingual application, i.e. an imprecise mapping of phonemes between training languages and target language
- Systems with more complex ANN architectures had decreased alignment performance
  - Despite having increased phoneme classification accuracies, not only on training data set, but also on evaluation data set in the target language
- The performance of all systems decreased in the second iteration of the bootstrapping process
  - Not only for phoneme classification accuracy, but also for cross-lingual phoneme classification accuracy and alignment results
  - Again, an imprecise mapping of phonemes could be the reason for this phenomenon

Further Research

- Adress the possible problems stated on the previous slide
- More careful mapping of phonemes between languages in the bootstrapping process
  - Employ more profound linguistic knowledge
  - Utilize data-driven approaches
- Choose training languages more carefully, i.e. by comparing lexical similarities or other linguistic distances to the target language
- Improve systems capability to handle multilingual data, e.g. by introducing modulation techniques

Thank you for listening!

Results: Comparison of MSE Scores

Experiment	$\bar{s}^{\text{MSE}}$	σ^{MSE}	$\tilde{s}^{\text{MSE}}$	$\bar{s}_{0.1}^{\text{MSE}}$
Monolingual FFNN (1)	0.1161	6.3992	**0.0028**	**0.0042**
Multilingual FFNN (1)	**0.1073**	4.5812	0.0058	0.0069
Multilingual TDNN (1)	0.1452	5.1452	0.0180	0.0196
Multilingual Stacked BiLSTM (1)	0.3180	8.6341	0.1639	0.1628
Multilingual FFNN (2)	0.1489	4.5837	0.0134	0.0183
Multilingual TDNN (2)	0.1616	3.3311	0.0579	0.0594
Multilingual Stacked BiLSTM (2)	0.7250	4.1788	0.6292	0.6084

Table: Comparison of the MSE scoring results between all experiments in the first and second iteration: total MSE score $\bar{s}^{\text{MSE}}$, as well as its standard deviation σ^{MSE}, median $\tilde{s}^{\text{MSE}}$ and trimmed mean (10%) $\bar{s}_{0.1}^{\text{MSE}}$.

Results: Comparison of Box Scores

Experiment	$\bar{s}^{\text{box}}$	σ^{box}	$\tilde{s}^{\text{box}}$	$\bar{s}^{\text{box}}_{0.1}$
Monolingual FFNN (1)	**0.4303**	0.1334	**0.44**	**0.4588**
Multilingual FFNN (1)	0.0709	0.0498	0.0652	0.0787
Multilingual TDNN (1)	0.0146	0.0253	0.0	0.0163
Multilingual Stacked BiLSTM (1)	0.0011	0.0088	0.0	0.0012
Multilingual FFNN (2)	0.0229	0.0326	0.0122	0.0254
Multilingual TDNN (2)	0.0044	0.0154	0.0	0.0049
Multilingual Stacked BiLSTM (2)	0.0009	0.0082	0.0	0.0010

Table: Comparison of the box scoring results between all experiments in the first and second iteration: mean box score $\bar{s}^{\text{box}}$, as well as its standard deviation σ^{box}, median $\tilde{s}^{\text{box}}$ and trimmed mean (10%) $\bar{s}^{\text{box}}_{0.1}$.

Results: Comparison of Overlap Scores

Karlsruhe Institute of Technology

Experiment	$\bar{s}^{\text{overlap}}$	σ^{overlap}	$\tilde{s}^{\text{overlap}}$	$\bar{s}^{\text{overlap}}_{0.1}$
Monolingual FFNN (1)	**0.6708**	0.1172	**0.6938**	**0.6998**
Multilingual FFNN (1)	0.4161	0.0978	0.4174	0.4360
Multilingual TDNN (1)	0.1133	0.0721	0.1024	0.1242
Multilingual Stacked BiLSTM (1)	0.0021	0.0112	0.0	0.0023
Multilingual FFNN (2)	0.1832	0.0878	0.1750	0.1981
Multilingual TDNN (2)	0.0181	0.0312	0.0035	0.0201
Multilingual Stacked BiLSTM (2)	0.0013	0.0112	0.0	0.0014

Table: Comparison of the overlap scoring results between all experiments in the first and second iteration: mean overlap score $\bar{s}^{\text{overlap}}$, as well as its standard deviation σ^{overlap}, median $\tilde{s}^{\text{overlap}}$ and trimmed mean (10%) $\bar{s}^{\text{overlap}}_{0.1}$.