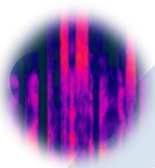
Zürcher Hochschule
für Angewandte Wissenschaften

# Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering

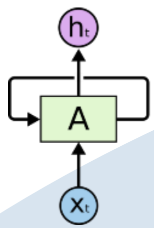
***8th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition, September 19-21, 2018, Siena, Italy (ANNPR'2018)***

*Thilo Stadelmann*, Sebastian Glinski-Haefeli, Patrick Gerber & Oliver Dürr

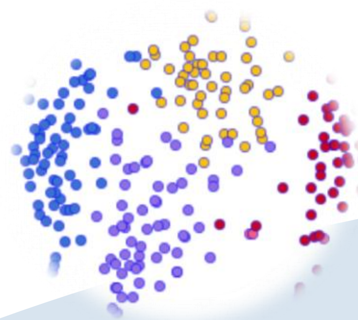
# Agenda



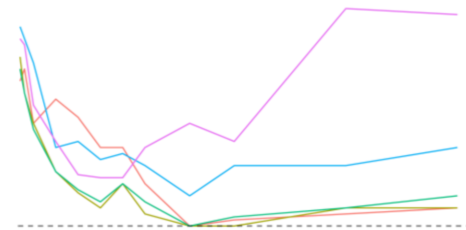
Motivation




Our approach

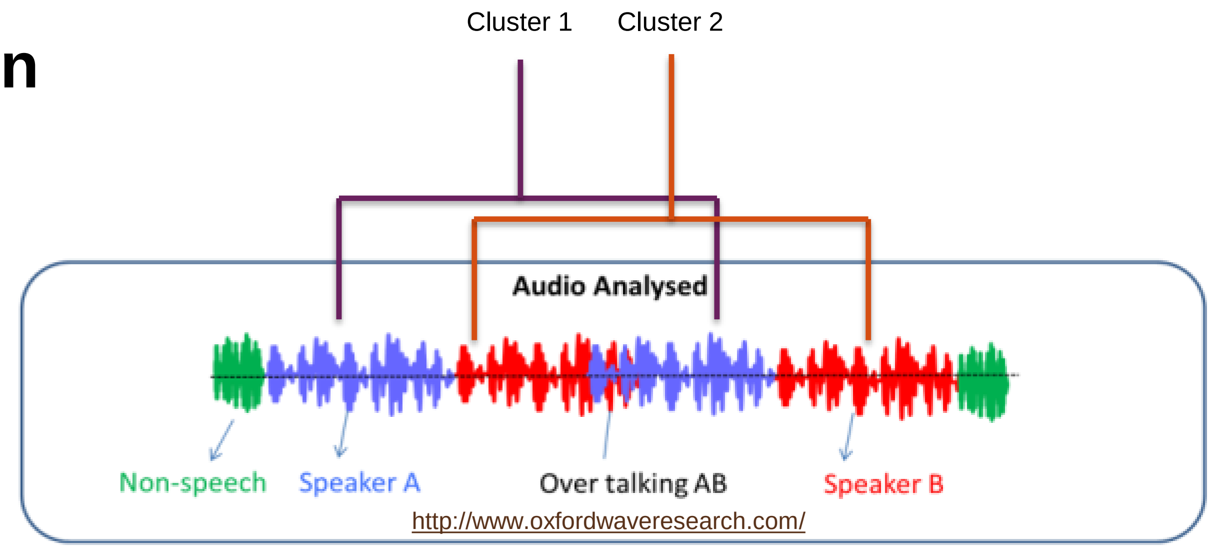


Experiments

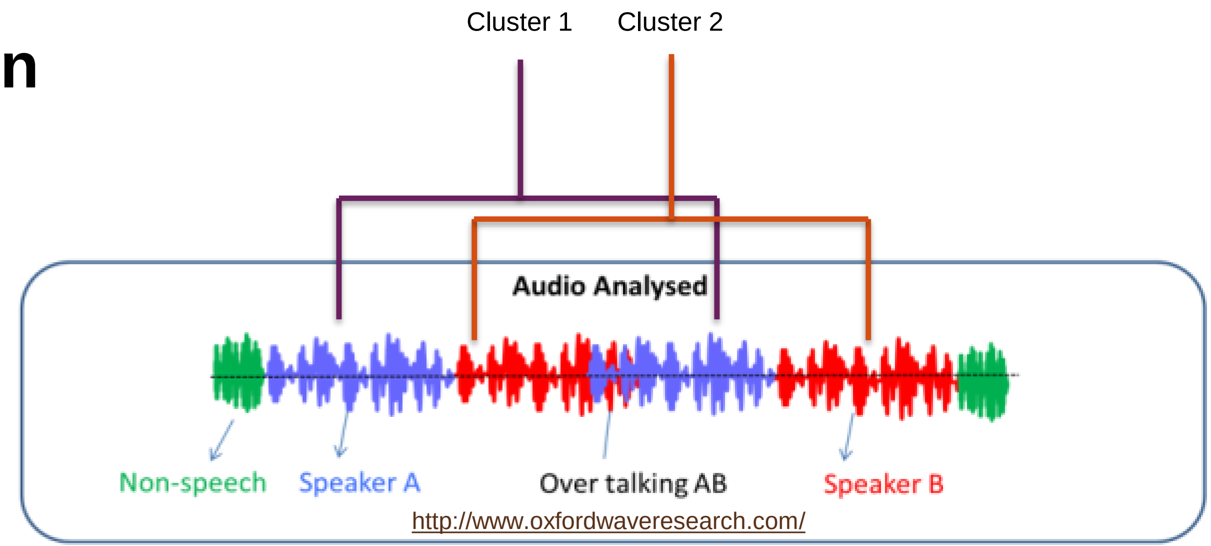


Learnings
&
future work

# Motivation


Cluster 1
Cluster 2
Audio Analysed
Non-speech
Speaker A
Over talking AB
Speaker B
http://www.oxfordwaveresearch.com/

# Motivation

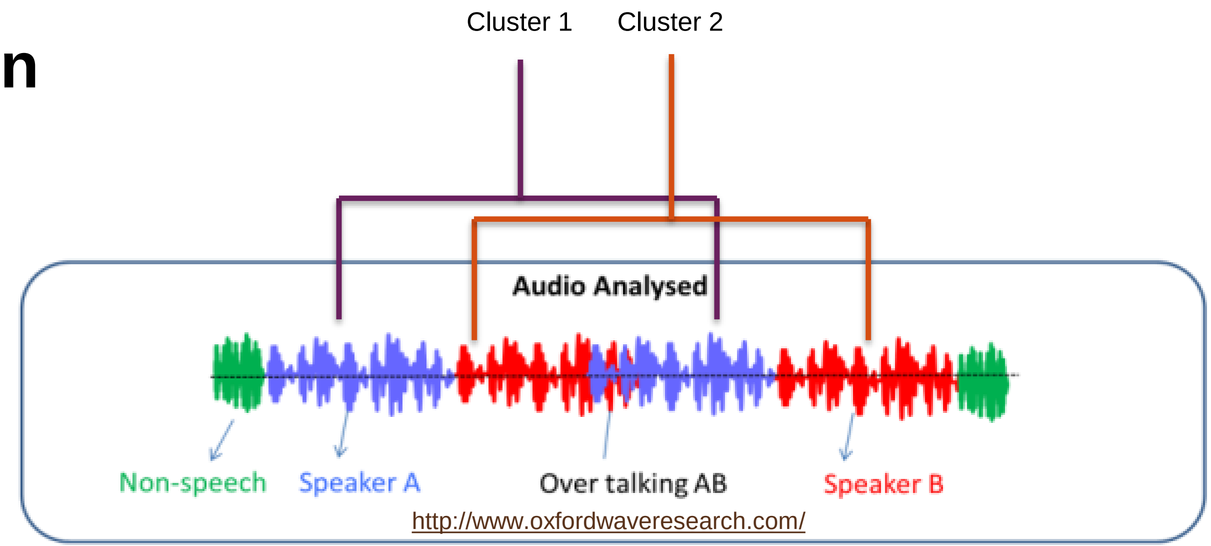



For the 630 training utterances, GMMs with 32 mixtures are built a priori, then an identification experiment is run for the 630 test utterances. It yields a satisfactory 0.5% closed set identification error.

[34]. Evaluations typically concentrate on data sets built from broadcast news/shows and meeting recordings, where diarization error rates ranging from 8% to 24% are reported [28][34][45]. These results are confirmed by more recent

Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

# Motivation




For the 630 training utterances, GMMs with 32 mixtures are built a priori, then an identification experiment is run for the 630 test utterances. It yields a satisfactory 0.5% closed set identification error.

[34]. Evaluations typically concentrate on data sets built from broadcast news/shows and meeting recordings, where diarization error rates ranging from 8% to 24% are reported [28][34][45]. These results are confirmed by more recent

The hypothesis of this paper is: the techniques originally developed for speaker verification and identification are not suitable for speaker clustering, taking into account the escalated difficulty of the latter task. However, the processing chain for speaker clustering is quite large – there are many potential areas for improvement. The question is: *where* should improvements be made to improve the *final* result?

Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

# Motivation: temporal context & voice prosody

The interpretation of our results has shown that it is the stage of modeling that bears the highest potential: the inclusion of temporal context information among feature vectors is what is crucially missing there. Furthermore, the inclusion

context vector. This corresponds to a syllable length of 130 ms and is found to best capture speaker specific sounds in informal listening experiments over a range of 32–496 ms (in intervals of 16 ms). Our context vector step is one orig-


Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

# Motivation: temporal context & voice prosody

The interpretation of our results has shown that it is the stage of modeling that bears the highest potential: the inclusion of temporal context information among feature vectors is what is crucially missing there. Furthermore, the inclusion

context vector. This corresponds to a syllable length of 130 ms and is found to best capture speaker specific sounds in informal listening experiments over a range of 32–496 ms (in intervals of 16 ms). Our context vector step is one orig-


Prosody

- *"use of **suprasegmental features** to convey sentence-level pragmatic meanings"* *
- *"those **elements** of speech that are not [elements of ] individual phonetic segments (vowels and consonants) but [...] **of syllables and larger units of speech**"* **

* Ladd (2008). *«Intonational phonology»*. Cambridge University Press.
** https://en.wikipedia.org/wiki/Prosody_(linguistics).
Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

# Our approach

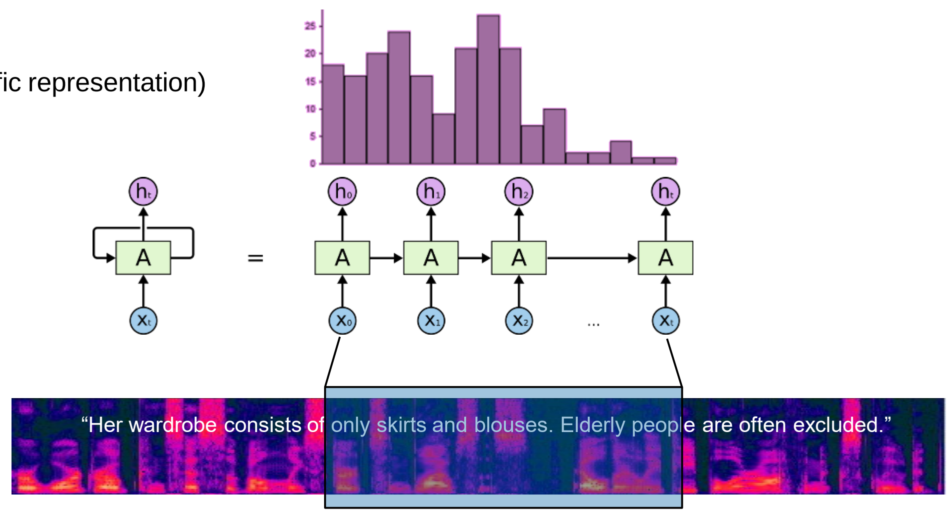
Idea

- **Leverage** on recent success of **deep learning** in audio processing
- Use **RNN for** its known **sequence learning** capabilities
- **Extract** speaker **embeddings** for new utterance from trained RNN

Output: Embedding (speaker-specific representation)

Model: Deep recurrent neural net

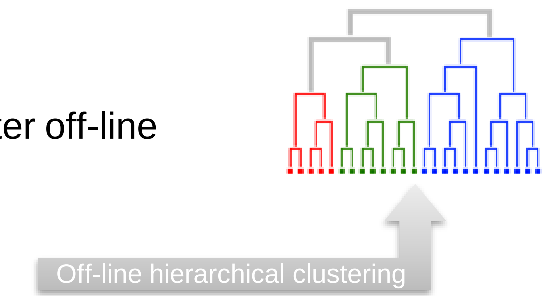
Input: Audio snippet



# Our approach

Idea

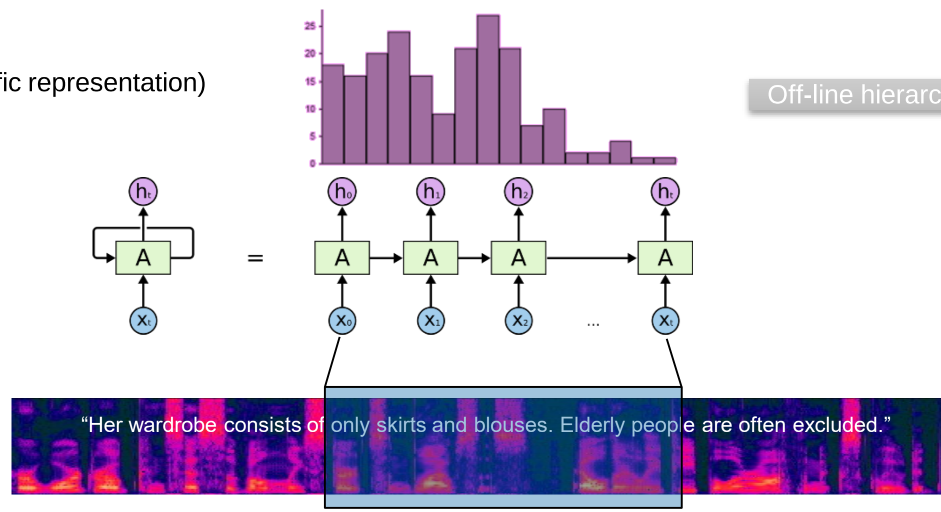
- **Leverage** on recent success of **deep learning** in audio processing
- Use **RNN for** its known **sequence learning** capabilities
- **Extract** speaker **embeddings** for new utterance from trained RNN → cluster off-line




Output: Embedding (speaker-specific representation)

Model: Deep recurrent neural net

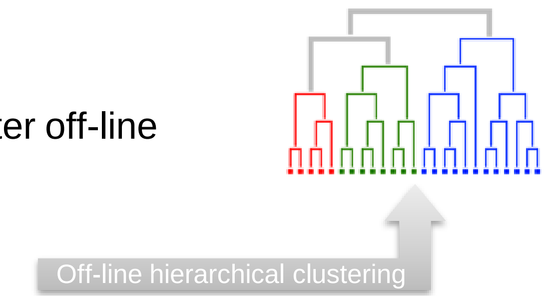
Input: Audio snippet



# Our approach

## Idea

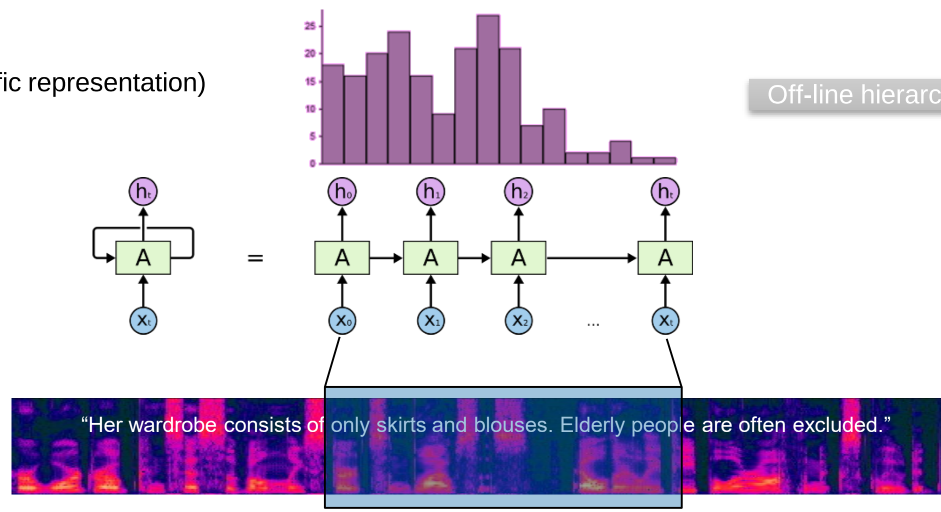
- **Leverage** on recent success of **deep learning** in audio processing
- Use **RNN for** its known **sequence learning** capabilities
- **Extract** speaker **embeddings** for new utterance from trained RNN → cluster off-line




Output: Embedding (speaker-specific representation)

Model: Deep recurrent neural net

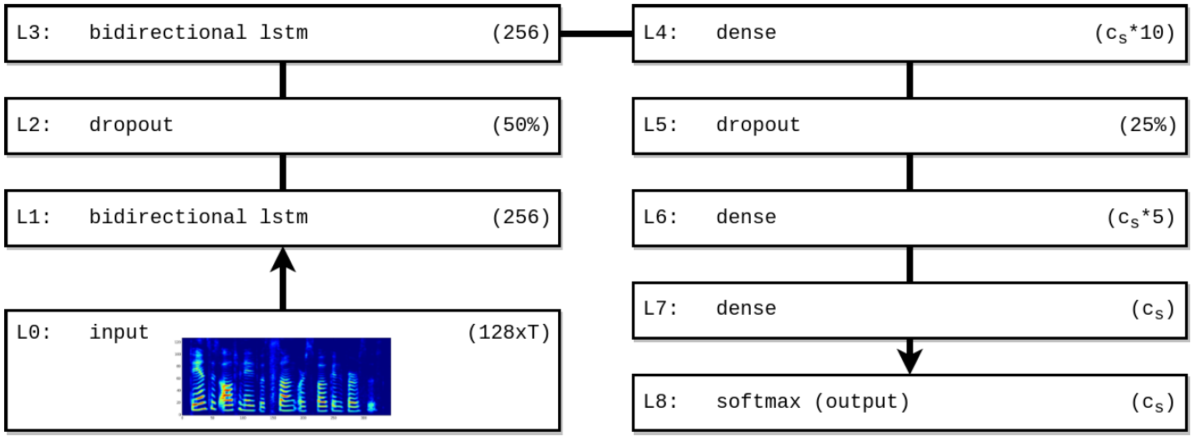
Input: Audio snippet



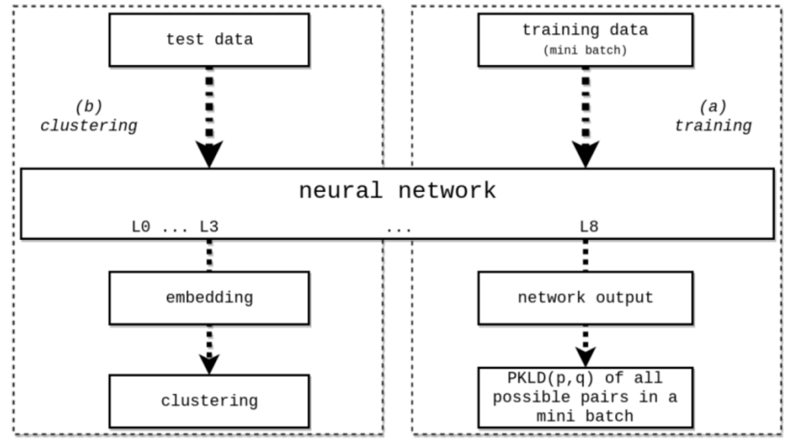

## Challenges

- **RNNs** known to be **hard to train**
- Additionally: **no natural training target** → need surrogate task with hopefully helpful loss

# Our approach: network architecture & training




&




## Learning target

- **`Lxx`** to output a **distribution** ($c_S$ = number of speakers in training set) that is similar for samples of the same speaker, dissimilar for different speakers

## Loss

- For all pairs $(p, q)$ of distributions in a mini batch:
  - **Pairwise Kullback-Leibler** distance between **same-speaker** pairs:
  - **Hinge** loss (with hyperparameter $margin$) between **different-speaker** pairs:
- (final loss gets symmetrized)

$$\mathrm{KL}(\mathbf{p} \parallel \mathbf{q}) = \sum_i^{c_s} p_i \log \frac{p_i}{q_i}$$

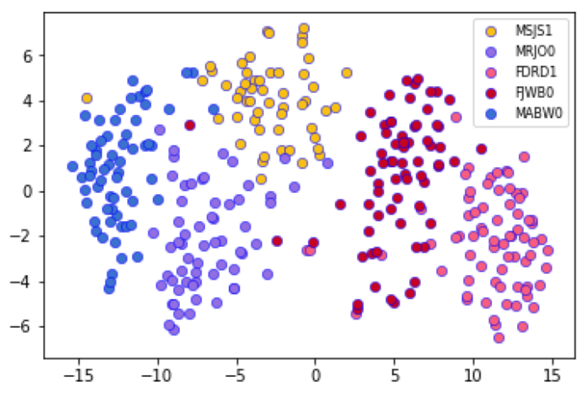
$$\mathrm{HL}(\mathbf{p} \parallel \mathbf{q}) = \max(0, \mathrm{margin} - \mathrm{KL}(\mathbf{p} \parallel \mathbf{q}))$$

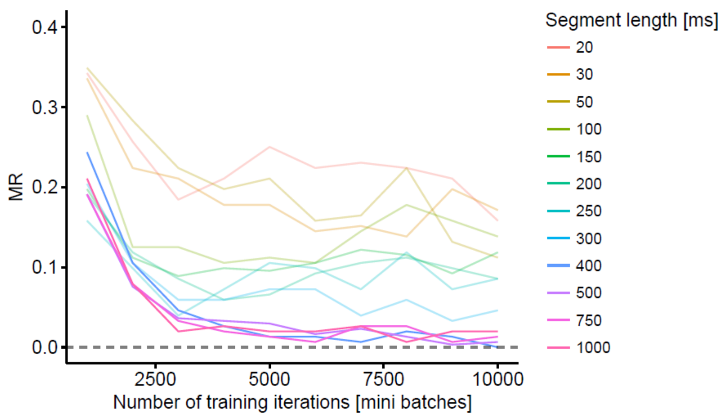
# Experiments

## Setup

- Based on Stadelmann & Freisleben (2009) for **comparability**: TIMIT (630 speakers, studio quality)
- **Signal processing**: mel-spectrograms (128 freq. bins)
- **Training** on 100 speakers (20% of these for validation): snippets of varying length (see below)
  Hyperparameters: standard Adam optimizer, $margin = 3$, 10'000 mini batches
- **Test** on distinct 40 speaker clustering test set: 1$^{st}$ utterance = 8 sentences, 2$^{nd}$ utterance = 2 sentences
  (Bug in code made intermediate experiments leave out 2 uncritical speakers, and made assignments of sentences to utterances random instead of lexicographic)
- **Clustering** using agglomerative hierarchical clustering, complete linkage and cosine distance of mean embeddings per utterance

## Intuitive hyperparameter justification of averaging & training time

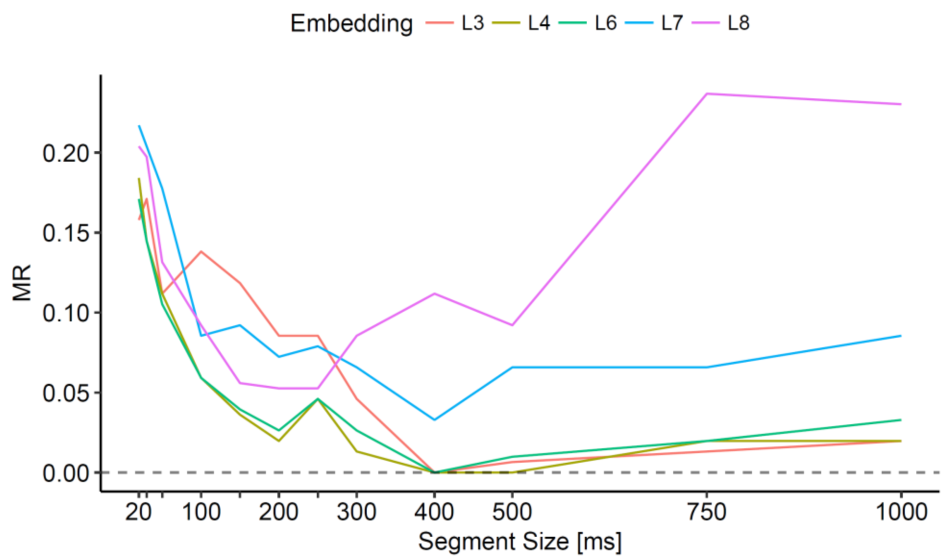


# Experiments: tracing prosodic information

Intermediate experiment

- Misclassification rate (MR) as a function of input segment length (~temporal context)

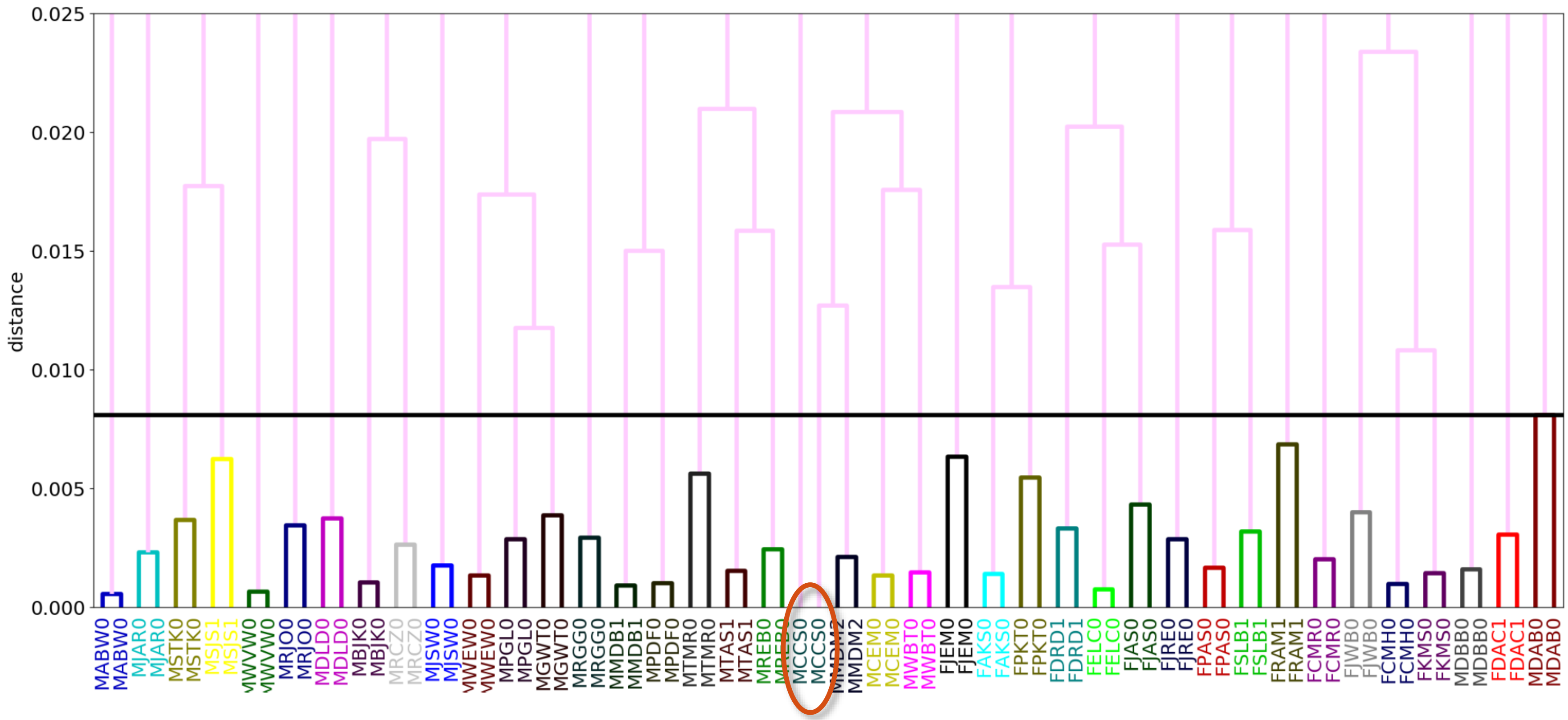



➔ All layers L3-L8 show a “sweet spot”

➔ Best performing layers have “sweet spot” around 400ms

➔ This is in the predicted range (on both axes) of Stadelmann & Freisleben (2009)

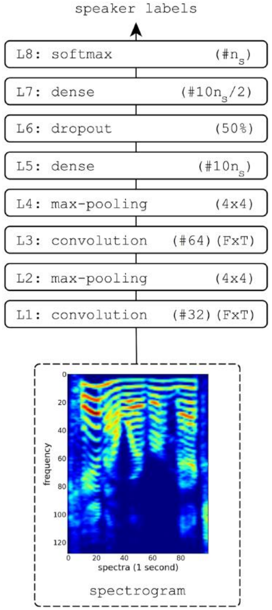
# Experiments: visual clustering performance




➔ Misclassification only for **`MCCS0`**

# Experiments: clustering performance vs. SotA

**CNN (MLSP'16)**

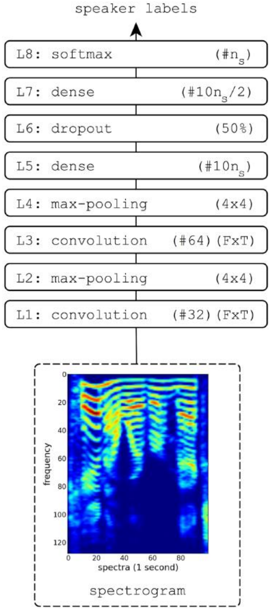



| Method | MR | MR (legacy) |
|---|---|---|
| **RNN /w PKLD** | **2.19%** $\left(\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}\right)$ | **4.38%** (average of 4 runs) |
| CNN /w PKLD [24] | - | **5%** |
| CNN /w cross entropy [23] | - | **5%** |
| $\nu$-SVM [40] | 6.25% | - |
| GMM/MFCC [40] | 12.5% | - |

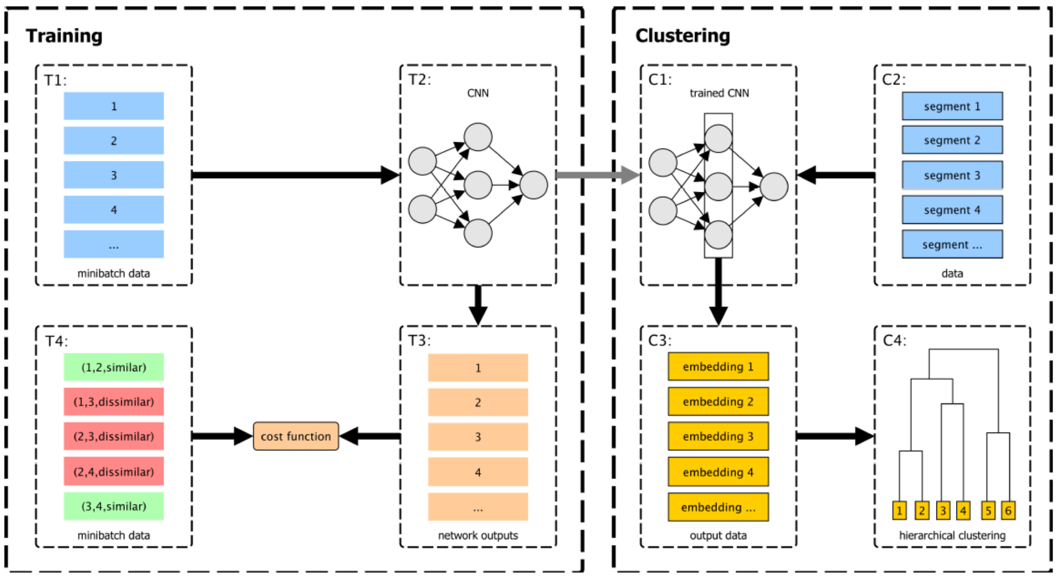

Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.

# Experiments: clustering performance vs. SotA

## CNN (MLSP'16)




## CNN & clustering-loss (MLSP'17)



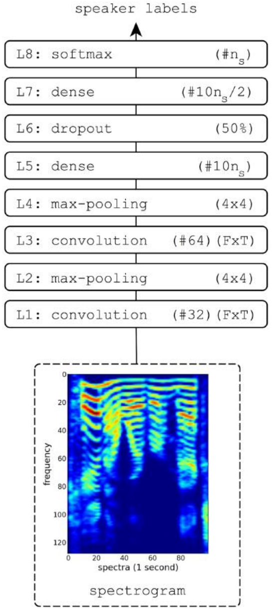

| Method | MR | MR (legacy) |
|---|---|---|
| **RNN /w PKLD** | **2.19%** $\left(\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}\right)$ | 4.38% (average of 4 runs) |
| CNN /w PKLD [24] | - | 5% |
| CNN /w cross entropy [23] | - | 5% |
| $\nu$-SVM [40] | 6.25% | - |
| GMM/MFCC [40] | 12.5% | - |

Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.

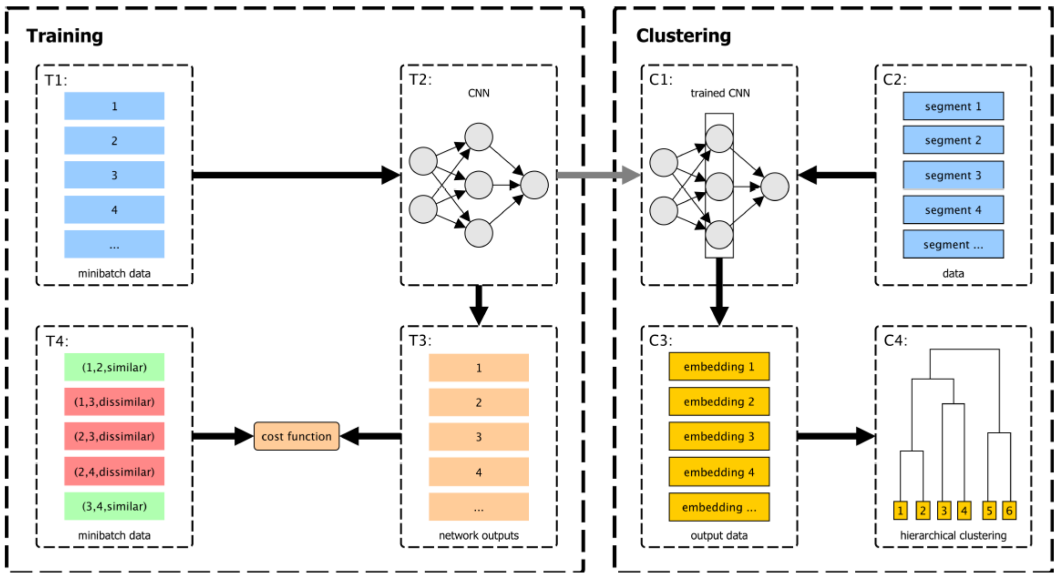
Zürcher Hochschule
für Angewandte Wissenschaften

# Experiments: clustering performance vs. SotA

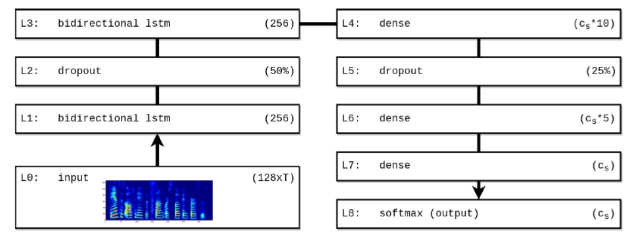
## CNN (MLSP’16)




## CNN & clustering-loss (MLSP’17)




## RNN & clustering-loss (ANNPR’18)




| Method | MR | MR (legacy) |
|---|---|---|
| **RNN /w PKLD** | **2.19%** $\left(\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}\right)$ | **4.38%** (average of 4 runs) |
| CNN /w PKLD [24] | - | **5%** |
| CNN /w cross entropy [23] | - | **5%** |
| $\nu$-SVM [40] | 6.25% | - |
| GMM/MFCC [40] | 12.5% | - |

Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP’2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP’2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR’2018.

Zürcher Hochschule
für Angewandte Wissenschaften

# Learnings & future work

«Pure» voice modeling seem largely solved

- RNN architecture is **very robust to hyperparameters** (different from earlier work)
- RNN model robustly exhibits *the predicted* **«sweet spot» for** the used **time information**
- Speaker clustering on clean & reasonably long input works **an order of magnitude better** (*as predicted*)
- Additionally, using a smarter clustering algorithm on top of embeddings makes **clustering on TIMIT as good as identification** (see ICPR'18 paper on dominant sets)

Future work

- Make models robust on **real-worldish data** (noise and more speakers/segments)
- Exploit findings for robust reliable **speaker diarization**
- **Learn** embeddings and the clustering algorithm **end to end** (we still pick embeddings from a lower layer, thus the surrogate task is not yet close enough to clustering despite PKLD)

swiss group for artificial intelligence and cognitive science

SGAICO

datalab

www.zhaw.ch/datalab

On me:

- Prof. AI/ML, head ZHAW Datalab, board SGAICO
- thilo.stadelmann@zhaw.ch
- 058 934 72 08
- https://stdm.github.io/
- Collaboration: datalab@zhaw.ch

➔ Happy to answer questions & requests.



Hibraj, Vascon, Stadelmann & Pelillo (2018). «Speaker Clustering Using Dominant Sets». ICPR'2018.
Meier, Elezi, Amirian, Dürr & Stadelmann (2018). «Learning Neural Models for End-to-End Clustering». ANNPR'2018.