Improving Children Speech Recognition through Feature Learning from Raw Speech Signal

S. Pavankumar Dubagunta[1,2], Selen Hande Kabil[1,2] and Mathew Magimai Doss[1]

[1]**Idiap Research Institute**, Martigny, Switzerland [2]**École polytechnique fédérale de Lausanne (EPFL)**, Switzerland

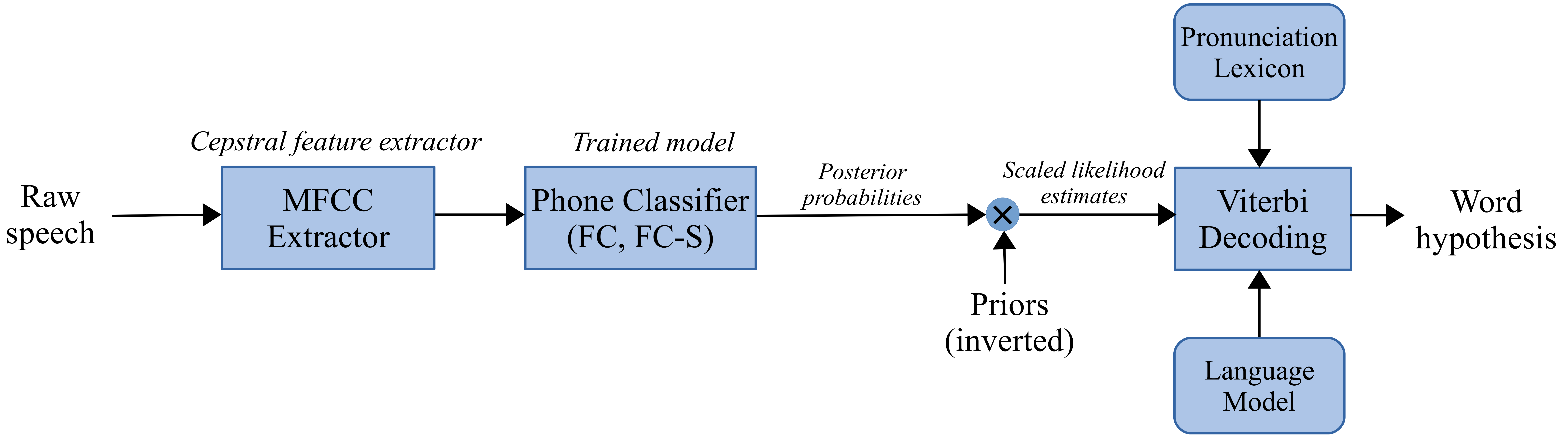
Funded under the project FLOSS by HASLERSTIFTUNG

Funded under the project SHISSM by FNSNF FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE

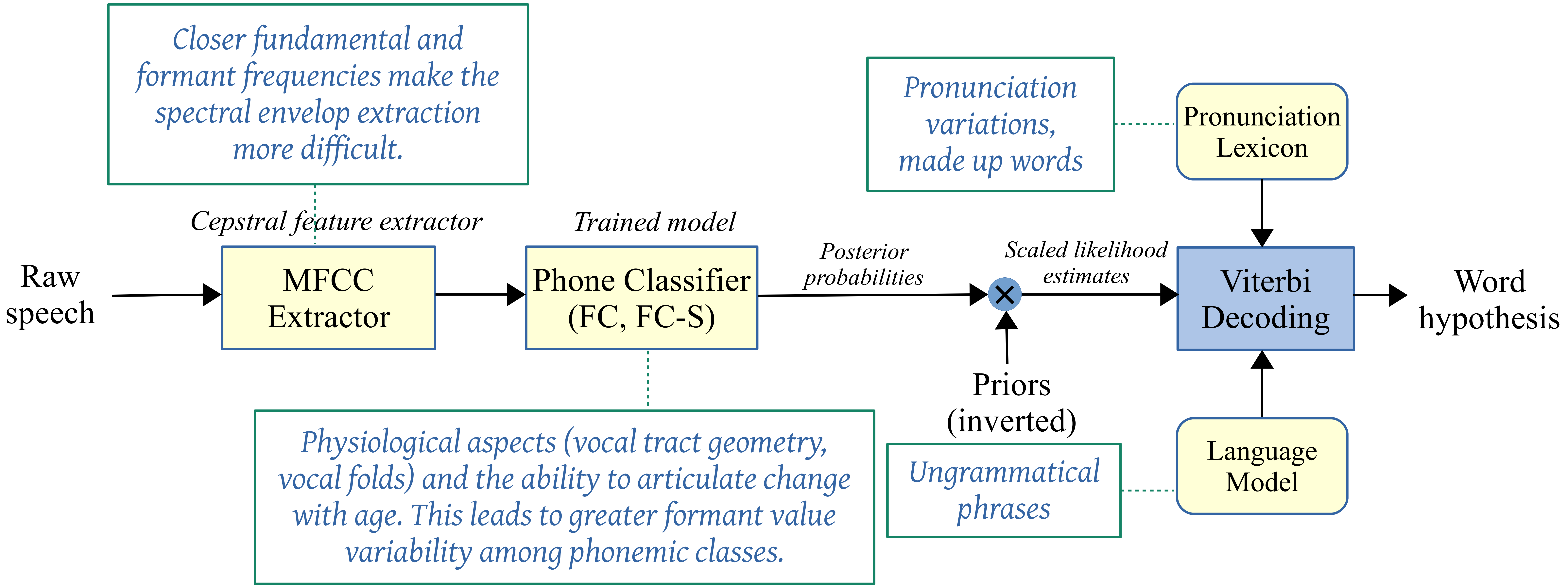
Overview of the paper

- Motivation: Challenges in children speech ASR
- Investigation: Jointly learning the features and the phone classifier
- Experimental setup and results
- Analysis
  - First convolutional layer filters as a spectral dictionary
  - Relevance analysis on the entire network
  - Transferability of adult speech representations to children speech

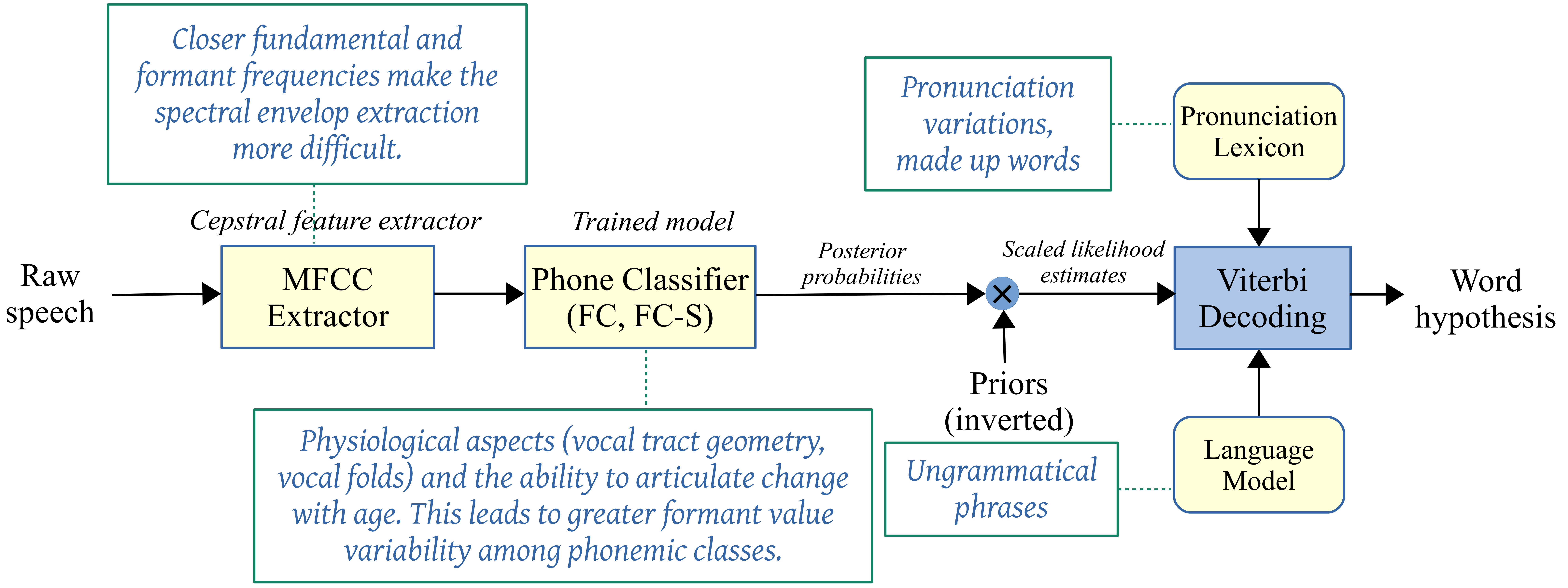
Automatic speech recognition and challenges with children speech



Automatic speech recognition and challenges with children speech

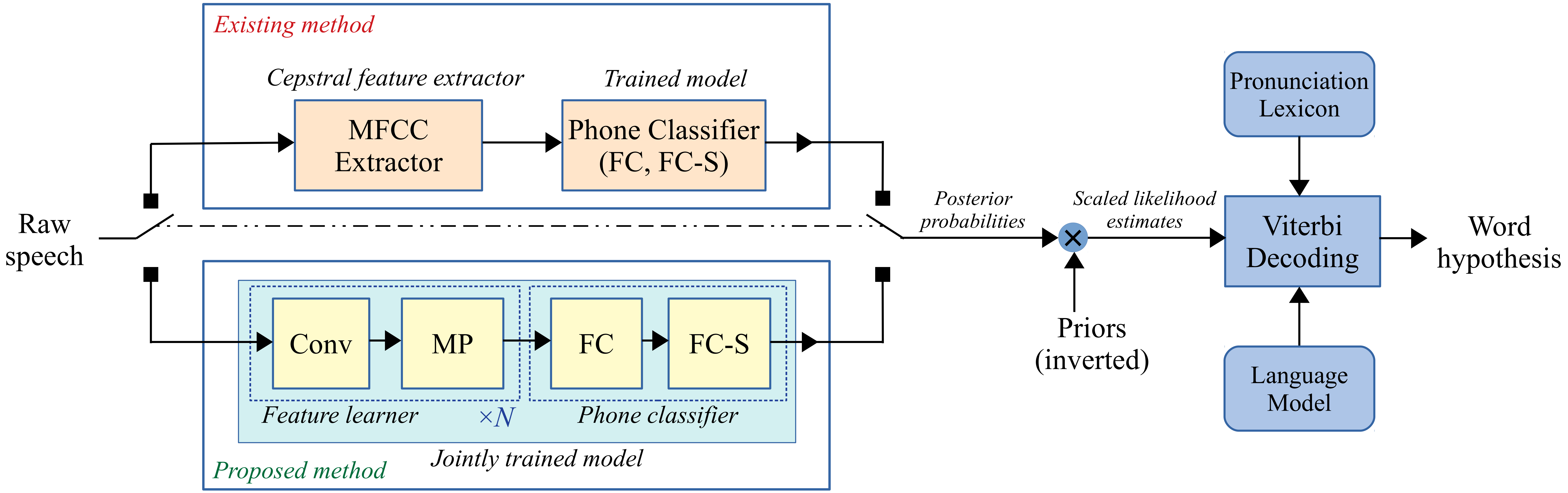


Automatic speech recognition and challenges with children speech



Here we address acoustic modelling.

Joint feature-classifier learning in hybrid HMM based ASR



FC: fully connected layer, FC-S: FC layer with softmax, Conv: convolutional layer, MP: max pooling.

D. Palaz, R. Collobert and M. Magimai.-Doss, “Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks”, in *Proc. Interspeech*, 2013.

Joint feature-classifier learning in hybrid HMM based ASR

Why this framework?

- It was shown to learn formant-like information from raw speech with minimal prior assumptions.
- *Hypothesis*: this should yield better children speech recognition than the conventional modelling of spectral envelop through source-system decomposition.

D. Palaz, R. Collobert and M. Magimai.-Doss, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks", in *Proc. Interspeech*, 2013.

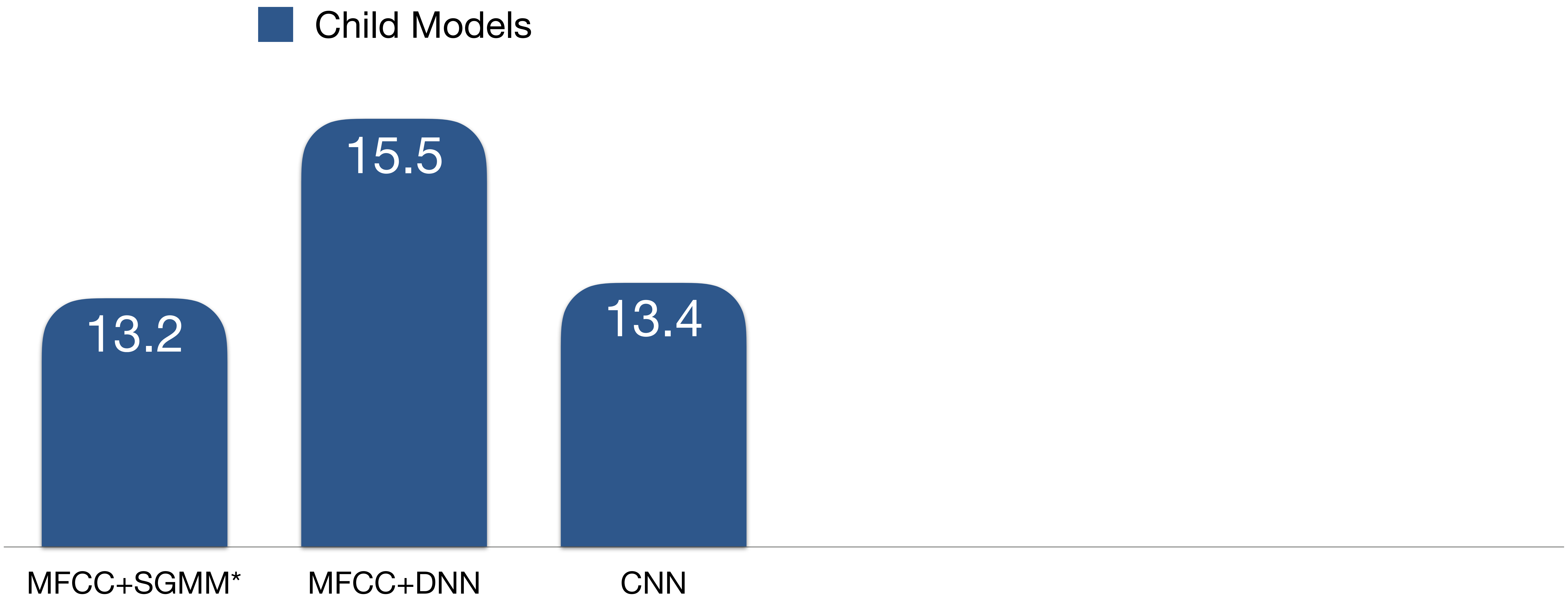
Data set

- Children speech data: PF-STAR corpus (training: 14.8 hours, testing: 4.7 hours).
- Two channel recordings of 158 child speakers in British English: we used both the channels.
- Pronunciation: Cambridge BEEP lexicon.
- 3-gram language model (LM) linearly interpolated from:
    - Training set LM, and
    - MGB-3 challenge data set LM.
- Adult speech data: WSJCAM0 corpus (training: 15.5 hours).

Experimental setup

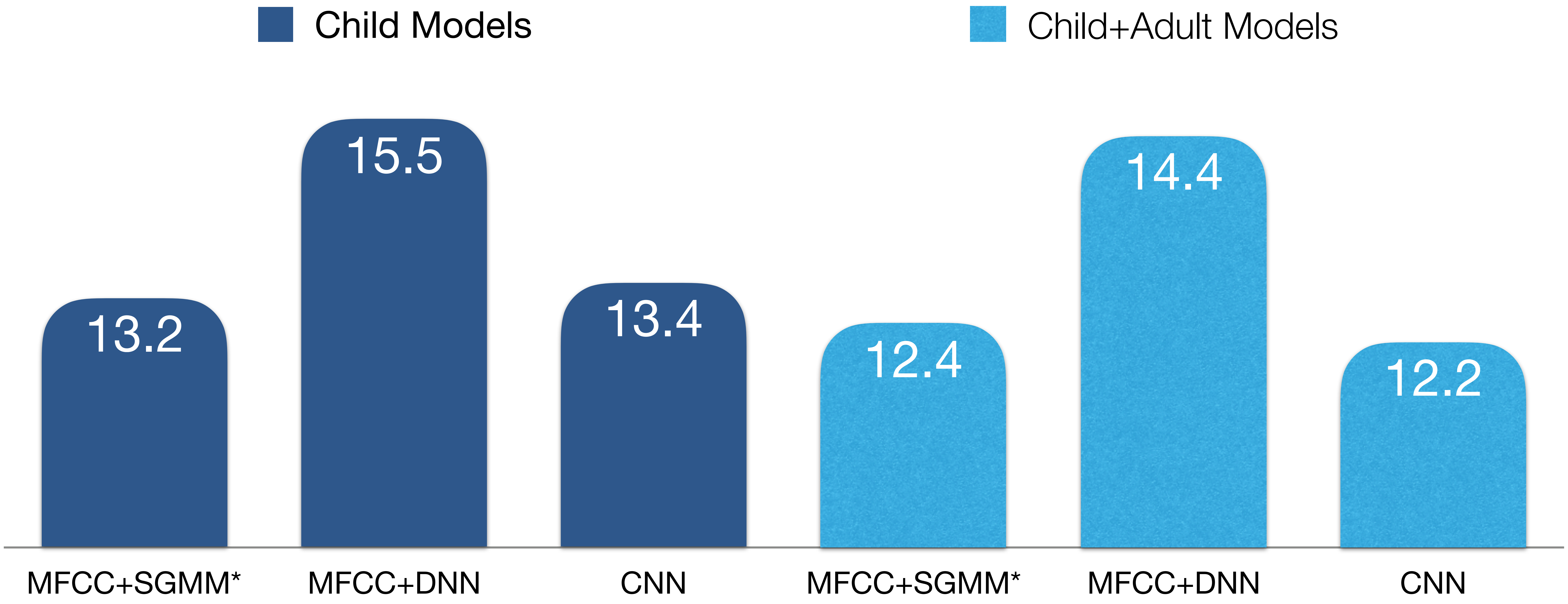
- Tools: HMMs using Kaldi, CNNs using Keras with Tensorflow backend.
- Training pipeline: monophone, triphone, LDA+MLLT, SAT with fMLLR, SGMM.
- CNN model architecture:
    - 3/4/5 convolutional layers, 1 fully connected layer.
    - 250ms input, operated by a 30 sample kernel.
- Conventional (DNN) systems:
    - Standard Mel frequency cepstral coefficient (MFCC) based features.
    - Models: 3 fully connected layers with rectified linear activations.
- Training was performed using cross-entropy loss, using stochastic gradient descent and dropout and a decaying learning rate.

Results: word error rates on near field child test



*Uses 3-pass decoding.

Results: word error rates on near field child test



*Uses 3-pass decoding.

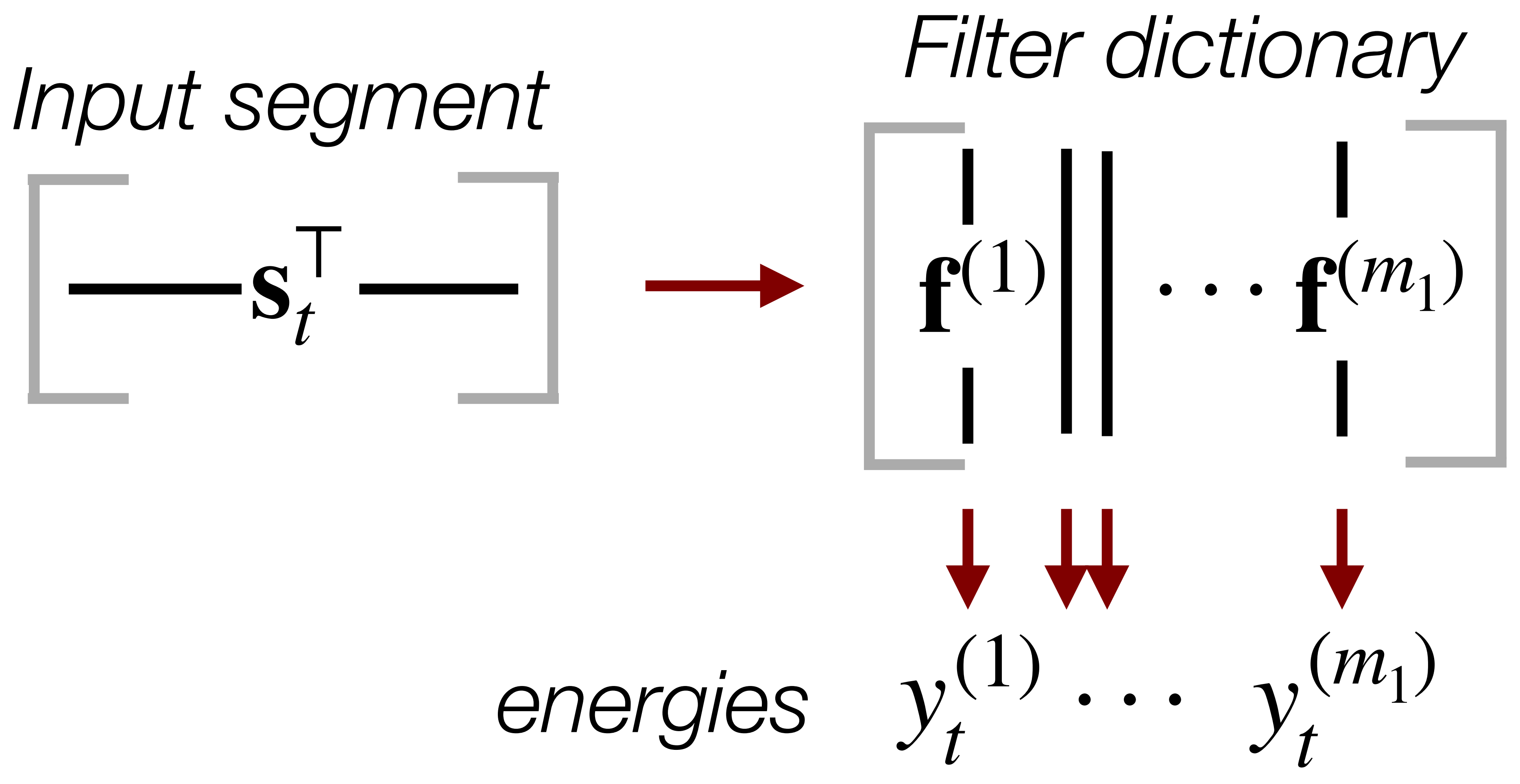
For far field test results, see the paper.

Overview

- Motivation: Challenges in children speech ASR ✓
- Investigation: Jointly learning the features and the phone classifier ✓
- Experimental setup and results ✓
- **Analysis**
    - **First convolutional layer filters as a spectral dictionary**
    - **Relevance analysis on the entire network**

Analysing the first convolutional layer

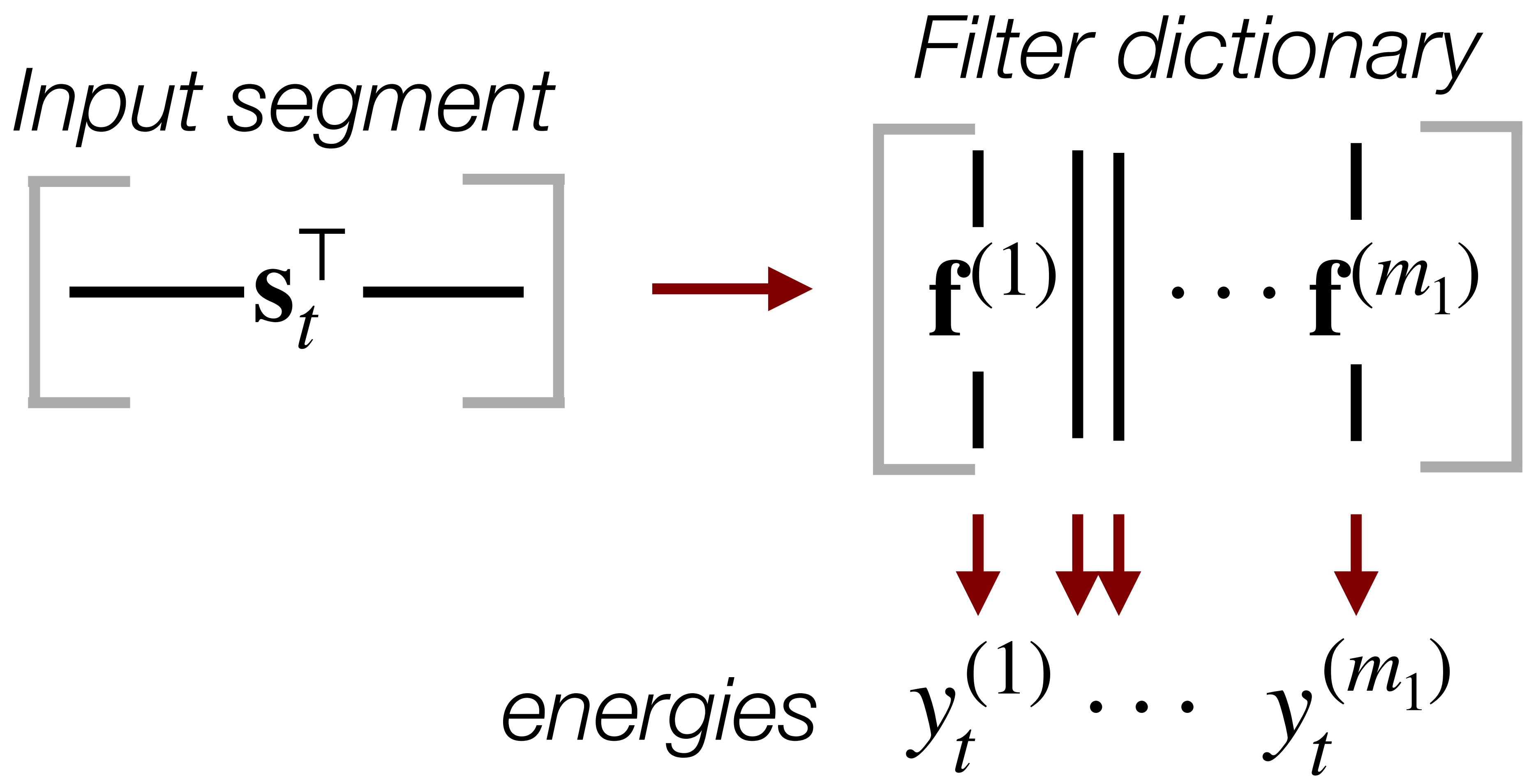
- The filters in the first convolution layer learn a spectral dictionary that discriminate phones[1].



[1]D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," Speech Communication, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.01.004

Analysing the first convolutional layer

- The filters in the first convolution layer learn a spectral dictionary that discriminate phones[1].



Different phones trigger different filter combinations. So energies are different for different phones.

[1]D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," Speech Communication, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.01.004

Analysing the first convolutional layer

- The filters in the first convolution layer learn a spectral dictionary that discriminate phones[1].

Input segment

$$\left[\ \text{—}\ \mathbf{s}_t^{\mathsf{T}}\ \text{—}\ \right]$$

Filter dictionary

$$\left[\ \mathbf{f}^{(1)} \ \Big|\ \Big|\ \cdots\ \mathbf{f}^{(m_1)}\ \right]$$

energies $y_t^{(1)} \cdots y_t^{(m_1)}$

We analyse the overall magnitude response to $\mathbf{s}_t$ *:*

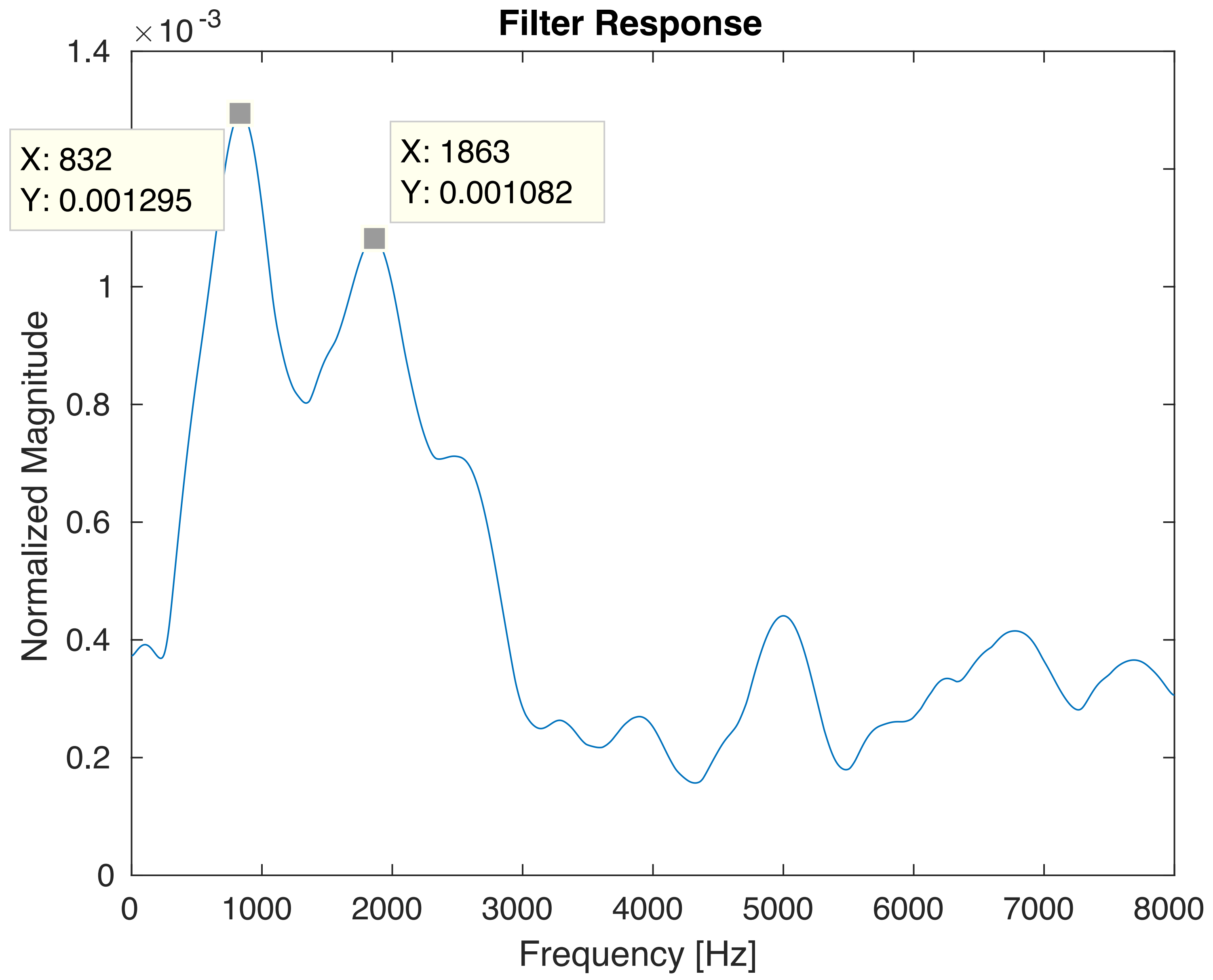
$$|\, y_t^{(1)}\mathbf{F}^{(1)} + y_t^{(2)}\mathbf{F}^{(2)} + \ldots + y_t^{(m_1)}\mathbf{F}^{(m_1)} \,|$$

Different phones trigger different filter combinations.
So energies are different for different phones.

[1]D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," Speech Communication, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.01.004

Analysing the first convolutional layer: experimental validation

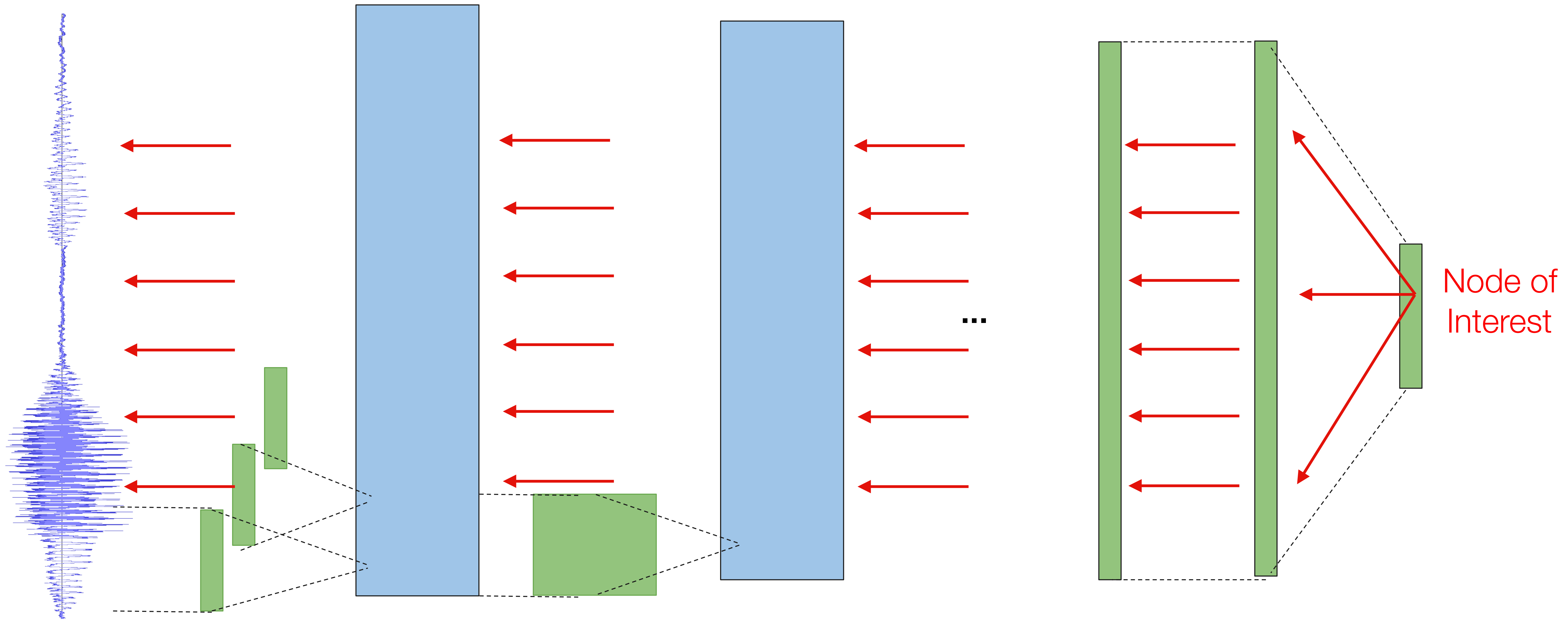
- Data: American vowel data set.
  - It consists of recordings of 12 vowels for each of its 150 speakers.
  - It contains annotated formant and F_0 values.
- We analysed on a standard subset of five vowels from four speakers (male, female, boy, girl), using 30ms short segments.
- We observed matching formant values across different vowels and speakers (consistent with the findings in *Palaz et al.*)
- This suggests that the first layer filters learn meaningful representations that discriminate phones.



Average filter response for a speech segment /er/ from a boy speaker using children speech CNN. Reference: F_1 = 614 Hz, F_2 = 1867 Hz.

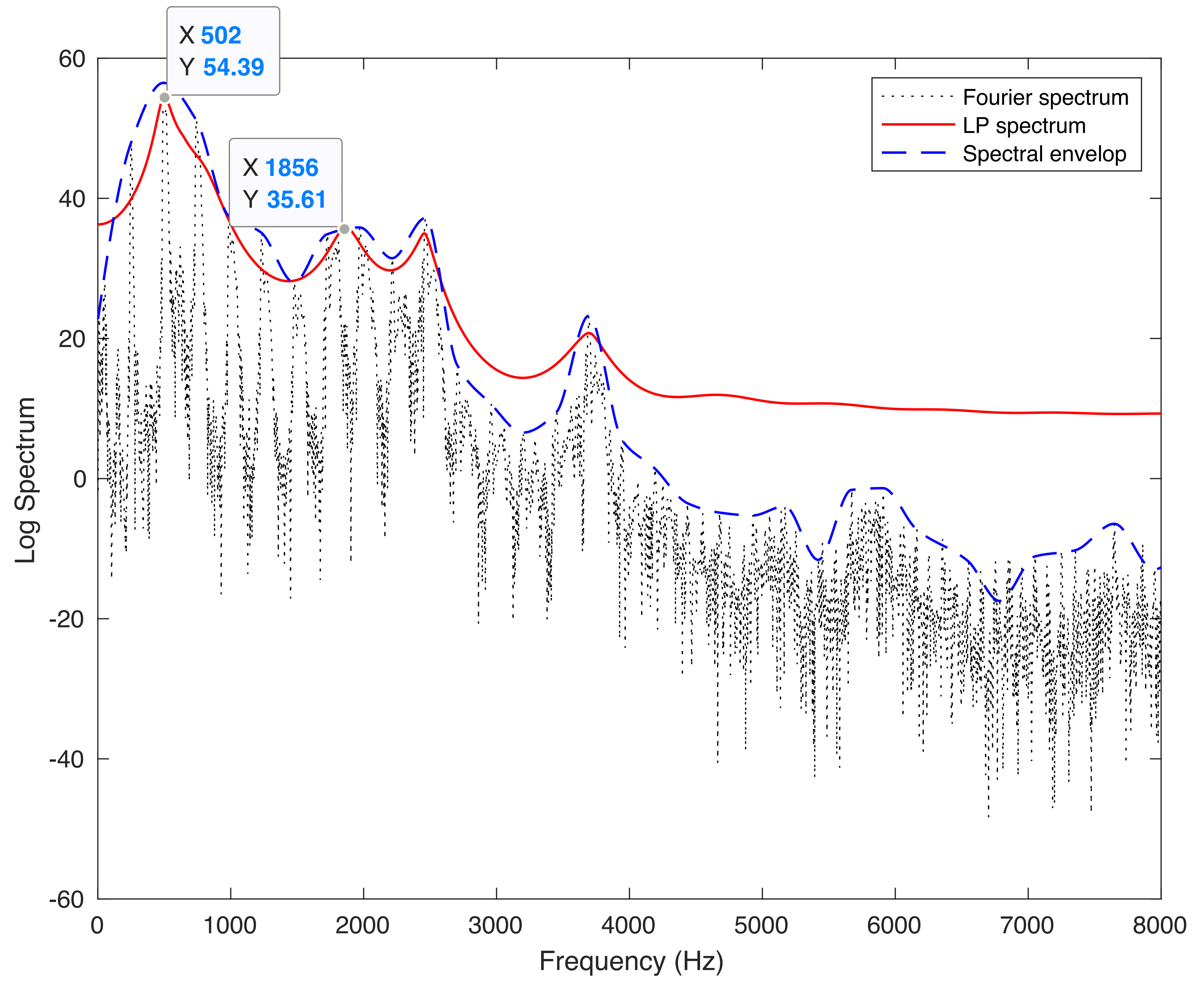
Analysing the entire network through relevance signals

- For a given input segment, the activation at a particular output node can be computed through forward pass.
- The gradient w.r.t. the node's output can be back-propagated to get a *relevance signal*.
- Relevance signal indicates the most informative input samples for the classification.
- Such methods are widely used in computer vision community.



Analysing the entire network through relevance signals: example

- Relevance signal can be analysed in terms of its spectral content[1].
- We computed relevance signals on 250ms children speech.
- We analysed their average linear prediction (LP) spectra through short-time processing.
- We observed that the estimated F_1 and F_2 values from the relevance signals are close to their references.



LP relevance spectrum for a speech segment /er/ from a boy speaker using children speech CNN. Reference: F_1 = 614 Hz, F_2 = 1867 Hz.

[1]H. Muckenhirn, V. Abrol, M. Magimai.-Doss, and S. Marcel, "Gradient-based spectral visualization of CNNs using raw waveforms," Idiap Research Institute, Tech. Rep. Idiap-RR-11-2018, Jul 2018. [Online]. Available from: http://publications.idiap.ch

Summary

- ***Children speech ASR can be improved through automatic feature learning***, instead of using the standard cepstral features.
- This may overcome the challenges in robustly extracting formant-related information from children speech.
- Augmenting children data with adult data could improve the systems further.
- Both the analysis of
    1. the first convolutional layer through the spectral dictionary interpretation and
    2. entire network analysis on gradient-based relevance signals

  showed that the CNNs learned information relevant to phone discrimination.

Thank you… Questions?

pavankumar.dubagunta@idiap.ch

https://www.linkedin.com/in/pavankumards

For full PDF, visit http://publications.idiap.ch or scan the QR code above.

Visit us at our posters

- 16:00 today at Poster Area B, *Learning voice source related information for depression detection*.
- 08:00 tomorrow at Poster Area A, *Segment level training of ANNs based on acoustic confidence measures for hybrid HMM/ANN speech recognition*.

idiap RESEARCH INSTITUTE

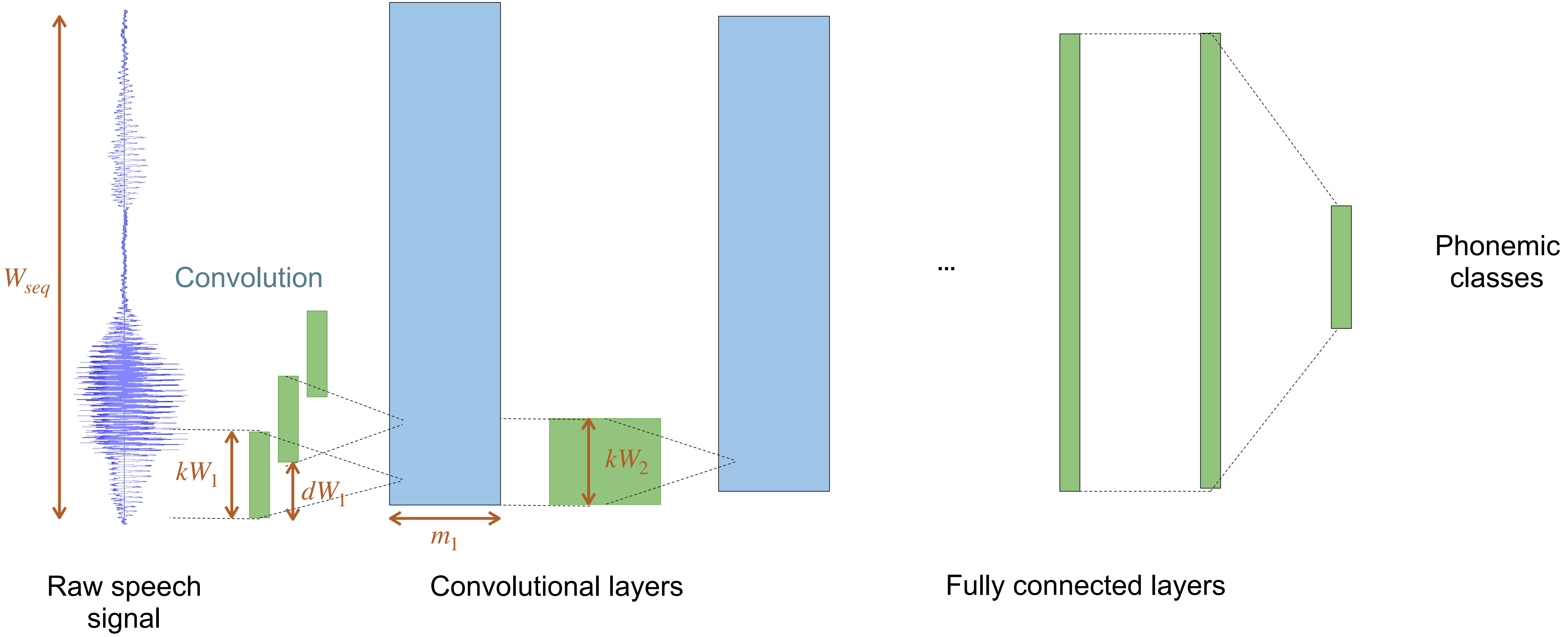
EPFL

HASLERSTIFTUNG

FNSNF
FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

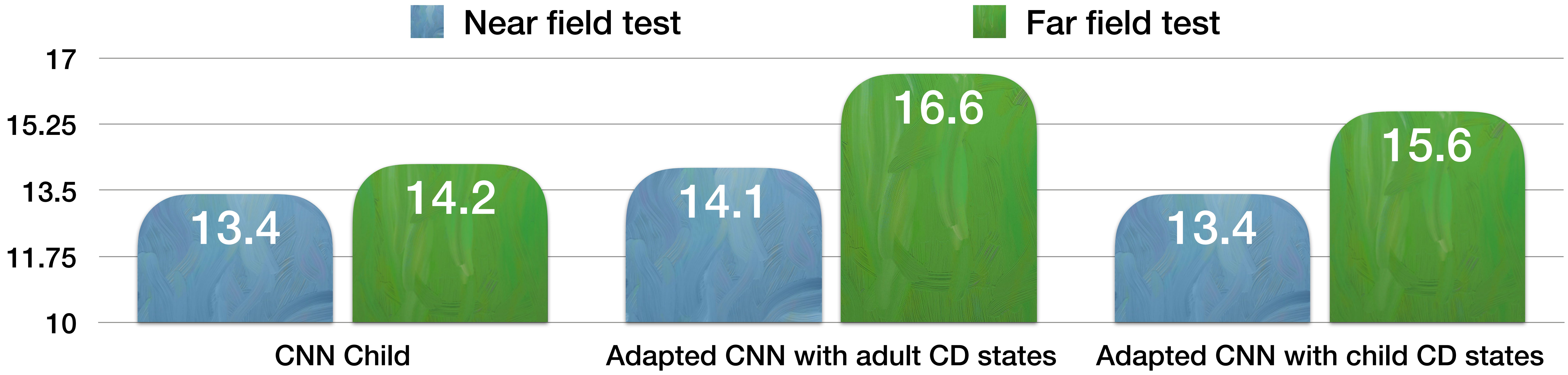
Additional Slides

Raw speech processing using CNNs



Transferability of adult feature embeddings to children speech

- We used the adult CNN parameters for children speech ASR — only the output layer was trained.
- This showed that the CNN feature representations learned from adult data are generalisable to ASR in children speech.
- However the context dependent (CD) state clustering may affect the performance.



Some other existing applications of raw speech modelling

- **LVCSR**: Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” in *Proc. Interspeech*, 2014, pp. 890–894.
- **VAD**: Rubén Zazo, Tara N. Sainath, Gabor Simko, and Carolina Parada, “Feature learning with raw-waveform CLDNNs for voice activity detection,” in *Proc. Interspeech*, 2016, pp. 3668– 3672.
- **Emotion recognition**: G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. ICASSP*, 2016, pp. 5200– 5204.
- **Spoofing detection**: H. Dinkel, N. Chen, Y. Qian, and K. Yu, “End-to-end spoofing detection with raw waveform CLDNNs,” in *Proc. ICASSP*, 2017, pp. 4860–4864.
- **Speaker verification**: H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, “Towards directly modeling raw speech signal for speaker verification using CNNs,” in *Proc. ICASSP*, 2018, pp. 4884–4888.
- **Gender identification**: S. H. Kabil, H. Muckenhirn, and M. Magimai.-Doss, “On learning to identify genders from raw speech signal using CNNs,” in *Proc. Interspeech*, 2018.
- **Classification of paralinguistic information**: B. Vlasenko, J. Sebastian, S. P. Dubagunta, and M. Magimai.-Doss, “Implementing fusion techniques for the classification of paralinguistic information,” in *Proc. Interspeech*, 2018.