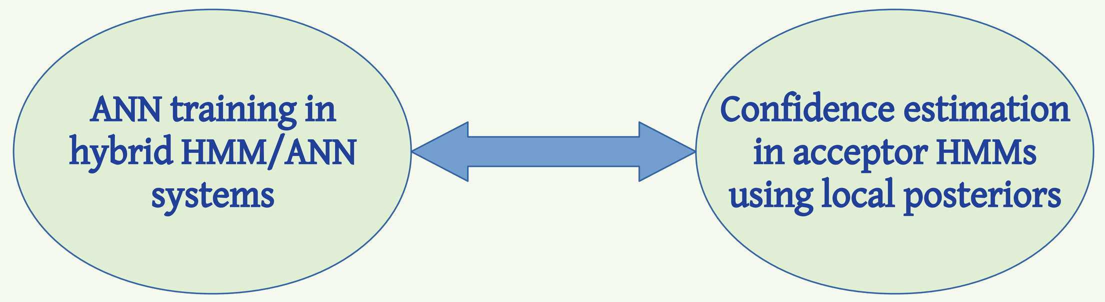
Segment-level training based on Confidence Measures for Hybrid HMM/ANN Speech Recognition

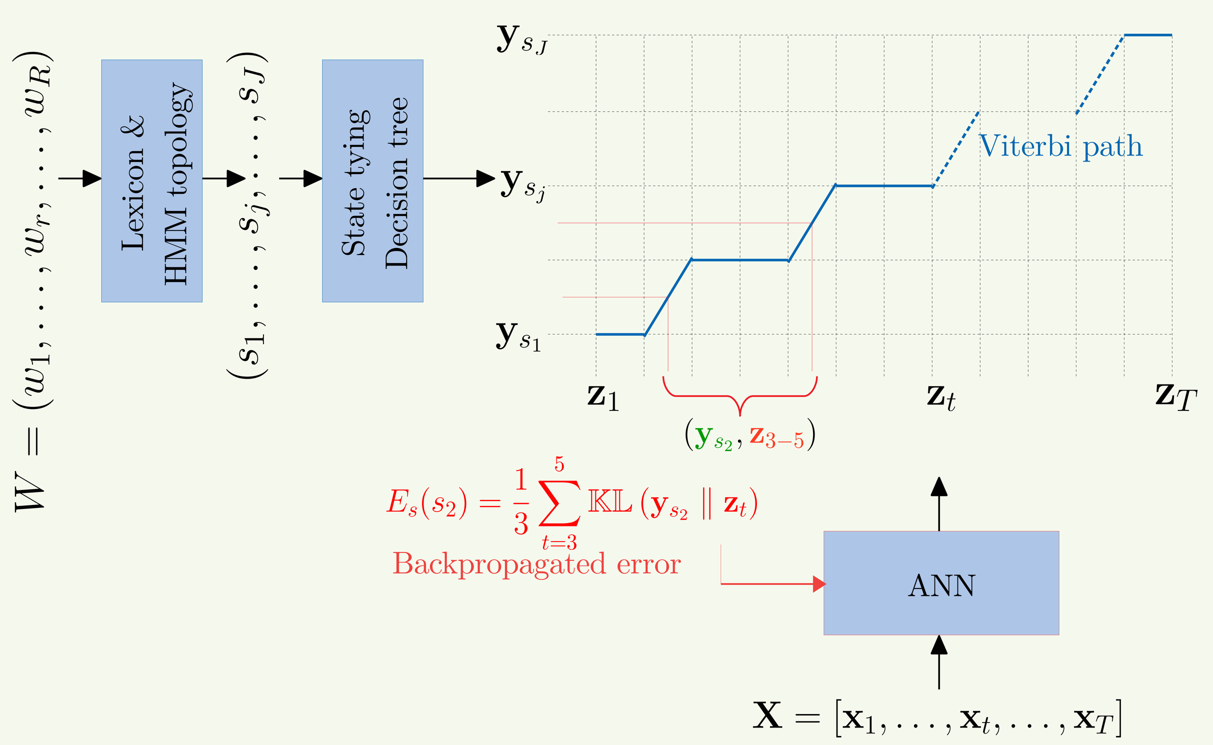
S. Pavankumar Dubagunta[1,2] and Mathew Magimai Doss[1]

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

① In this Work



④ Proposed Segment-level ANN Training



- When the segments represent HMM states, a state-level error $E_s(s_j)$ can be minimised:

$$E_s(s_j) = -CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \mathbb{KL}\left(\mathbf{y}_{s_j} \| \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}$$

- When the segments represent phone units, a phone-level error $E_{ph}(ph_k)$ can minimised:

$$E_{ph}(ph_k) = \frac{1}{N_{ph_k}} \sum_{n=1}^{N_{ph_k}} E(s_{j+n}) \text{ ; where } ph_k \text{ comprises } N_{ph_k} \text{ states}$$

- Priors $P(a^d)$ are estimated from the state segment counts instead of frame label counts.

⑤ Data Sets and Experimental Setup

	AMI	Mediaparl German	Mediaparl French	TIMIT
Training hours	77.3	14.5	16.1	3.1
Phone set count	176	57	38	48
Vocabulary size	52.5k	16.7k	12.4k	48
LM order	3-gram	2-gram	2-gram	2-gram
Features	fMLLR+spk-iVec	MFCC		
Tools	Kaldi+Keras/Tensorflow			
Alignments	From HMM-SGMM systems			
Training	E_f, E_s or E_{ph}, followed by sMBR			

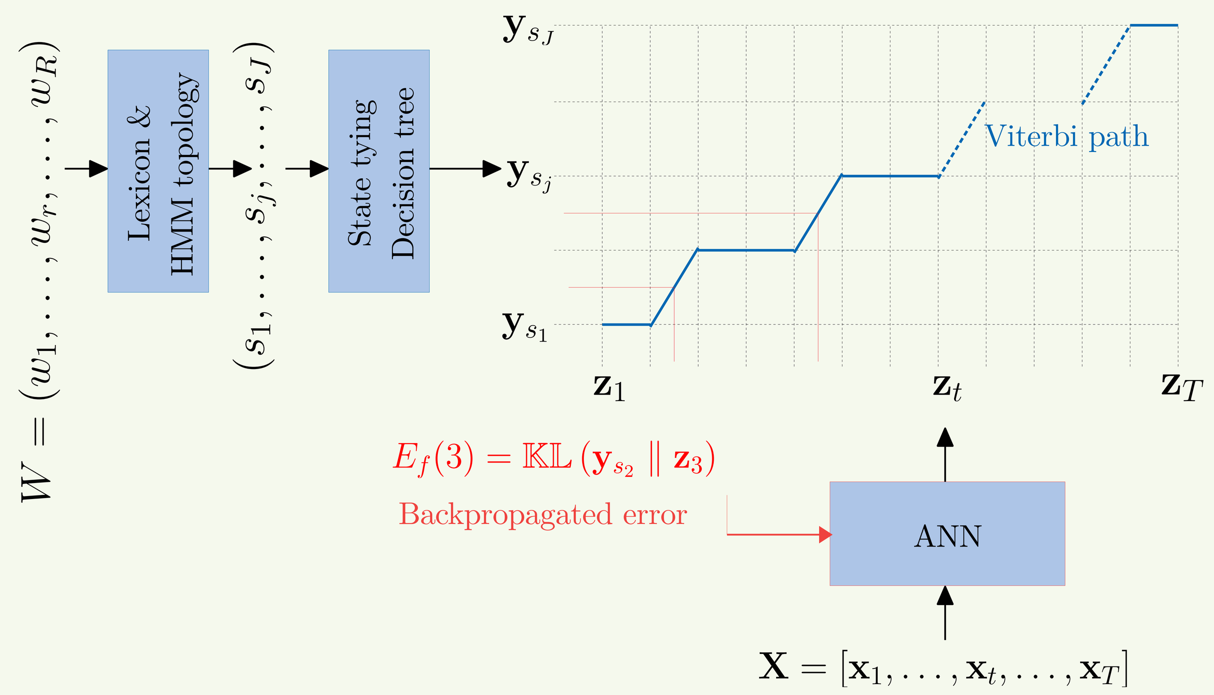
② Background: Conventional Hybrid HMM/ANN Systems

- In hybrid hidden Markov model (HMM) based speech recognition, the *scaled likelihood* of an acoustic observation $\mathbf{x}_t$ given a HMM state q_t at time t, labelled l^i, is estimated as:

$$\frac{p(\mathbf{x}_t|q_t = l^i)}{p(\mathbf{x}_t)} = \sum_{d=1}^{D} \frac{p(\mathbf{x}_t|a^d)}{p(\mathbf{x}_t)} P(a^d|q_t = l^i) = \sum_{d=1}^{D} \frac{\overset{\text{ANN}}{P(a^d|\mathbf{x}_t)}}{\underset{\text{priors}}{P(a^d)}} \underset{\text{decision tree}}{P(a^d|q_t = l^i)}$$

- Conventionally, given the *segmentation*, the artificial neural network (ANN) is trained using one hot encodings of the targets and minimising frame level cross-entropy. This can be expressed as:

$$E_f(t) = \mathbb{KL}\left(\mathbf{y}_{s_j} \| \mathbf{z}_t\right) \overset{(s_j = l^j) \mapsto a^{d'}}{=\!=} \mathbb{KL}\left(\delta_{d'} \| \mathbf{z}_t\right) = -\log\left(P(a^{d'}|\mathbf{x}_t)\right) \qquad (1)$$



⑥ Results

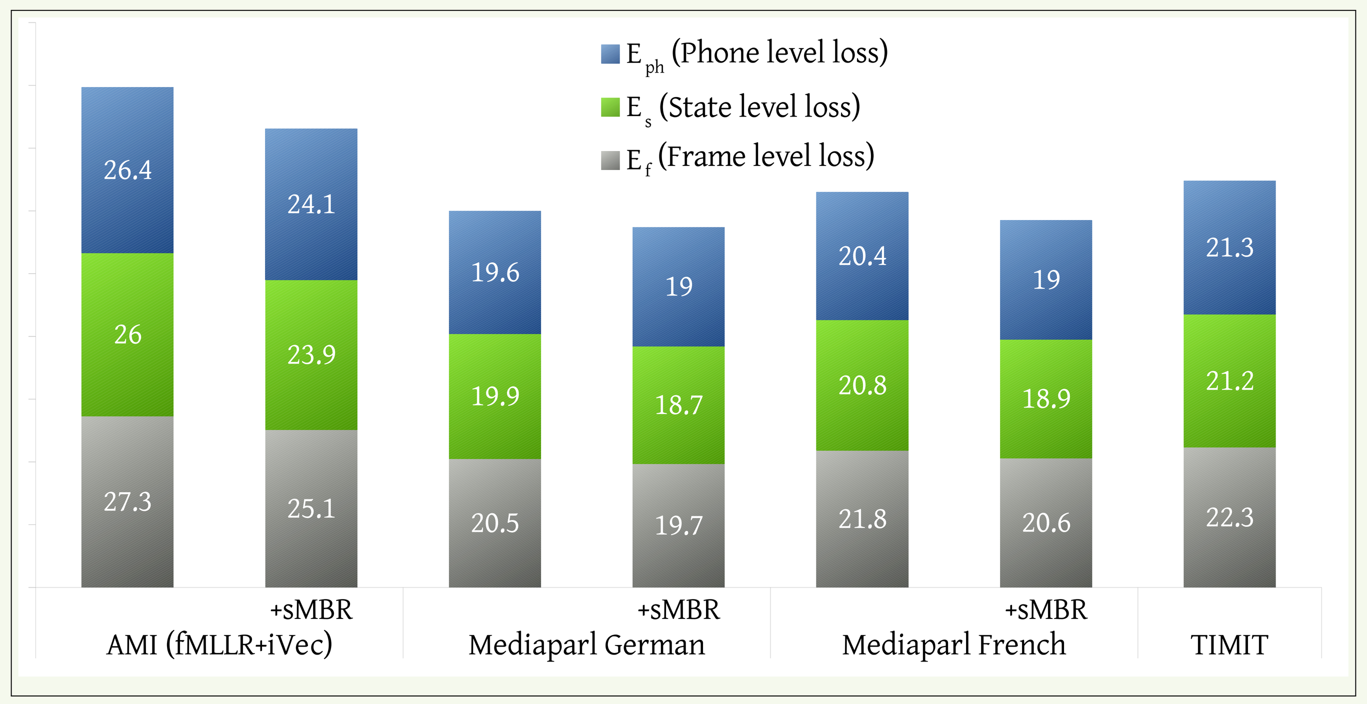


Figure: Word error rate on AMI and Mediaparl data sets, phone error rate on TIMIT

Acknowledgements

HASLERSTIFTUNG

This work was funded by the Hasler foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS) and was entirely carried out at the Idiap Research Institute.

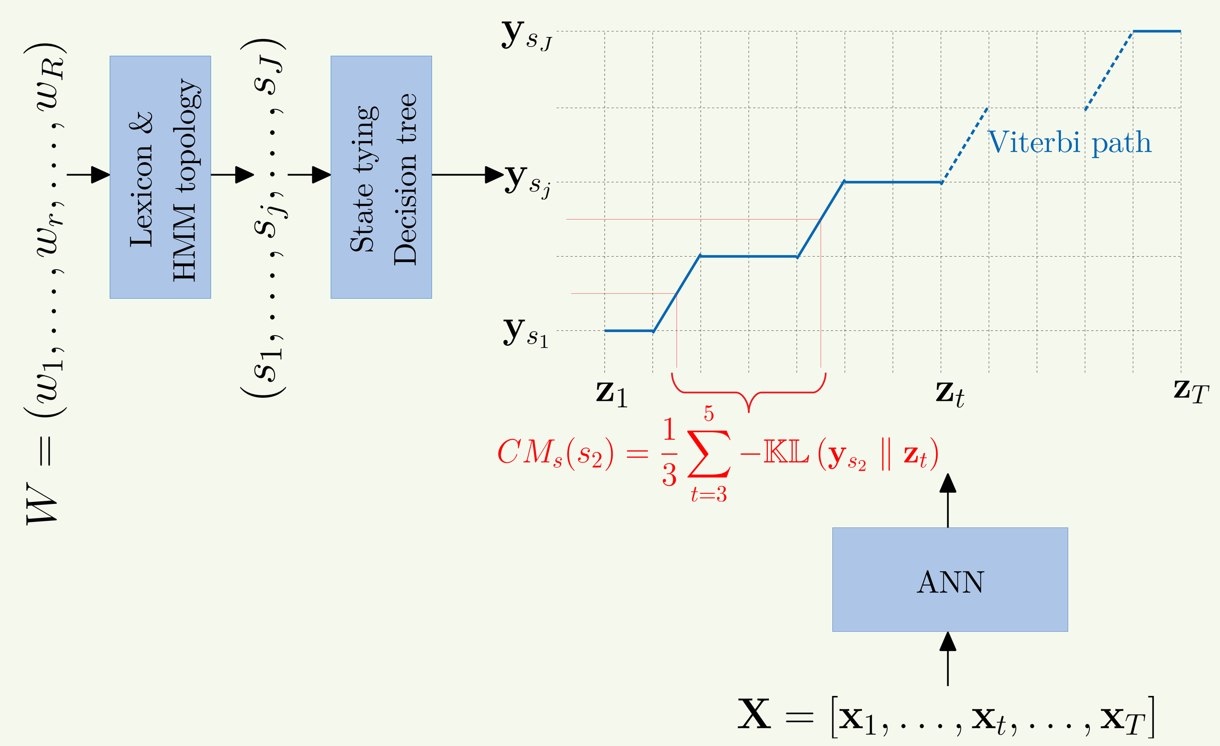
③ Confidence Estimation using Local Posteriors

- Given an alignment between $\mathbf{X}$ and W and the local posterior probability estimates, a confidence measure $CM(s_j)$ can be estimated by rescoring the segment of s_j at $t = \{b(s_j), \dots, e(s_j)\}$ as

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(q_t = l^j|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1} \overset{l^j \mapsto a^{d'}}{=\!=} \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(a^{d'}|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1}$$

- Based on (1), we can express the state level confidence $CM(s_j)$ estimation as a matching of $\mathbf{Y}$ and $\mathbf{Z}$ with a local cost based on Kullback-Leibler (KL) divergence:

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} -\mathbb{KL}\left(\mathbf{y}_{s_j} \| \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}$$



⑦ Analysis: Effect of Silence Duration on the Training

- Here we show how increasing the silence duration affects training using the three cost functions.
- Silence was artificially added at the beginning and end of each training and test utterance.

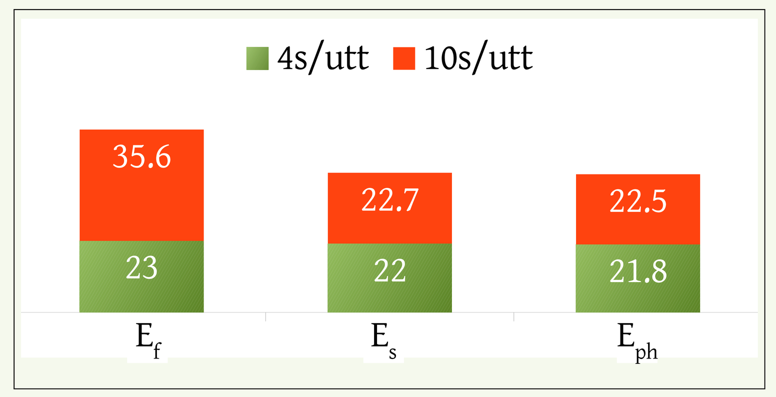


Figure: Phone error rate on TIMIT

⑧ Conclusion

The proposed linguistic-segment-level training of ANNs based on confidence measures

- yields better systems than using frame-level cross-entropy.
- adds to the efficacy of further sequence discriminative training.
- improves robustness to duration variations in the training data set.

Get in touch? pavankumar.dubagunta@idiap.ch https://linkedin.com/in/pavankumards

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