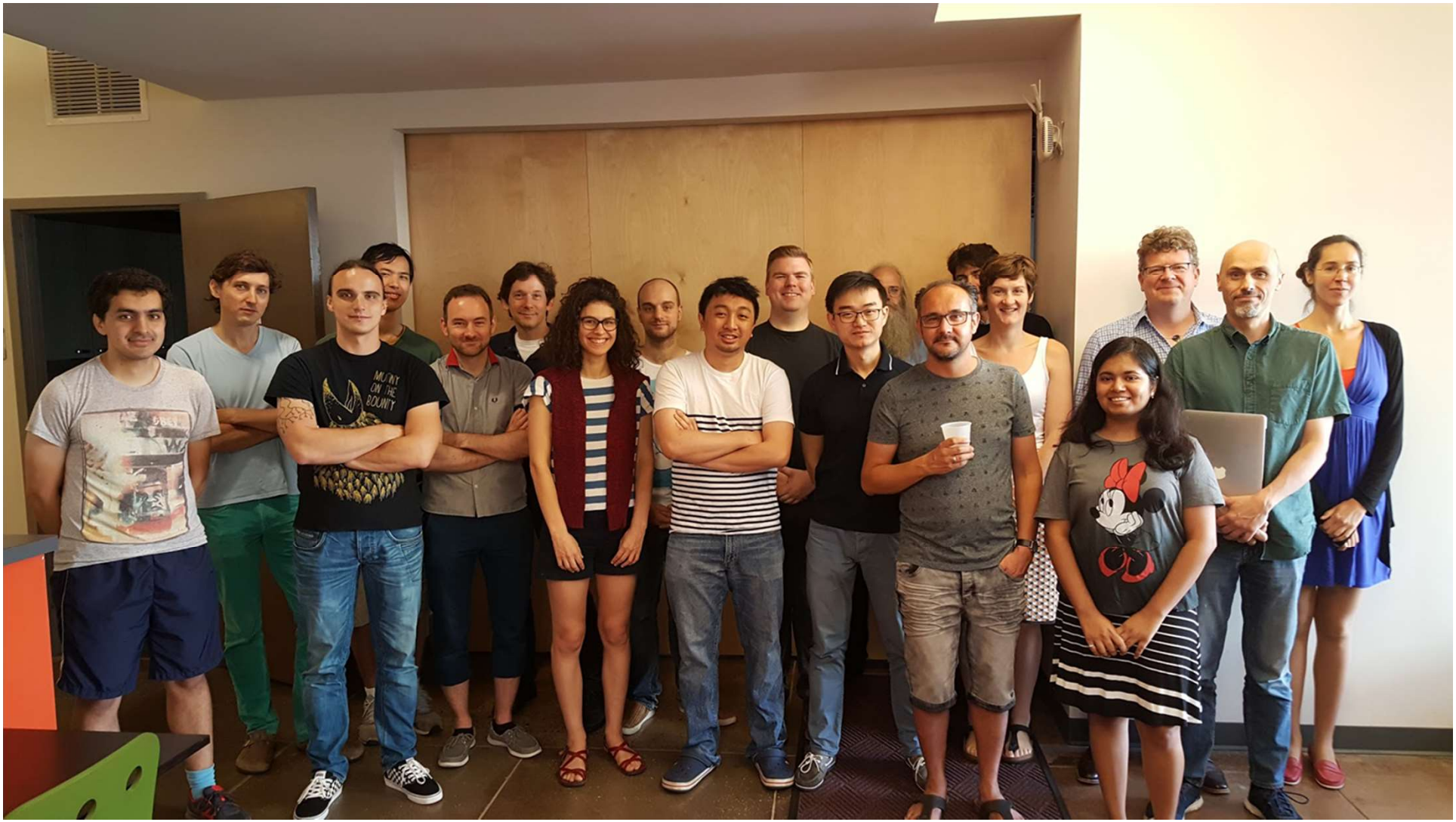
The Speaking Rosetta Stone - Discovering Grounded Linguistic Units for Languages without Orthography

JSALT 2017 CMU - Final Presentation

Emmanuel Dupoux, Odette Scharenborg, Graham Neubig, Laurent Besacier, Mark Hasegawa-Johnson, Alan Black, Florian Metze, Sebastian Stüker, Lucas Ondel, Pierre Godard, Markus Müller, Shruti Palaskar, Francesco Ciannella, Philip Arthur, Elin Larsen, Danny Merkx, Liming Wang, Mingxing Du, Rachid Riad


MOUNT
ON THE
BOUNTY

Schedule for this morning

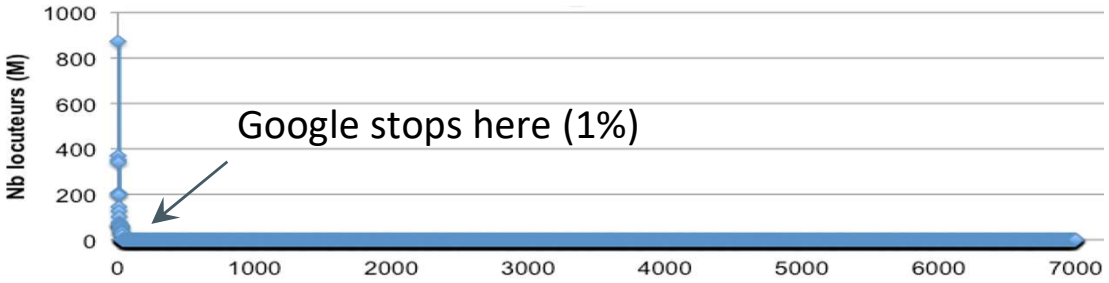
- Introduction
  15 mins
- XMNT: a general architecture for end-to-end experiments 20 mins
- Bottom-up unit discovery
  30 mins
- Q&A
  10 mins
- Break
  15 mins
- End2end cross-modal unit discovery
  30 mins
- Constructing a TTS system in an unwritten language 30 mins

Coming up next…

Introduction
(what did we do and why)

Starring (in order of appearance): Emmanuel Dupoux, Laurent Besacier

Rationale



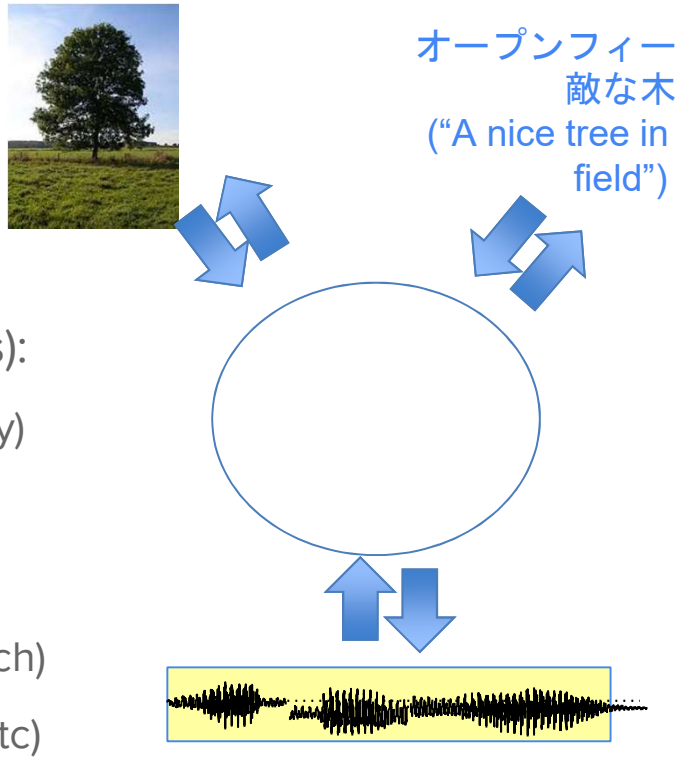
- Speech and Language Technology (SLT) requires large quantities of text
  - e.g. Microsoft ASR system (Xiong et al, 2016)
    - acoustic modeling: 2000+ hours of orthographically transcribed speech
    - language modeling: 350+ Mwords of text
- Long tail: too expensive/impractical for the majority of the world languages
  - orthography not stable or not used in everyday life
  - eg. Moroccan Arabic (21M), Nigerian languages: Igbo (25M), etc.
- 4 years olds are efficient at speech processing *before* they can read and write

➔ ***Engineering Challenge****:* can we build useful SLT without *any* textual resources?

➔ ***Scientific Challenge****:* can we build algorithms that learn languages like infants do?

http://www.vistawide.com/languages/top_30_languages.htm

Goals

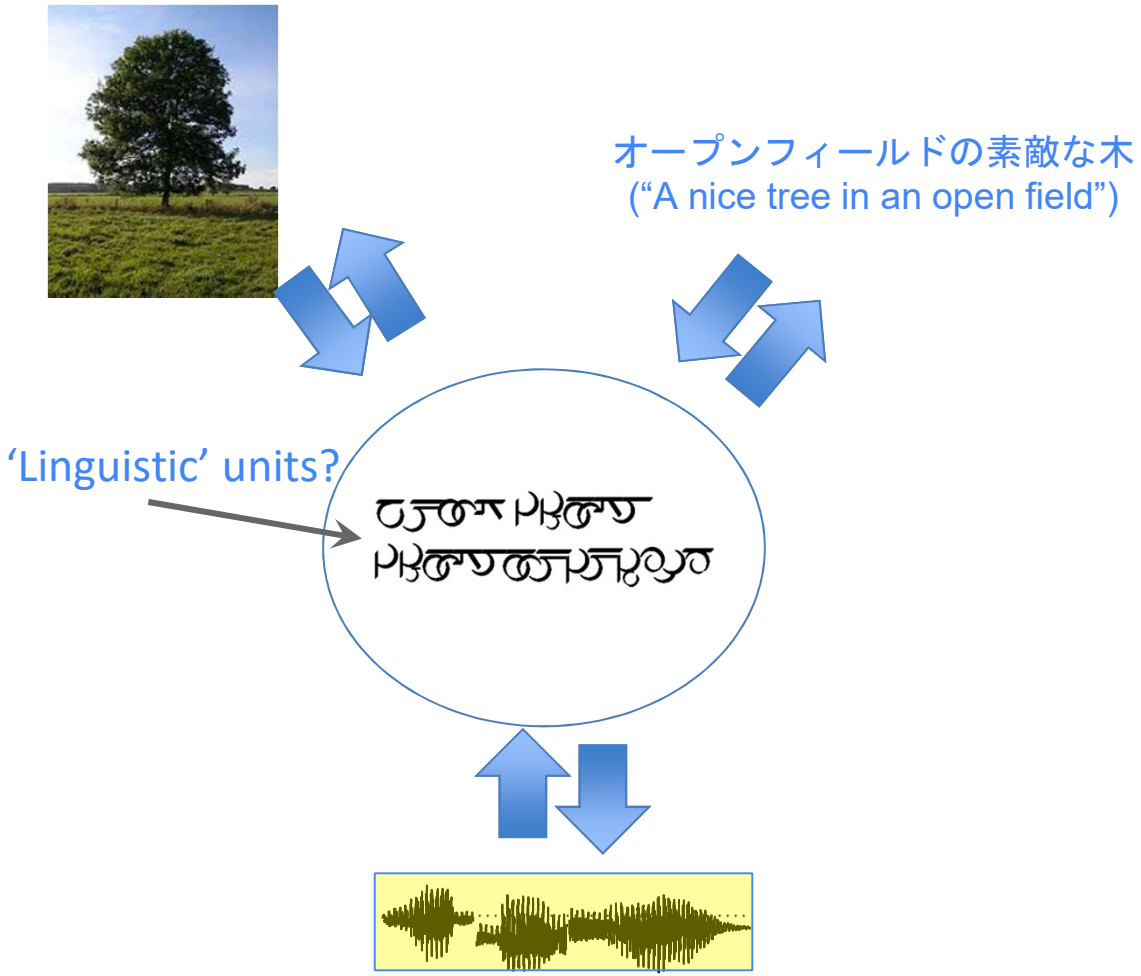


- Ultimate goal
  - Technological challenge (end-to-end tasks):
    - image or document retrieval (speech query)
    - spoken captioning of images
    - speech translation (both ways)
    - spoken document summarization (in speech)
    - full spoken dialogue system (SIRI, Alexa, etc)
  - Scientific challenge (models of autonomous language acquisition):
    - grounded learning (Roy & Pentland, 2002; ACORNS, Boves et al, 2007...)
    - grammar induction (Siskind, 1996; Kwiatkowski et al 2012..)
    - predictive models of phonetic and lexical learning in infants

Main research question:

Do we need intermediate symbolic units ?

If yes, to do what?

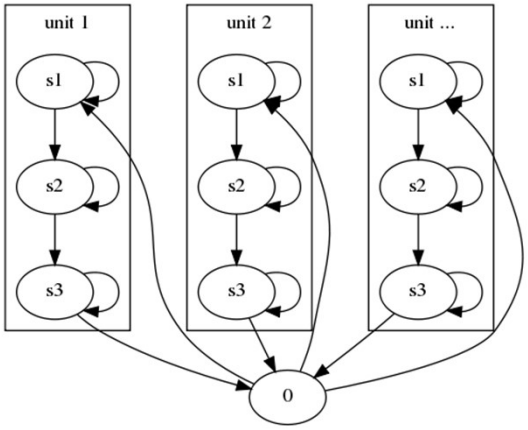


Strategy

- Approach
  - Concentrate on 4 tasks
    - two with symbolic units: unit discovery, speech synthesis
    - two end-to-end: speech2image, speech2translation
  - Study the interactions between the symbolic and end-to-end tasks
- Goals for this workshop
  - Construct a core set of open-source baseline systems and datasets
  - Conduct proof of concept experiments & test new ideas

Baseline systems (1): linguistic unit discovery

- Phone discovery
  - Task: *given speech, discover discrete phoneme-like units*
  - How: Clustering
    - DP-GMM (Chen et al. 2015)
    - **DP-HMM** (Lee & Glass; Ondel et al., 2016)
  - Plus: Input feature learning
    - Speaker adaptation (VTLN, fMMLR; Heck et al 2017.)
    - **Universal Bottleneck or articulatory features** (Frantisek et al., 2011, 2014; Deng, 1997; Metze & Waibel, 2002; Black et al 2012)
    - **Weakly supervised NNs** (Autoencoders: Badino et al, 2014; Renshaw et al 2015; Siamese networks: Synnaeve et al. 2014)

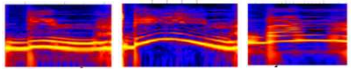


- Evaluation

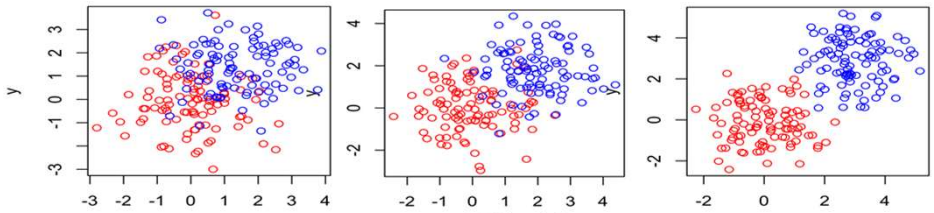
Minimal pairs ABX task

A B X

ba_{T1} ga_{T1} ga_{T2}



$$\theta(A,B) := \frac{1}{m(m-1)n} \sum_{a \in A} \sum_{b \in B} \sum_{x \in A \setminus \{a\}} \left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x)=d(b,x)} \right), \quad m = |A|, \; n = |B|$$



ABXerr=30% (1.6σ)

ABX=.20 (2σ)

ABX=.05 (2.4σ)

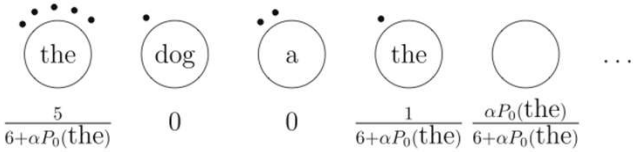
Schatz et al, 2013;2014

Properties

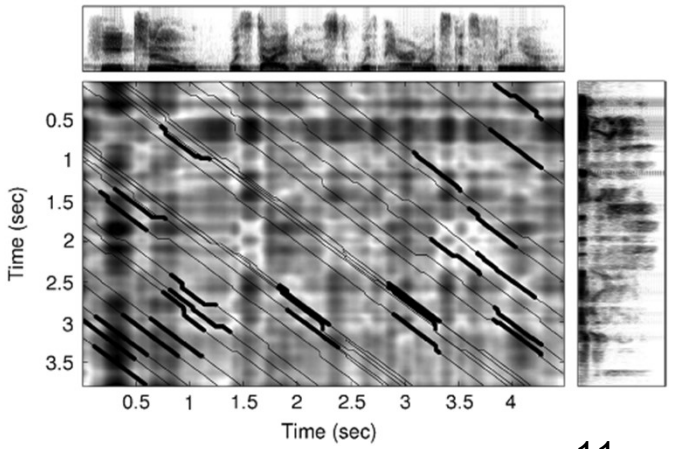
- no training
- model free
- representation independent (continuous, discrete, probabilistic)
- predicts the outcome of unsupervised clustering
- statistically stable

Baseline systems (1): linguistic unit discovery

- Word discovery
  - Task: *given speech, segment word-like units, cluster them, and use them to parse input*
  - How (from text or lattices):
    - **Non Parametric Bayesian Models** (Goldwater et al. 2006; Neubig et al, 2010; Heiman et al, 2013; 2014)
  - How (from speech):
    - **DTW based** (Park & Glass, 2008; Jansen & van Durme, 2011)
    - Word embeddings+clustering (Kamper et al. 2017)

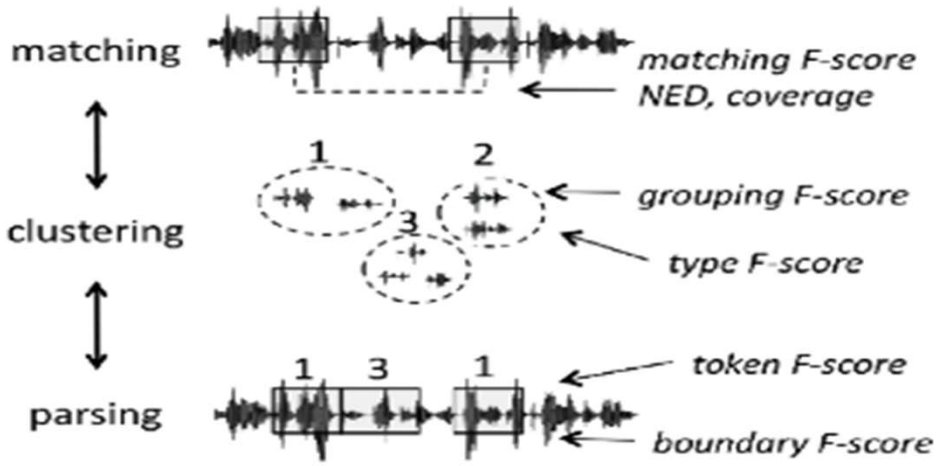


From Goldwater et al. 2009



From Park & Glass, 2008

- Evaluation
  - evaluating separately the three logical steps: matching, clustering, parsing
  - based on phone labels



Baseline systems (2): TTS

- Task: *given a string of phonemes, produce a plausible waveform*
- How: Parametric Speech Synthesis (HMM-MFCC based) (Tokuda et al, 2000; CMU Clustergen; Black 2006)

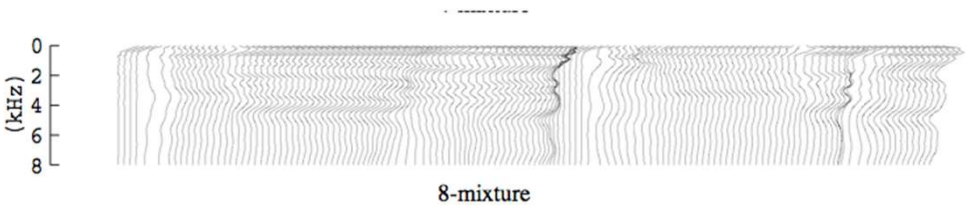


Figure 1: Generated spectra for a sentence fragment "kiNzokuhiroo."

From Tokuda et al (2000)

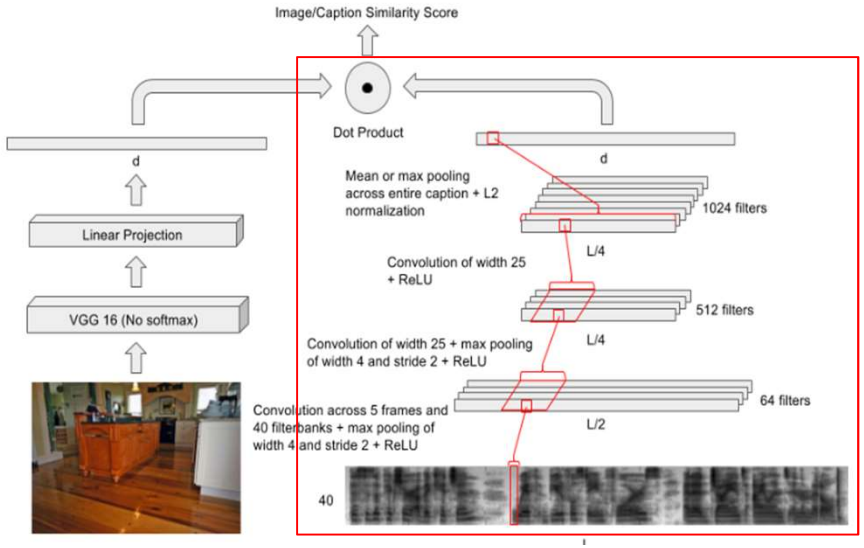
- Evaluation: Mel Cepstral

$$10/ln(10) * \sqrt{2 \sum_{i=1}^{24} (mc_i^t - mc_i^p)^2}$$

- Distortion measure on held out data
- Used for voice conversion and speech synthesis
- Correlates with perceptual judgments

Baseline systems (3): Speech2Image retrieval

- Task: *given speech, retrieve the image it is describing (or vice versa)*
- How: two encoders, learn a ‘similar’ embedding for the two modalities (Harwath&Glass 2015,2016,2017; Chrupala et al. 2017)

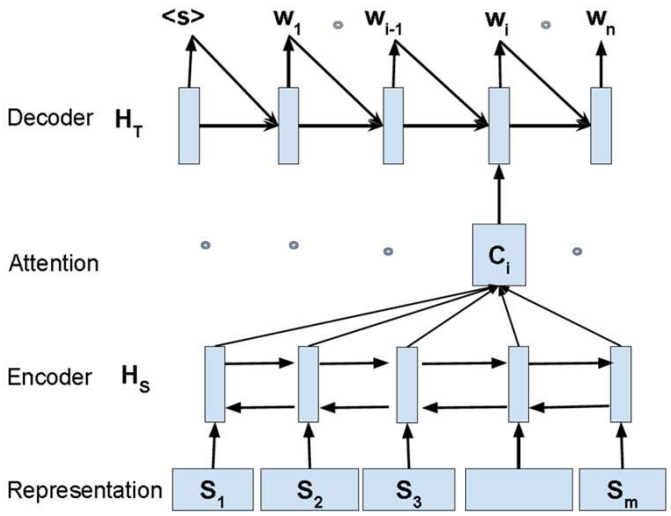


- Evaluation
  - in a held out test set
  - Recall R@1, R@5, R@10

- Can be applied to speech2translation retrieval

Baseline systems (4): Speech translation

- Task: *given speech, generate the text translation*
- How: sequence encoder, attention, sequence decoder (Bahdanau et al, 2015, Duong,et al 2016; Berard et al 2016)



- Evaluation
  - Bleu score (modified precision of transcribed output compared to reference ones; on held out data)
  - Bleu1 (unigram) for words
  - Bleu4 (4-grams) for phonemes or letters

- Can be applied backwards (from text, generate phones, pseudophones or even speech frames)

[Laurent]

Datasets

- Overview

Data Set	Lang.	Size	# speech utterances	aligned translations	aligned images	#spkrs
Flickr8k	En	62h	40 000	yes (Japanese - MT)	yes	183
Mboshi	Mb	5h	5 157	yes (French - human)	no	3
Mscoco	En	600h	616 767 (tts)	yes (Japanese - MT)	yes	8 (tts voices)
CGN	Du	64h	-	no	no	324

Datasets

- AMT recordings obtained from **Flickr8k** - 40k spoken captions available online
  - https://groups.csail.mit.edu/sls/downloads/
  - D. Harwath and J. Glass, *"Deep multimodal semantic embeddings for speech and images"* in IEEE ASRU, Scottsdale, Arizona, USA, December 2015
- We added Japanese translations (Google MT) for all captions (+ tokenization)

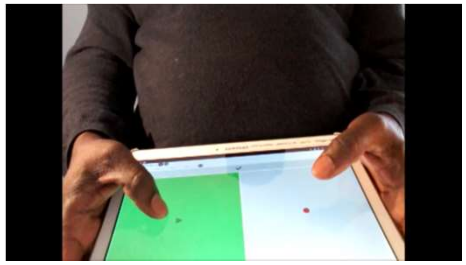
-A brown and white dog is running through the snow
-A dog is running in the snow
-A dog running through snow
-A white and brown dog is running through a snow covered field
-The white and brown dog is running over the surface of the snow

-茶色の白い犬が雪の中を流れています。
-犬は雪の中を走っています
-雪の中を走っている犬。
-白く茶色の犬が雪で覆われた畑を流れています。
-白い茶色の犬と茶色の犬が雪の表面を走っています。

+

Datasets

- **Mboshi** is spoken in Congo-Brazzaville - documented by the BULB project
  - G. Adda & al. *"Breaking the unwritten language barrier: The Bulb project"* in Proceedings of SLTU, Yogyakarta, Indonesia, 2016
  - collected with Lig_Aikuma mobile app: https://lig-aikuma.imag.fr
- Corpus built from translated reference sentences and from a Mboshi dictionary
- (Human) French translations aligned to the speech utterances
- Language rarely written - linguists have defined a nonstandard grapheme form (close to language phonology) - forced alignments between speech & transcripts



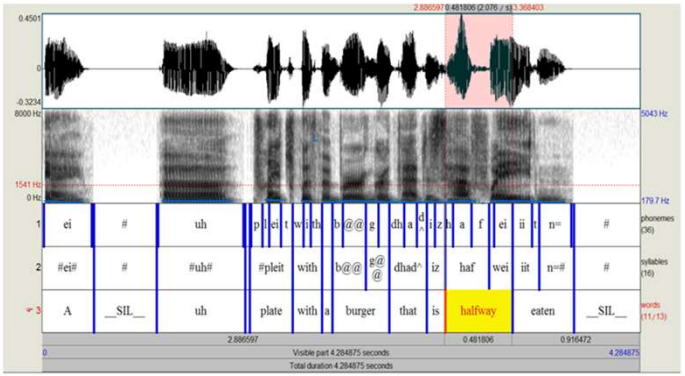
táá ya bísí adí la kándzá lá iné || notre père a une grande paillotte
letsɛngɛ líibhámá l' ebhóyá || le sol durcit pendant la saison sèche

+ WAV

Datasets

- Augmentation of **Mscoco** with (Voxygen) TTS Speech
  - W. Havard, L. Besacier & O. Rosec, *"SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO"*, GLU Workshop of Interspeech, Stockholm, Sweden, 2017
  - https://persyval-platform.imag.fr/perscido/web/DS80/detaildataset
- Disfluencies and speed perturbation to add variability
- WAV paired with JSON containing timecode of each word/syllable/phoneme
- We added Japanese translations (Google MT) for all captions (+ tokenization)

+



A plate with a burger that is halfway eaten

中途半端に食べられたハンバーガー付きプレート。

Tools

I. XMNT: A general architecture for end-to-end experiments

Research questions

I. Using out-of-domain languages to help unit discovery: *(almost) zero-shot adaptation*
II. Synergies between unit discovery and end-to-end tasks: *where are symbols coming from and what are they good for?*
III. Using unit discovery to construct speech synthetizers: *TTS without T*
IV. Using multimodal information to improve standard ASR: *this can be useful even for 'rich' languages*

Coming up next…

I. XMNT: a general architecture for end-to-end experiments

Starring (in order of appearance): Graham Neubig, Danny Merkx, Liming Wang, and Mingxing Du

Overview of tools

ZR (and others)

XNMT

オープンフィールドの
素敵な木

text enc

text dec

image enc

alignment/attention model

word enc

phone dec

speech enc

Word segmentation

Adaptation

AF+clus

VBHMM

features

Eval (ABX, etc)

TTS

meaning

word

Phonetic

Work in Grounded Speech

(Your host for the next 10 minutes: Graham Neubig)

An Attentional Model for Speech Translation Without Transcription

Long Duong,[12] **Antonios Anastasopoulos,**[3] **David Chiang,**[3] **Steven Bird**[14] and **Trevor Cohn**[1]
[1]Department of Computing and Information Systems, University of Melbourne

Pyramidal LSTM!

ConvNet!

Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Ron J. Weiss[1], *Jan Chorowski*[1], *Navdeep Jaitly*[2*], *Yonghui Wu*[1], *Zhifeng Chen*[1]
[1]Google Brain

Unsupervised Learning of Spoken Language with Visual Context

Different ConvNet!

David Harwath, Antonio Torralba, and James R. Glass
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Representations of language in a model of visually grounded speech signal

Recurrent Highway Network!

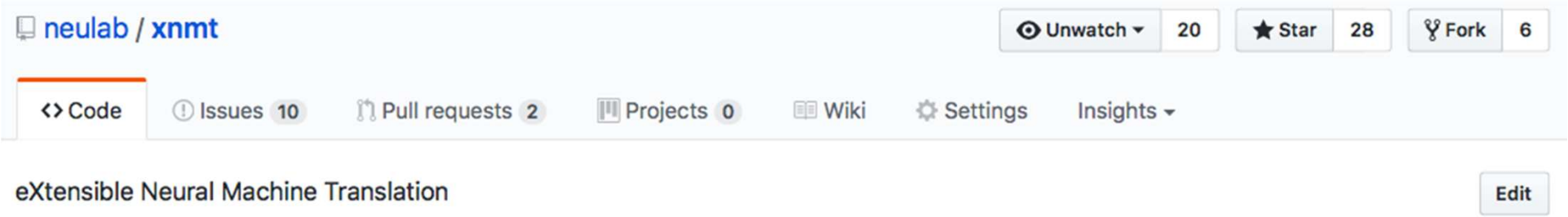
Grzegorz Chrupała
Tilburg University

Lieke Gelderloos
Tilburg University

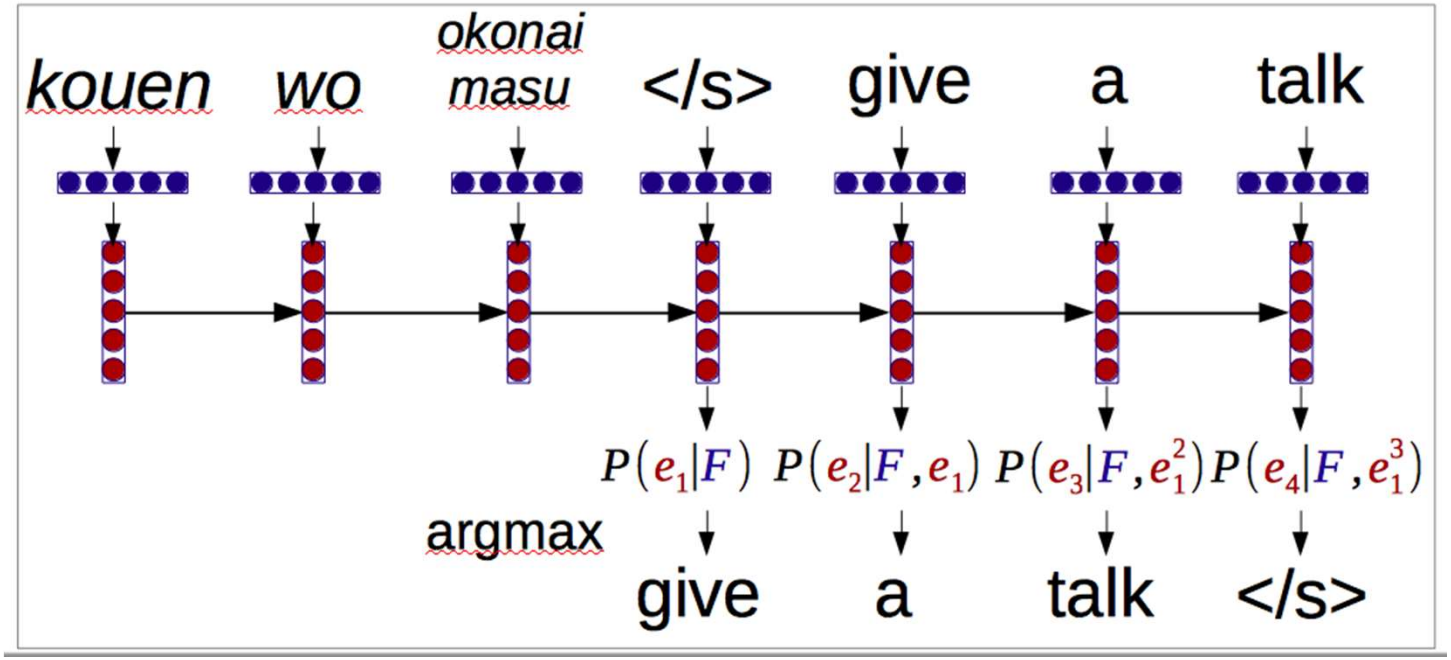
Afra Alishahi
Tilburg University

eXtensible Neural Machine Translation (XNMT)

- A **flexible platform for experimental design** in sequence-to-sequence models
- Designed to make it easy to:
  - **Compare models** with state-of-the-art components
  - **Implement new ideas**
  - Specify and **run experiments**
- http://github.com/neulab/xnmt



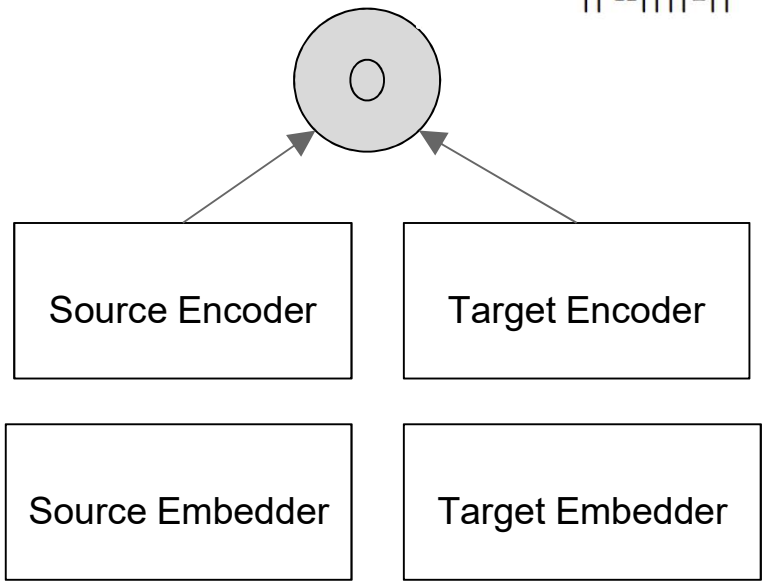
Translation Models



- Train w/ maximum likelihood of output sentence

Retrieval Models

Similarity: $S_{ij} = \frac{\mathbf{u}^T \mathbf{i}}{||\mathbf{u}||||\mathbf{i}||}$



$$\sum_{u,i} \left(\sum_{u'} \max[0, \alpha + d(u,i) - d(u',i)] + \sum_{i'} \max[0, \alpha + d(u,i) - d(u,i')] \right)$$

- Train w/ max margin objective and negative sampling

XNMT Model Consists Of

- Task (Translation, Retrieval)
- Input Type (Text, Speech, Image)
- Output Type (Text, Speech, Image)
- Input-side Encoder (LSTM, CNN, Pyramidal, Highway Network, Linear)
- Output-side:
  - Decoder (for translation)
  - Encoder (for Retrieval; LSTM, CNN, Pyramidal, Highway Network, Linear)
- Evaluation Metric (BLEU, WER, TER, Recall)

Components are pluggable!
Different models are just different components!

Speech-to-text Translation (Duong+15)

- Task (**Translation**, Retrieval)
- Input Type (Text, **Speech**, Image)
- Output Type (**Text**, Speech, Image)
- Input-side Encoder (LSTM, CNN, **Pyramidal**, Highway Network, Linear)
- Output-side:
  - **Decoder (for translation)**
  - Encoder (for retrieval; LSTM, CNN, Pyramidal, Highway Network, Linear)
- Evaluation Metric (**BLEU**, WER, TER, Recall)

Speech-to-image Retrieval (Harwath+16)

- Task (Translation, **Retrieval**)
- Input Type (Text, **Speech**, Image)
- Output Type (Text, Speech, **Image**)
- Input-side Encoder (LSTM, CNN, **Pyramidal**, Highway Network, Linear)
- Output-side:
    - Decoder (for translation)
    - Encoder (for retrieval; LSTM, CNN, Pyramidal, Highway Network, **Linear**)
- Evaluation Metric (BLEU, WER, TER, **Recall**)

How to Specify Models?

- Flexible and easily extensible YAML format configuration file

```yaml
train:
  default_layer_dim: 512
  restart_trainer: True
  trainer: Adam
  learning_rate: 0.0002
  lr_decay: 0.5
  dev_metrics: bleu
  training_corpus: !BilingualTrainingCorpus
    train_src: examples/data/train.ja
    train_trg: examples/data/train.en
    dev_src: examples/data/dev.ja
    dev_trg: examples/data/dev.en
  corpus_parser: !BilingualCorpusParser
    src_reader: !PlainTextReader {}
    trg_reader: !PlainTextReader {}
  model: !DefaultTranslator
    src_embedder: !SimpleWordEmbedder
      emb_dim: 512
    encoder: !LSTMEncoder
      layers: 1
    attender: !StandardAttender
      hidden_dim: 512
      state_dim: 512
      input_dim: 512
    trg_embedder: !SimpleWordEmbedder
      emb_dim: 512
    decoder: !MlpSoftmaxDecoder
      layers: 1
      mlp_hidden_dim: 512
```

Result of the JSALT Workshop

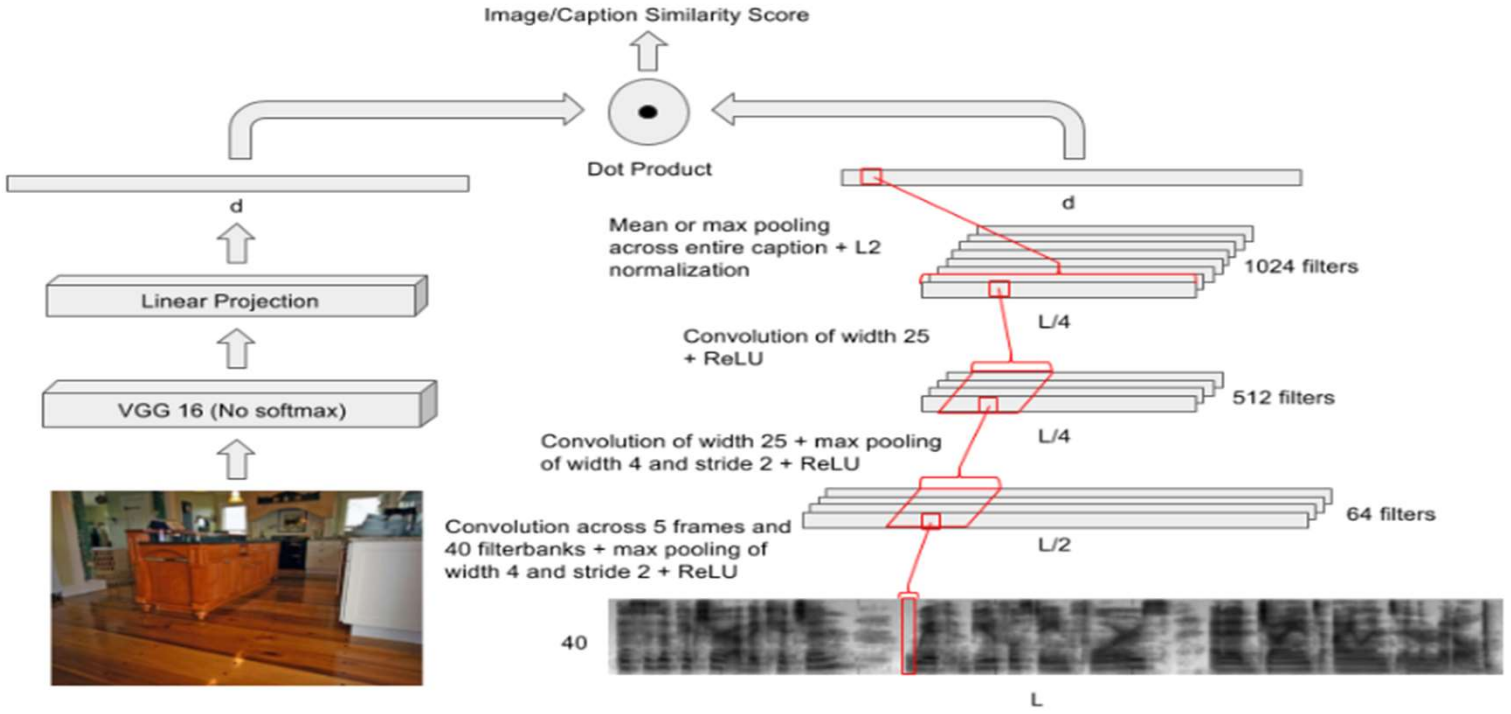
(Graham Neubig)

- Exhaustive YAML configuration file format
- Retrieval functionality
- HTML/data dump reporting functionality
- **Speech-oriented encoders**

The “Harwath Encoder:” CNN for Speech

(Your host for the next 5 minutes: Liming Wang)

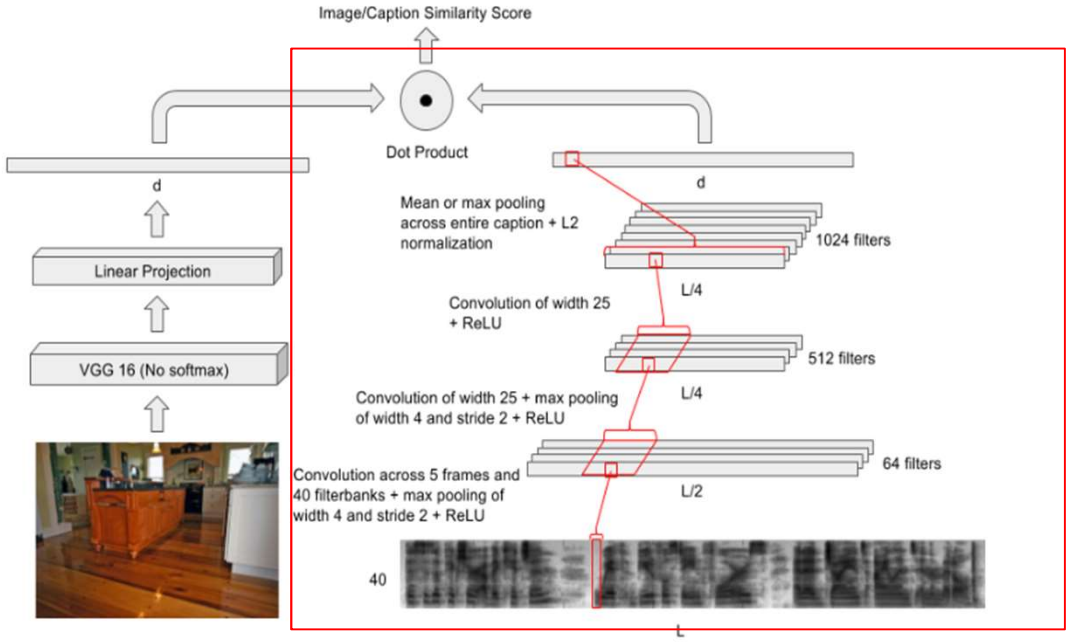
- Original model of Harwath and Glass 2016 NIPS paper
- Goal: use image context to extract better speech unit (phones, words, phrases, etc.)



Speech-to-Image Retrieval with CNN

(Liming Wang)

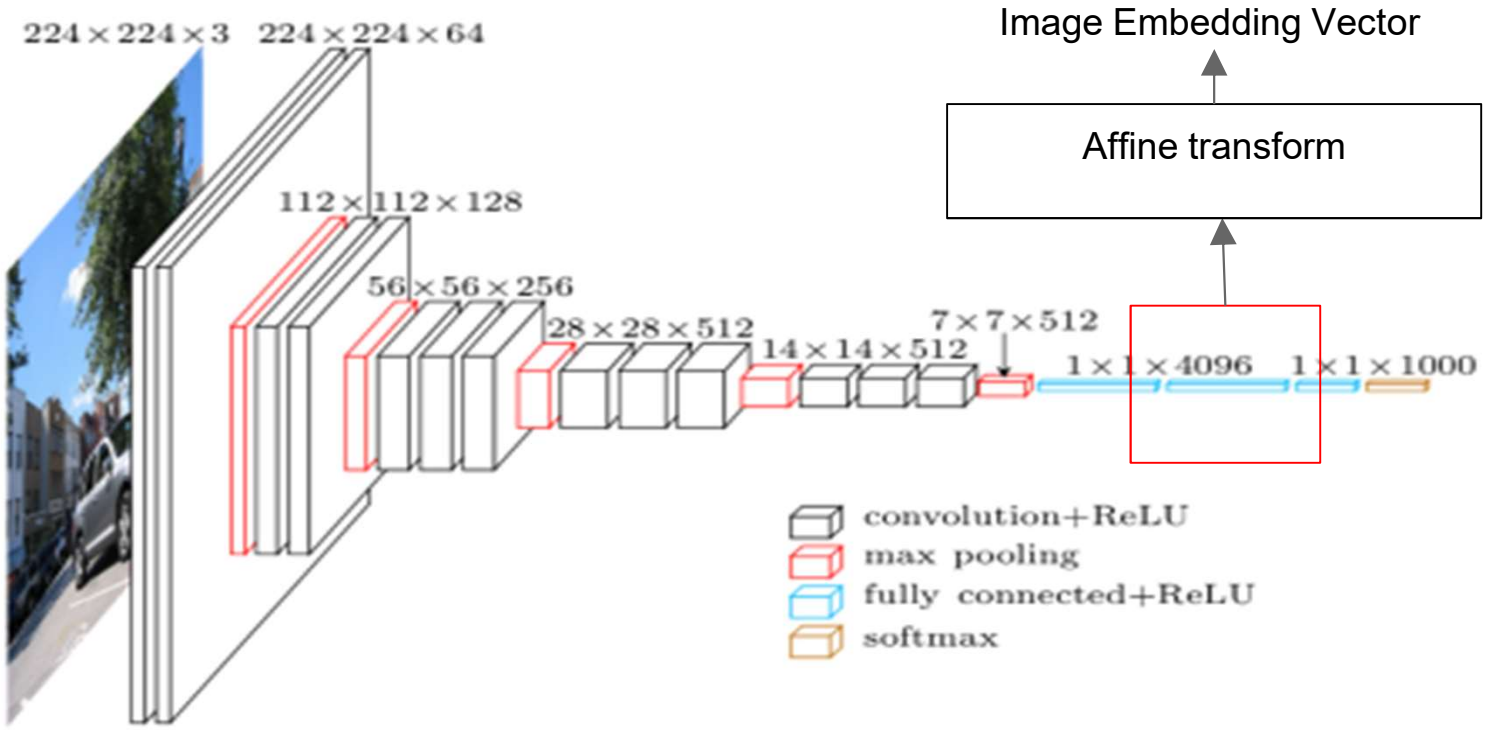
- Harwarth speech encoder: CNN



Speech-to-Image Retrieval with CNN

(Liming Wang)

- Harwath Image Encoder: Pretrained VGG 16 feature + Affine transform



Replicating Harwath and Glass 2016

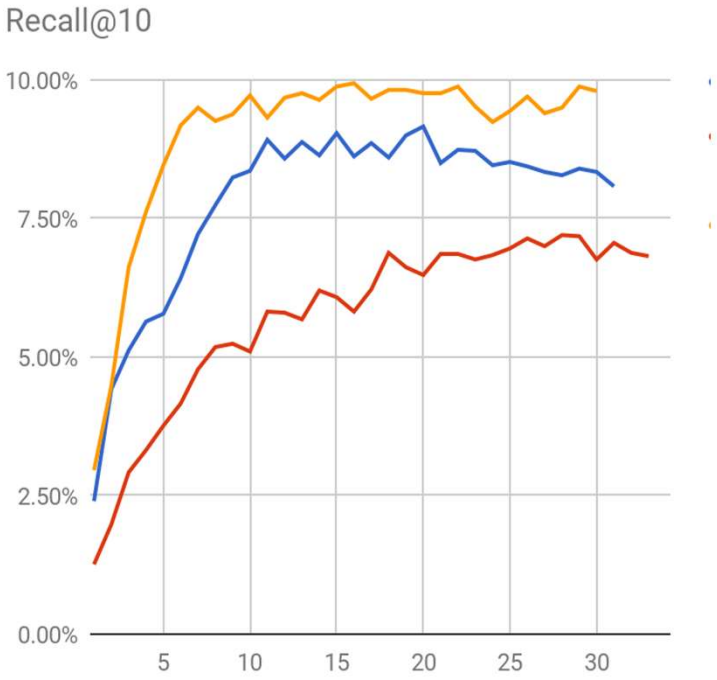
(Your host for the next 5 minutes: Danny Merkx)

- Speech input: mean normalised filterbank spectrograms
- Issues: - flickr is a smaller dataset than the one used by H&G
- Padding and truncation
- naive normalisation
- Dynet can accept variable length input
- Mean and variance normalisation using forced alignments
- Compare MFCCs and Fbanks

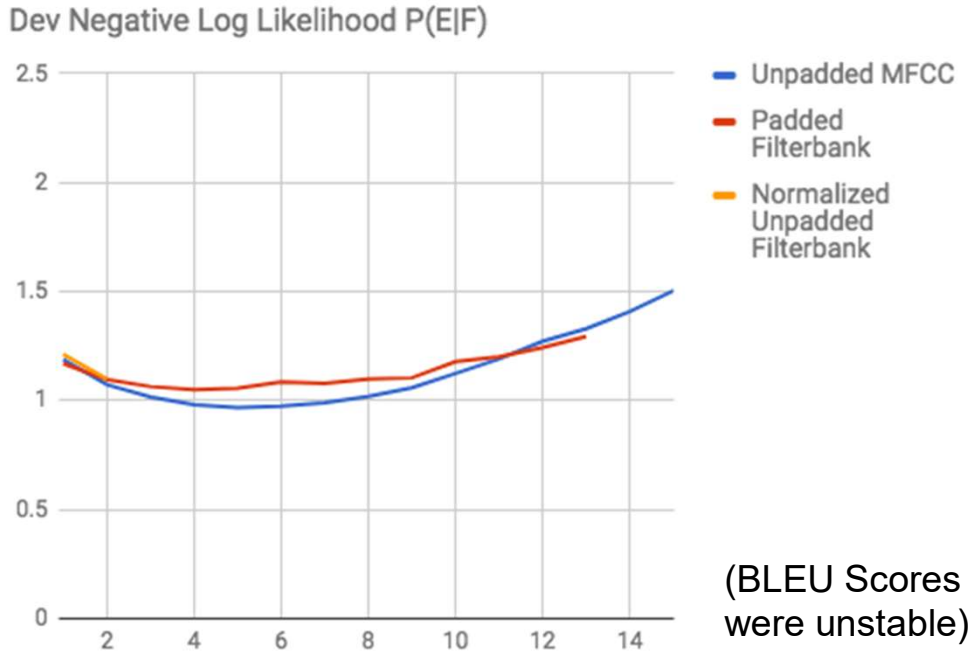
Effect of Features on Retrieval

(Danny Merkx)

Retrieval



Translation

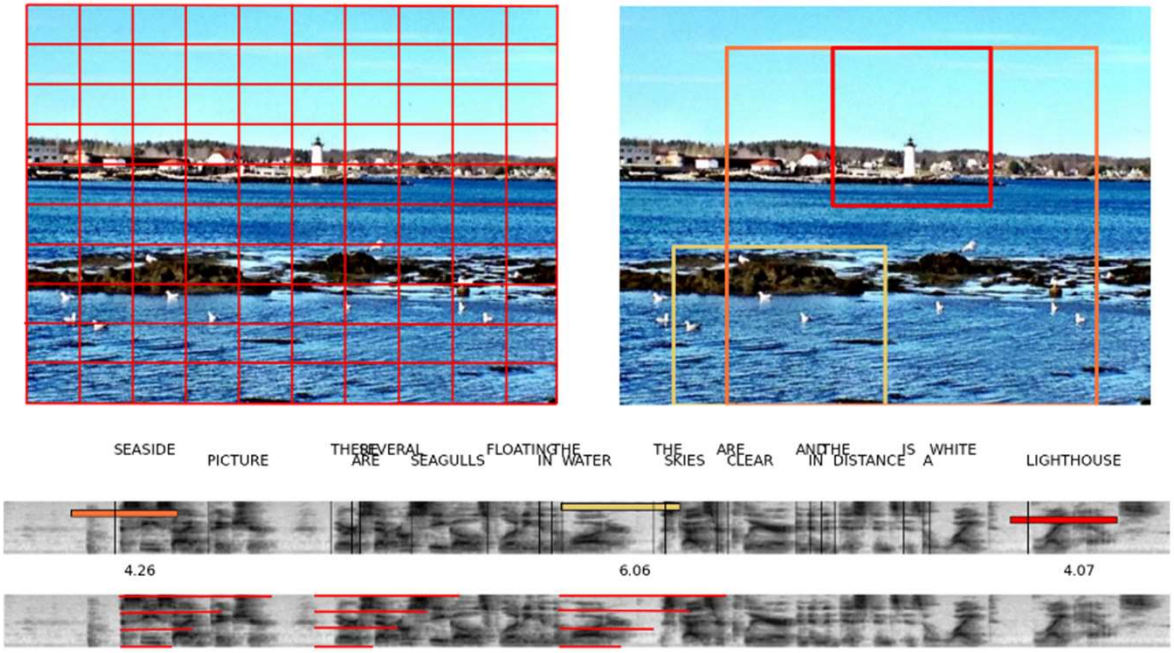


(BLEU Scores were unstable)

Segmented speech2image

(Danny Merkx)

- goal: encode speech and image segments and align them.
- use a model trained on full images and speech to embed segments



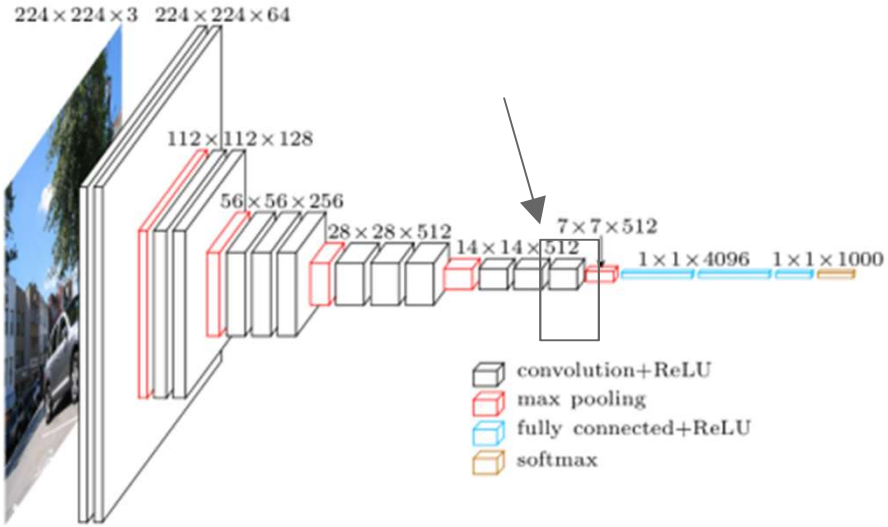
Segmented speech2image

(Danny Merkx)

- cluster the new embeddings and get affinity scores for the speech and image clusters

$$\text{Affinity}(\mathcal{I}, \mathcal{A}) = \sum_{\mathbf{i} \in \mathcal{I}} \sum_{\mathbf{a} \in \mathcal{A}} \mathbf{i}^\top \mathbf{a} \cdot \text{Pair}(\mathbf{i}, \mathbf{a})$$

- new idea: use vgg features directly instead of segmenting

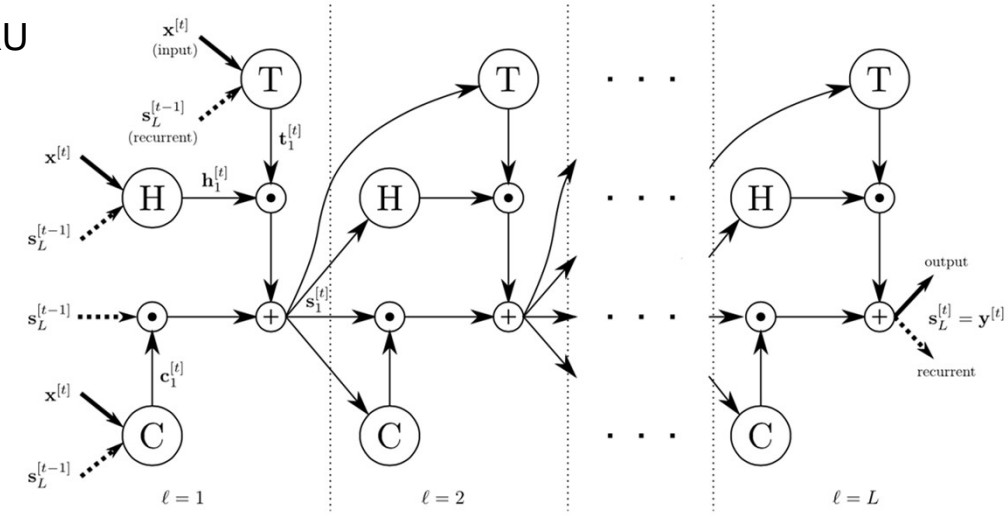


Beyond CNN

(Your host for the next 5 minutes: Mingxing Du)

Recurrent Highway network

- A generalization of recurrent neural network.
- Reported to outperforme LSTM and GRU
- Had baseline we can compare

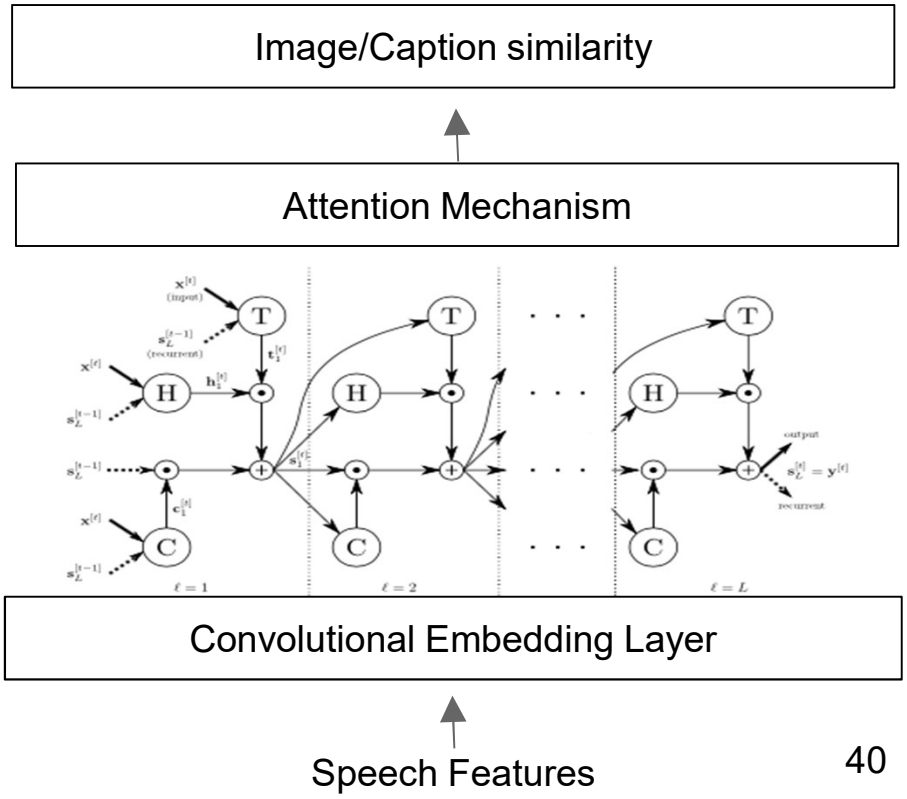


One example of recurrent highway layer

Recurrent Highway Speech Encoder (Mingxing Du)

Model by Chrupala et al. 2017

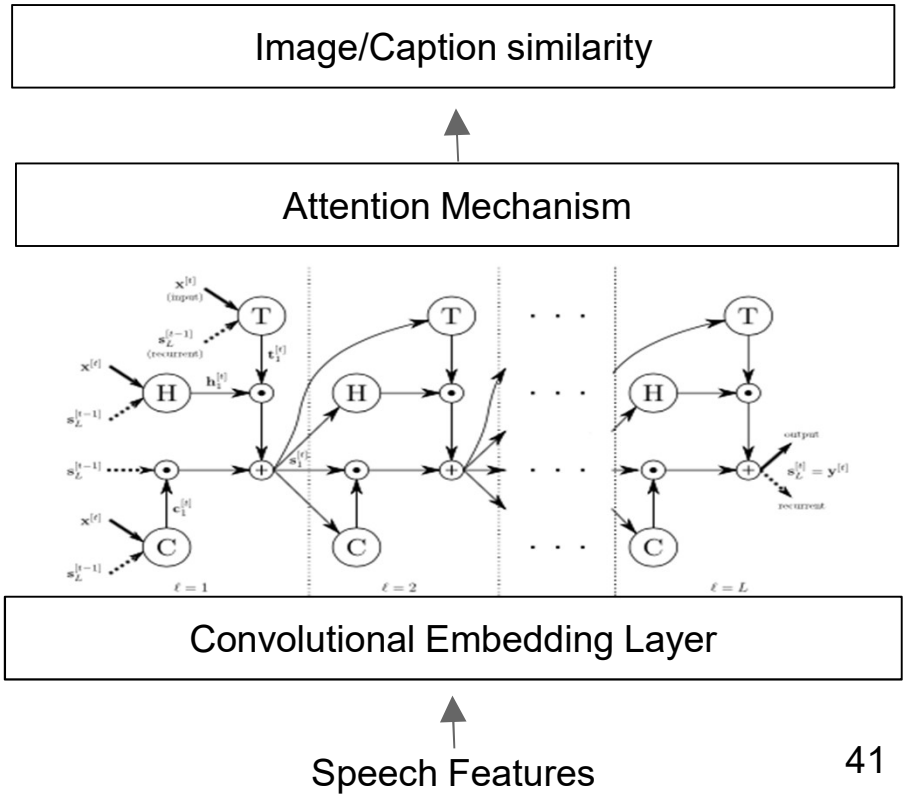
1. Main component: Recurrent Highway Network by Zilly et al. 2015
2. Input embedding: convolutional layer
3. Simplified attention mechanism
4. Residual operation in each of hidden RHN layers



Recurrent Highway Speech Encoder (Mingxing Du)

$$\text{Output} = \sum_t \alpha_t \mathbf{x}_t$$

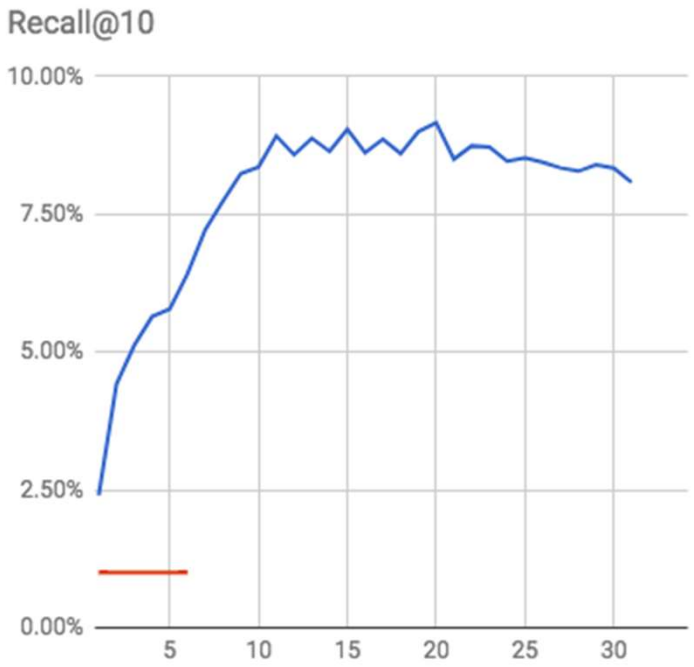
$$\alpha_t = \frac{\exp(\mathbf{U}\tanh(\mathbf{W}\mathbf{x}_t))}{\sum_{t'}\exp(\mathbf{U}\tanh(\mathbf{W}\mathbf{x}_{t'}))}$$



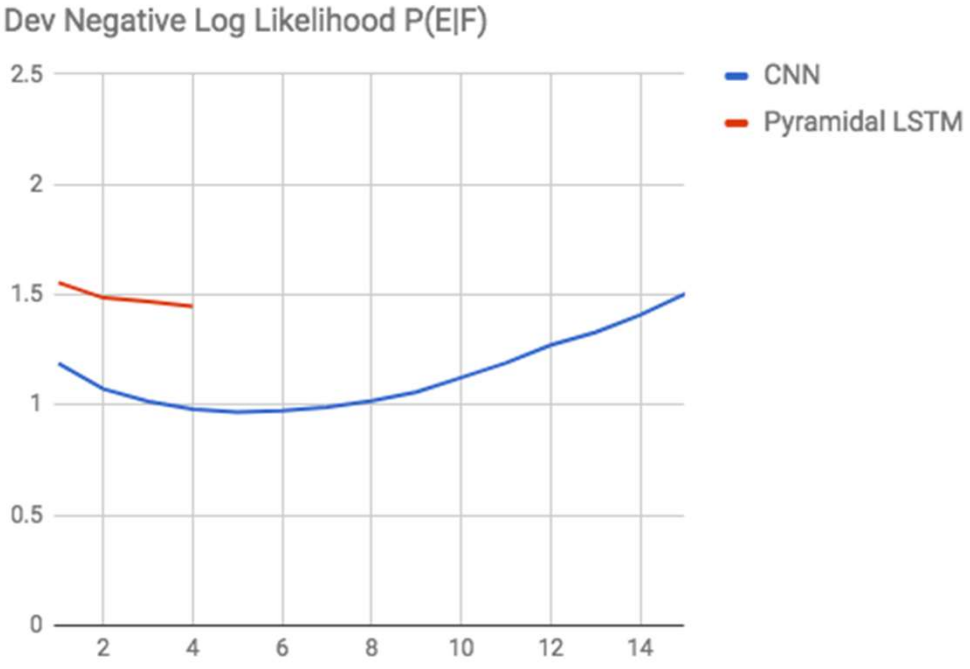
Experiment Results

(Liming Wang & Mingxing Du)

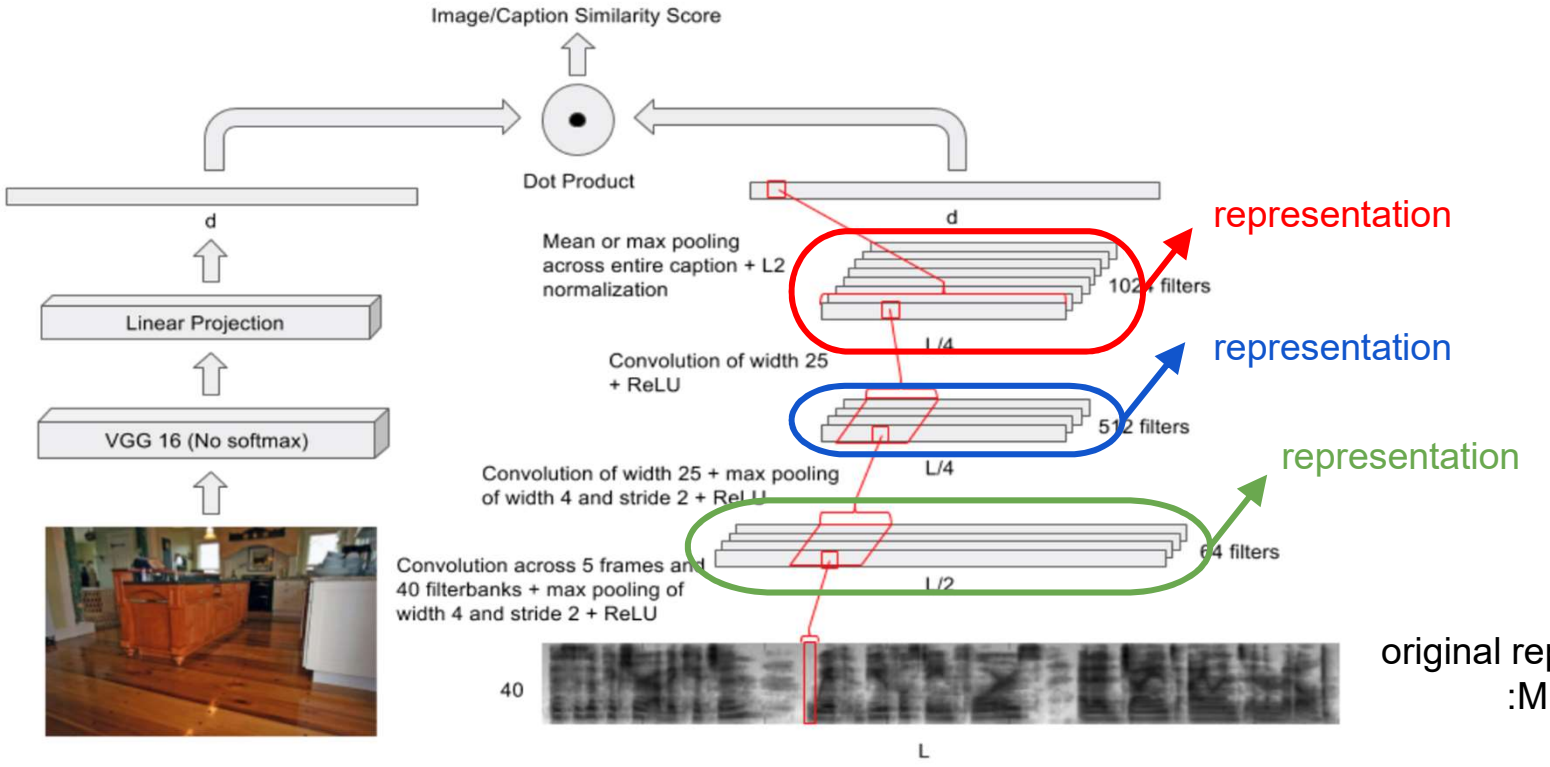
Retrieval



Translation



Hidden states as speech representation



Coming up next…

II. Discovering acoustic units in an unwritten language using (almost) zero-shot adaptation

Starring (in order of appearance): Lucas Ondel, Markus Müller, Odette Scharenborg, Francesco Ciannella

II. Discovering acoustic units in an unwritten language using (almost) zero-shot adaptation

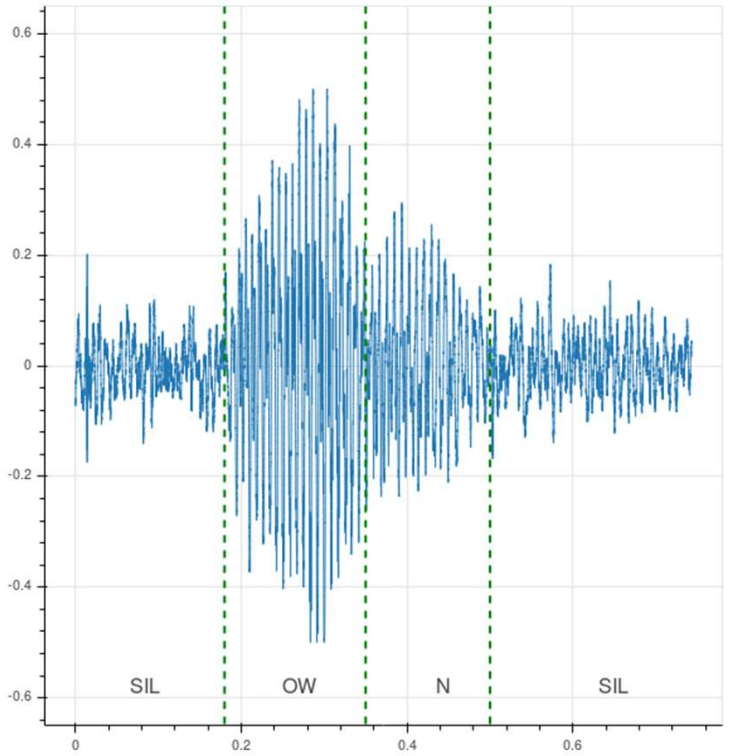
オープンフィールドの
素敵な木

II.A - Using prior information for AUD

(your host for the next 10 minutes: Lucas Ondel)

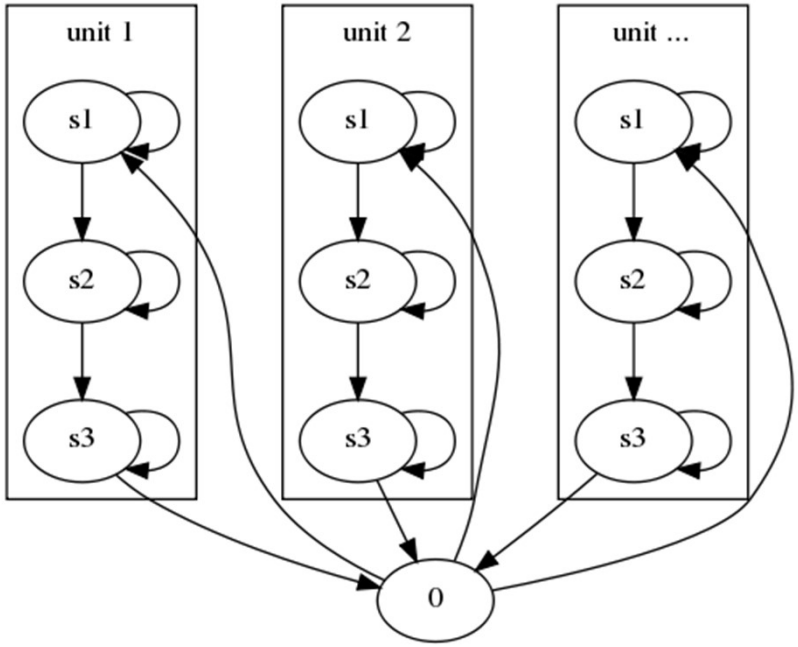
AUD in a nutshell:

- Find the segmentation of the AUD
- Cluster the segments into pseudo-phones
- Find the appropriate number of pseudo-phones needed to describe the target data.



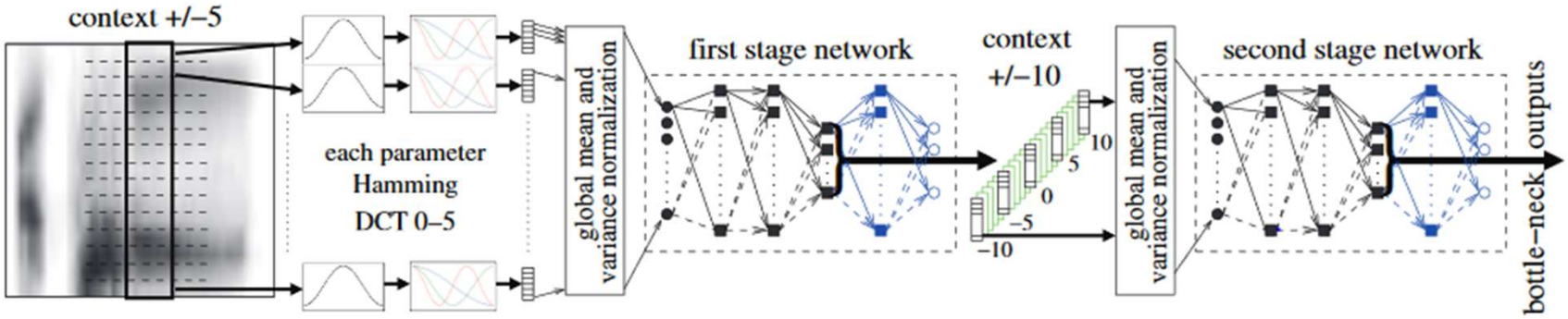
Prior work: Bayesian Phone-Loop

- Similar to standard ASR HMM but trained without supervision
- 3-states HMM to model the duration of the pseudo-phones
- GMM emissions to model the trajectory of the pseudo-phones
- (Non-informative) Prior over the parameters to regularize the training



Empirical Prior

MFCC (and similar) features bare too much variability. We can use well resourced languages to develop “universal” features:



Bayesian Informative Prior I

Standard Bayesian Inference

$$p(\boldsymbol{\eta} \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \boldsymbol{\eta})\, p(\boldsymbol{\eta})}{p(\mathbf{X})}$$

Non-informative prior

Bayesian Informative Prior II

New idea: use the prior to inject information into the AUD algorithm

$$p(\boldsymbol{\eta} \mid \mathbf{X}^l) = \frac{p(\mathbf{X}^l \mid \boldsymbol{\eta}) p(\boldsymbol{\eta})}{p(\mathbf{X}^l)}$$

Train the "prior" on labeled data

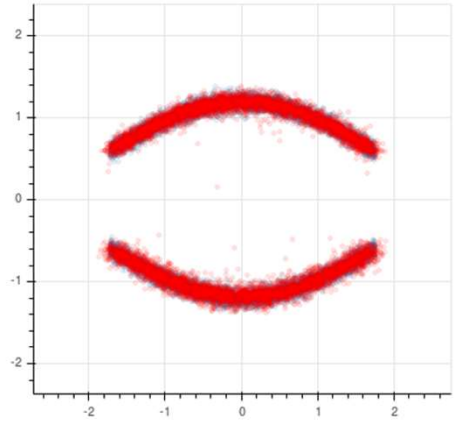
$$p(\boldsymbol{\eta} \mid \mathbf{X}^u) = \frac{p(\mathbf{X}^u \mid \boldsymbol{\eta}) p(\boldsymbol{\eta} \mid \mathbf{X}^l)}{p(\mathbf{X}^u)}$$

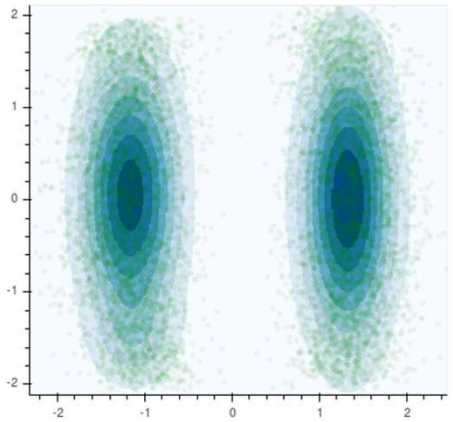
Bayesian inference with informative prior

Bayesian Informative Prior III

The more complex our model is the more information is embedded into the “prior”:

$$p(\boldsymbol{\eta} \mid \mathbf{X}^l) = \frac{p(\mathbf{X}^l \mid \boldsymbol{\eta}) p(\boldsymbol{\eta})}{p(\mathbf{X}^l)}$$





- Train the “prior” on labeled data using a Structured Variational AutoEncoder
- Unsupervised retraining on the target (MBOSHI) language

Results

Word Boundary Accuracy on MBOSHI and TIMIT as prior database

Inputs	Precision	Recall	Fscore
true phones	30.4	100.0	46.6
AUD (MFCC)	23.9.	81.0	36.9
AUD (MBN)	25.2	80.2	38.2
AUD informative prior (MBN) no adaption	26.0	68.2	37.6
AUD informative prior (MBN)	28.0	72.3	40.4
SVAE AUD informative prior (MBN)	**28.7**	**74.2**	**41.4**

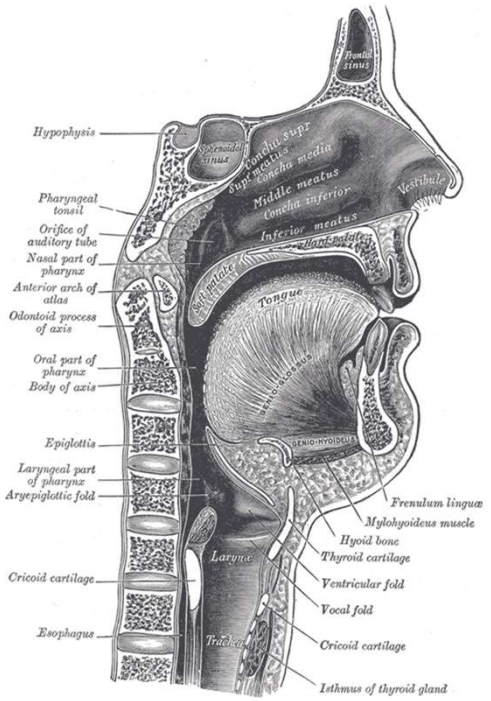
Conclusion & Future work

- We have shown that unsupervised learning of speech can be improved by using information extracted from well resourced languages
  - Multilingual Bottleneck Features
  - Informative Bayesian Prior
- Behind the proof of concept:
  - Informative Bayesian Prior build on top of several languages
  - Improving the model to embed more information
- More to be about the units by Laurent & co.

II.B Universal articulatory features

(your host for the next 10 minutes: Markus Müller)

- Each language has its own phone inventory
- Total number of phones produced by humans is limited
  - Anatomy of human vocal tract
- Phone: Bundle of articulatory features (AF)
- AFs describe targets of articulators in vocal tract
  - Atomic units to describe human articulation



(from Gray's Anatomy 1918 edition)

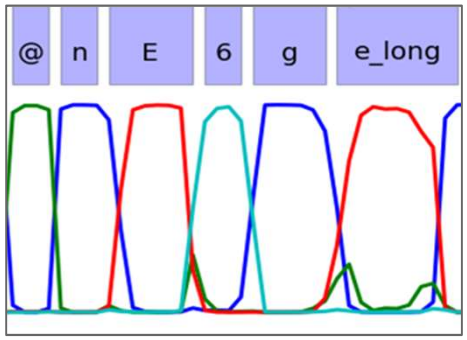
Unsupervised Phone Discovery (UPD)

3 step method

- Step 1: Segmentation
  - Use BLSTM-based approach to detect phone boundaries
- Step 2: AF detection
  - Automatic recognition of AFs for each segment
- Step 3: Clustering
  - Cluster segments into a phone inventory based on AFs

Articulatory features based UPD

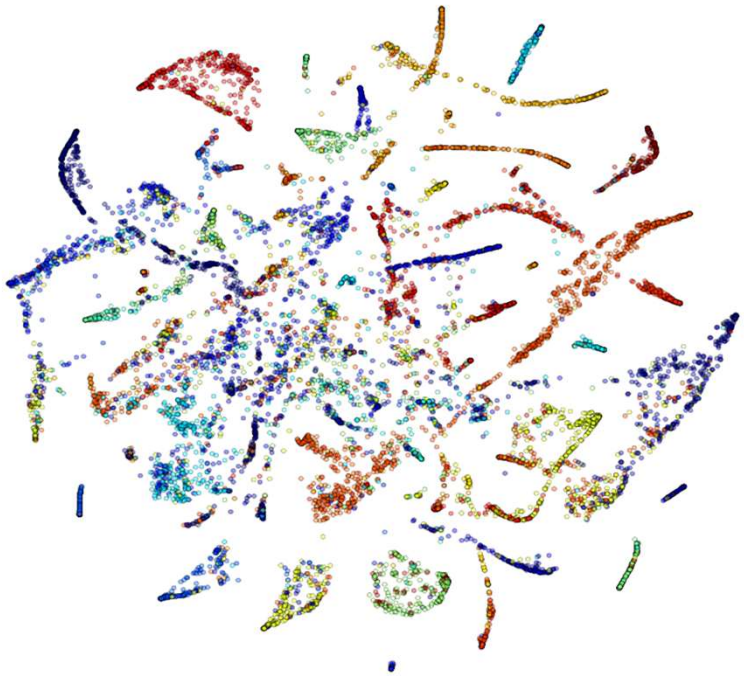
- Trained multilingual on Euronews corpus (TV broadcast news)
  - Using data from French, German, Turkish
- Mapped phone labels to AFs
  - Used only inner third of frames for each phone
  - Less coarticulation artifacts, articulators more stable
- 7 AF types (3 for consonants, 4 for vowels)
- Design decisions
  - Clustering algorithm / method (e.g. k-Means)
  - Algorithm parameters (e.g. number of classes for k-Means)
- Number of classes (phones) is unknown

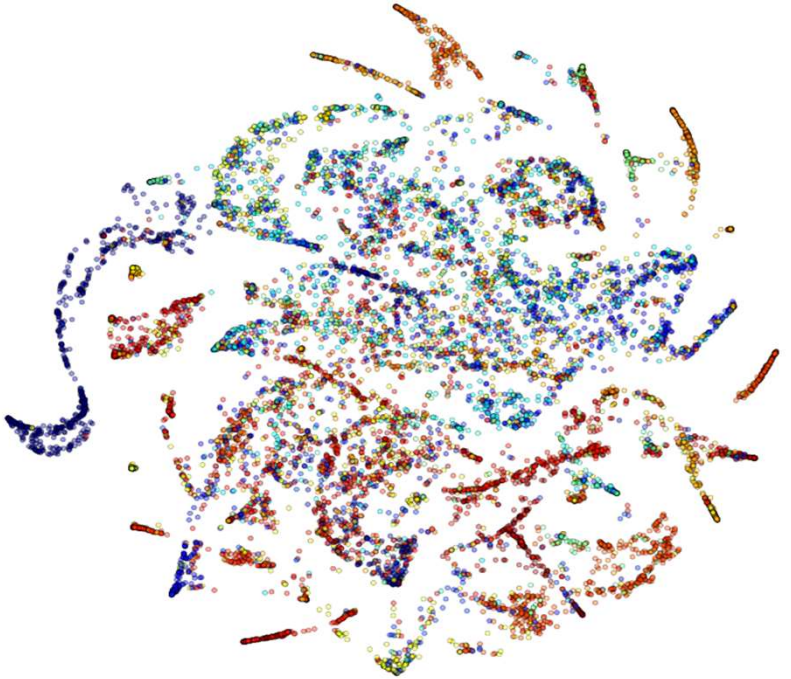


Example of detected AF

Visualizing extracted AFs

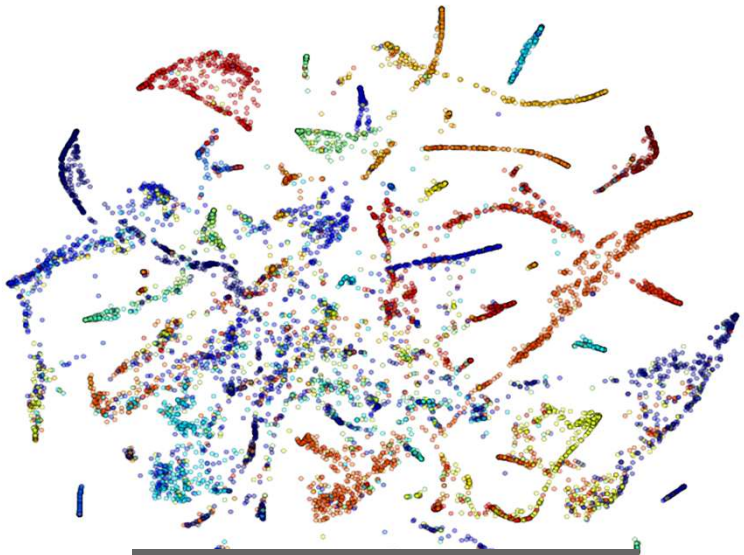
- t-distributed Stochastic Neighbor Embedding, colored by phone identity



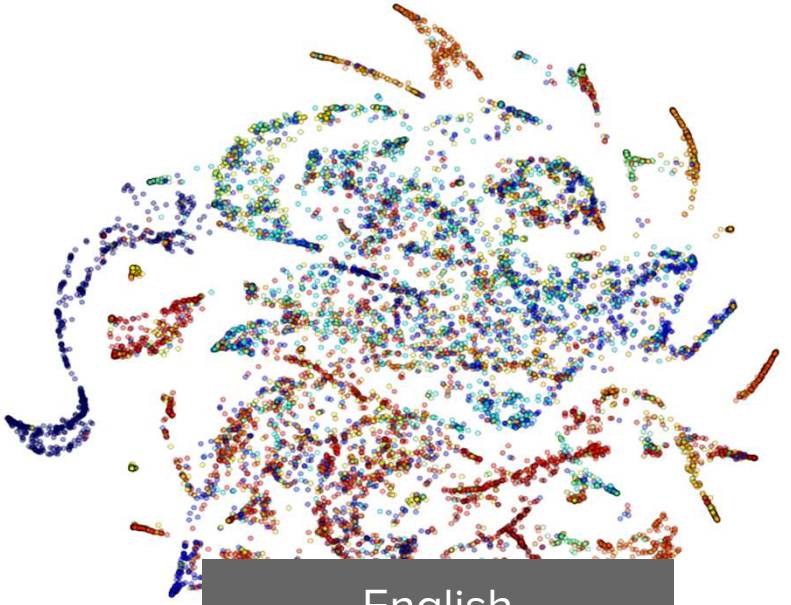


Visualizing extracted AFs

- t-distributed Stochastic Neighbor Embedding, colored by phone identity



German
(seen during training)



English
(unseen during training)

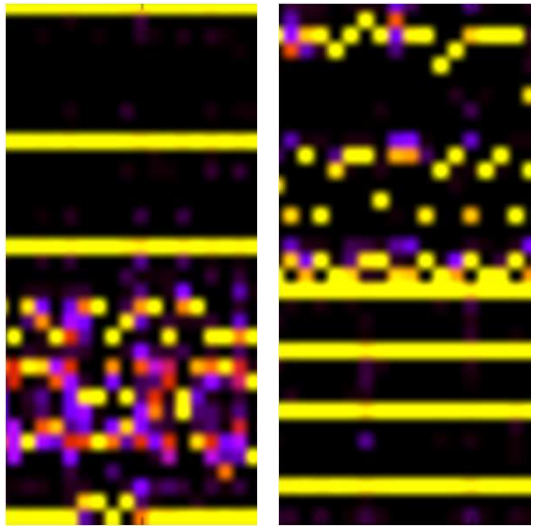
Clustering segments based on AFs

- 2 major classes
  - **Vowels**
  - **Consonants**
  - (Noises)
- Separate major classes first
- Cluster vowels / consonants separately

Language	Precision	Recall	Fscore
German Vwl	0.88	0.94	0.91
German Cns	0.88	0.96	0.92
English Vwl	0.57	0.91	0.70
English Cns	0.85	0.89	0.87

Example AF vectors

Vowels Consonants



UPD: Unsupervised evaluation

- Evaluation of derived units for Mboshi using MCD score
- Unsupervised measure to evaluate whole pipeline
  - Build TTS system based on discovered units
  - Compute Mean Cepstral Distortion (MCD) Score

Method	MCD Score
Phone forced align.	5.25
Phone forced align. + text UWD	5.26
UPD	5.78
UPD + textUWD	5.72

Concluding remarks

- Improve crosslingual performance of AF detectors
  - Include more languages to cover wider variety of sounds
- Evaluate additional clustering methods
  - e.g. Dirichlet process mixture models
  - Clustering based on mapping AFs to IPA symbols
- Use recognized AFs on top of given clustering to determine phone identity

II.C - Cross-language definition of units

(your host for the next 10 minutes: Odette Scharenborg & Francesco Ciannella)

Background:

- Same-language ASR is pretty good
- Cross-language ASR is pretty bad
- Maybe we can improve cross-language ASR through adaptation?

Aim: build an ASR system for a low-resource language through cross-language adaptation of an ASR system of a high-resource language

Assumptions

- We have a ‘description’ of the phone(me) inventory of the low-resource language
- We have some unlabeled speech data in the low-resource language
- We know the language family
- We have an ASR system trained on a well-resourced language of the same language family, although.....

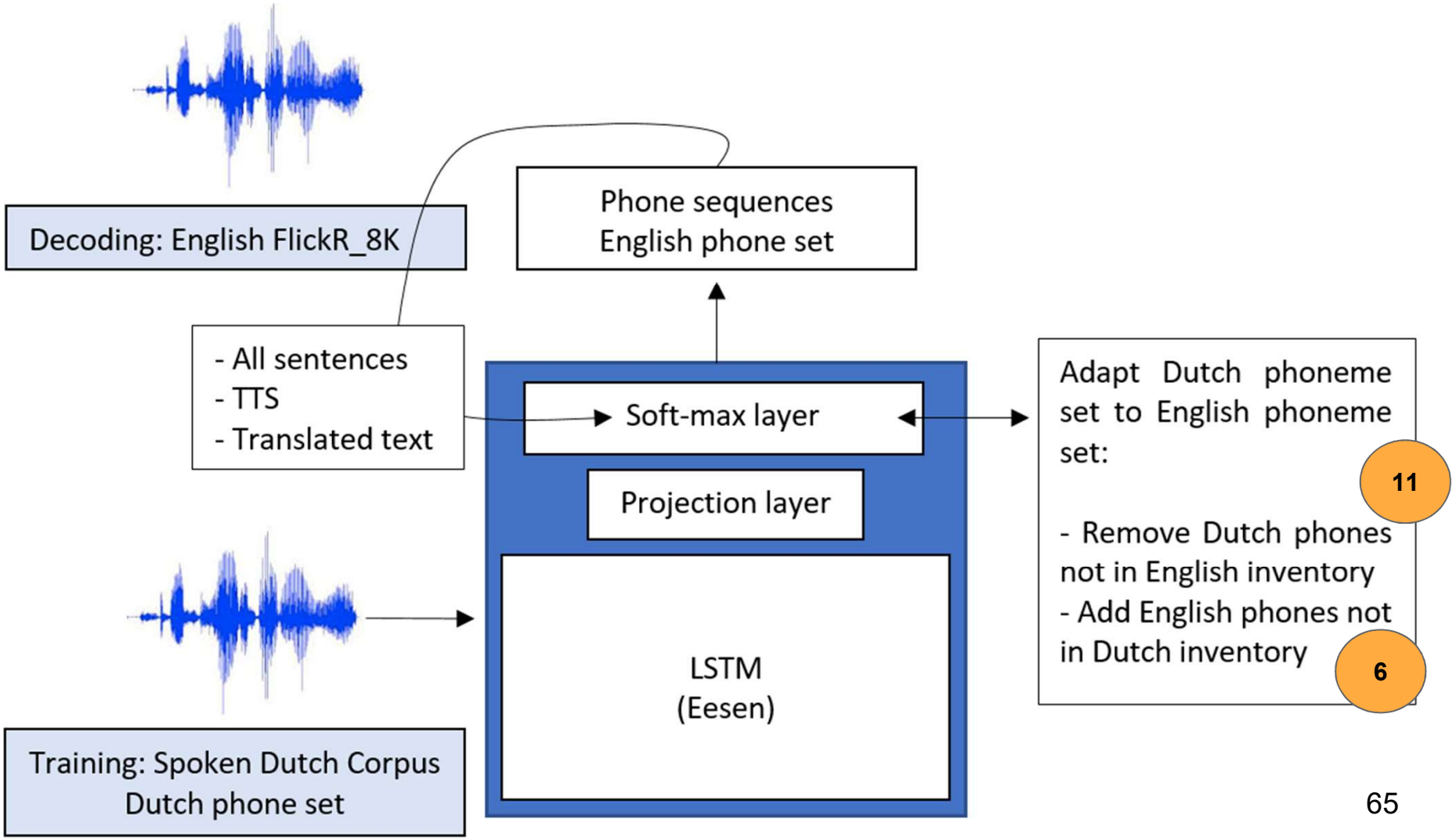
Languages

For now:

- Low-resource language: 3660 English sentences from FlickR_8K
- High-resource language: Dutch

(Near) Future:

- Low-resource language: Mboshi
- High-resource language: Dutch

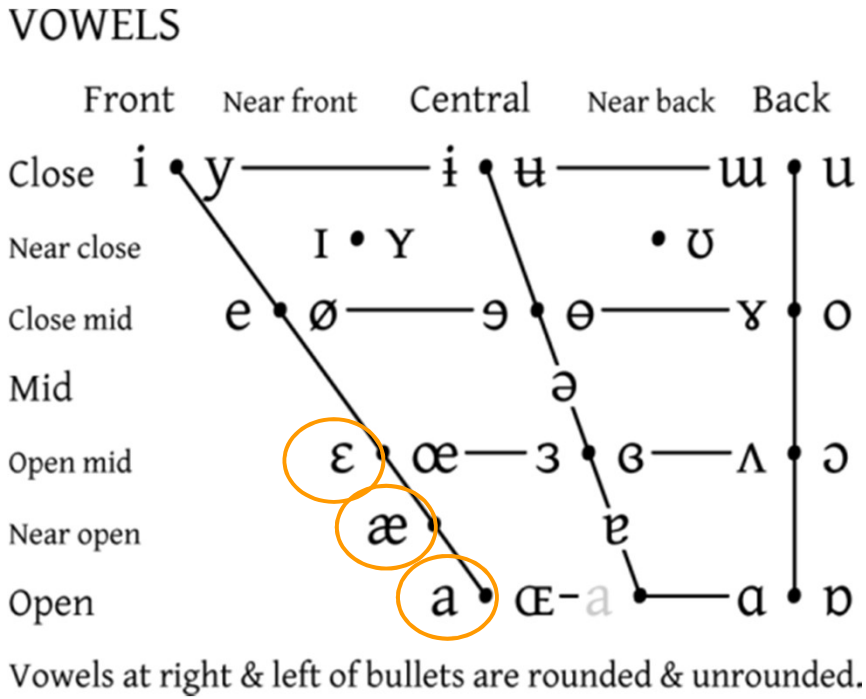

Decoding: English FlickR_8K
Phone sequences
English phone set
- All sentences
- TTS
- Translated text
Soft-max layer
Projection layer
LSTM
(Eesen)
Adapt Dutch phoneme set to English phoneme set:
11
- Remove Dutch phones not in English inventory
- Add English phones not in Dutch inventory
6
Training: Spoken Dutch Corpus
Dutch phone set

Creating new English phones

$$\vec{V}_{|\varphi|,L2} = \vec{V}_{|\varphi|,L1:1} + 0.5\,(\vec{V}_{|\varphi|,L1:2} - \vec{V}_{|\varphi|,L1:3})$$

Adapting the softmax layer through extrapolation using the above formula

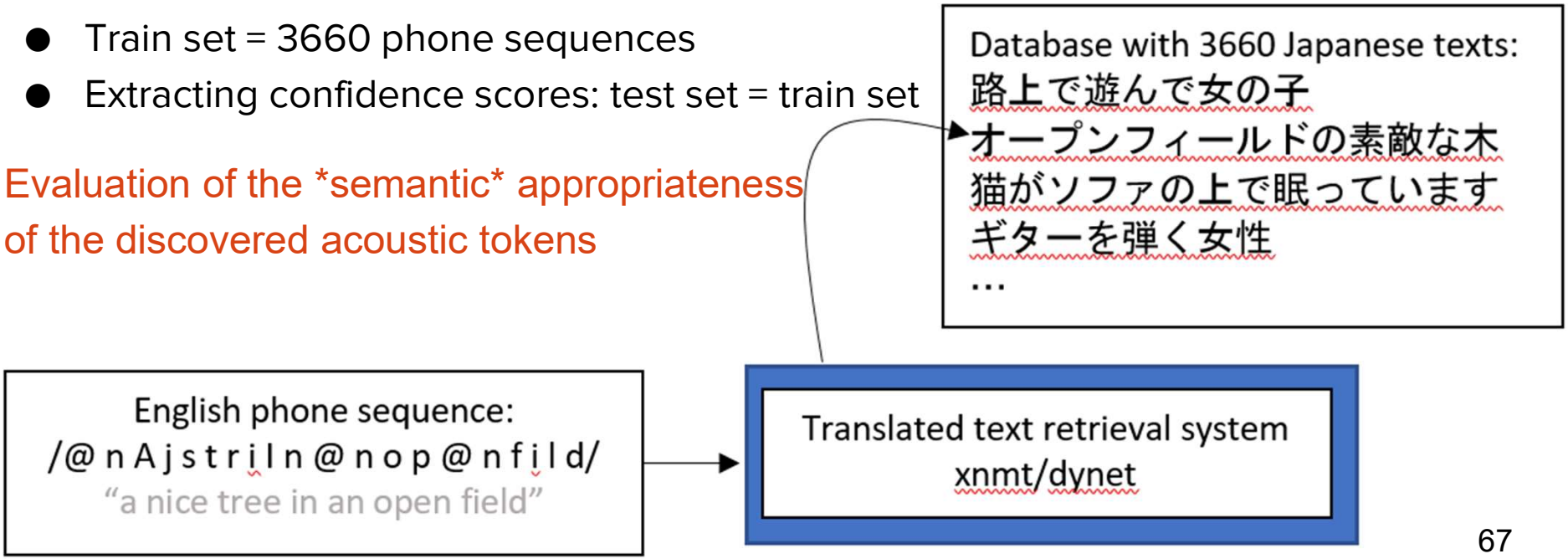
Extrapolation from existing Dutch phones: e.g. /ae/ = halfway between /E/ and /a/



Getting confidence scores from the translation retrieval task

- Train set = 3660 phone sequences
- Extracting confidence scores: test set = train set

Evaluation of the *semantic* appropriateness of the discovered acoustic tokens



Results: Translated Text Retrieval (%)

- Baseline system:
  - Train set: 3660 gold standard phone transcriptions
  - Test set == train set

	Recall@1	Recall@5	Recall@10
Gold standard Baseline - original phone set	23.80	44.81	55.33
Gold standard Baseline - adapted phone set	10.41	21.23	27.16
Dutch-based system + adaptation, Iteration 0	23.69	44.29	54.26
+ Retraining on all, Iteration 1	9.95	21.37	28.14

Results: Token Error Rates (%)

- Test == train set 3660 FlickR

	Retraining (TER)	Number of utts used in retraining
Dutch-based system + adaptation, Iteration 0	72.6%	n/a
+ Retraining, Iteration 1 - X% best TTS	71.7%	1831/3660
+ Retraining, Iteration 1 - X% best Translation Retrieval	71.8%	1986/3660
+ Retraining, Iteration 1 - X% best ASR	71.8%	2263/3660
+ Retraining, Iteration 1 - All	71.7%	3660

Concluding remarks

- Acoustic, phoneme speech tokens can be discovered through adaptation
- Tokens seem to be ‘semantically’ meaningful

Future work

- Evaluation on independent sets
- Visualisation of the DNN hidden layers:
  a. Which phones were correctly learned?
  b. Comparison of the phone space in the soft-max layers
- Adaptation to a real low-resource language
- ….

Coffee and tea break

Back in 9.99 minutes....

Coming up next…

III. Synergies between unit discovery and end2end tasks

Starring (in order of appearance): Rachid Riad, Pierre Godard, Laurent Besacier, Philip Arthur

III. Synergies between cross-modal End2End and unit discovery

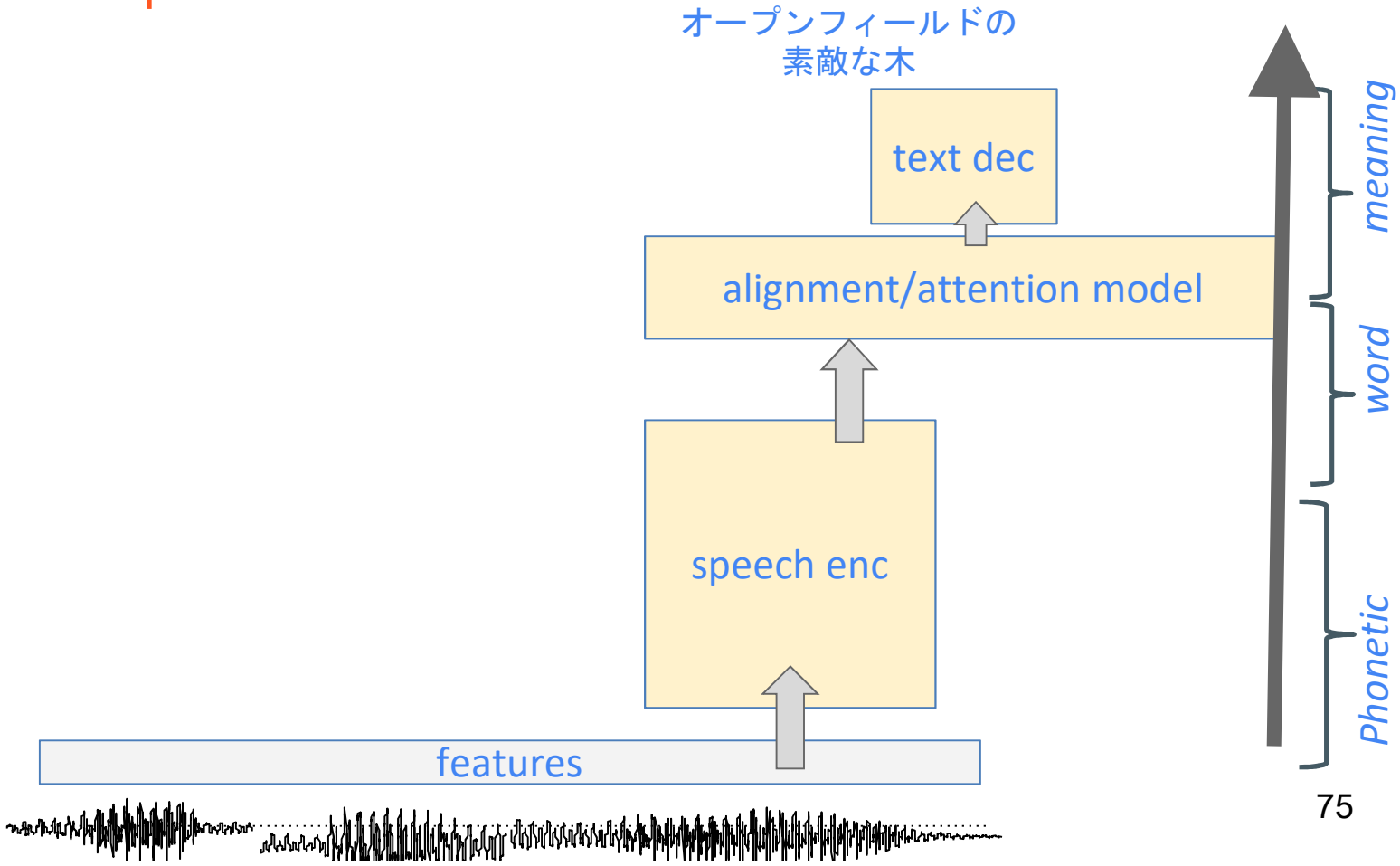
A. End2End speech-to-text architecture

B. Spoken term discovery through attention

C. Phone unit discovery via embedding extraction

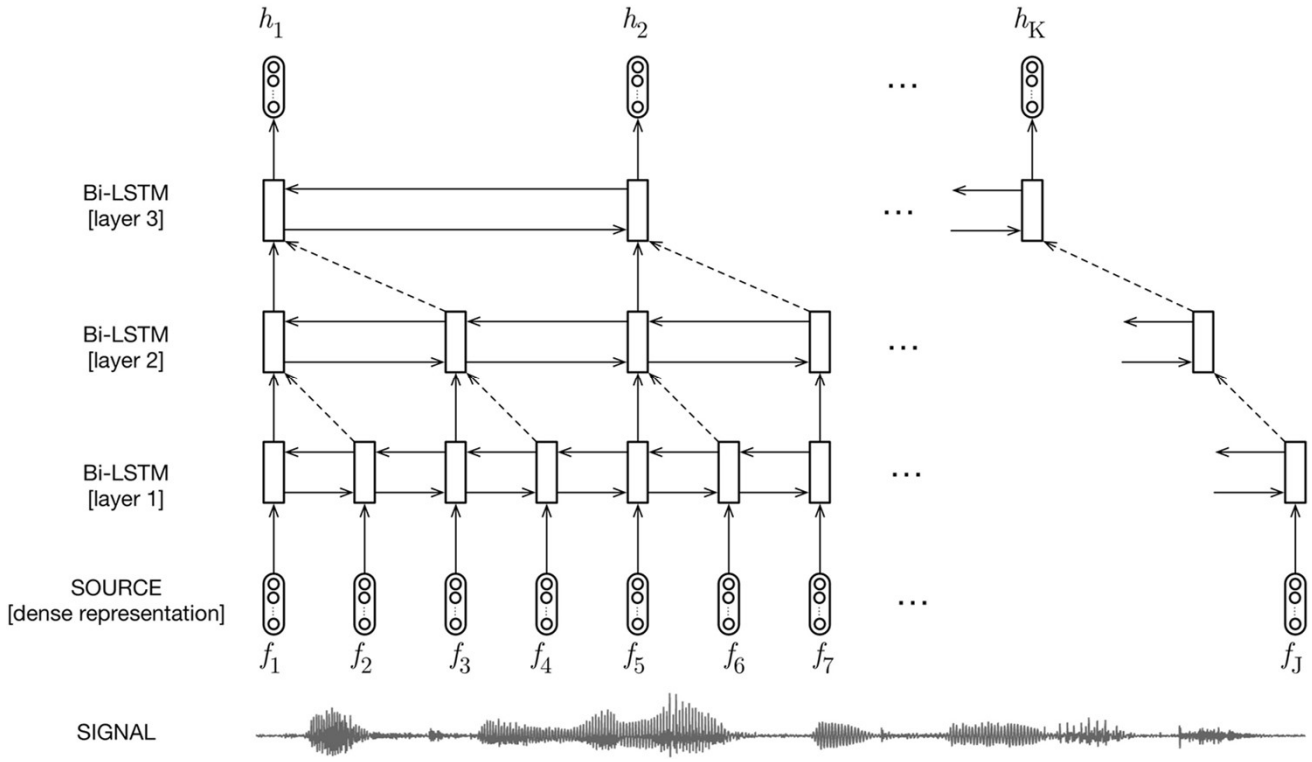
D. Segmentation with Reinforcement Learning

(A) End2End speech-to-text architecture



(A) Speech encoders

Pyramidal BiLSTM encoder

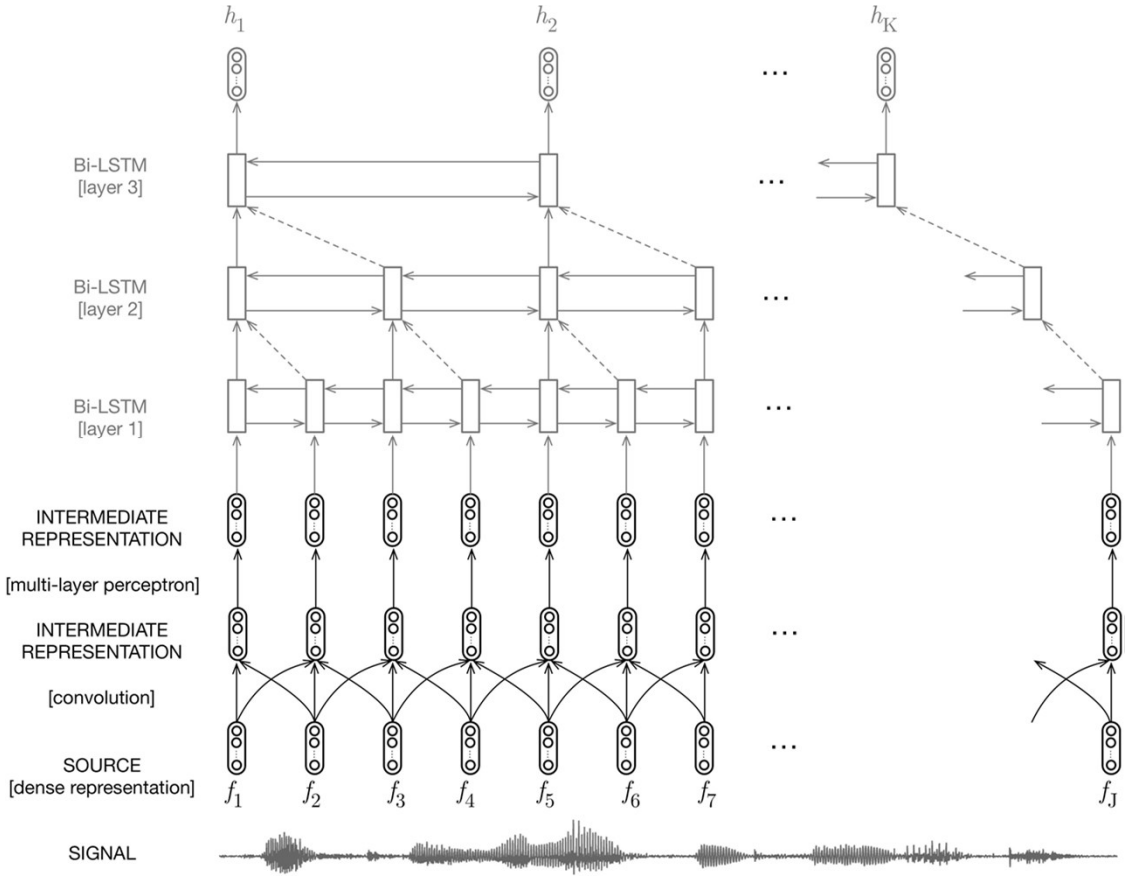


References:

- Listen, Attend and Spell (2015), Chan et al.
- Listen And Translate, a proof of concept (2016), Berard et al.

(A) Speech encoders

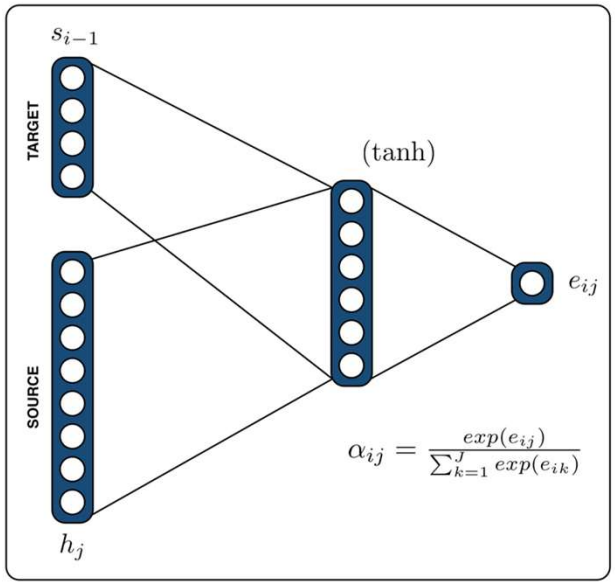
Pyramidal BiLSTM encoder with convolutional layers

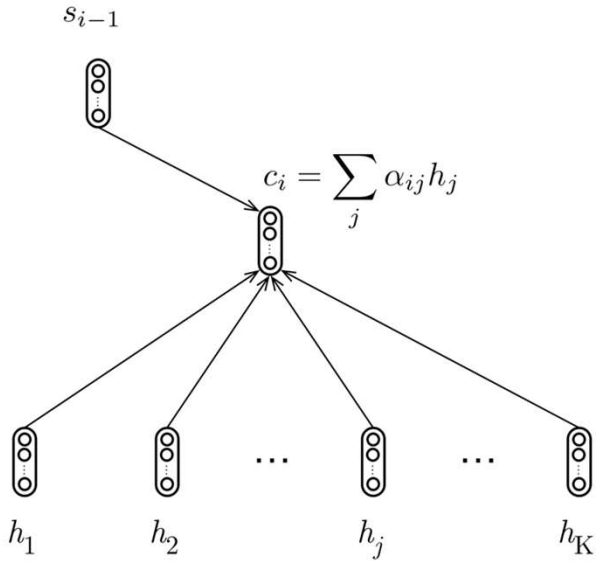


Intuition:
Add intermediate representation

Avoid to go directly to word-like representation

(A) Attention mechanism

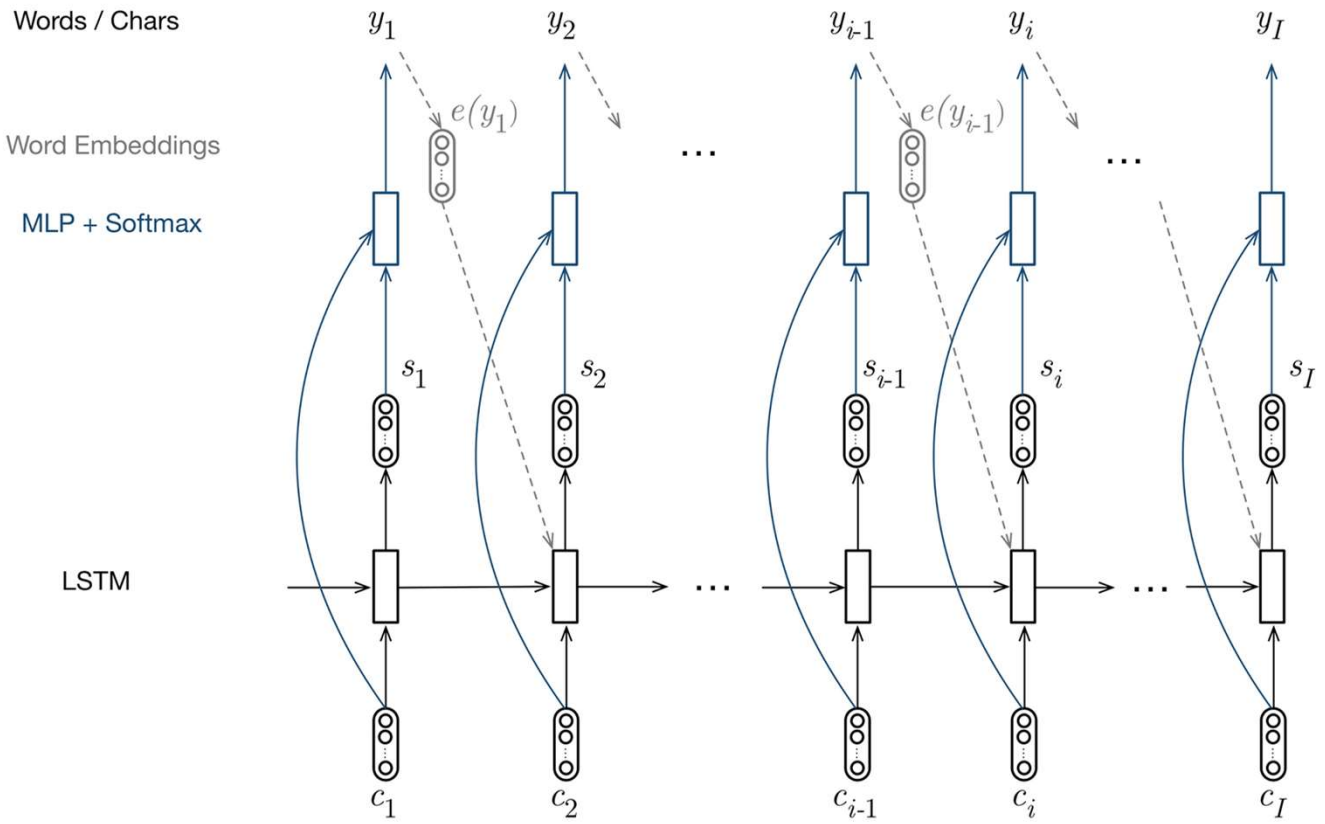




"Neural machine translation by jointly learning to align and translate."

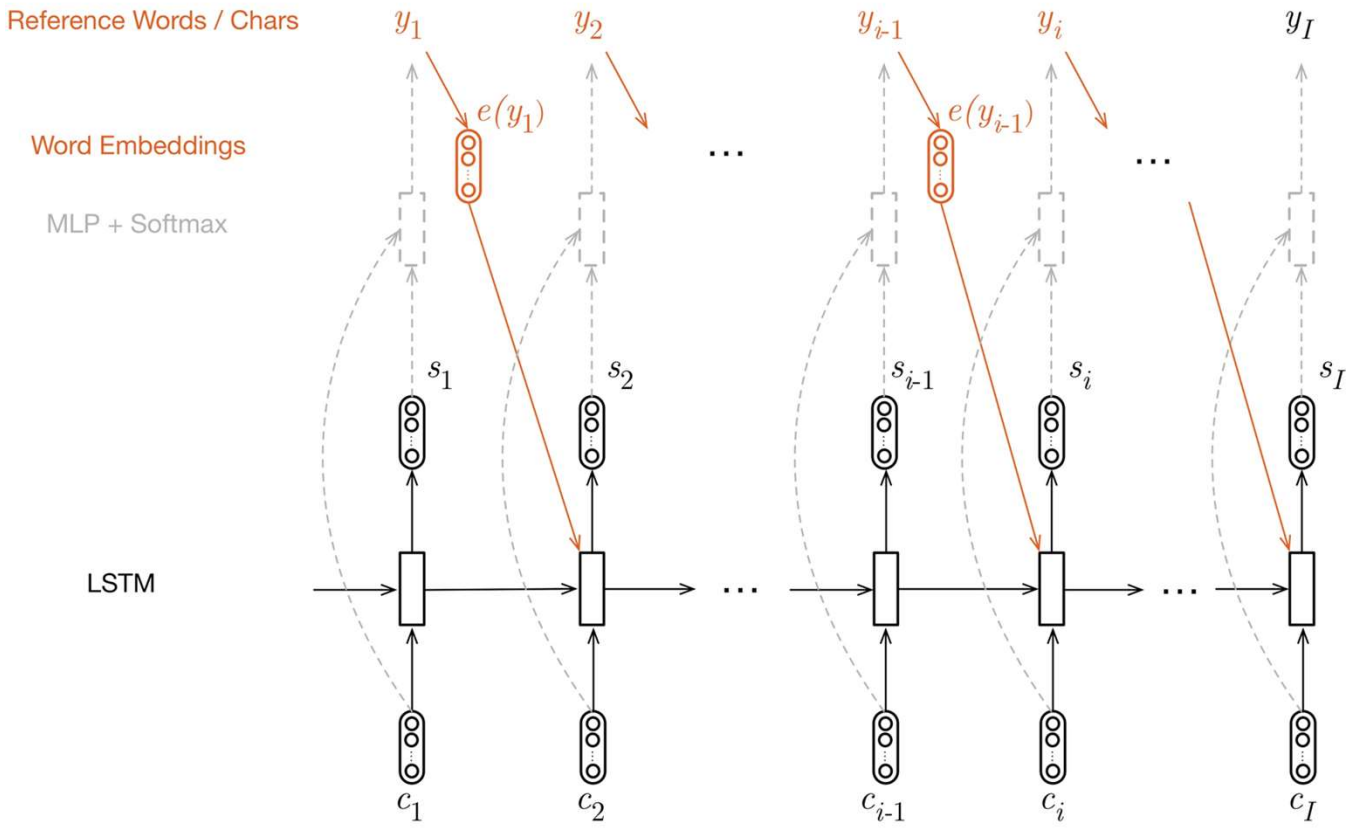
Bahdanau, D., Cho, K., Bengio, Y. (2014).

(A) Decoder



- MLP + Softmax to predict the current output word / char
- LSTM input:
  - context vectors from the attention mechanism
  - previous ref target word / char during training
  - previous predicted word / char during test

(A) Forced Decoding



- we want to evaluate an unsupervised auxiliary task (eg. source segmentation using attention matrices)
- we can use all the training data
- during decoding, we force the previous reference words (or characters) to the LSTM current input.

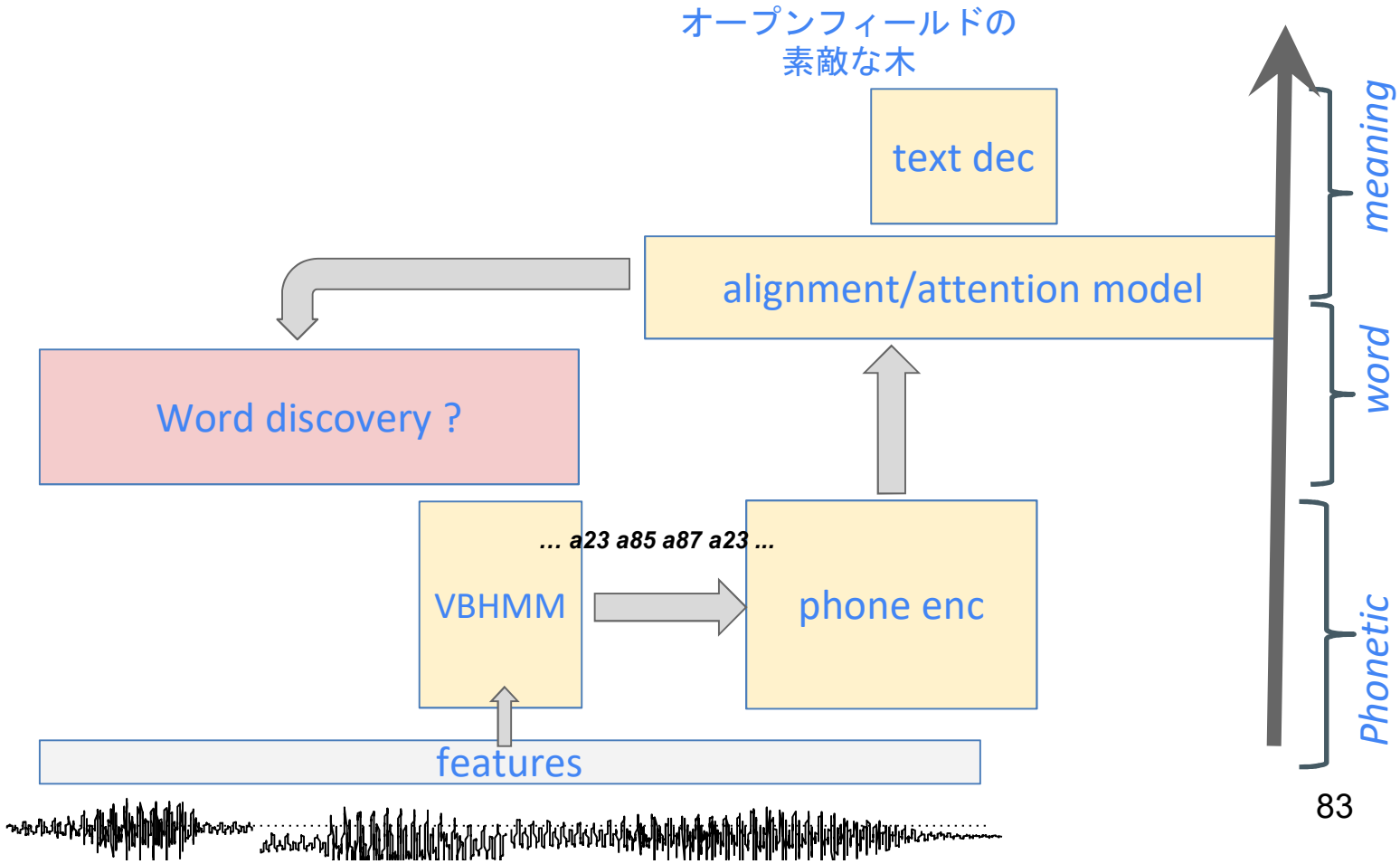
(A) Did the model learn to translate end2end from speech ?

Dataset DEV	Input	Target	BLEU4	CER
Flickr8k Train~30h	Speech English	Characters English	17.74	73.37
Flickr8k Train~30h	Speech English	Characters Japanese	30.99	66.53
Bulb Train~4h	Speech Mboshi	Characters Mboshi	56.91	35.29
Bulb Train~4h	Speech Mboshi	Characters French	22.36	79.54

(A) Did the model learn to translate end2end from speech ?

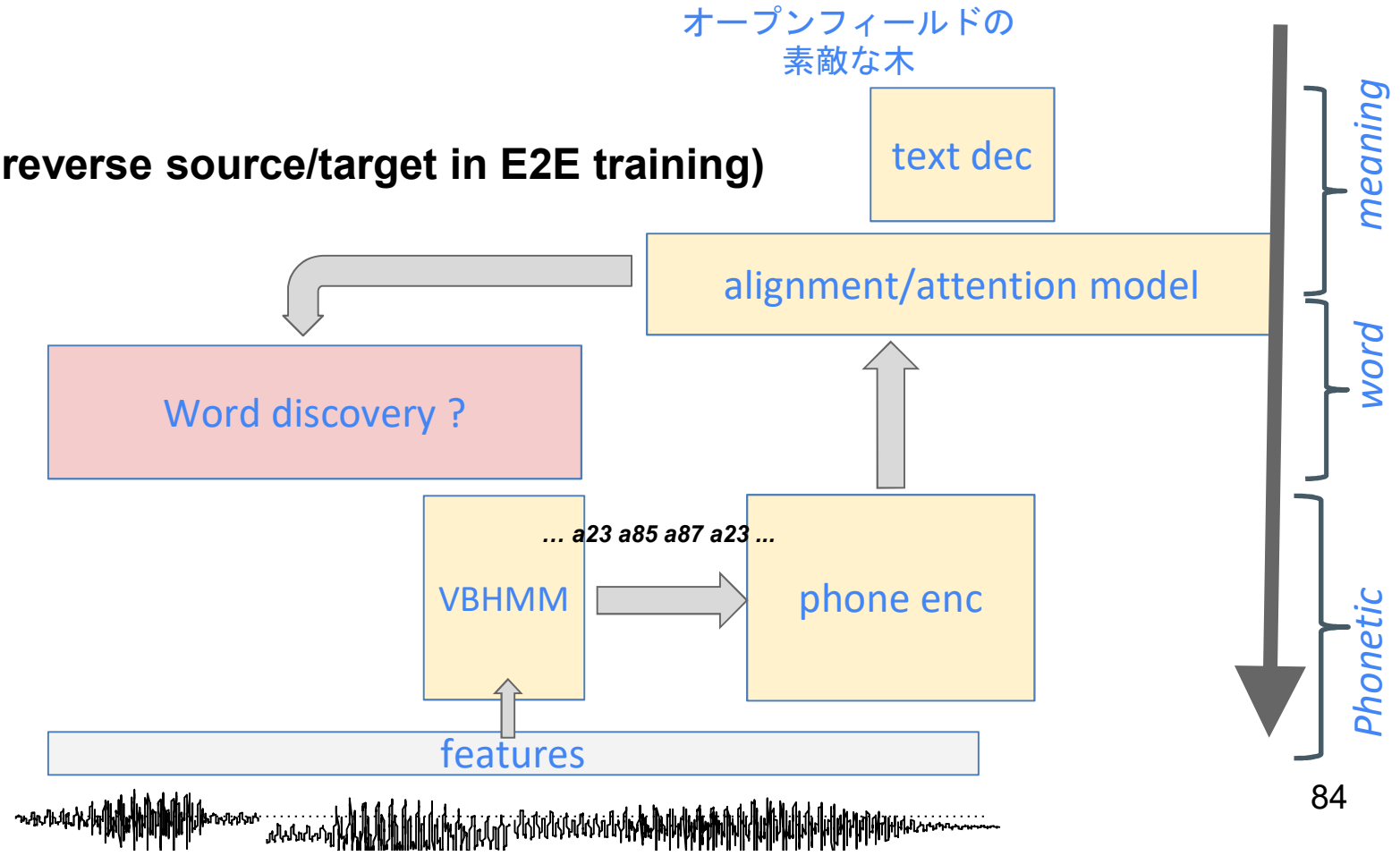
Dataset TEST	Input	Target	BLEU4	CER
Flickr8k Train~30h	Speech English	Characters English	12.71	74.81
Flickr8k Train~30h	Speech English	Characters Japanese	25.36	71.04
Bulb Train~4h	Speech Mboshi	Characters Mboshi	39.53	51.39
Bulb Train~4h	Speech Mboshi	Characters French	12.28	89.82

(B) Spoken term discovery through attention



(B) Spoken term discovery through attention

(possible to reverse source/target in E2E training)



(B) Spoken term discovery through attention

- Train End2End Mboshi-French NMT in both directions
- Full corpus *forced-decoded* after model training
- Many-to-one encouragement
  - Higher temperature factor T (>1)
  - Smoothing attention matrix
  - Long Duong & al. *An attentional model for speech translation without transcription.* In Proceedings of NAACL- HLT, 2016.

$$\alpha_{ij} = \frac{exp(e_{ij}/T)}{\sum_k exp(e_{ik}/T)}$$

$$\alpha_{ij} = \frac{\alpha_{ij-1} + \alpha_{ij} + \alpha_{ij+1}}{3}$$

- "Hard segmentation" by aligning source symbol in position *i* with target word in position *j* such that: *j= argmax_i (aij)*
- Evaluation
  - Train/Dev partition: 4643 utt / 514 utt
  - Comparison with two baselines (UPD+UWD pipeline and segmental DTW)
  - Scoring with zero-resource challenge metrics (spoken term discovery metrics)

(B) Spoken term discovery through attention

- Word **boundary** detection results: *true* phones and *pseudo* phones (speech!)

Input	Systems	P	R	F
true phones	Bayesian (dpseg) (Goldwater & al., 2009)	**68.2**	**82.6**	**74.7**
true phones	Attention (fr-mb)	51.7	67.9	58.7
true phones	Attention (mb-fr)	41.1	52.1	45.9
speech	Segmental DTW Baseline (Jansen & Van Durme, 2011)	27.3	12.0	16.6
speech (UPD)	Bayesian (dpseg)	25.3	**77.5**	38.1
speech (UPD)	Attention (fr-mb)	**36.6**	43.2	**39.6**
speech (UPD)	Attention (mb-fr)	34.7	43.6	38.6

(B) Spoken term discovery through attention

- Word **types** discovery results: *true* phones and *pseudo* phones (speech!)

Input	Systems	P	R	F
true phones	Bayesian (dpseg) (Goldwater & al., 2009)	**21.4**	**28.2**	**24.3**
true phones	Attention (fr-mb)	13.1	22.9	16.7
true phones	Attention (mb-fr)	4.9	9.9	6.6
speech	Segmental DTW Baseline (Jansen & Van Durme, 2011)	3.1	1.7	2.2
speech (UPD)	Bayesian (dpseg)	1.9	3.0	2.3
speech (UPD)	Attention (fr-mb)	**3.5**	**6.4**	**4.5**
speech (UPD)	Attention (mb-fr)	3.1	6.3	4.2

(B) Spoken term discovery through attention

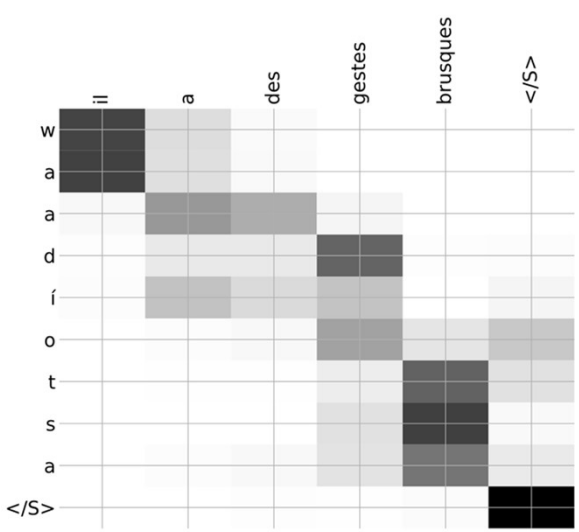
- Effect of UPD method on the word **boundary** detection performance (last minute!)

UPD	Systems	P	R	F
none	Segmental DTW Baseline (Jansen & Van Durme, 2011)	27.3	12.0	16.6
HMM-MFCC	Bayesian (dpseg)	25.3	**77.5**	38.1
HMM-MFCC	Attention (fr-mb)	36.6	43.2	39.6
HMM-MFCC	Attention (mb-fr)	34.7	43.6	38.6
SVAE HMM MBN	Bayesian (dpseg)	29.3	66.4	40.7
SVAE HMM MBN	Attention (fr-mb)	**44.0**	38.5	**41.1**
SVAE HMM MBN	Attention (mb-fr)	40.4	41.6	41.0

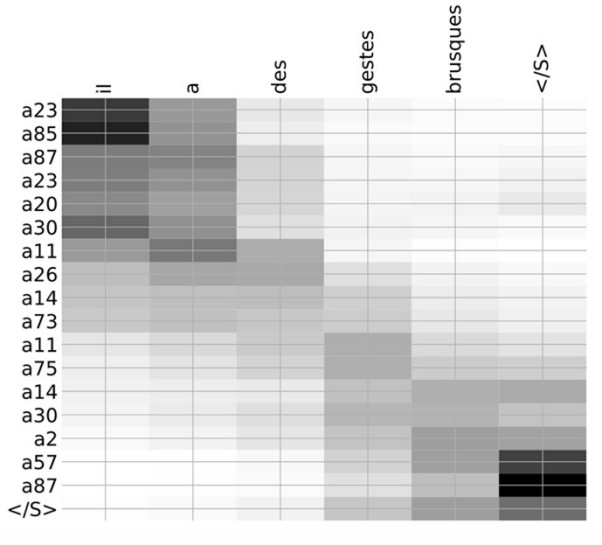
(B) Example of attention matrices

more (5k) matrices on https://github.com/JSALT-Rosetta/Illustrations/tree/master/AttentionMatrices

true phones



pseudo phones

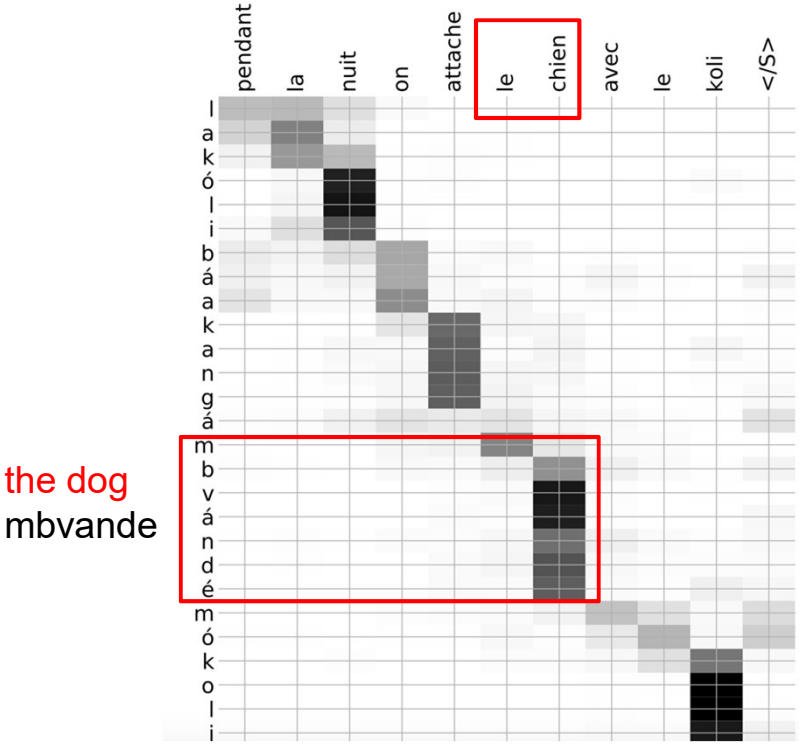


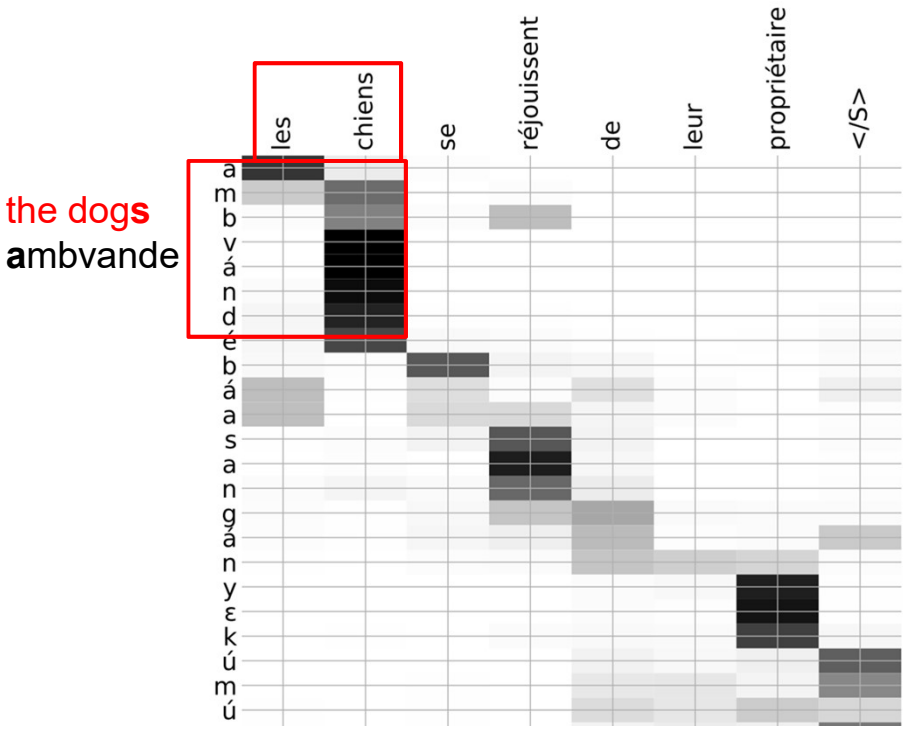
speech

soon!

(B) More attention matrices

more (5k) matrices on https://github.com/JSALT-Rosetta/Illustrations/tree/master/AttentionMatrices

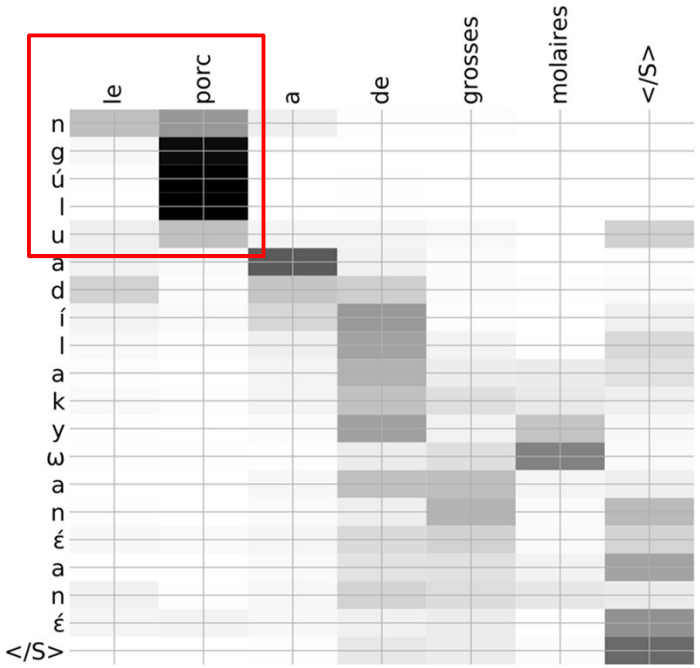




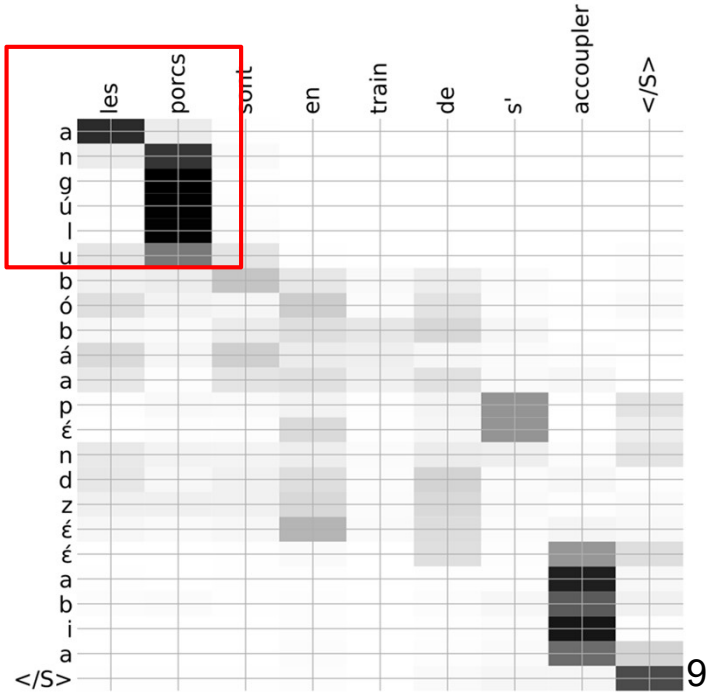
(B) More attention matrices

more (5k) matrices on https://github.com/JSALT-Rosetta/Illustrations/tree/master/AttentionMatrices

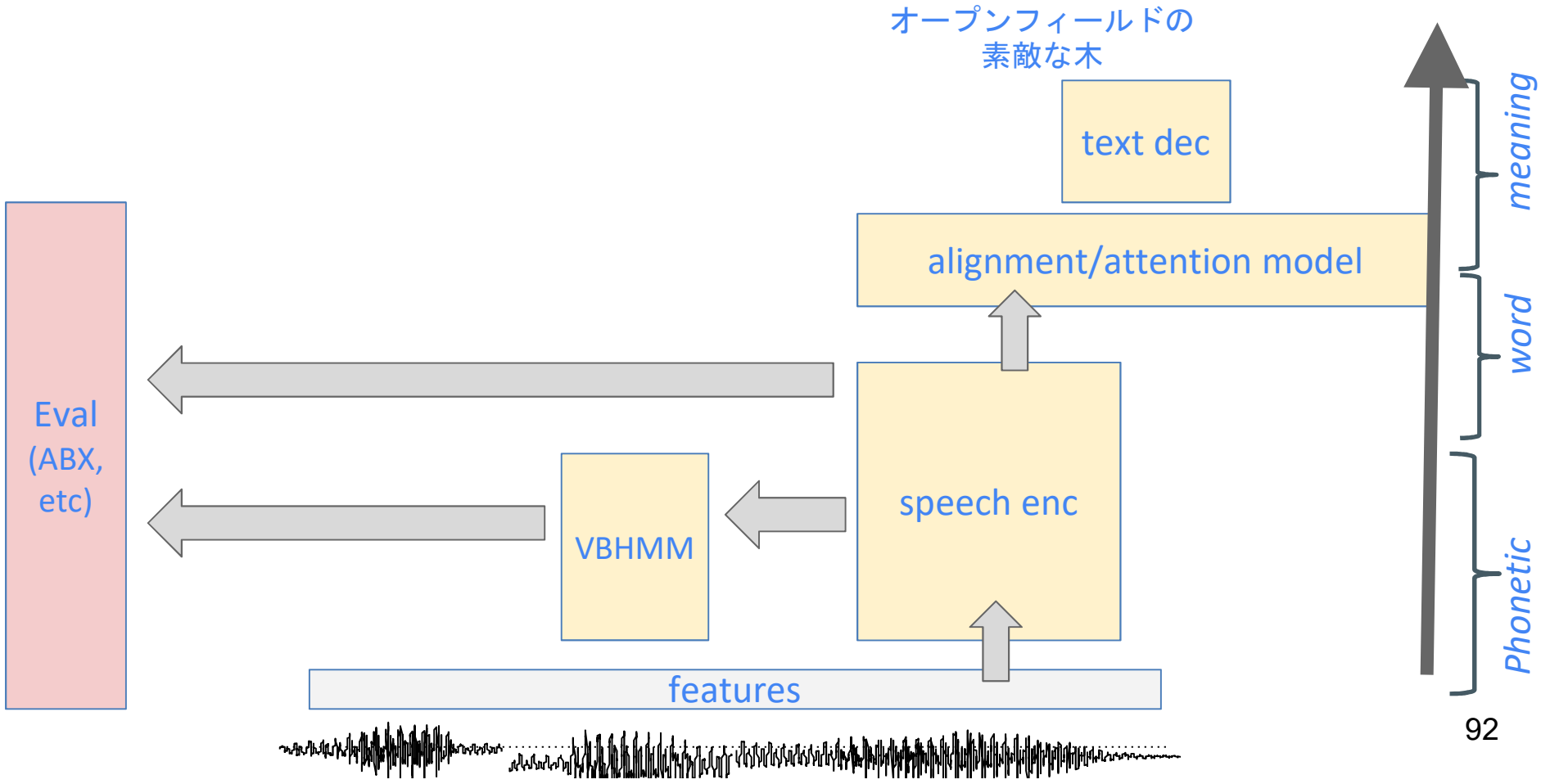
the pig
ngulu



the pig**s**
angulu



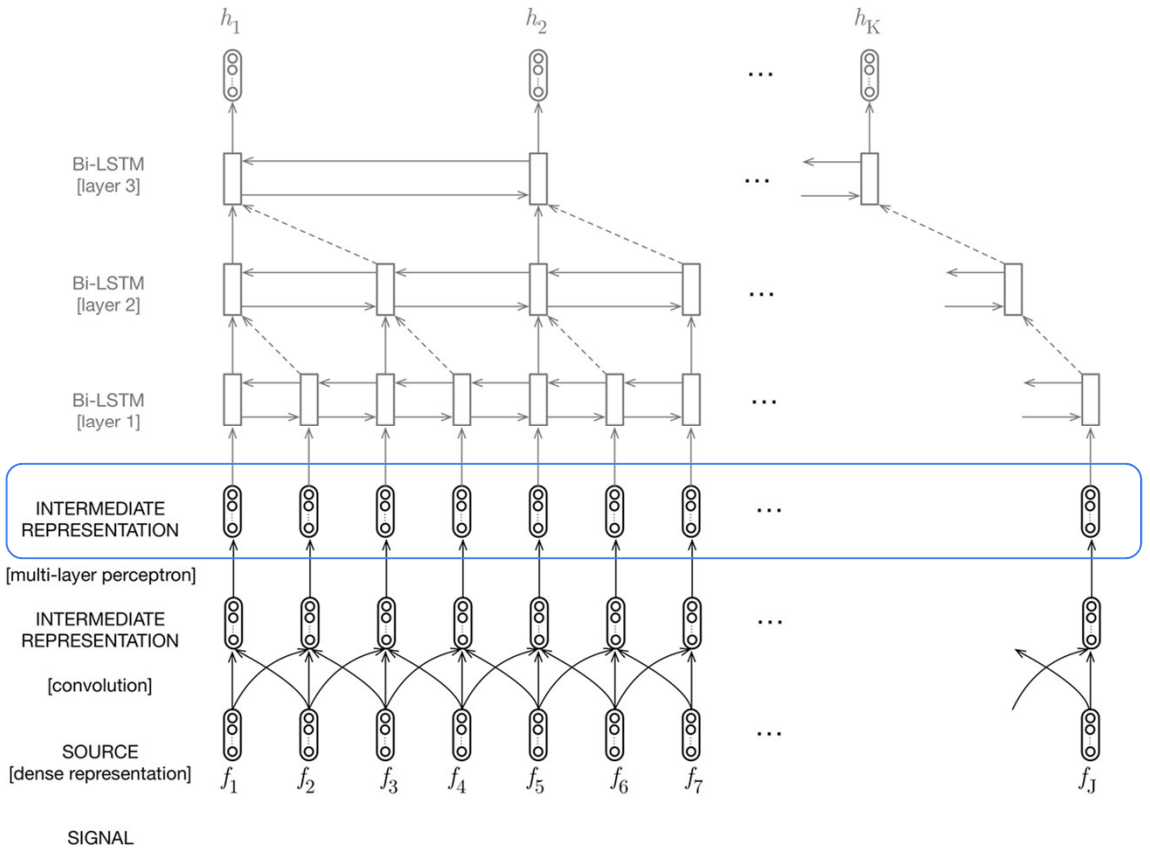
(C) UPD with End2End embeddings



(C) Extraction from XNMT

Distant supervision

Extraction of learned features:
Replace the MFCC for unit discovery and discriminability in the ABX tasks

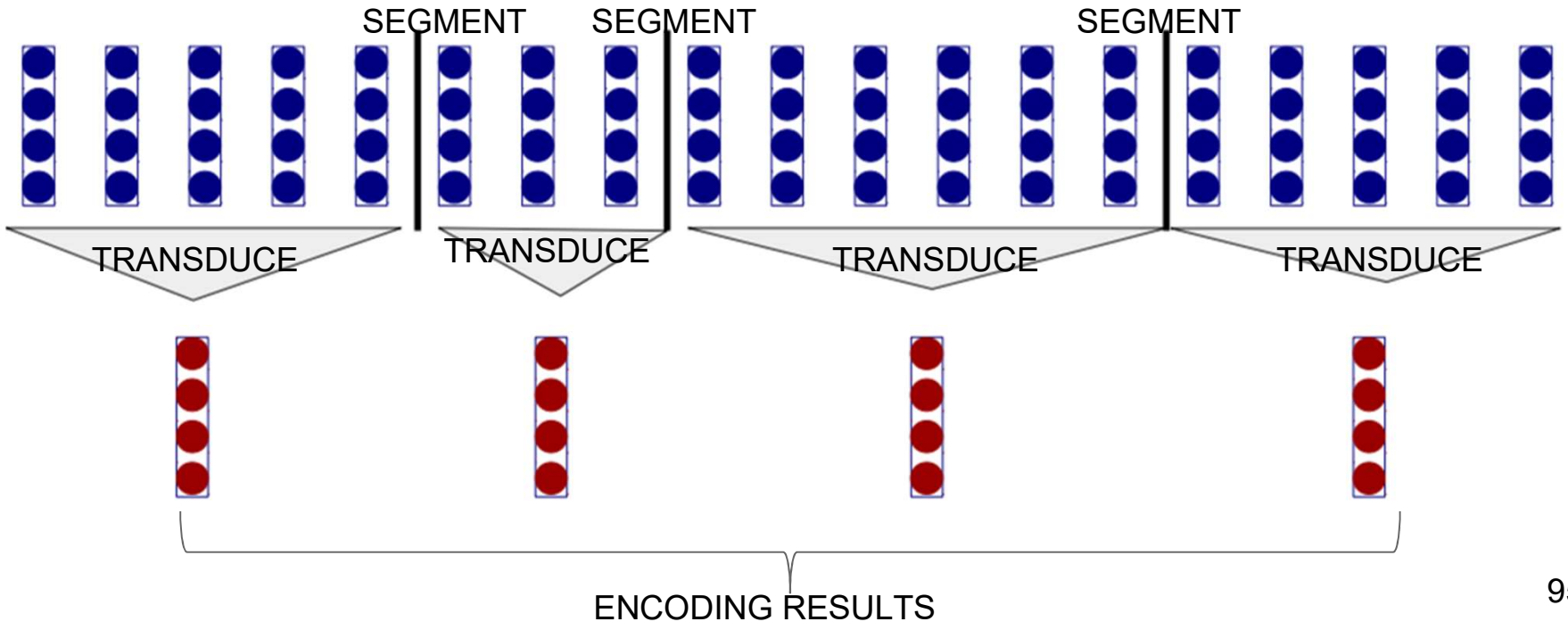


(C) Comparative results (preliminary)

Comparing all our features on **FlickR** with the ABX task

	ABX errors on phonemes		ABX errors on talkers
	within talker	across talkers	across phonemes
MFCC	10.4	28.3	36.9
Posteriogram	12.1	29.2	43.7
Articulatory features	24.2	31.1	43.6
SVAE HMM MBN	29.8	33.7	45.2
speech2text	incoming	incoming	incoming
speech2transaltion	incoming	incoming	incoming

(D) End2End Segmentation w/ Reinforcement Learning



(D) End2End w/Reinforcement Learning

- Segmentation Probability

P(segment | s_{1..i}) = softmax(RNN(s_{1..i})))

segment = action { READ, SEGMENT, DELETE }

- Encoding of the embedding:
  - Using the LSTM encoder.
  - Each time step, an action is determined.
  - For each "SEGMENT" action, the transduce() function is invoked.

(D) End2End w/Reinforcement Learning

- Segment Transducer:
  - Function to reduce [vector] -> vector
  - Can be implemented in several ways: RNN, CNN, Average, Sum, Tail, Downsampling, etc

- Consist of 2 parts:
  - Encoder -> Optional. If we want to learn parameter for transducing, LSTM can be used.
  - Transformer -> Applied after the encoding. It is basically a "reduce" function to reduce sequence of vector to a single vector

(D) End2End w/Reinforcement Learning

- We use the REINFORCE objective to learn the parameter of the segmentation.

$$REINFORCE(f,e) = \lambda * \sum_{i=1}^{n} \left(\log(P_{MLE}(f,e)) * \sum_{j=1}^{|f|} \log(P(\text{segment}|f_{1..j})) \right)$$

(D) End2End w/Reinforcement Learning

Challenges:

- Currently learning is not so stable, yet the best learning scheme is still a very interesting question.
- If learning is not well, then all the actions will be SEGMENT, i.e., no segmentation at all

Possible Remedy:

- Semi supervisely learning the segmentation probability with real segmentation.
- Delay learning of the segmentation until original model is good enough.
- Calculate average(word/frame) + penalize if #segments are very different.

Coming up next…

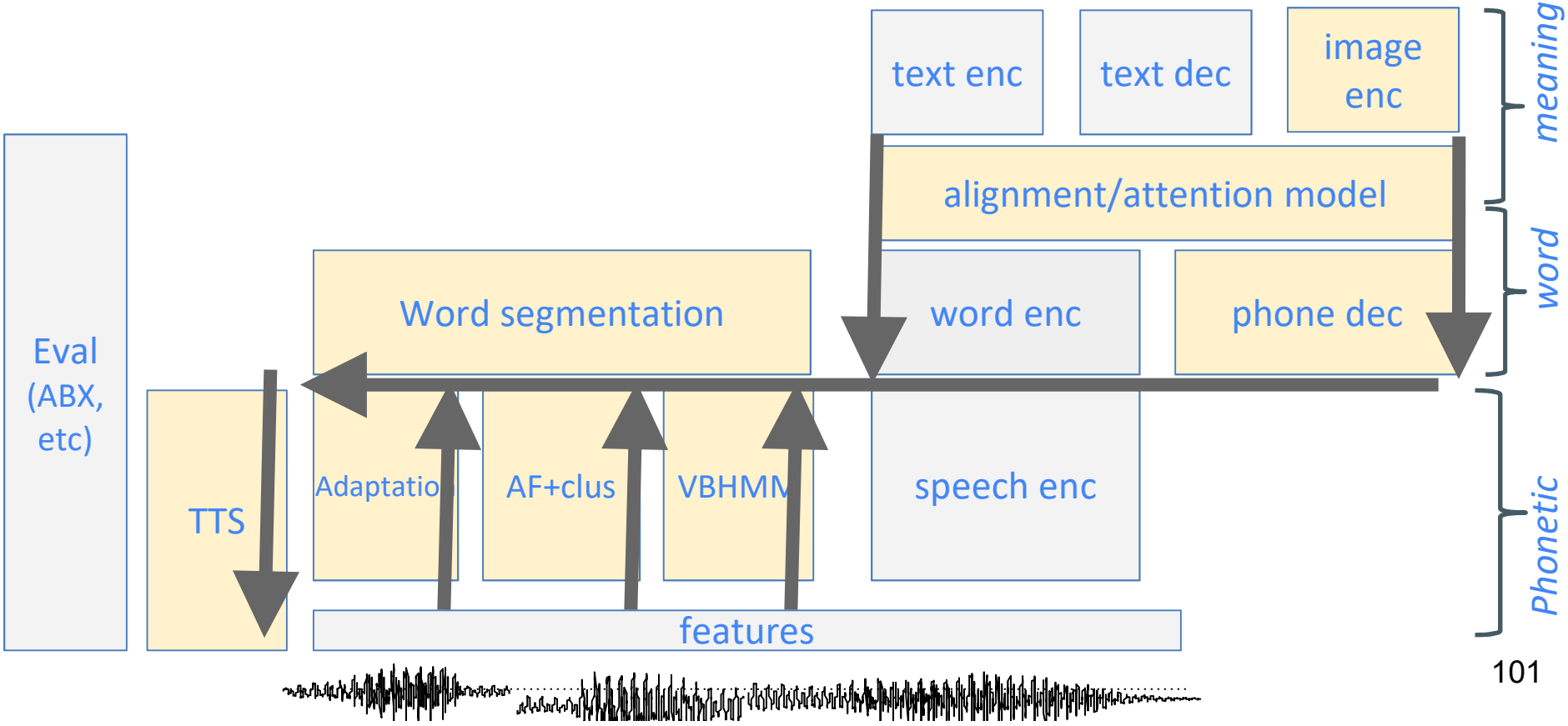
IV. Constructing a TTS system in an unwritten language

"TTS without T"

Starring (in order of appearance): Alan Black, Mark Hasegawa-Johnson, Laurent Besacier, Odette Scharenborg

IV. Constructing a TTS system in an unwritten language

オープンフィールドの
素敵な木



IV.A - Speech synthesis using segmental units

(your host for the next 8 minutes: Alan Black)

Speech synthesis is one medium we need to generate

We still (probably) need some form of symbolic sequential tokens

These could be “**phonemes**”, “**syllables**”, “**words**” and “**phrases**”

Two uses of synthesis in this project:

- Can we generate speech from unit sequences
- Can we evaluate how good unit inventories are

IV.A - Clustergen Parametric Synthesis

CMU Clustergen Parametric Speech Synthesizer [Black 2006]

- Predict sequences of (mcep) frames from contextualized phones sequences
- Built from phrases/words/syllables/phones/phonetic features plus recordings

Festvox Voice Building Tools (http://festvox.org)

- Lexicons, data sets, LTS tools, ML modeling techniques for
- prosodic, spectral, duration etc.
- Produce single “voice file” deployable on Android device
- Used for large number of languages (by multiple groups)

IV.A - Clustergen for evaluation metric

Input: waveform file plus symbolic sequences of “words” or “phones”

Output: Simple synthesizer and distortion measure on held out data

Clustergen:

- Single CART model on spectral features (no real prosody)
- Built-time: about 20 mins for 1 hour of speech

IV.A - Synthesis Objective Measure

MCD: Melcepstral distortion [Toda et al 2004]

Output: Simple synthesizer and distortion measure on held out data

$$10/ln(10) * \sqrt{2 \sum_{i=1}^{24} (mc_i^t - mc_i^p)^2}$$

- Used for voice conversion and speech synthesis
- Typical range 3.00 - 10.00++ (small is better)
- 0.08 considered significant
- Good for measuring improvements
- Not so good as absolute measure
- 0.12 improvement typical for doubling database size

IV.A - Clustergen for Synthesis

Input: waveform file plus symbolic sequences of “words” or “phones”

Output: Simple synthesizer and distortion measure on held out data

Clustergen rf3 [Black and Muthukumar 2015]

- 4 models for F0, spectrum, excitation and duration
- Random Forests of CARTs
- Optimize Segment Boundaries for prediction improvements
- Typically 0.3 improvement over base voices (more for “bad” “smaller” voices)
- Build-time: about 5-6 hours for 1 hour of speech (10x over base voice)

IV.A - Synthesis for unwritten languages

Outstanding Issues :

- People don’t deliver in consistent “reading” style
- Often you have multiple, rather than single speaker
- Not as good recordings in poor environment
- Other speech features may be important (words, phrases)
- Other other speech features may be important (stress, tone)
- Typical only a small amount of data of poor quality

IV. B - im2ph: speech from images

(your host for the next 8 minutes: Mark Hasegawa-Johnson)

1. Image Representation: 14x14 = 196 vectors per image, calculated from the last CNN layer of the VGG ImageNet image classifier
2. Translation from Image Vectors to Output Phones: xnmt sequence-to-sequence neural machine translation, PyramidalLSTM encoder, StandardAttender, MlpSoftmaxDecoder
3. Examples from flickr8k
4. Examples from MSCOCO
5. Numerical results of three experiments

Image representation: CNNFEAT

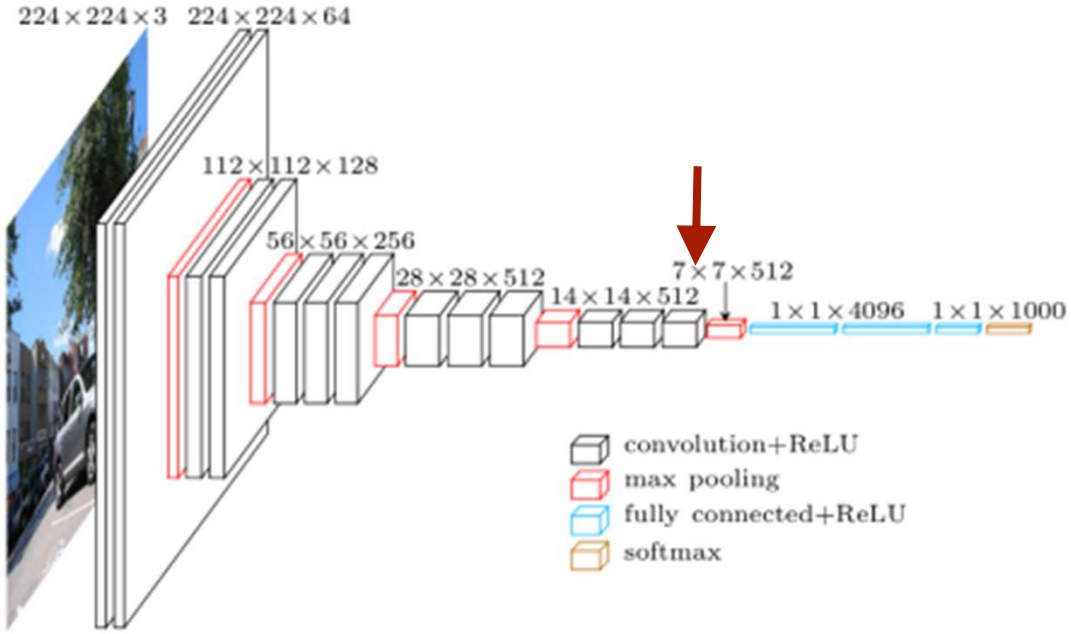


Figure copied without permission from Simonyan & Zisserman, 2014.

- ImageNet = 500+ images/noun of each of the nouns in WordNet.
- VGG = 13-layer CNN + 2-layer FCN, trained on 14m images, covering the 1000 most numerous nouns, 92.7% top-5 test accuracy.
- **CNNFEAT: 196 feature vectors/image, 512d/vector, from the last CNN layer. Each receptive field covers about 40x40 pixels in the original 224x224 image.**
- VGGFEAT (used later in today's talk, not right now): 1 vector/image, 4096d/vector, from penultimate FCN layer

im2ph = machine translation (xnmt)

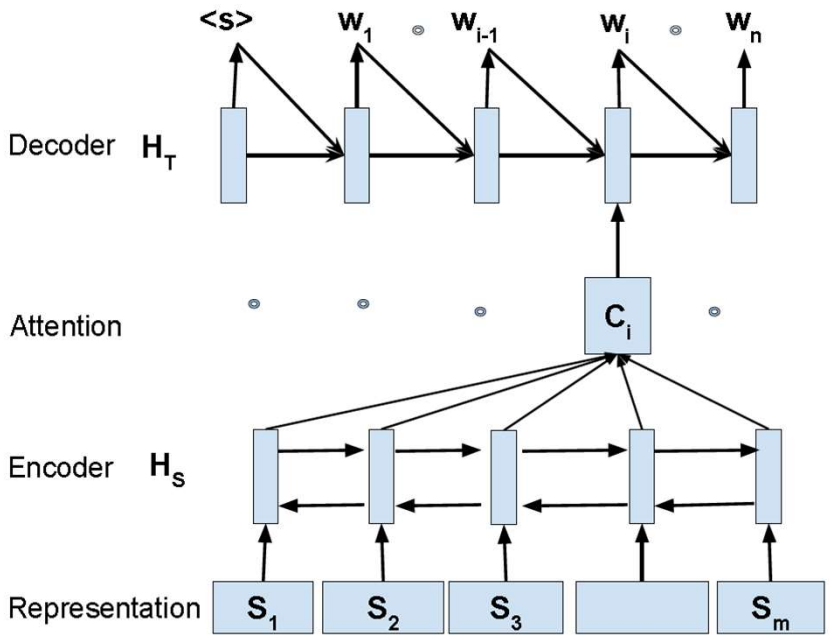
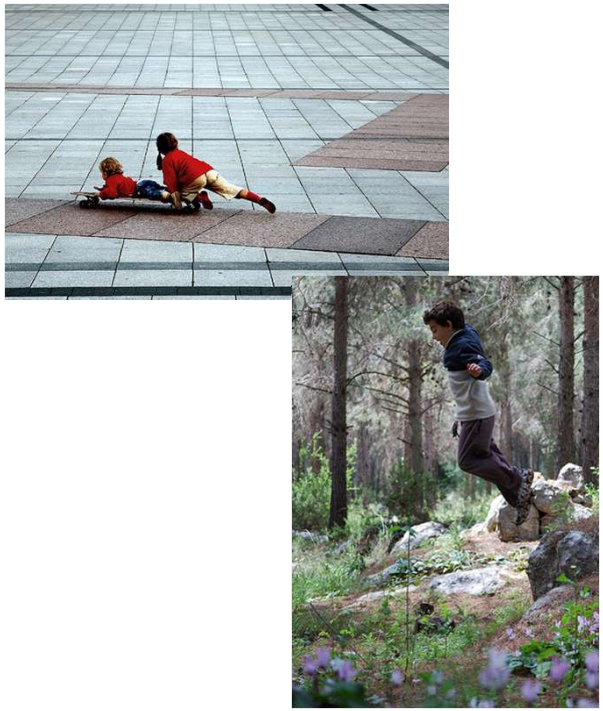


Figure copied without permission from Duong, Anastasopoulos, Chiang, Bird & Cohn, NAACL-HLT 2016.

- "Representation:" 196 vectors/image
- "Encoder:" PyramidalLSTM with one 128d state vector. Sequence is row-wise raster scan of the image.
- "Attention:" StandardAttender, 128d input, 128d state vector, N hidden nodes
- "Decoder:" MlpSoftmaxDecoder, 3 layers, 1024d hidden vectors
- Output vocabulary: synthetic phones (MSCOCO), force-aligned phones (flickr8k), or acoustic unit discoveries (both)

flickr8K



Images and Reference Texts: Hodosh, Young & Hockenmaier, 2013. Waveforms: Harwath and Glass, 2015

- Reference 1: “The boy +um+ laying face down on a skateboard is being pushed along the ground by +laugh+ another boy.” (synthetic version)
- Reference 2: “Two girls +um+ play on a skateboard +breath+ in a court +laugh+ yard.” (synthetic version)
- Hypothesis (128d attender): SIL +BREATH+ SIL T UW M EH N AA R R AY D IX NG AX R EH D AE N W AY T SIL R EY S SIL
- Hypothesis (64d attender): SIL +BREATH+ SIL T UW W IH M AX N W AO K IX NG AA N AX S T R IY T SIL
- Reference 1: “A boy +laugh+ in a blue top +laugh+ is jumping off some rocks in the woods.” (synthetic version)
- Reference 2: “A boy +um+ jumps off a tan rock.” (synthetic version)
- Hypothesis (128d attender): SIL +BREATH+ SIL EY M AE N IH Z JH AH M P IX NG IH N DH AX F AO R EH S T SIL
- Hypothesis (64d attender): SIL +BREATH+ SIL EY Y AH NG B OY W EY R IX NG AX B L UW SH ER T SIL IH Z R AY D IX NG AX HH IH L SIL

MSCOCO

Images and Reference Texts: MSCOCO

- Reference 1: “A group of men enjoying the beach, standing in the waves or surfing.”
- Reference 2: “A group of people standing on a beach next to the ocean.”
- Hypothesis (64d attender): # @ g r uu p @ v p ii p l= s t a n d i ng o n @ b ii ch # #

- Reference 1: “A, a black and white photo of a fire hydrant near a building.”
- Reference 2: “Aa, a fire hydrant that is out next to a house.”
- Hypothesis (64d attender): # @ p @@ s n= w oo k i ng @ t @ m e d^ l= d au n @ n d @ r e d f ai r h ai d r @ n t # #

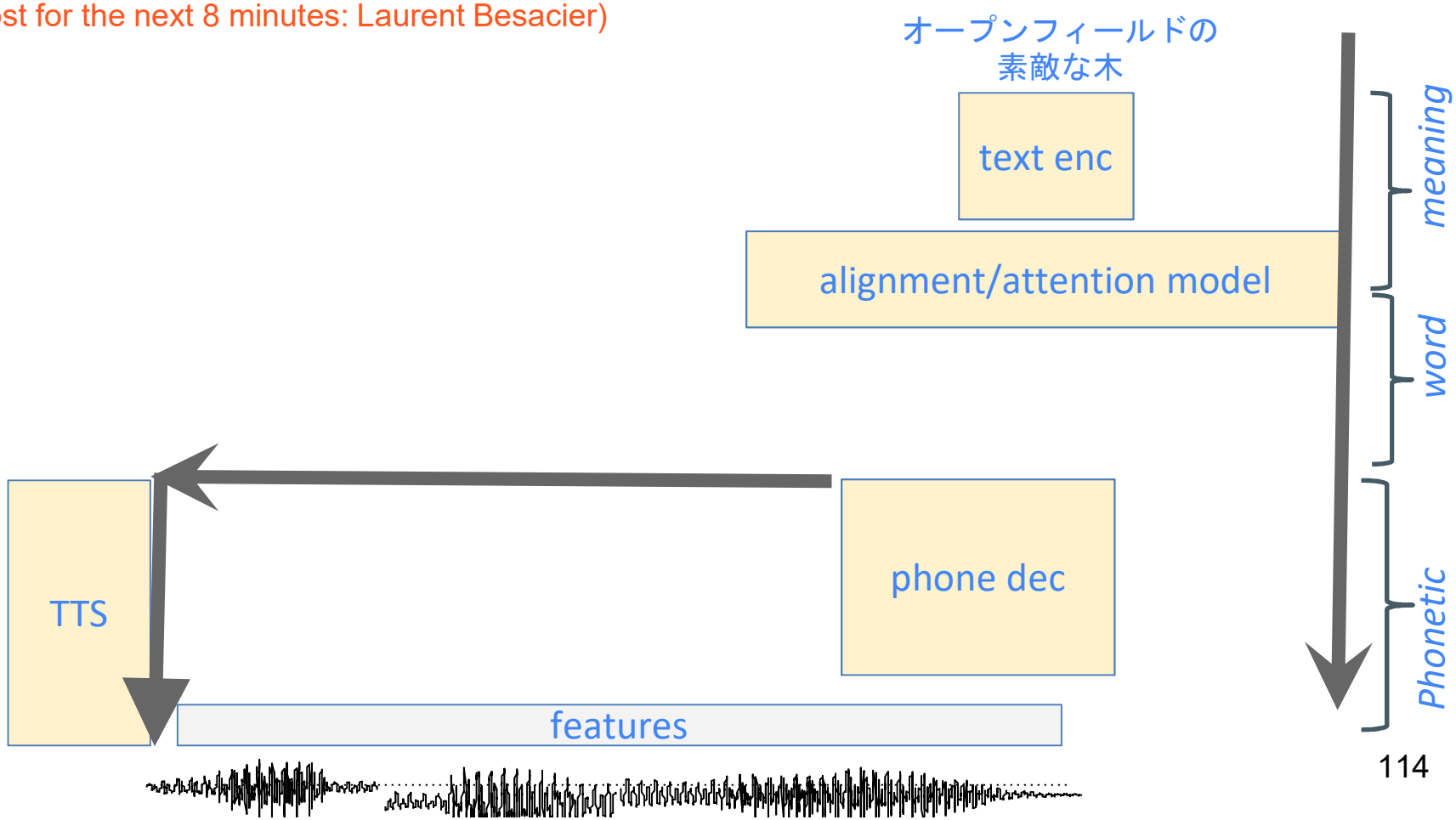
Numbers and Caveats

Dataset	Architecture	Dev 1-ref BLEU	Dev XENT	Dev PER
Flickr8k	PyramidalLSTM, Attender128, 3-MlpSoftmax1024	16 (54/22/12/6)	0.46 bits/phn	
Flickr8k	PyramidalLSTM, Attender64, 3-MlpSoftmax1024	16 (54/22/12/6)	0.46 bits/phn	86%
MSCOCO 8k	PyramidalLSTM, Attender64, 3-MlpSoftmax1024	16 (59/24/11/7)	0.89 bits/phn	

Caveat: This uses BLEU w.r.t. a single reference per image. It should measure BLEU w.r.t. 5 references per image (5 human captions). That measure has not yet been computed.

IV.C - Speech from foreign text

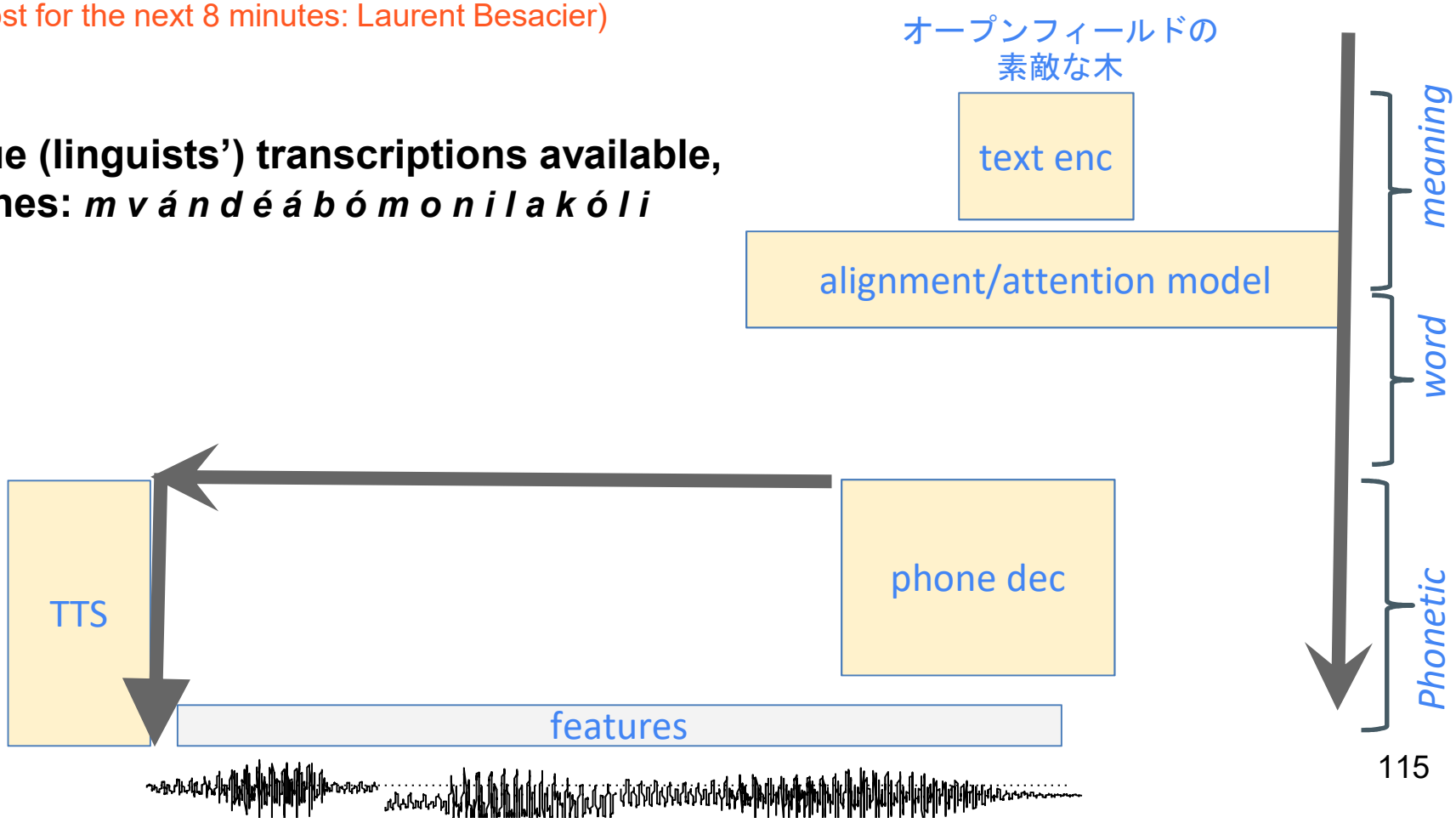
(your host for the next 8 minutes: Laurent Besacier)



IV.C - Speech from foreign text

(your host for the next 8 minutes: Laurent Besacier)

If true (linguists') transcriptions available,
phones: *m v á n d é á b ó m o n i l a k ó l i*

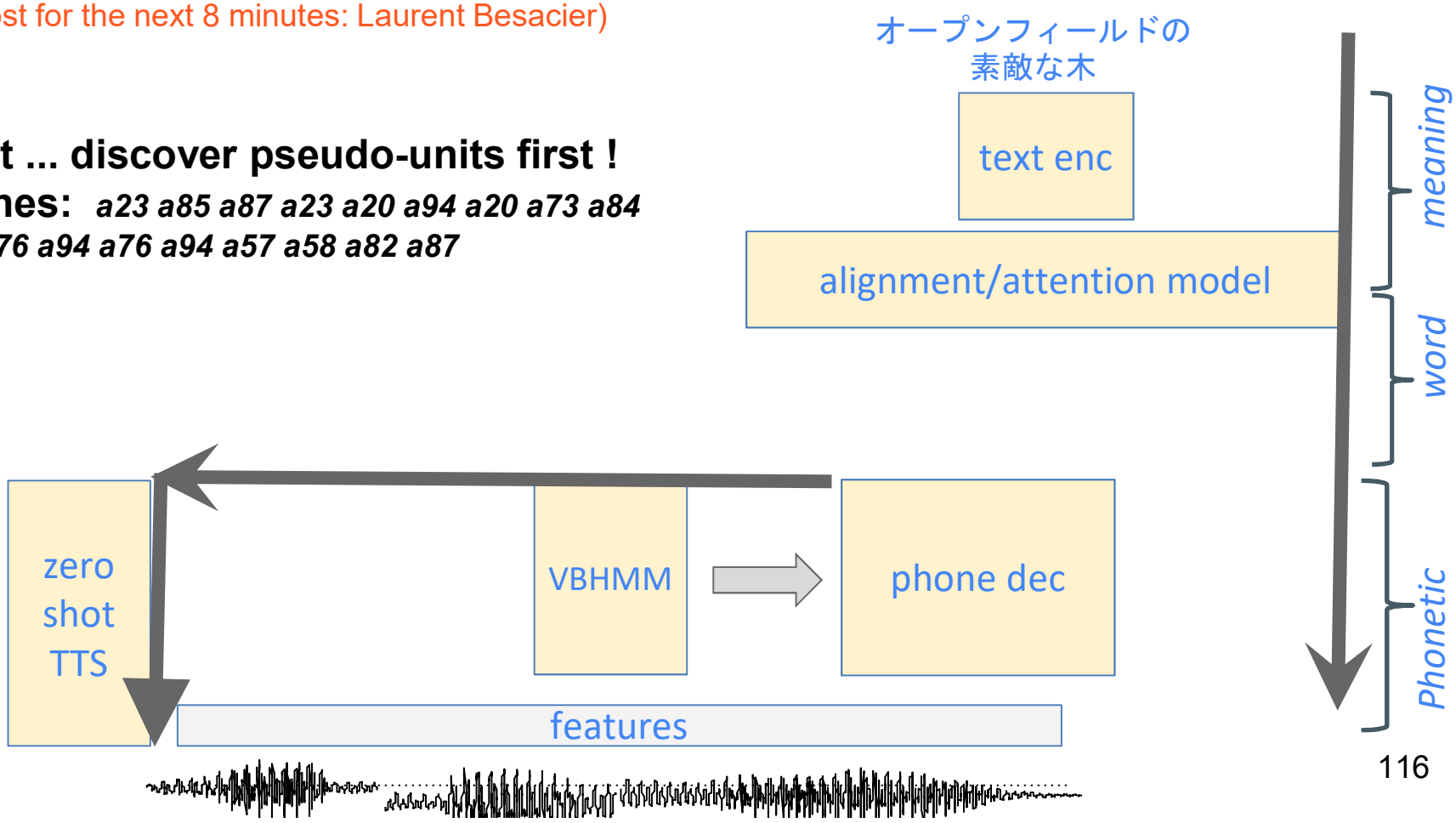


IV.C - Speech from foreign text

(your host for the next 8 minutes: Laurent Besacier)

If not ... discover pseudo-units first !

phones: ***a23 a85 a87 a23 a20 a94 a20 a73 a84 a53 a76 a94 a76 a94 a57 a58 a82 a87***



Speech from foreign text

- **French text to Mboshi phones translation demo**
  - Train an attention-based Neural Machine Translation (NMT) from French text to Mboshi phones (or to Mboshi pseudo phones discovered through UPD)
  - Train/Dev partition: 4643 utt / 514 utt (small corpus!)
  - Did not add yet monolingual (target-target) data (cf NMT4low_resource findings...)
- NMT architecture and parameters
  - Source and target embeddings: 32
  - Standard BLSTM encoder / LSTM decoder
  - Training using Adam, learning rate 0.001, batch size 32
- Evaluation
  - Report results for true phones (topline) and pseudo phones (speech!)
  - Scoring with BLEU4 (on characters!)

Speech from foreign text

- Translation performance (BLEU4)

dataset	true phones (topline)	pseudo phones (speech)
train	48.80	19.12
dev	31.95	8.32

- Example of mboshi outputs (true phones and pseudo phones) from dev

ref: **m v á n d é á** b ó m o n i l a **k ó l i**
hyp: **m v á n d é á** m i s á á o **k ó l i**

ref: **b á n a b ó b á a** m i i g h á **m b í a**
hyp: **b á n a b ó b á a** d z á á **m b í a**

ref: **a23** a85 **a87 a23 a20** a94 a20 a73 a84 a53 **a76 a94 a76 a94** a57 **a58 a82 a87**
hyp: **a23** **a87 a23 a20** a73 a26 a24 a41 a94 **a76 a94 a76 a94** **a58 a82 a87**

ref: **a23 a87 a23** a83 a98 a71 a26 a41 a61 a12 a66 **a90 a94** a35 a54 a73 a2 **a57 a87**
hyp: **a23 a87 a23** a13 a38 a26 a35 a26 a24 a35 a94 **a90 a94** a76 a94 **a57 a87**

Speech from foreign text

- Let’s listen to a few WAV files examples
- https://github.com/JSALT-Rosetta/Illustrations/tree/master/TTS/mboshi/
  - Zero Shot TTS demo
  - source_text2target_speech pipeline

IV.D - Cross-language units

(your host for the next 4 minutes: Odette Scharenborg)

Background: adaptation of a cross-linguistic ASR system to discover acoustic units in a low-resource language

Goal: find best acoustic units, i.e., they should be good

- semantically; allow good translated text retrieval ← earlier presentation
- acoustically ← this presentation

TTS used for the evaluation of the *acoustic* appropriateness of the phones

TTS system / Methodology

- **Training material:** same training set as ASR adaptation, 3660 utterance (3.5 hours), English, multispeaker from Flickr_8K
- **Input to TTS:**
  - Self-labelled phone strings from the cross-language phone system
  - Audio given to cross-language phone system
- **Build synthesizer** from data
  - Resynthesize each sentence and calculate MCD
- **Output of TTS:** ordered list of the phone sequences of their distortion ➔
- Best ones used for retraining the adaptation model

This can be better than simple ASR alignment score as synthesis cares much more about segmental boundaries, duration, phrasing, prosody etc.

Initial Results

- Oracle phones (gold standard): MCD 8.67
- Iteration 0 (Dutch base + adaptation): MCD 9.06
- Iteration 1 … (Still running)

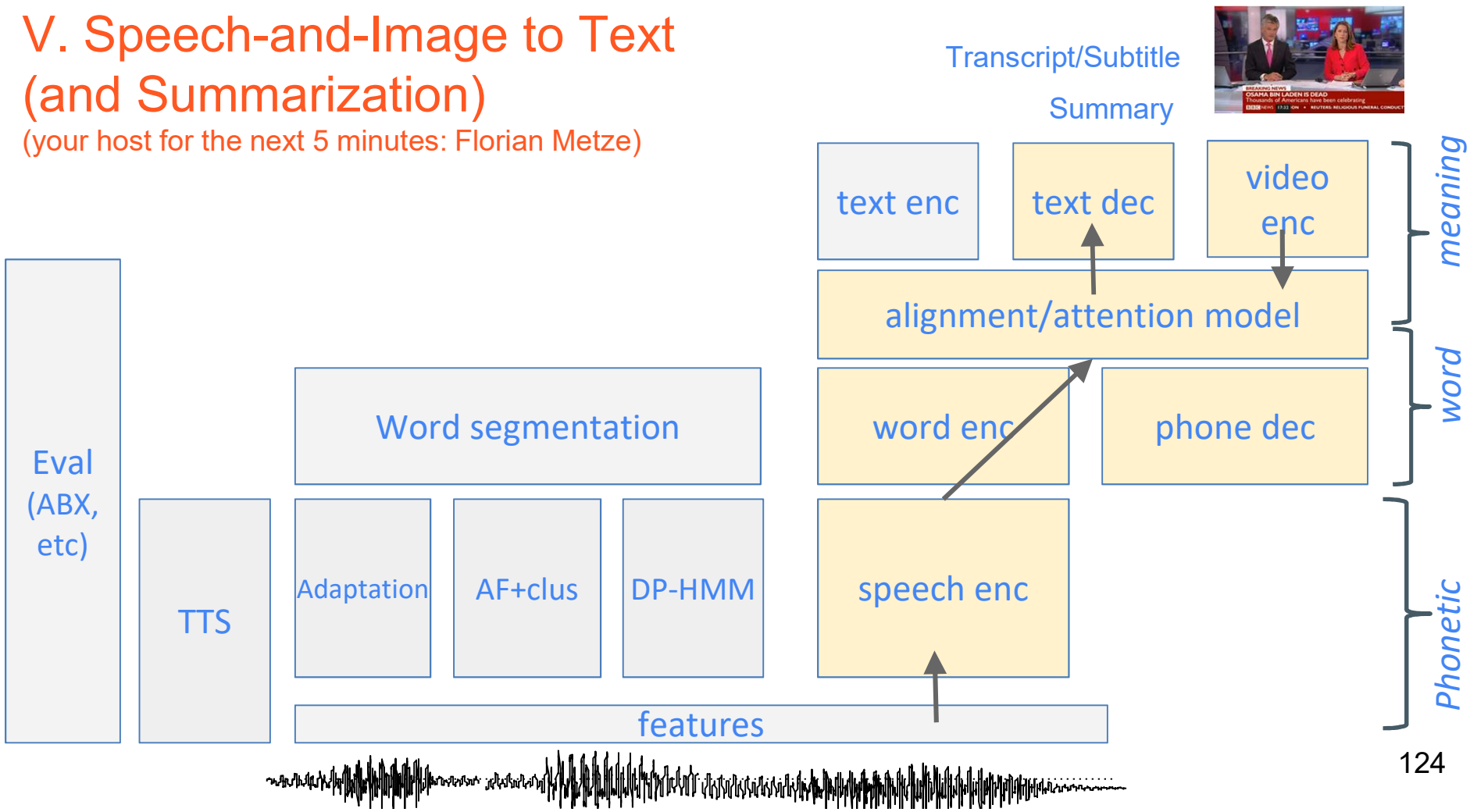
Coming up next…

V. Speech-and-Image to Text (and Summarization)

Starring (in order of appearance): Florian Metze, Shruti Palaskar

V. Speech-and-Image to Text (and Summarization)

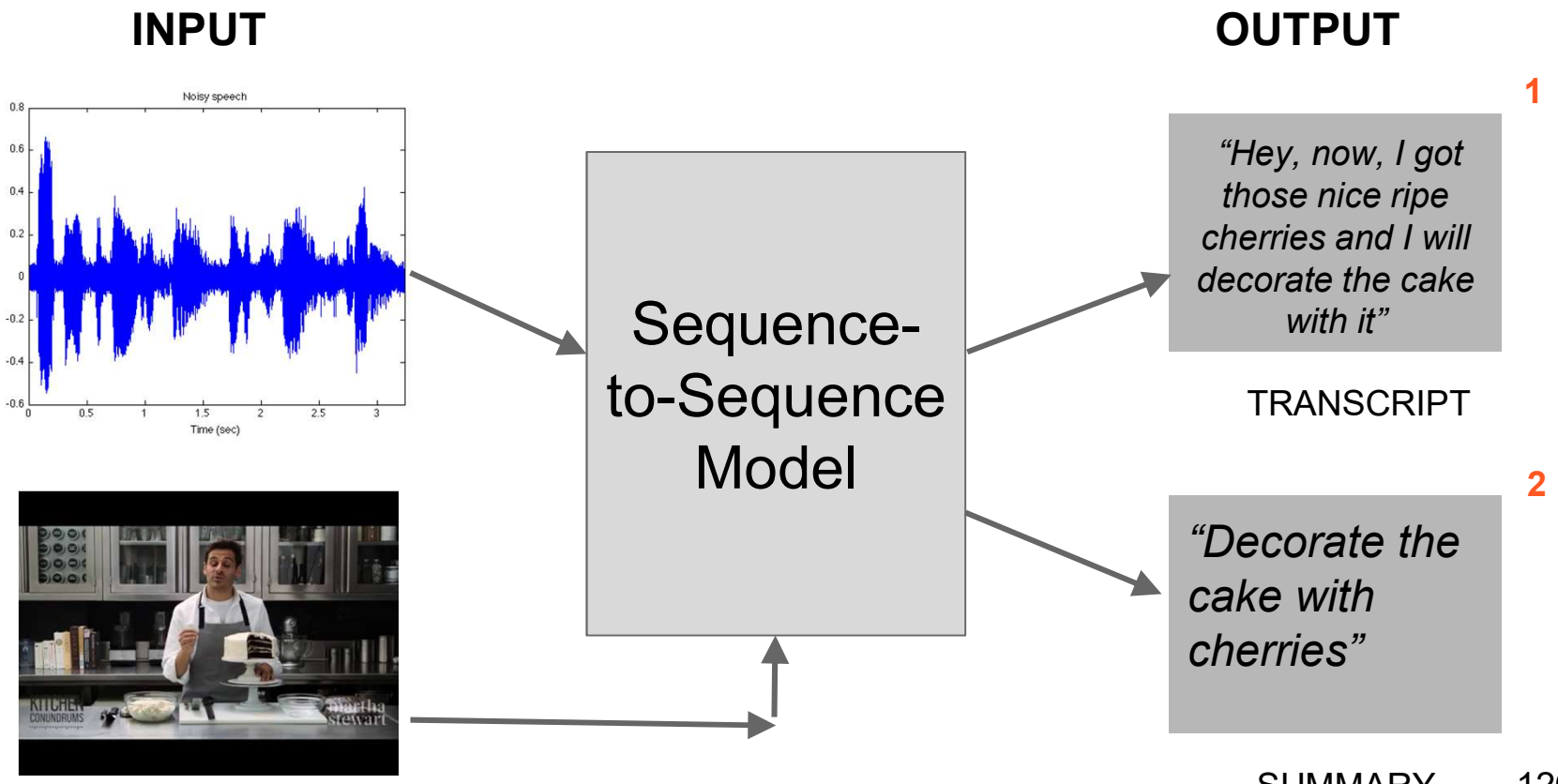
(your host for the next 5 minutes: Florian Metze)



Dataset: “How-to” videos corpus

- “How-to” dataset of instructional videos
- “Utterance” is 8-10 seconds, on average 18 words
- Over 480 hours of videos with **subtitles** [1]
  - 90h align well with audio (**transcripts**) [2]
  - 360h sub-titled with abstractive **summarization**
- Visual features like object/ place detection, or action recognition provide **context** vector
  - Assume context to be static over utterance

V. Using video as side-information in ASR



Baselines and Project Goals

(Your host for the next 5 minutes: Shruti Palaskar)

1. **Goal 0**: "Traditional" **AM+LM ASR Baseline** using CTC and Char RNNLM

   HMM-DNN and WFST+Word-RNNLM [Miao & Metze '16, Gupta et al. '17]

   "**It works**": <20% WER on 4h test set, >5% AV improvement

   Separate training and adaptation of AM and LM (both help!)

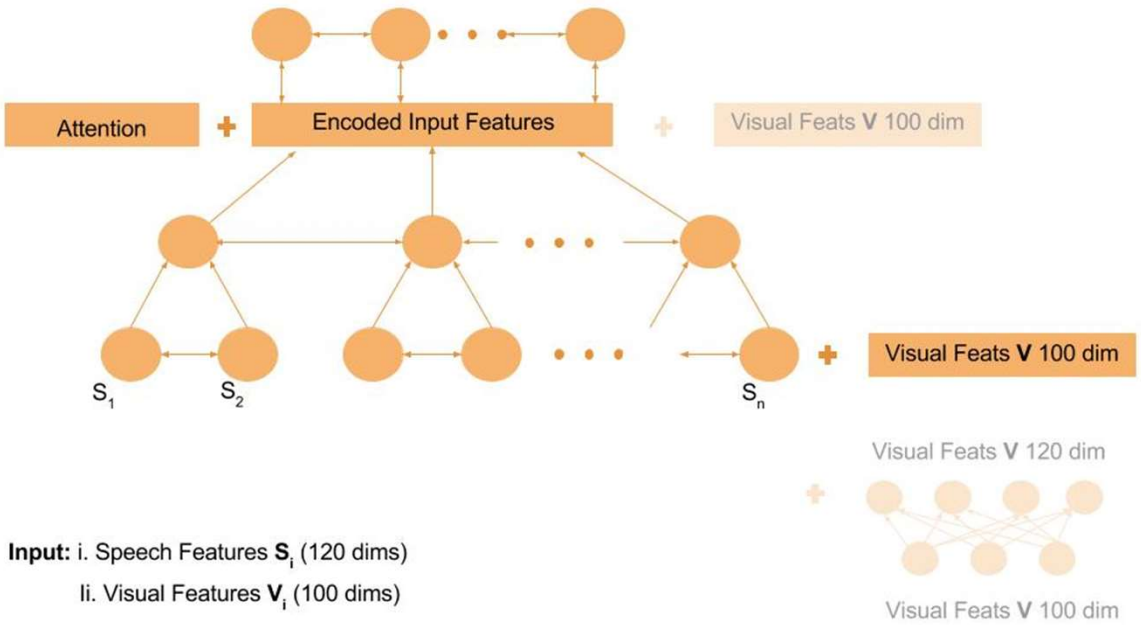
2. **Goal 1:** Seq-2-Seq (Audio-Visual) **Speech Recognition** (90 hrs)

   Explore multiple ways to adapt Seq2Seq model to context

3. **Goal 2:** Seq-2-Seq Audio-Visual **Speech Summarization** (up to 480 hrs)

   Leverage large amounts of data

Proposed Model: Seq-2-Seq with Attention



Input: i. Speech Features $\mathbf{S}_i$ (120 dims)

Ii. Visual Features $\mathbf{V}_i$ (100 dims)

BiLSTM cells in encoder and decoder

ADVANTAGES

1. **Pyramidal encoder** - more effective for longer utterances
2. Capable of jointly learning to recognize and summarize
3. Directly go to characters instead of phones
4. End-to-End approach - AM+LM modeled together!

Results: Seq-2-Seq Speech Recognition (Audio only)

- Why?
  - Speech only baseline
  - Does the model work?
  - Compare results on a well established task - WSJ
- 10 hours WSJ data
- 71 phone units

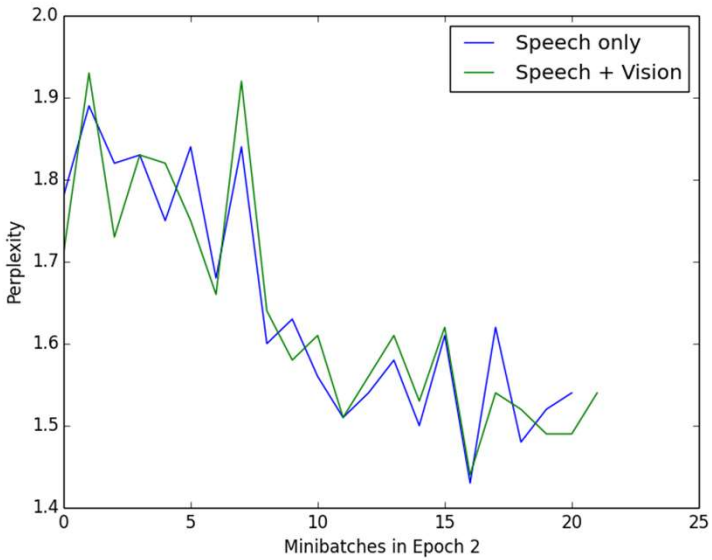
- 45 hours of “how-to” data
- 87 phone units

Data	Model	Train PPL	Val PPL	PER
WSJ	SGD, LR=0.01	~1.4	3.46	-
	SGD, LR=0.1	~1.09	1.65	17.25%
How-to	SGD, LR=0.05	~2.1	3.12	-
	SGD, LR=0.1	~3.5	4.67	-
	Adam, LR=0.0002	~1.45	2.02	29.02%

[Using OpenNMT toolkit]

* parameters not optimized fully

Does adding visual features help? Yes!



Training perplexity in Epoch 2

- Character model
- 43 character units

	Audio only	Audio-Visual
Epoch 1	2.06	2.14
Epoch 2	1.69	1.67
CER	39.43%	**35.69%**

Validation Perplexity on Audio only & Audio-Visual

Adaptation Technique: Concatenation at input

[Using OpenNMT toolkit]

Conclusion and Future Work

- We are fusing speech-to-text and image-to-text techniques to improve speech recognition and summarization
- We are getting it to work!
  - Try different adaptation & fusion techniques, scale up to 480h, improve feature extraction
- Compare Seq2Seq models on recognition and summarization task
- This should really work nicely in low resource scenarios, to bootstrap recognition in a new language (e.g. one without orthography), allow sharing across languages, etc.

Coming up next…

Wrap up and perspectives

Starring (in order of appearance): Odette Scharenborg, Emmanuel Dupoux

Scientific main results

- (Almost) zero-shot adaptation possible
- Successful first attempt to discover units using an attention matrix
- TTS is a useful tool in the evaluation of discovered units of different types
- Synergies between end-to-end systems and unit discovery
- Several working proof-of-concept end-to-end demos
- Pipeline of metrics to systematically evaluate the different systems
- Summarization of speech and image without going through text is possible
- Ideas for new models and improvements

Deliverables

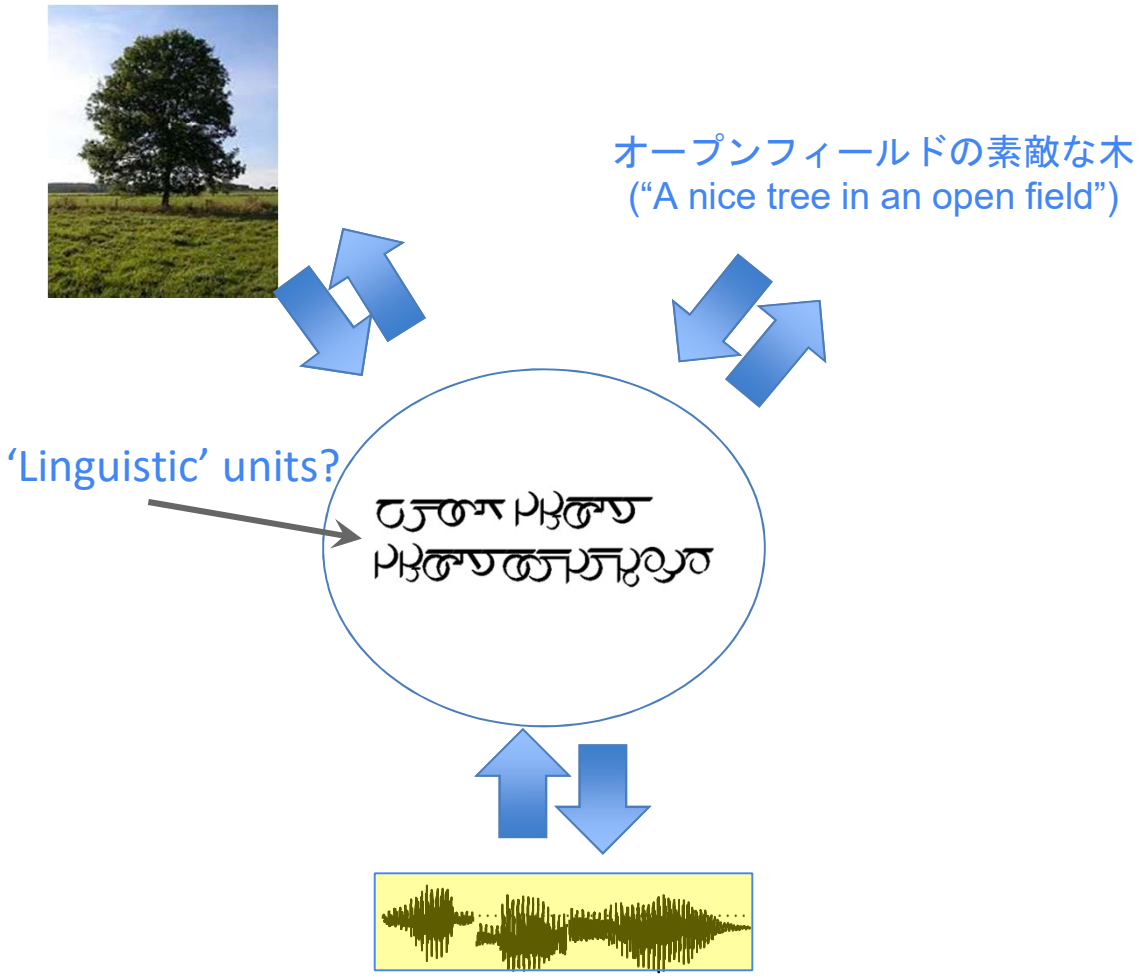
- Datasets
  - Trimodal flickrR (images, speech, (automatic) japanese translation)
  - SpeechCOCO (images, TTS, (automatic) japanese translation)
  - Mboshi (... under discussion)

- Github: https://github.com/JSALT-Rosetta/wiki
  - software (XNMT base, yaml, recipes), documentation, results
  - access to feature representations and Rosetta website

Main research question

Do we need intermediate symbolic units ?

If yes, to do what?



Coming up next

- JSALT report, conference and journal papers
- Satellite workshop “Grounding Language Understanding” @Interspeech August 25, 2017 (http://www.speech.kth.se/glu2017/)
- Watch out for a possible special session “Grounded Language Representation Learning” at ICASSP 2018
- Follow up meeting in Europe in December 2017/Jan 2018
  - finish up the experiments
  - continue the buildup of tools
  - task force on scalable datasets collections in unwritten languages/infant speech data
  - prepare a zero-resource Challenge 2019
  - new collaborations and ideas for grant proposals

Final words

A big thanks to the sponsors and organisers for making this wonderful event possible!

Especially: Sanjeev, Florian, Alan, Jae!

ありがとうございました
ευχαριστώ
አመሰግናለሁ
շնորհակալություն
ধন্যবাদ
Danke
Дзякуй
ကျေးဇူးတင်ပါတယ်
merci
謝謝
გმადლობთ
આભાર
धन्यवाद
תודה
bedankt
...

Thanks!