# Speech technology systems for unwritten languages

Dr. Odette Scharenborg

*Associate Professor & Delft Technology Fellow*
*Multimedia Computing Group*
*O.E.Scharenborg@tudelft.nl*
*OScharenborg*

TU Delft

# In collaboration with…

Emmanuel Dupoux, 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

TU Delft

# Acknowledgements


- 2017 Jelinek Summer Workshop on Speech and Language Technology (JSALT) at Carnegie Mellon University, Pittsburgh

- Organised by Johns Hopkins University and Carnegie Mellon University

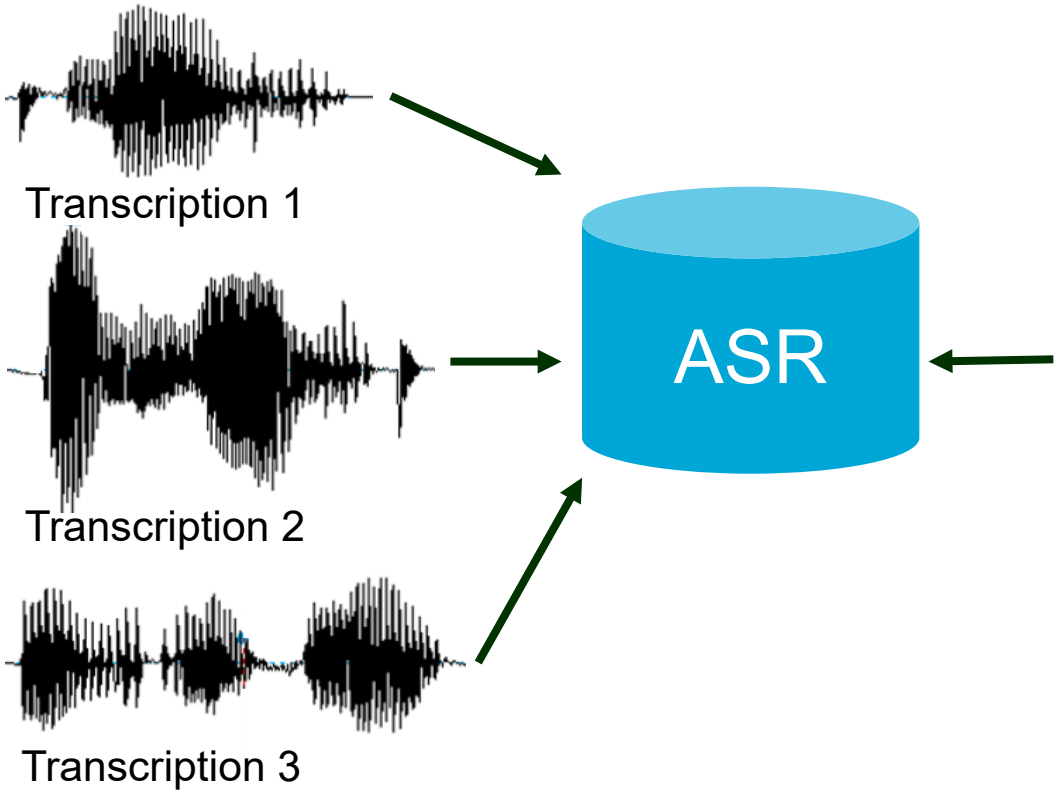
- Grants from Google, Microsoft, Amazon, Facebook, and Apple

- Computing: Blacklight system of XSEDE at the Pittsburgh Supercomputing Center

- More information:

Scharenborg et al. (2020). Speech technology for unwritten languages. IEEE/ACM Transactions on Audio, Speech and Language Processing, 28, 964-975. DOI:10.1109/TASLP.2020.2973896.

# High-resource speech and language technology

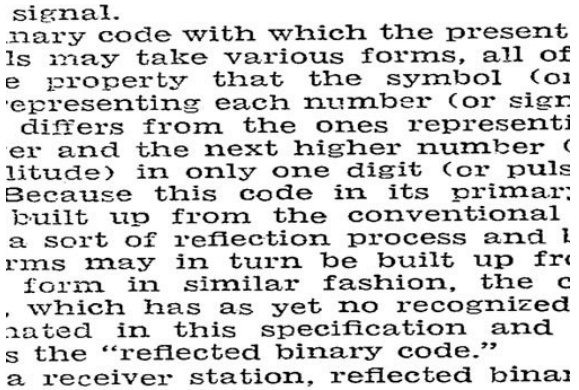



*Training audio data*

*Training text data*

# For many languages in the world….

Transcription



signal.
nary code with which the present
ls may take various forms, all of
e property that the symbol (or
epresenting each number (or sign
differs from the ones representi
er and the next higher number (
litude) in only one digit (or puls
Because this code in its primary
built up from the conventional
a sort of reflection process and
rms may in turn be built up fr
form in similar fashion, the c
which has as yet no recognized
nated in this specification and
s the "reflected binary code."
a receiver station, reflected binar

*Training audio data*

*Training text data*

# An estimated half of human languages…

- How to build an ASR system for low-resource languages and languages without orthography?
- Cannot use a high-resource language system for low-resource language → Cross-language ASR is pretty bad

# … but any 4-year-old child demonstrates

- Learn a language communication system **before** learning to read and write
- From raw sensory signals
- With limited human supervision

# Different technological approaches

- **Zero-resource approach:** discover acoustic/linguistic units of the low-resource language from the raw speech data

- **Multi-language approach:** try to create universal or cross-linguistic ASR systems by training on multiple languages

# … but children

- Auditory information

- Visual information

# New technological approaches

- Use visual information (images) to discover word-like units from the speech signal, using speech-image associations [Hawarth & Glass, 2015; Harwarth et al. 2016; Chrupala et al., 2017]

⇨ "The Speaking Rosetta Stone": one step further

→ Add translations in a high-resource language:

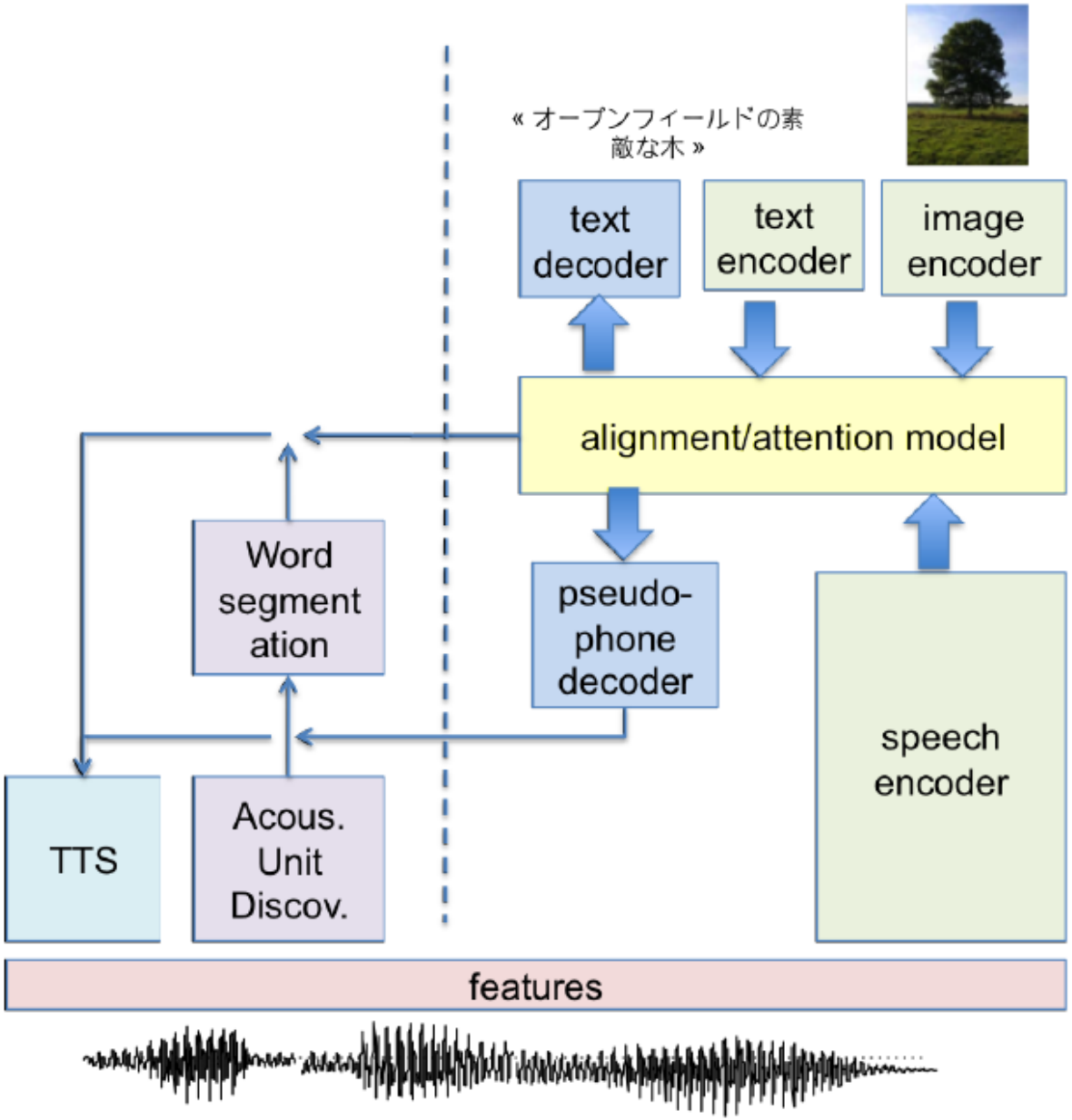
- Parallel data between speech from an unwritten language and translations in another language can easily be collected, often done by field linguists

# “The Speaking Rosetta Stone” project

1. Discover intermediate symbolic units
2. Investigate their role in building SLT systems

**4 tasks:**

- Symbolic units:
  - Unit discovery
  - Speech synthesis
- End-to-end tasks without the need for explicit symbolic units
  - Speech2image
  - Speech2translation




"A nice tree in an open field"

# Trimodal dataset: FlickR_8K

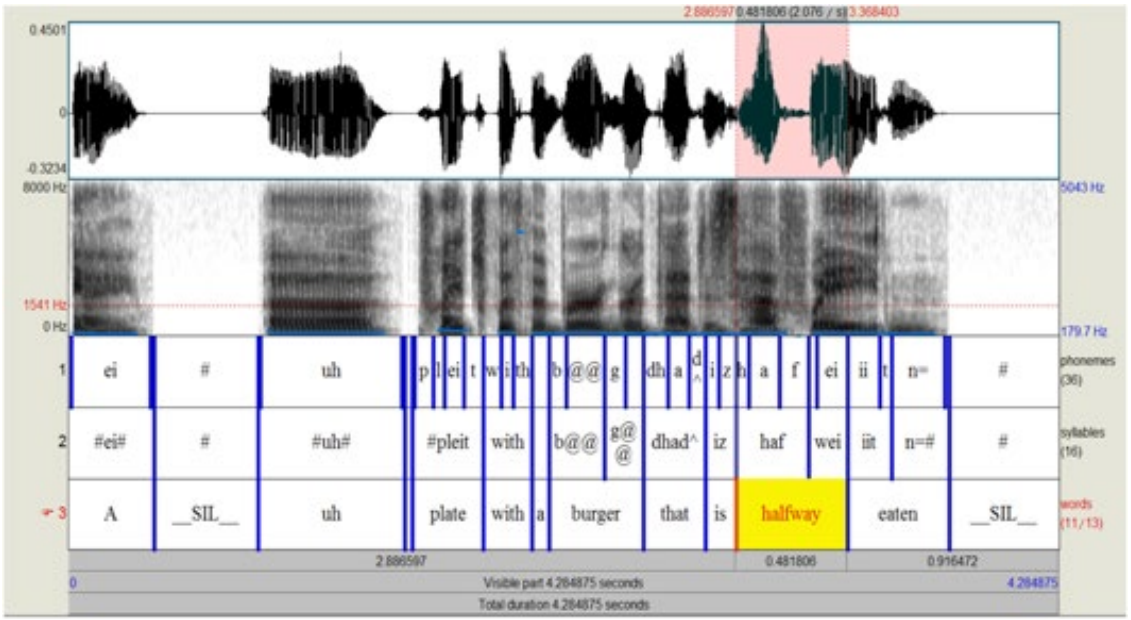
- AMT recordings - 40k spoken captions available online [Harwath & Glass, 2015]
- JSALT 2017: Japanese translations (Google MT) for all captions (+ tokenization)

-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。

# Trimodal dataset: Augmented MSCOCO-synthetic

- 120K images, 5 captions/image (=text) + TTS in English (8 different voices; = 600h of speech) + 600K Japanese translations (Google) + syllable and phonetic alignments of speech + Japanese tokenization and POS [Havard, Besacier & Rosec, 2017]



# ‘Unwritten’ languages databases

- Mboshi: 5K spoken utterances + linguists’ transcriptions + phonetic alignments + French translations
- Basaa: 3K spoken utterances + linguists’ transcriptions + phonetic alignments + French translations [Adda et al., 2016]

Bantu languages from Sub-Saharan Africa

# End-to-end sytem: XNMT

- Sequence-to-sequence neural machine translation toolkit [G. Neubig]
- Greatly improved during the course of this project
- Reads in a sequence of (variable-length) inputs
- Generates a different sequence of (variable-length) output
- Available via: https://github.com/neulab/xnmt/

# Tasks with symbolic units

**Unit discovery**

→ Discover phoneme-like acoustic units from raw speech using out-of-domain languages to help unit discovery through (almost) zero-shot adaptation

- Use Bayesian models for unsupervised pseudo-phone discovery: information from high-resource languages is integrated by means of an informative prior [Ondel et al., 2018]
- Use data from multiple languages for universal articulatory features and phoneme inventory discovery [Mueller et al., 2017]

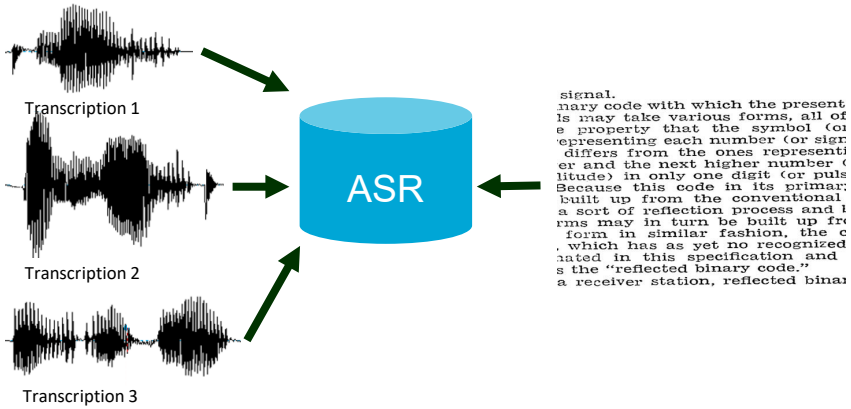
## Cross-language definition of acoustic units

Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system [Scharenborg et al., 2017, 2018]

# Assumptions

- Unlabeled speech data in the low-resource language
- ‘Description’ of the phone(me) inventory of the low-resource language from field linguistics
- ASR system trained on a well-resourced language of the same language family, although…..

*high-resource language*




*low-resource language*

Ø ɑ
ɘ ʌ
ϴ ʊ

# Results Dutch to English

- Limited improvement in phone error rate after retraining the acoustic units on the self-labelled utterances

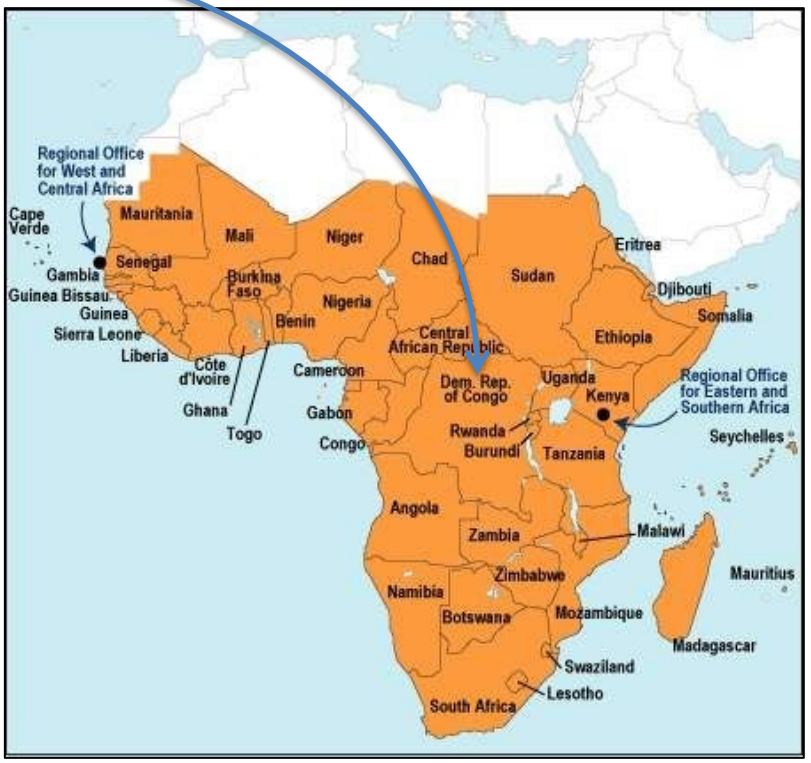
Encouraging …

# Next steps

- Does the proposed methodology work on an unrelated language pair: Dutch to Mboshi?
  (an actual low-resource language!)

# Dutch and Mboshi

- Dutch: Germanic language, 550K+ words, ~64h
- Mboshi: Bantu language, 2.3K utterances, ~2h



# Step 1: Train FF-DNN for Dutch

- Input: 40-dimensional FBank features
- 6 fully-connected hidden layers with 1024 units trained using sigmoid nonlinearities
- Output: one unit with soft-max activation function for each Dutch phone
- Frame accuracy Dutch on Dutch: 71.7%

# Step 2: Cross-language definition of acoustic units

1. Remove phones of Dutch not present in Mboshi from the soft-max layer

2. Add Mboshi phones not present in Dutch, by adapting the softmax layer using linguistic knowledge

# Initialization of the ‘missing’ Mboshi phones using linguistic knowledge

| L2 phone | Mapping | | | L2 phone | Mapping | | |
|---|---|---|---|---|---|---|---|
| | L1:1 | L1:2 | L1:3 | | L1:1 | L1:2 | L1:3 |
| $^{m}b^{v}$ | b | .5(m-v) | b | $^{n}d$ | d | n | d |
| $b^{v}$ | b | v | b | $p^{f}$ | p | f | p |
| $^{ŋ}g$ | g | ŋ | g | ß | b | v | b |
| $^{m}b$ | b | m | b | $^{m}w$ | w | m | w |

Linear extrapolation:

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

# Step 3: Retraining with self-labels

1. Run a free phone recognition pass on the Mboshi speech data

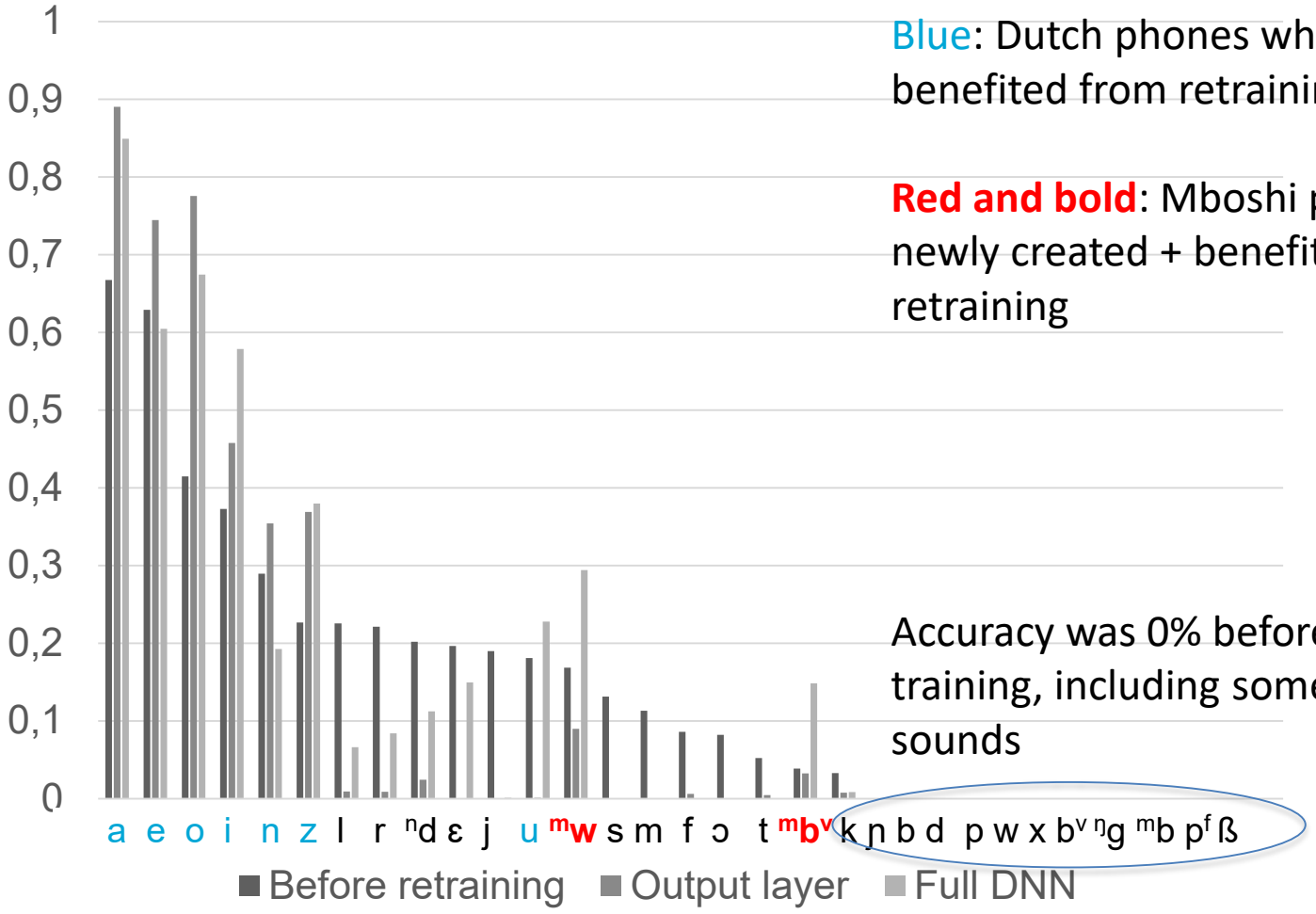
2. Retrain the model with these self-labels:
   - Only the output layer
   - The full model

# Results: frame classification accuracy

- Evaluation on ‘gold standard’
- Training with self-labels

| Retraining | Train | Test |
|---|---|---|
| Baseline | 31.3 | 32.2 |
| Output layer only | 36.4 | **37.9** |
| Full model | 36.4 | 37.6 |

# Results: per phoneme category




Blue: Dutch phones which benefited from retraining

**Red and bold**: Mboshi phones, newly created + benefited from retraining

Accuracy was 0% before and after training, including some Dutch sounds

# Conclusions

- Improvement > 5% after retraining with Mboshi self-labels
  - Approach is usable for unrelated language pairs
  - No large differences between training methods
  - Mixed results for individual phonemes, *why?*

# Tasks with symbolic units (contd.)

**"TTS without T"**

- Generate speech from unit sequences (because language without orthography)
- Evaluate quality of units
- CMU Clustergen Parametric Speech Synthesizer [Black 2006]

**Input:** waveform file plus symbolic sequences of "words" or "phones"

**Output:** Simple synthesizer and distortion measure on held out data

Issues:

- People don't deliver in consistent "reading" style
- Often you have multiple, rather than a single speaker
- Typically poor recordings in poor environment
- Typically only a small amount of data

# Proof-of-concept: Image2speech

Generate spoken captions for images
[Hasegawa-Johnson et al., 2017]

- Flickr8k:

Ref #1: A boy +laugh+ in a blue top +laugh+ is jumping off some rocks in the woods.

Ref #2: A boy +um+ jumps off a tan rock.

Network: 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.

*(A man is jumping in the forest)*

# Example 2

- MSCOCO:

Ref #1: A, a black and white photo of a fire hydrant near a building.

Ref #2: Aa, a fire hydrant that is out next to a house.

Network: # @ 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 # #

*(A person working at a metal down, and a red fire hydrant)*

# Example 2

- MSCOCO:

Ref #1: A, a black and white photo of a fire hydrant near a building.

Ref #2: Aa, a fire hydrant that is out next to a house.

Network: # **@ 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 # #

*(**A person working at a metal down**, and a red fire hydrant)*

# PoC: Foreign-text-2-speech (NMT + TTS)

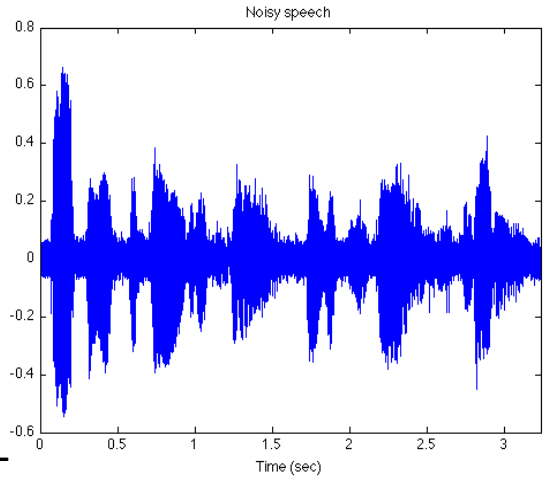
Translate French words (text) into Mboshi phones which were either true phones or pseudo-phones obtained via AUD [L. Besacier]

Example: NMT was trained to generate true Mboshi phones (not truly zero-shot)

- French: ils sont devenus vieux
- *(English: they became old)*
- Mboshi (ref): b á a á b h i á i d u n ú
- Mboshi (hyp): b á a á b h i á i d u n ú

# Speech and image to text (and summarization)

Use multi-modal information consisting of speech and videos to improve standard (supervised) ASR [Palaskar et al., 2018]




INPUT

Sequence-to-Sequence Model

"Hey, now, I got those nice ripe cherries and I will decorate the cake with it"

TRANSCRIPT

*"Decorate the cake with cherries"*

SUMMARY

# End-to-end task: Speech-to-translation

- Translation from raw speech without any intermediate transcription [Bérard et al., 2011]
- BLEU4 scores at character level on development sets: [L. Besacier & P. Godard]
  - FlickR English-to-Japanese: 30.99%
  - Mboshi-to-French: 22.36%

→ End-to-end models are able to encode some regularities in the speech signal in order to decode predictable sequences of characters in a target language

**Attention-based system:** [L. Besacier & P. Godard]

- Word segmentation from a pseudo-phone sequence obtained using AUD
- Word boundary detection performance:

| System | Precision | Recall | F |
|---|---|---|---|
| Segmental DTW [Jansen & Van Durme, 2011] | 31.9 | 13.8 | 19.3 |
| Attention (French to Mboshi) | 36.5 | 46.1 | 40.7 |
| Attention (Mboshi to French) | 36.3 | 46.6 | 40.8 |

# End-to-end task: Speech-to-image

- Retrieve an image given a spoken caption [Hawarth & Glass, 2015; Harwarth et al. 2016; Chrupala et al., 2017]

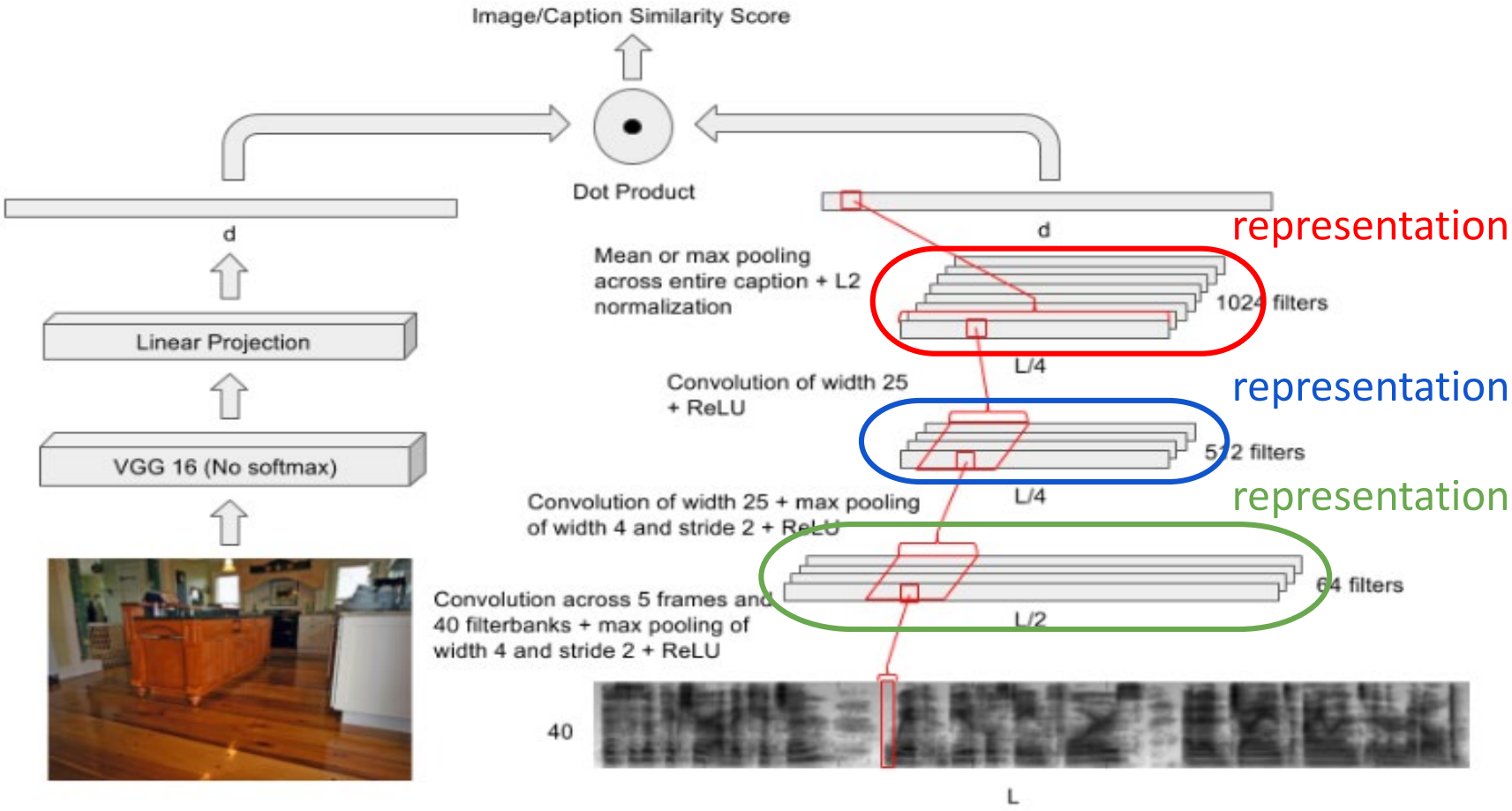
→ Unsupervised discovery of speech units using multi-modal input ~ children

- Implementation based on [Harwarth et al., 2016]

- Four types of acoustic features were compared

# The “Harwath Encoder:” CNN

Use image context to extract better speech unit (phones, words, phrases, etc.)



# Results

- During workshop [Scharenborg et al., 2018]

| Feature type | Recall@1 | Recall@5 | Recall@10 |
|---|---|---|---|
| Mel-filterbank | 0.96 | 4.70 | 8.56 |
| Multilingual Bottleneck | **1.30** | **5.30** | **9.94** |
| AUD (one epoch) | 0.12 | 0.44 | 1.12 |
| Cochleagram | 0.08 | 0.50 | 1.04 |

- After workshop [D. Merkx]

| Feature type | Recall@1 | Recall@5 | Recall@10 |
|---|---|---|---|
| Mel-filterbank | **1.52** | **6.28** | **10.40** |
| Multilingual Bottleneck | 1.30 | 5.30 | 9.94 |
| MFCC | **7.34** | **21.8** | **32.1** |

# Scientific main results

- (Almost) zero-shot adaptation possible: Discovered units are meaningful as shown by their usefulness in upstream tasks
- Synergies between end-to-end systems and unit discovery
- First attempt to discover spoken terms from speech using an attention matrix
- TTS is a useful tool in the evaluation of discovered units of different types
- Several working proof-of-concept end-to-end demos
- Summarization of speech and videos 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

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

# Concluding remarks

- The “Speaking Rosetta” JSALT 2017 project: “Unsupervised multi-modal language acquisition”

- It is possible to build useful SLT systems without any textual resources in the language for which the SLT is built, in a way that is similar to that of how infants learn a language