Aalto University
School of Electrical
Engineering

Improved subword modeling for WFST-based speech recognition

Peter Smit, Sami Virpioja, Mikko Kurimo

Aalto University, Department of Signal Processing and Acoustics

August 23, 2017

Research questions

- How to do sound WFST modeling for subwords?
- How to reconstruct words from subwords?
- What is a good subword vocabulary?
  - Size of vocabulary?
  - Segmentation method?

How big is your vocabulary?

	# Word forms
WSJ small LM	5.000
WSJ big LM	20.000
Native English Speaker	20.000 – 35.000
CMU dict	134.000

How big is your vocabulary?

	# Word forms
WSJ small LM	5.000
WSJ big LM	20.000
Native English Speaker	20.000 – 35.000
CMU dict	134.000
Finnish Adult	>1.000.000
Finnish Text Collection	>4.000.000

Is a big vocabulary a problem?

- Current systems do support vocabularies >4M

But:

- Out of vocabulary problems
- Data sparsity – valid words might only appear once
- Dimensionality problems (e.g. RNNLM input/output layers)

Subword modeling

- Split words into smaller units
- Reduces vocabulary size
- Split either knowlegde-driven (e.g. grammatical morphs) or data-driven (e.g. Morfessor)

Subword marking and reconstruction

Style (abbreviation)	Example
boundary tag (<w>)	<w> two <w> slipp er s <w>
left-marked (+m)	two slipp +er +s
right-marked (m+)	two slipp+ er+ s
left+right-marked (+m+)	two slipp+ +er+ +s

Subword marking and reconstruction

Style (abbreviation)	Example
boundary tag (<w>)	<w> two <w> slipp er s <w>
left-marked (+m)	two slipp +er +s
right-marked (m+)	two slipp+ er+ s
left+right-marked (+m+)	two slipp+ +er+ +s

- two <w> slipp er s <w>
- Vocab size $V + 1$

Subword marking and reconstruction

Style (abbreviation)	Example
boundary tag (<w>)	<w> two <w> slipp er s <w>
left-marked (+m)	two slipp +er +s
right-marked (m+)	two slipp+ er+ s
left+right-marked (+m+)	two slipp+ +er+ +s

- +two slipp +er +s
- Vocab size $2V$

Subword marking and reconstruction

Style (abbreviation)	Example
boundary tag (<w>)	<w> two <w> slipp er s <w>
left-marked (+m)	two slipp +er +s
right-marked (m+)	two slipp+ er+ s
left+right-marked (+m+)	two slipp+ +er+ +s

- two slipp +er +s
- Vocab size $4V$

Subword problems

- Restricting output of decoder to be valid (don't start or end a sentence halfway a word)

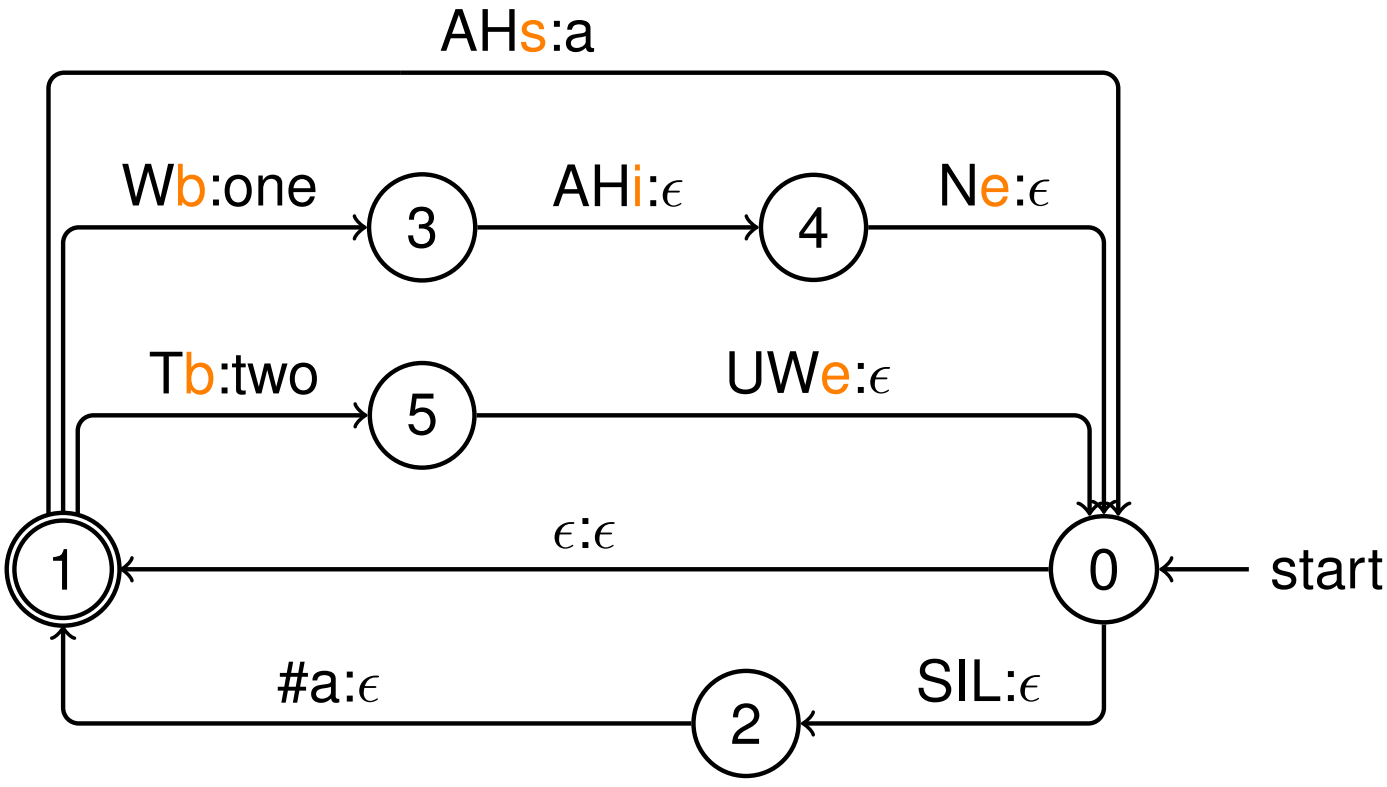
  two slip+ +per+ +s

  +two slip+ per+ +s

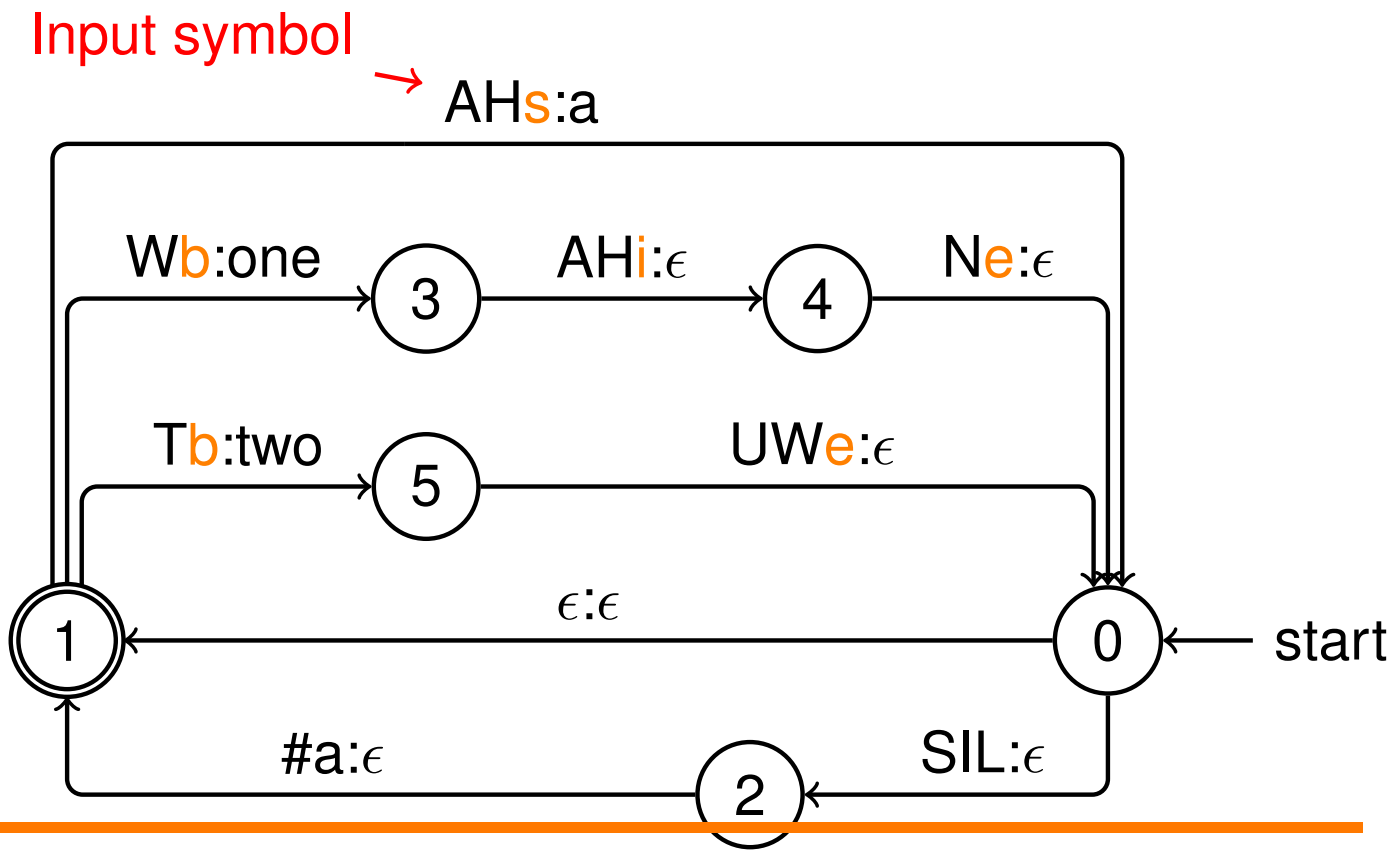
- Word-position dependent phonemes

- Longer contexts are needed in language modeling

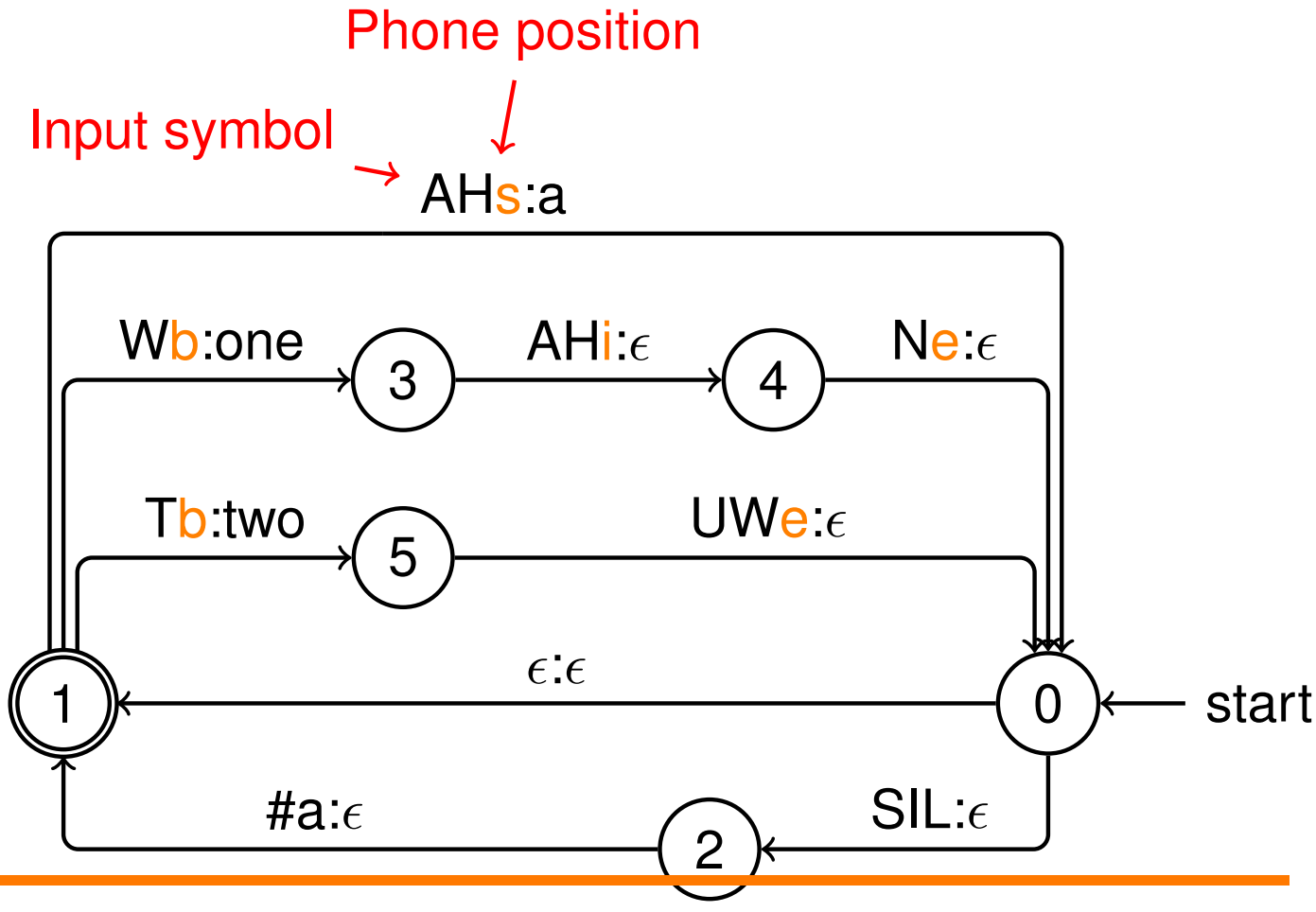
Original Lexicon FST (kaldi)



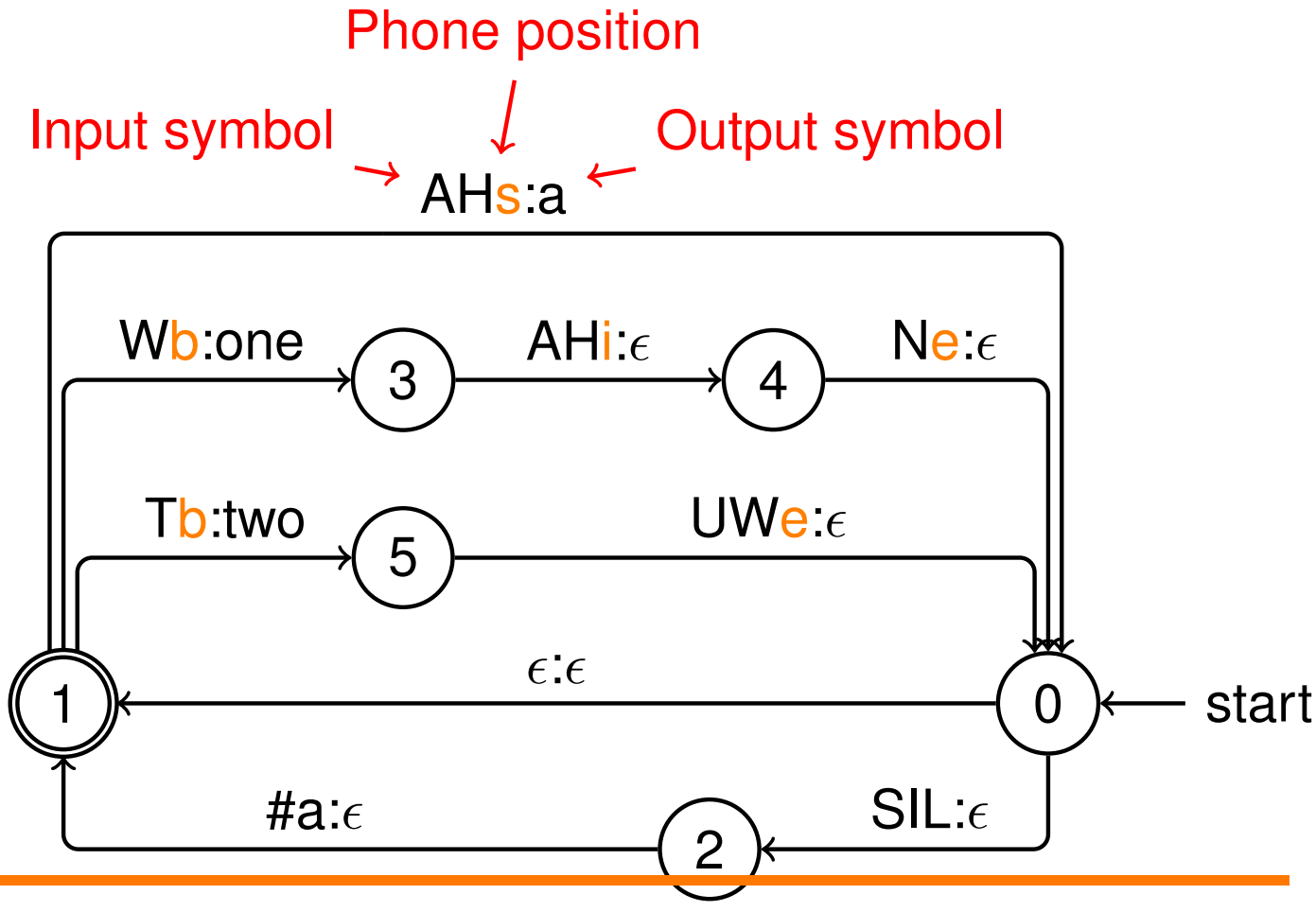
Original Lexicon FST (kaldi)



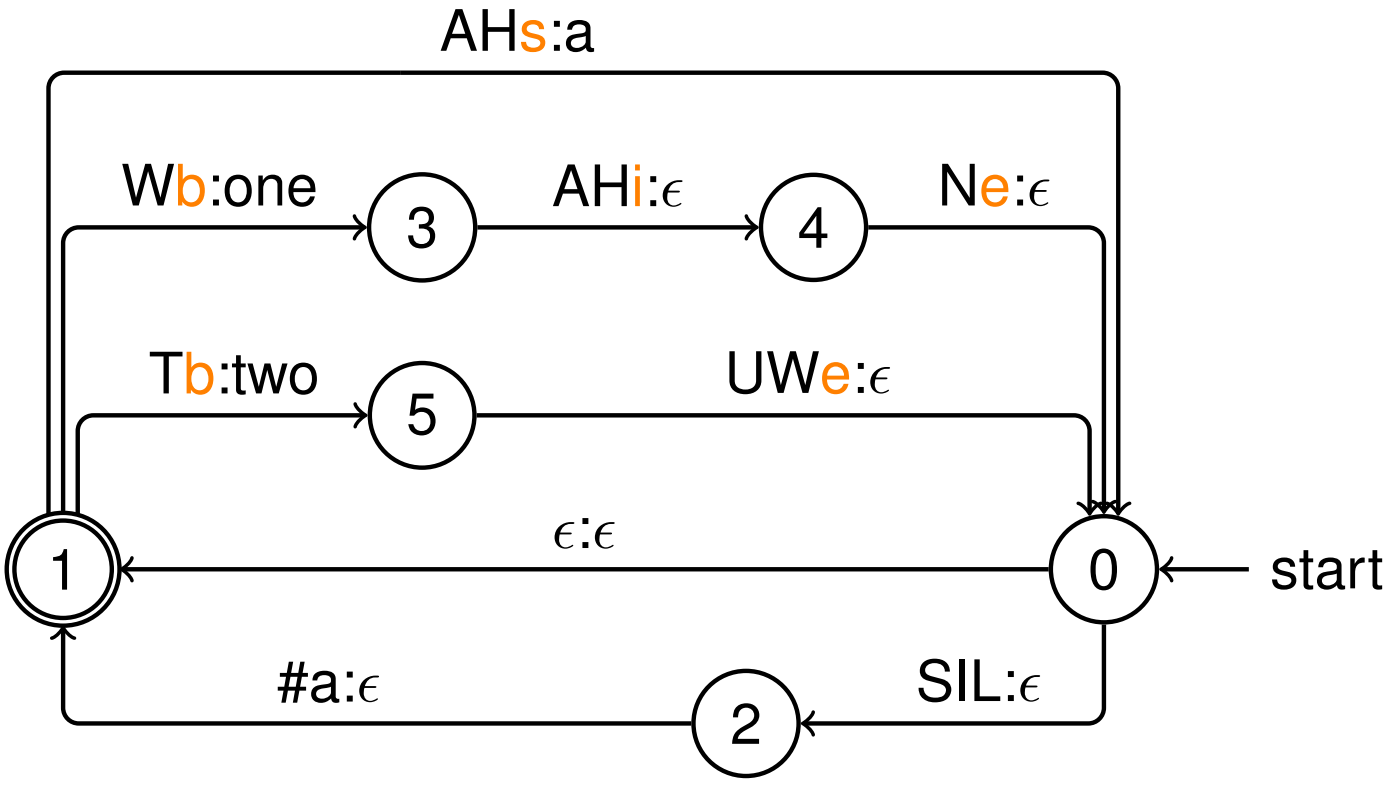
Original Lexicon FST (kaldi)



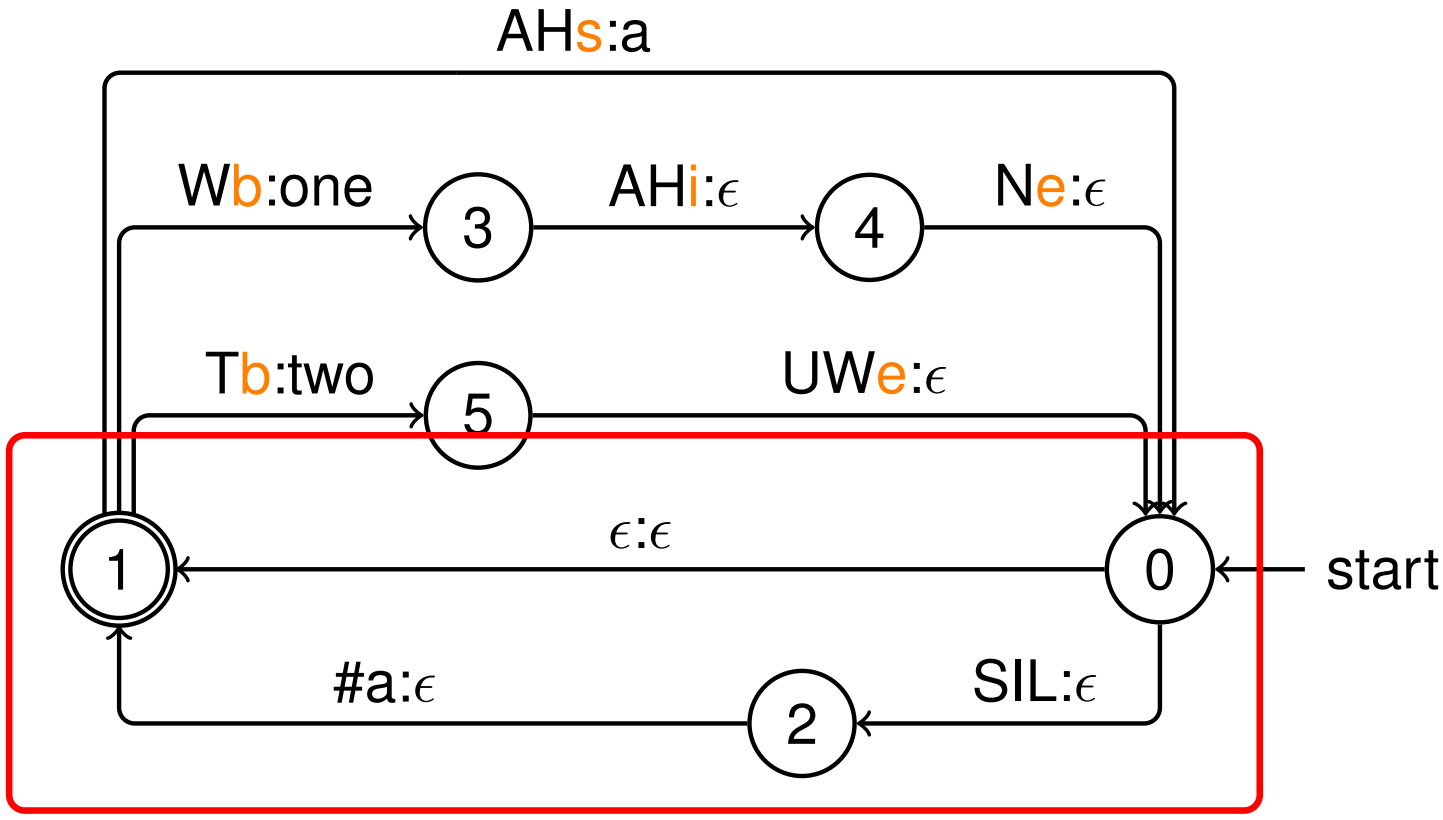
Original Lexicon FST (kaldi)



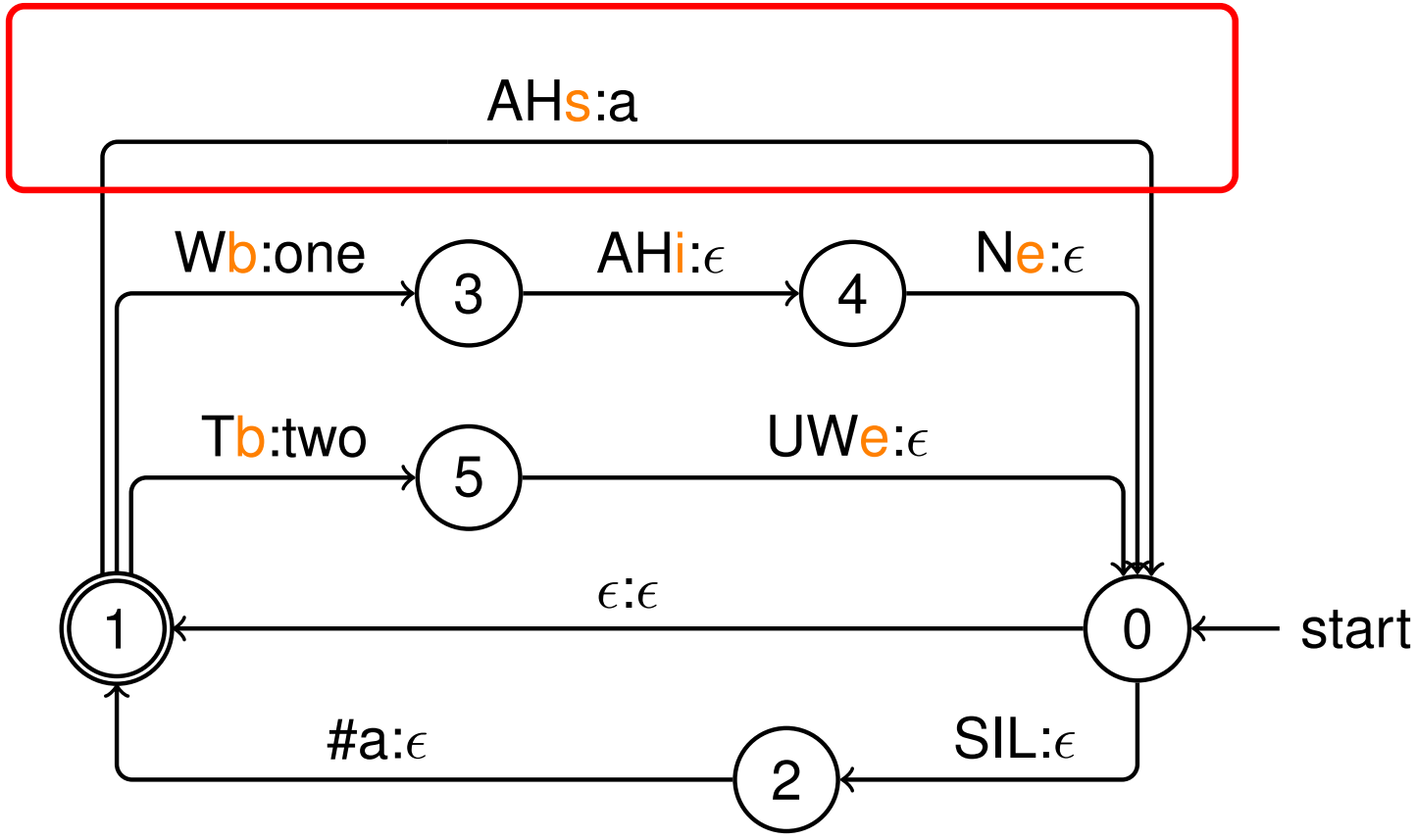
Original Lexicon FST (kaldi)



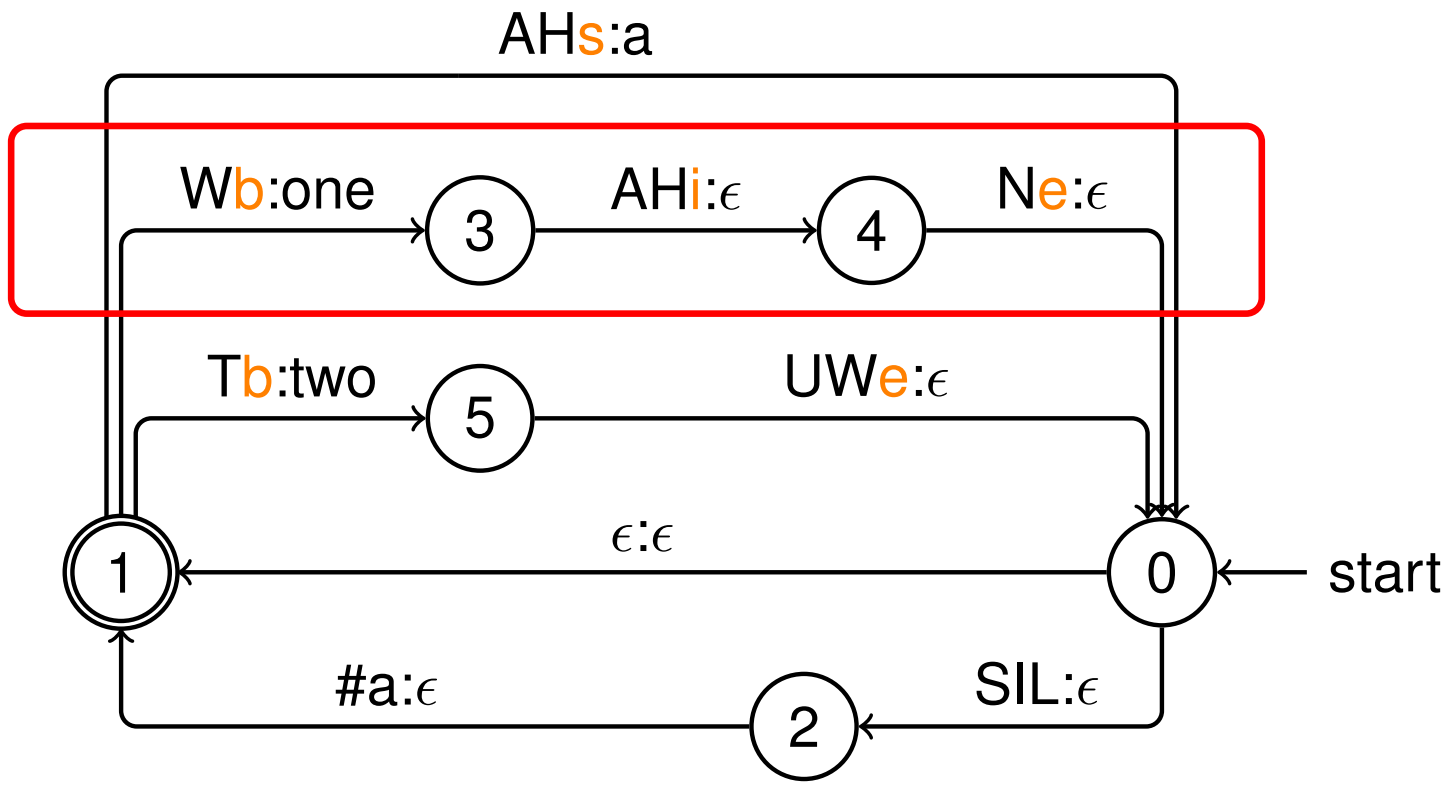
Original Lexicon FST (kaldi)



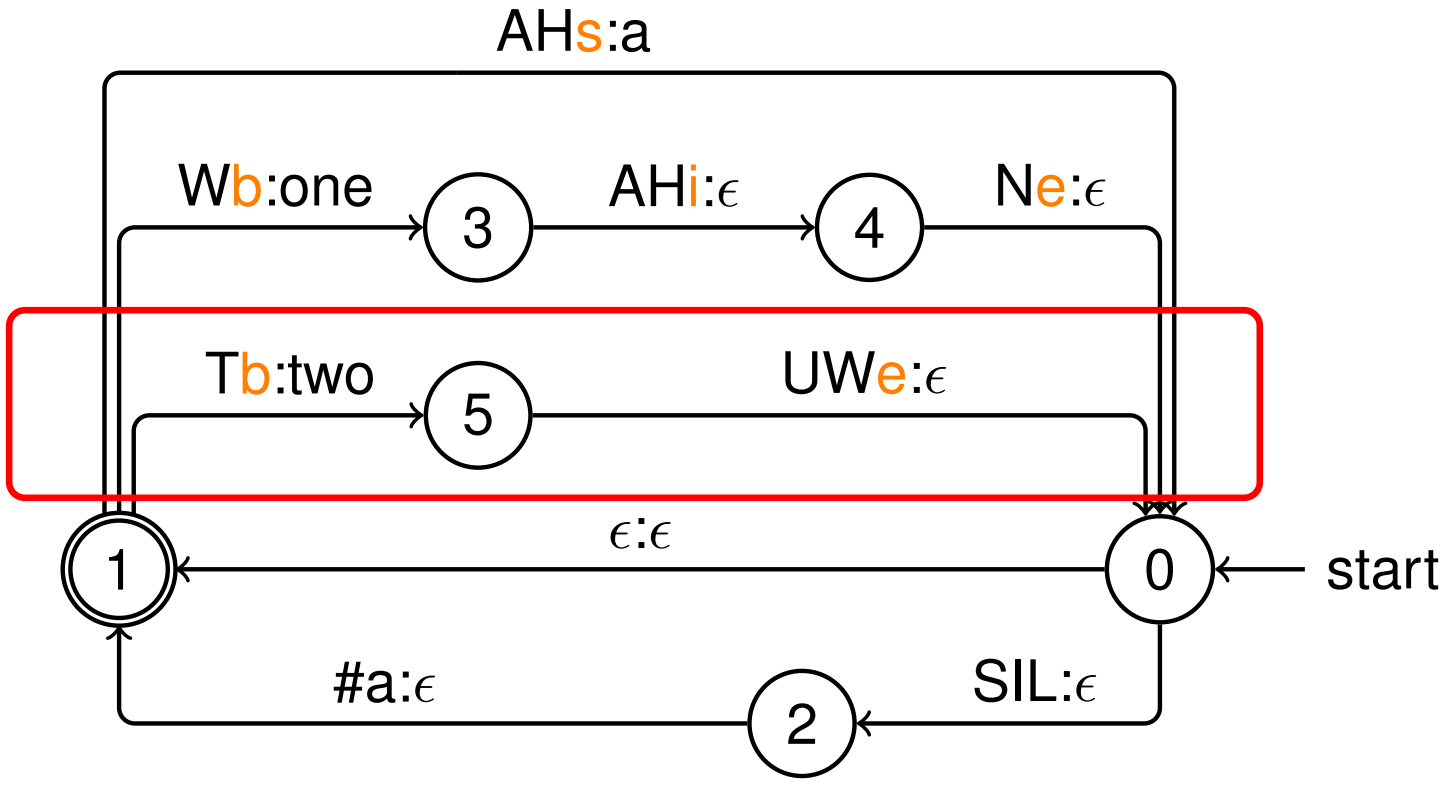
Original Lexicon FST (kaldi)



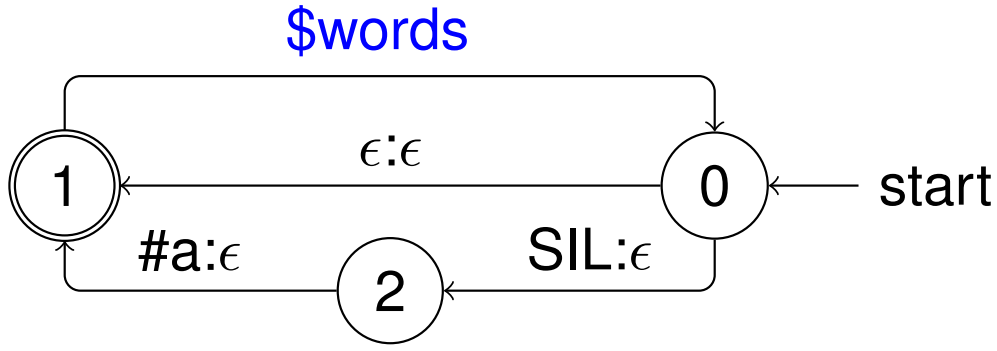
Original Lexicon FST (kaldi)

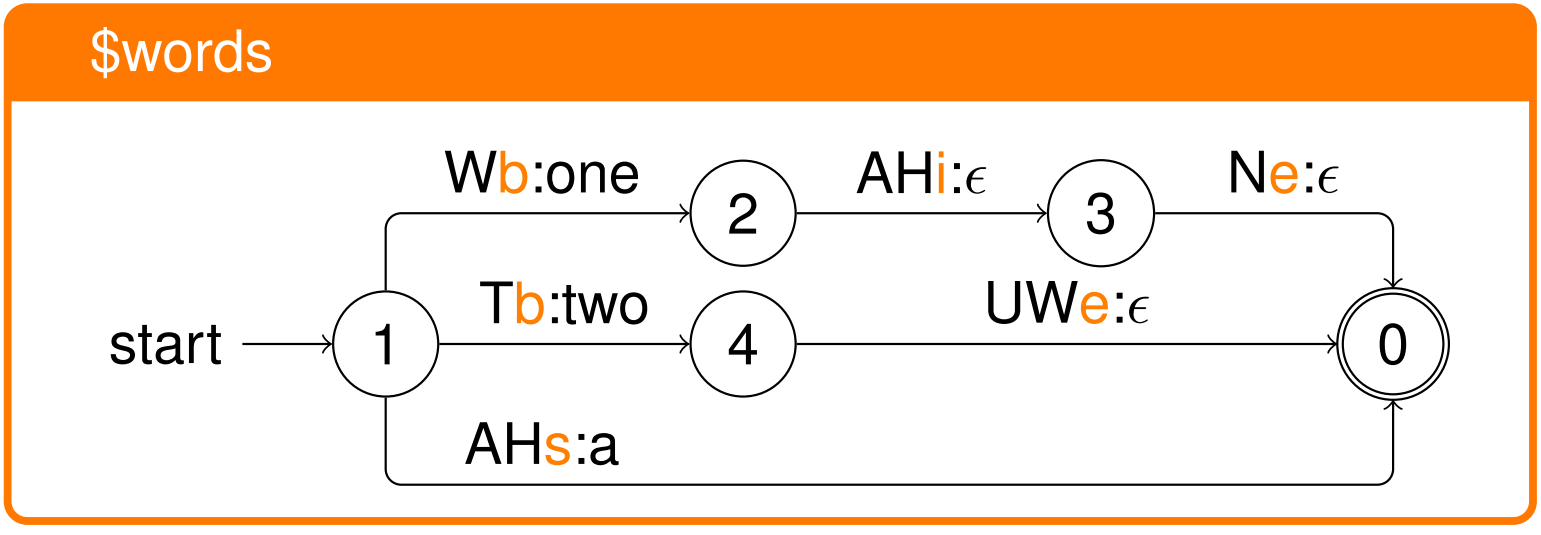


Original Lexicon FST (kaldi)

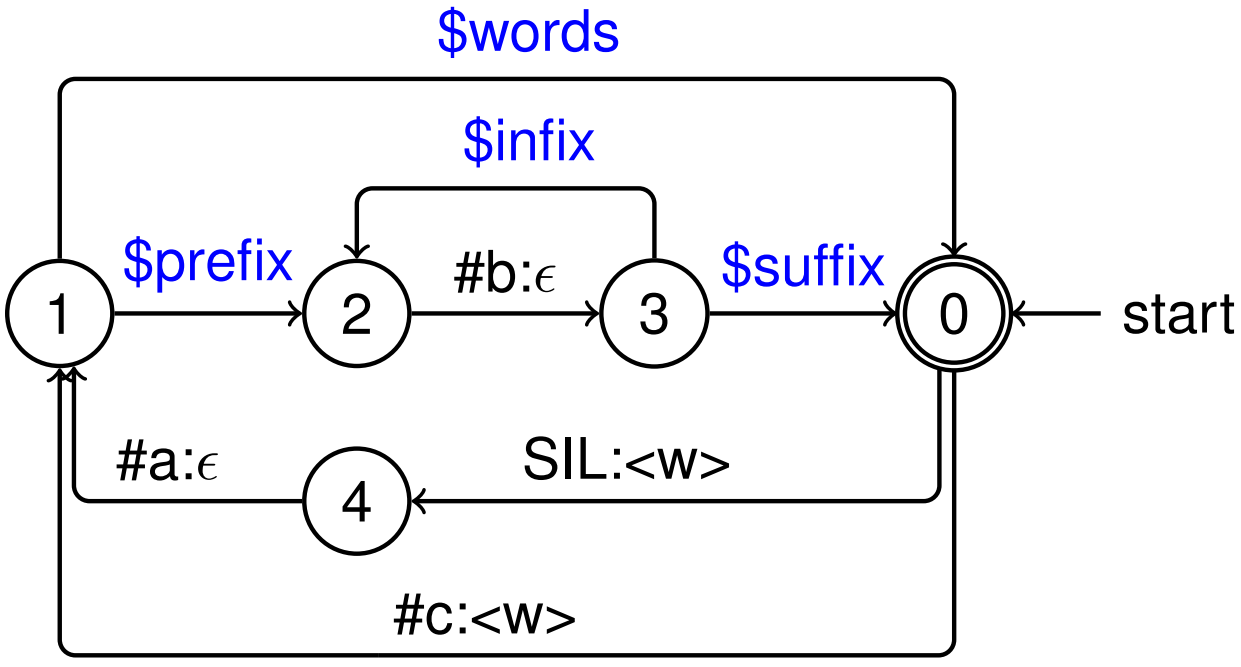


Original Lexicon FST (kaldi)





Subword Lexicon FST



Replace FST's <w>: <w> two <w> slipp er s <w>

$words

two	Tb UWe
slipp	Sb Li lHi Pe
er	ERs
s	Zs

$prefix

two	Tb UWi
slipp	Sb Li lHi Pi
er	ERb
s	Zb

$suffix

two	Ti UWe
slipp	Si Li lHi Pe
er	ERe
s	Ze

$infix

two	Ti UWi
slipp	Si Li lHi Pi
er	ERi
s	Zi

Replace FST's m+: two slipp+ er+ s

$words

two	Tb UWe
s	Zs

$prefix

slipp+	Sb Li IHi Pi
er+	ERs

$suffix

two	Ti UWe
s	Ze

$infix

slipp+	Si Li IHi Pi
er+	ERi

Experiment Setup

- AM: Finnish, Kaldi, TDNN, 150 hours, 425 speakers, clean read data (SPEECON)
- LM: Variable-order n-gram, Finnish Text Collection, 150M tokens, 4M word forms
- Test1: READ, SPEECON, clean, read, 20 speakers, 1 hours
- Test2: NEWS, Broadcast news, 5-10 speakers, 5 hours
- More experiments in the paper

Results – Different marking styles

Word Error Rate (%) devset
Morfessor segmentation

Optimized vocabulary size

	NEWS	READ
Word	23.73	8.60
Naive +m	24.11	9.70
Naive +m+	25.45	9.10
Proposed <w>	**22.89**	6.62
Proposed +m+	22.96	**6.55**
Proposed +m	23.47	7.12
Proposed m+	23.79	7.24

Results – Different segmentation methods

Word Error Rate (%) devset

Subword vocab size	News 5k	News 10k	News 15k
Morfessor	23.02	22.82	**22.79**
Greedy Unigram	23.06	22.93	23.02
Byte Pair Encoding	23.18	23.17	23.17

- Only minor differences between segmentation methods

Comparison previous results on Finnish and Estonian datasets

	Eval-sets		
	Word	**Subword**	**Previous best**
et-bn-ak	**17.48**	18.28	
et-bn-er	8.36	**7.70**	8.2 (Alumäe, 2014)
fi-news	25.49	**24.98**	28.9 (Varjokallio, 2017)
fi-phone	14.07	**12.79**	21.88 (Varjokallio, 2013)
fi-read	11.11	**9.44**	13.3 (Alumäe, 2014)

Recap

- Subword modeling is beneficial (or required) for languages with large vocabulary.
- In a WFST-based decoder the lexicon FST can be modified such that only valid subword sequences are allowed and position-dependent-phones are preserved.
- Experimental results show up to 23% improvement over word modeling, 28% improvement over naive subword modelling.
- The optimal marking style for subwords depends on language or dataset.
- Only small differences in performance between segmentation methods

Automatic Construction of the Finnish Parliament Speech Corpus

André Mansikkaniemi, Peter Smit and Mikko Kurimo

Aalto University, School of Electrical Engineering
Department of Signal Processing and Acoustics

August 24, 2017

Aalto University
School of Electrical
Engineering

Aalto system for the 2017 Arabic multi-genre broadcast challenge

Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, Mikko Kurimo

Aalto University, Department of Signal Processing and Acoustics

December 20, 2017 - ASRU - pending review

Aalto University
School of Electrical
Engineering

Character-based units for Unlimited Vocabulary Continuous Speech Recognition

Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, Mikko Kurimo

Aalto University, Department of Signal Processing and Acoustics

December 20, 2017 - ASRU - pending review

Code

- Code released under open source license
- `github.com/aalto-speech/subword-kaldi`