Acoustic Data-Driven Subword Modeling for End-to-End Speech Recognition

Wei Zhou, Mohammad Zeineldeen, Zuoyun Zheng, Ralf Schlüter, Hermann Ney
zhou@cs.rwth-aachen.de
Interspeech 2021

Overview

1. Introduction

2. Acoustic Data-Driven Subword Modeling (ADSM)

3. Experiments

4. Conclusion

Introduction

End-to-end automatic speech recognition (ASR)

- great simplicity and state-of-the-art performance
- **subwords**: most common label units

Text-based subword modeling approaches

- **byte pair encoding (BPE)** [Sennrich & Haddow+ 16]: deterministic segmentation of words
  - split all words in the text corpus into single characters
  - merge pairs of units based on frequency
- **WordPieceModel (WPM)** [Schuster & Nakajima 12]: similar as BPE
  - subword merging based on the likelihood of the text data
- **unigram language model (ULM)** [Kudo 18]: probabilistic segmentation
  - EM training: marginal likelihood over all within-vocabulary segmentations of the text data
  - iterative vocabulary refinement and model training
  - subword regularization: draw samples of segmentation variants based on the trained ULM

No consideration of the underlying acoustic signal: key of ASR

Introduction

Automatic label learning from an acoustic perspective

- well studied for classical ASR systems [Bacchiani 99]
- but not fully addressed in end-to-end ASR

Acoustic-based subword methods

- **pronunciation-assisted subword modeling (PASM)** [Xu & Ding$^+$ 19]
  - pronunciation lexicon: acoustic structure of subword units
  - text corpus: post-processing for final labels (**no acoustic data involved**)
- **GramCTC** [Liu & Zhu$^+$ 17] and **latent sequence decompositions (LSD)** [Chan & Zhang$^+$ 17]
  - expose the ASR model to various segmentations in training
  - jointly learn an acoustic-based sequence decomposition within a fixed vocabulary
  - vocabulary: most frequent *n*-gram characters in the transcription
  - **not aim at acoustic-oriented subword modeling**

Propose: Acoustic Data-Driven Subword Modeling (ADSM)

- **fully acoustic-oriented label design and learning process**
- combine most advantages of the aforementioned methods
- acoustic-structured subword units
- acoustic-matched target sequence for further ASR training

Acoustic Data-Driven Subword Modeling (ADSM)

Notation

- $\vec{a}$: sequence of subwords a from vocabulary V
- $S(w) = \{\vec{a} : w\}$: set of allowed segmentations of word w using $a \in V$.

ADSM Initialization

- pronunciation lexicon: grapheme-to-phoneme (G2P) pairs
- V: all subword units from those G2P pairs
  - **acoustic structure**: graphemic representation of phonemes
- $S(w)$: all possible segmentation of w using $a \in V$
  - largely relaxed quality requirement of G2P alignment
- further discriminate subwords at word end: $a_{\_}$ vs. a, e.g. "a b l e_"
  - different acoustic property [Le & Zhang+ 19]
  - reconstruction of word

  $\rightarrow$ **V and $S(w)$**

Acoustic Data-Driven Subword Modeling (ADSM)

ADSM Repeatable Iteration:

Step 1. vocabulary refinement

- given $S(w)$ and V
  - training utterance (X, W): acoustic feature sequence and corresponding word sequence
  - $S(W)$: the set allowed subword sequences A for the full utterance W

- model θ training
  - **extended marginal likelihood in ULM + further dependency on the acoustic input**

$$\mathcal{L}(\theta) = -\log \sum_{A \in S(W)} p(A \mid X; \theta)$$

Acoustic Data-Driven Subword Modeling (ADSM)

Step 1. vocabulary refinement (continue)

- extended connectionist temporal classification (CTC) training as GramCTC
  - **marginalize over all CTC alignments of all allowed subword decomposition of W**

$$\mathcal{L}(\theta) = -\log \sum_{A \in S(W)} p(A \mid X; \theta) = -\log \sum_{A \in S(W)} \sum_{y_1^T : A} p'(y_1^T \mid h_1^T; \theta) = -\log \sum_{A \in S(W)} \sum_{y_1^T : A} \prod_{t=1}^{T} p'(y_t \mid h_1^T; \theta)$$

  - $h_1^T = f_\theta^{\text{enc}}(X)$: encoding (optional subsampling)
  - y_1^T: blank ϵ-augmented CTC alignment sequence
  - CTC collapsing function $B(y_1^T) = A$
  - p': defined over $V \cup \{\epsilon\}$

- **learn most probable segmentation of each utterance in an acoustic data-driven manner**

Acoustic Data-Driven Subword Modeling (ADSM)

Step 1. vocabulary refinement (continue)

- Viterbi aligning with trained model θ

$$\tilde{A} = B\Big(\underset{y_1^T : A \in S(W)}{\arg\max} \frac{p'(y_1^T \mid h_1^T; \theta)}{q^\lambda(y_1^T)} \Big) = B\Big(\underset{y_1^T : A \in S(W)}{\arg\max} \prod_{t=1}^{T} \frac{p'(y_t \mid h_1^T; \theta)}{q^\lambda(y_t)} \Big)$$

  - q: prior distribution (marginalize p' over the training data)
  - $\lambda \in [0, 1]$: smoothness of the overall model
    - increasing λ: more segmentation variants of each word in the alignment

- **forced alignment + weight-filtering $\rightarrow$ refined $\tilde{S}(w)$ and $\tilde{V}$**
  - for each w: gather all subword decomposition variants $\vec{a}$ in alignment with counts
  - normalize counts to weights w.r.t. occurrence of w
  - filter out $\vec{a}$ with weight less than threshold μ: remaining $\vec{a} \rightarrow \tilde{S}(w) \rightarrow \tilde{V}$

Acoustic Data-Driven Subword Modeling (ADSM)

Step 2. subword merging

- major idea of BPE and WPM: merge subword units based on certain criterion
  - avoid too long sequence with many small units
  - **spelling and context dependency modeling**

- **enhance $\tilde{S}(w)$ and $\tilde{V}$ with subword merging**
  - for each $\vec{a} \in \tilde{S}(w)$: merge any two neighboring units $\rightarrow$ all possible new sequences
    e.g. $\vec{a} = (a_1, a_2, a_3, a_4) \longrightarrow (a_1a_2, a_3, a_4), (a_1, a_2a_3, a_4), (a_1, a_2, a_3a_4)$
  - new labels in $\tilde{V}$ and new sequences in $\tilde{S}(w)$: original $\vec{a}$ always kept
  - **merged units: retain acoustic structure**

Repeat iteration with enhanced $\tilde{S}(w)$ and $\tilde{V}$

- vocabulary refinement: increase subsampling in f_θ^{enc} by 2

Acoustic Data-Driven Subword Modeling (ADSM)

ADSM Finalization

- **vocabulary refinement + word-count-filtering** $\rightarrow S_{\text{final}}(w)$ **and** V_{final}
  - w occurs less than k times: only take single best $\vec{a}$ based on weights
  - vocabulary size $|V_{\text{final}}|$: controlled by prior scale λ, weight-filtering μ and k jointly
- V_{final}: **acoustic-structured ADSM labels**
- final forced alignment: **acoustic-matched target sequence for further ASR training**
  - acoustically most probable decomposition of each utterance

Word	Initialization	Vocab-refinement	Subword-merging	Finalization
able	a b l e_ a b le_ a ble_	a ble_	a ble_ able_	a ble_
word	w o rd_ w or d_ wo r d_ wo rd_	w or d_	w or d_ w ord_ wor d_	w or d_ w ord_

Acoustic Data-Driven Subword Modeling (ADSM)

Text segmentation without audio

- needed for training subword LM on extra text data
- words in $S_{\text{final}}(w)$: draw samples of $\vec{a}$ based on weights
- words not in $S_{\text{final}}(w)$
  - train a simple n-gram LM on $S_{\text{final}}(w)$
  - best-score segmentation among all possible variants (V_{final}): acoustic preference

Experiments

- 960h LibriSpeech corpus [Panayotov & Chen$^+$ 15]
- ADSM setup
  - initialization: official Librispeech lexicon
  - 6×512 BLSTM + max-pooling layers for subsampling (initial factor 2)
  - vocabulary refinement: 25 full epochs (about 1 week on a single GTX-1080-Ti-GPU)
  - prior scale $\lambda = 0.3$, weight-filtering $\mu = 0.05$, word-count-filtering $k = 20$

- 1 iteration + finalization: **5k ADSM labels**
- clear reduction of $|V|$ and $|S(w)|$
  - **specific acoustic probable decomposition**
- decreasing $\text{len}(\vec{a})$: **learn larger units**
  - 5k BPE: $\text{len}(\vec{a}) = 3.2$
  - 5k PASM: $\text{len}(\vec{a}) = 5.7$
  - phoneme: $\text{len}(\text{pronunciation}) = 6.5$

Step		$\|V\|$	Average	
			$\|S(w)\|$	$\text{len}(\vec{a})$
Initialization		2k	51.7	8.1
1 Iteration	vocab-refinement	1k	1.2	5.4
	subword merging	21k	6.4	5.2
Finalization		**5k**	**1.1**	**4.7**

$|S(w)|$: average number of segmentation variants per word
$\text{len}(\vec{a})$: average length of all subword sequences in complete $S(w)$

Experiments

Model	Subword	dev WER[%]		test WER[%]	
		clean	other	clean	other
CTC	PASM	9.0	21.2	8.9	21.5
	BPE	9.5	20.0	9.5	20.9
	ADSM	**8.7**	**20.0**	**8.7**	**20.6**
RNN-T	PASM	5.3	13.2	5.4	13.6
	BPE	5.6	13.2	5.9	14.0
	ADSM	**5.0**	**12.6**	**5.2**	**12.8**
Attention	PASM	4.9	13.5	5.2	14.5
	BPE	4.9	13.0	5.1	13.6
	ADSM	**4.8**	**12.8**	**5.0**	**13.5**

Subword	"bachelor"	"password"	"together"
PASM	b a ch elor_	p a s s w or d_	togethe r_
BPE	bac hel or_	pas sword_	together_
ADSM	b a chel or_	p a ss w ord_	to g e ther_

- further end-to-end ASR
  - CTC [Graves & Fernández+ 06]
  - monotonic RNN-T [Tripathi & Lu+ 19]
  - LSTM-based attention model [Zeyer & Bahar+ 19]
- word error rate (WER) without external language model
- **ADSM clearly outperforms both BPE and PASM in all cases**
- **ADSM suitable for both time-sync. and label-sync. models**
  - acoustically more logical segmentation
  - acoustically more balanced sequence length (label size): spelling and context modeling

Experiments

Analysis: subword CTC + 4-gram word-LM

Importance of both acoustic structure and label size

Model	Subword	dev WER[%]		test WER[%]	
		clean	other	clean	other
CTC	PASM	9.0	21.2	8.9	21.5
	BPE	9.5	20.0	9.5	20.9
	ADSM	8.7	20.0	8.7	20.6
+ word-LM	PASM	4.1	10.4	4.3	10.9
	BPE	4.7	11.2	4.8	11.9
	ADSM	4.1	10.2	4.6	11.0

- idealized context modeling
  - spelling: perfectly defined in dictionary
  - cross-word context: word-LM
- both acoustic-based subwords (ADSM and PASM): similarly good and outperform BPE
- PASM: most degradation without LM
  - longest sequence (smaller label units): no merging
  - disadvantage of too long sequence length for end-to-end ASR

Conclusion

- ADSM: a fully acoustic-oriented subword modeling approach
  - acoustic-based label design and learning: more consistent with ASR
  - combine advantages of several subword methods into one pipeline
  - acoustic-structured subword units
  - acoustic-matched target sequence for further ASR training
- ADSM labels: evaluated for different end-to-end ASR approaches on Librispeech corpus
  - CTC, RNN-T and attention models
  - clearly outperform both BPE and PASM in all cases
- ADSM is suitable for both time-sync. and label-sync. models
  - acoustically more logical segmentation
  - acoustically more balanced sequence length (label size)

Thank you for your attention

Any questions ?

References

[Bacchiani 99] M. A. U. Bacchiani.
Speech Recognition System Design Based on Automatically Derived Units.
Ph.D. thesis, USA, 1999.

[Chan & Zhang+ 17] W. Chan, Y. Zhang, Q. V. Le, N. Jaitly.
Latent Sequence Decompositions.
In *Int. Conf. on Learning Representations (ICLR)*, 2017.

[Graves & Fernández+ 06] A. Graves, S. Fernández, F. J. Gomez, J. Schmidhuber.
Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.
In *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 369–376, 2006.

[Kudo 18] T. Kudo.
Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates.

In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 66–75, 2018.

[Le & Zhang+ 19] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, M. L. Seltzer.
From Senones to Chenones: Tied Context-Dependent Graphemes for Hybrid Speech Recognition.
In *IEEE ASRU*, pp. 457–464, 2019.

[Liu & Zhu+ 17] H. Liu, Z. Zhu, X. Li, S. Satheesh.
Gram-CTC: Automatic Unit Selection and Target Decomposition for Sequence Labelling.
In *Proc. Int. Conf. on Machine Learning (ICML)*, Vol. 70, pp. 2188–2197, 2017.

[Panayotov & Chen+ 15] V. Panayotov, G. Chen, D. Povey, S. Khudanpur.
Librispeech: An ASR corpus based on public domain audio books.
In *Proc. ICASSP*, pp. 5206–5210, 2015.

[Schuster & Nakajima 12] M. Schuster, K. Nakajima.
Japanese and korean voice search.
In *Proc. ICASSP*, pp. 5149–5152, 2012.

References

[Sennrich & Haddow+ 16] R. Sennrich, B. Haddow, A. Birch.
Neural Machine Translation of Rare Words with Subword Units.
In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[Tripathi & Lu+ 19] A. Tripathi, H. Lu, H. Sak, H. Soltau.
Monotonic Recurrent Neural Network Transducer and Decoding Strategies.
In *IEEE ASRU*, pp. 944–948, 2019.

[Xu & Ding+ 19] H. Xu, S. Ding, S. Watanabe.
Improving End-to-end Speech Recognition with Pronunciation-assisted Sub-word Modeling.
In *Proc. ICASSP*, pp. 7110–7114, May 2019.

[Zeyer & Bahar+ 19] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, H. Ney.
A Comparison of Transformer and LSTM Encoder Decoder Models for ASR.
In *IEEE ASRU*, pp. 8–15, 2019.