

# On the Difficulty of Segmenting Words with Attention

Ramon Sanabria Hao Tang Sharon Goldwater

School of Informatics

The University of Edinburgh

{r.sanabria, hao.tang}@ed.ac.uk, sgwater@inf.ed.ac.uk

## Abstract

Word segmentation, the problem of finding word boundaries in speech, is of interest for a range of tasks. Previous papers have suggested that for sequence-to-sequence models trained on tasks such as speech translation or speech recognition, *attention* can be used to locate and segment the words. We show, however, that even on monolingual data this approach is brittle. In our experiments with different input types, data sizes, and segmentation algorithms, only models trained to predict phones from words succeed in the task. Models trained to predict words from either phones or speech (*i.e.*, the opposite direction needed to generalize to new data), yield much worse results, suggesting that attention-based segmentation is only useful in limited scenarios.<sup>1</sup>

## 1 Introduction

Word segmentation is the task of finding word boundaries in speech. The task has a wide range of applications, including documenting under-resourced languages (Dunbar et al., 2017) and bootstrapping speech recognizers (Juang and Rabiner, 1990). It is often the first step to a variety of unsupervised speech tasks (Chung and Glass, 2018; Chung et al., 2018; Baeviski et al., 2021) and to the NLP pipeline for languages with no whitespace between words.

While unsupervised speech segmentation has been studied (Kamper et al., 2017; Räsänen et al., 2015), in many cases a parallel data source may be available, such as transcriptions, translations, or images. Previous researchers have suggested that it is possible to extract word segments from the attention map created by training an end-to-end sequence-to-sequence model on such parallel data (Palaskar and Metze, 2018; Boito et al., 2017, 2019, 2020; Godard et al., 2018). However,

<sup>1</sup>code available in the following link [https://github.com/ramonsanabria/insights\\_2021](https://github.com/ramonsanabria/insights_2021)

Palaskar and Metze’s evaluation was non-standard<sup>2</sup> and the others (which we refer to collectively henceforth as BOITOEAE) used models with text translations on the *input* side—decoding to phones or phone-like units—which means the trained models cannot be applied to segment novel (untranslated) sequences. In addition, the interpretation of attention as alignments in other areas of NLP has been questioned (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019).

Prompted by the prior work, we set out to study word segmentation with attention in sequence-to-sequence models, aiming to better understand when and how attention can be interpreted as alignments and in what settings it can be used to segment words. Our main experiments follow BOITOEAE in performing and evaluating word segmentation on the same data used to train the sequence-to-sequence model. However, instead of training translation models as in BOITOEAE, we train models to perform speech recognition: a well-studied task with a simpler (monotonic) alignment structure. This setting is similar to forced alignment, where both the speech and the transcription are given, and the goal is to discover the hidden alignments. Aside from this potential use case, this setting is useful for analysis because it abstracts away from the need to generalize to a novel test set. If the attention is not able to provide acceptable word boundaries in this setting, then the approach is unlikely to succeed in other, more difficult, settings.

We perform experiments on both the low-resource Mboshi dataset from BOITOEAE and a much larger English dataset, MuST-C (Di Gangi et al., 2019). We study models trained in both directions with a variety of input-output types (*e.g.*, phones, speech frames) and different postprocessing strategies to extract alignments from the atten-

<sup>2</sup>For their ASR model, they reported mean frame error relative to a forced alignment. Positive and negative errors cancel, so a small mean error does not imply correct boundaries.

tion weights. We find that although the particular configuration used by BOITOEAE works well, in most configurations the word segments extracted from the attention map are poor, even when using a larger dataset or model size. In particular, we did not get good results from any of the configurations that can be applied to a novel test set or to speech frames (rather than phone-like units). We conclude that even in the simple monotonic case studied here, attention typically does not provide clear word-level alignments.

## 2 Problem Setting

We consider an  $n$ -sample dataset  $S = \{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n, \mathbf{z}_n)\}$ , each of which is a triplet  $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ . As an example,  $\mathcal{X}$  is the set of sequences of speech frames,  $\mathcal{Y}$  is the set of sequences of words, and  $\mathcal{Z}$  is the set of sequences of segments, where a segment is a triplet  $(s, t, w)$  that indicates the start time  $s$ , the end time  $t$ , and the word  $w$ .

The goal is to learn a function  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$  given *only*  $S|_{xy} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ , i.e., discovering the alignments without observing them. Formally, we aim to find  $f$  that minimizes  $\sum_{i=1}^n \ell(f(\mathbf{x}_i, \mathbf{y}_i), \mathbf{z}_i)$  using only  $S|_{xy}$ , for some loss function  $\ell$  that evaluates the quality of the segmentation. Note that the evaluation of  $f$  is on the set  $S$ . We could evaluate  $f$  on a test set, but, in general, generalization is *not* involved in this setting.<sup>3</sup>

The setting is general, subsuming many tasks. When  $\mathcal{X}$  is the set of speech utterances and  $\mathcal{Y}$  is the empty set, this is the usual unsupervised word segmentation. The set  $\mathcal{Y}$  can be images or translations, grounding words from other modalities (Harwath et al., 2018). In this work, we focus on  $\mathcal{Y}$  being transcriptions, i.e., we have a forced alignment task.

## 3 Word Segmentation with Attention

Our pipeline, due to BOITOEAE, consists of two steps: generating an attention map from a sequence-to-sequence model, followed by postprocessing to convert the map into an alignment.

### 3.1 Sequence-to-Sequence Models

Below is a review of sequence-to-sequence models. Readers should refer to, for example, Luong et al. (2015) for a detailed exposition.

<sup>3</sup>We do report generalization results in the Appendix, for completeness, though these do not change our main story.

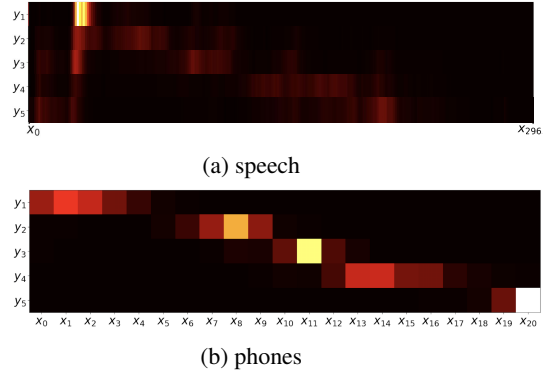


Figure 1: Example attention maps for sequences using (a) speech or (b) phones as input. Note that for space reasons, the maps are shown transposed: each row shows the attention for a single output timestep.

Given a speech utterance  $\mathbf{x} = x_1x_2 \dots x_T$  or simply  $x_{1:T}$  and its transcription  $\mathbf{y} = y_{1:K}$ , a sequence-to-sequence model learns a function of  $\mathcal{X} \rightarrow \mathcal{Y}$ . An encoder Enc and a decoder Dec take  $\mathbf{x}$  and  $\mathbf{y}$  as input and produce their respective hidden vectors  $h_{1:T} = \text{Enc}(x_{1:T})$  and  $q_{2:K} = \text{Dec}(y_{1:K-1})$ .

The attention map

$$\alpha_{t,k} = \frac{\exp\left(W_a \begin{bmatrix} h_t \\ q_k \end{bmatrix}\right)}{\sum_{i=1}^T \exp\left(W_a \begin{bmatrix} h_i \\ q_k \end{bmatrix}\right)} \quad (1)$$

is computed with a weight matrix  $W_a$ . Examples are illustrated in Figure 1. In the early stages of the project, we experimented with dot product attention and the results were similar. Finally, the probability of the label is computed as

$$p(y_k | y_{1:k-1}, x_{1:T}) = \text{softmax}\left(W \begin{bmatrix} c_k \\ q_k \end{bmatrix}\right) \quad (2)$$

where  $c_k = \sum_{t=1}^T \alpha_{t,k} h_t$ . The model is trained to maximize the probability

$$p(y_{1:K} | x_{1:T}) = p(y_1 | x_{1:T}) \prod_{k=2}^K p(y_k | y_{1:k-1}, x_{1:T}).$$

We emphasize that, in our setting,  $\mathbf{x}$  and  $\mathbf{y}$  are always given, also known as teacher forcing (Lamb et al., 2016), and we are interested in the attention map  $\alpha$ , not how well the model maps  $\mathbf{x}$  to  $\mathbf{y}$ . Note that the assignments to  $\mathbf{x}$  and  $\mathbf{y}$  can be swapped since both are given. For example, we can align word transcriptions to phonetic transcriptions, or

vice versa. However, the choice of directionality has two implications. First, for each output  $y_k$ , some parts of  $\mathbf{x}$  will have high attention weights, whereas for parts of  $\mathbf{x}$ , there may be no  $y_k$  with high weights. This asymmetry affects the choice of postprocessing method, as described below. Second, typically only one direction will be feasible if we want to apply the trained model to new (unannotated) data. While our experiments here focus on segmenting the training set, we would ideally like to find a method that can also work on new data.

### 3.2 Postprocessing

We explore three types of postprocessing to obtain the alignments from the attention map  $\alpha$ . The first of these was introduced by Boito et al. (2017); the others are novel.

**Hard Assignment** This approach aligns each output symbol  $y_k$  to the input that has the highest attention weight, i.e., to  $x_{t_k}$ , where  $t_k = \arg \max_t \alpha_{t,k}$ . Due to the attention map asymmetry noted above, this approach is only applied when the transcribed words are on the input side; otherwise some phones (or speech frames) may not be aligned to any word. This method hypothesizes a word boundary between two output symbols if they are aligned to different words; otherwise they are considered part of the same word.

BOITOEAL mainly use this method to train *translation* models (French input words; Mboshi output phones), but Boito et al. (2020) also present monolingual results (Mboshi input words; Mboshi output phones), which we compare to below.

**Thresholding** When the attention weight is higher than a threshold  $\tau_{\text{onset}}$ , then we hypothesize a start of a word segment. When the attention weight is lower than a threshold  $\tau_{\text{offset}}$ , then we hypothesize an end of a word segment. Thresholds are set by exhaustive search using F-score on the development set as the search metric. Thresholding can generate multiple segments for a given output, which is not desirable for our setting. However, in an automatic speech translation setup, such behavior can be helpful in some language pairs.

**Segmental Assignment** Since we know that word segments are contiguous chunks of speech, this constraint should be baked into the postprocessing. In particular, we find a sequence  $(s_1, t_1), \dots, (s_K, t_K)$  such that the sum of attention weights that each segment covers, i.e.,

$\sum_{t=s_k}^{t_k} \alpha_{t,k}$ , is maximized, while respecting the connectedness constraint, i.e.,  $s_{k+1} = t_k + 1$ . This can be achieved by finding the maximum weighted path in a graph with edges as word segments and weights of the edges as the attention weights a segment covers. See a detailed description in Appendix A.1 and (Tang et al., 2017).

## 4 Experiments

Most of our experiments are conducted on the Mboshi dataset. It contains 4616 short read-speech utterances for training (3 seconds/6 words on average; 4.5h in total), with a vocabulary of 6638 words, and 514 utterances for development. Mboshi is a Bantu Language with no orthography, and the speech is transcribed at the word level using a phonetic orthography designed by linguists. We regard the basic units in the transcriptions as phones.<sup>4</sup> We do not use the French translations of the utterances.

For experiments with speech, we use Kaldi to extract speech features, with a 25 ms window shifted by 10 ms. Each acoustic frame consists of 40-dimensional log mel features and 3-dimensional pitch features. The acoustic feature vectors are used directly, we do no clustering or acoustic unit discovery.

We use a 1-layer bidirectional LSTM encoder and a 1-layer unidirectional LSTM decoder,<sup>5</sup> with 0.5 dropout on the encoder and a 256-dimensional hidden layer.<sup>6</sup> Further hyperparameter details are in Appendix A.2.

As noted above, our main questions do not require testing generalization, so except where otherwise noted (to compare to previous work), we evaluate all models on the training set. Results on the development set, which do not change the story, are reported in Appendix A.3. We report precision, recall, and F-score of the hypothesized word boundaries. When transcribing phones to words, a

<sup>4</sup>We use the term *phone* rather than *phoneme* both here and with reference to the MUST-C data set (below), to avoid making any commitments about the underlying cognitive/linguistic form, which the term *phoneme* implies. For the Mboshi data especially, we are not sure if these commitments hold. However, like phonemic transcriptions, the transcriptions we work with assume a single pronunciation for each word type (effectively, dictionary lookup of pronunciations).

<sup>5</sup>Boito et al. (2020) showed that LSTMs worked better than Transformer or CNN models with their framework.

<sup>6</sup>The number of layers, dropout, and dimensions were tuned on the development set. The values we found best are the same ones Boito et al. (2020) reported, except their hidden layer size is 64. It is unclear why dropout helps in this setting, since without generalization, there is no concern of overfitting, but we did find a small benefit.

Table 1: Word boundary scores on Mboshi for models predicting phones from words ( $w \rightarrow p$ ), as in Boito et al. (2020), using Hard or Segmental assignment. <sup>†</sup>Models trained on Train+Dev. \*Results from Boito et al. (2020), averaging five attention maps.

		P	R	F	OS
$w \rightarrow p$	Hard	92.3	83.2	87.5	-9.8
$w \rightarrow p$	Seg	93.5	93.5	93.5	0.0
$w \rightarrow p^{\dagger}$	Hard	95.5	85.7	90.4	-10.2
$w \rightarrow p^{\dagger*}$	Hard	92.9	92.1	92.5	-

word boundary must be hypothesized at the exact place to be counted as correct. In later experiments, particularly for speech, we follow the The Zero Resource Speech Challenge (Dunbar et al., 2017) evaluation and use a 30ms tolerance window, *i.e.*, the hypothesized boundary is counted as correct if it falls within 30 ms of the correct boundary. Similar to BOITOE, we use force alignments extracted with a Kaldi (Povey et al., 2011) GMM-HMM model as ground truth word boundaries. We also report the amount of over-segmentation, defined as  $(N_h - N_{ref})/N_{ref}$ . If the quantity is positive, the model hypothesizes too many boundaries; if the quantity is negative, the model hypothesizes too few boundaries.

#### 4.1 Predicting Phones From Words

We begin by confirming the positive results from previous work, following BOITOE in training a model to predict phones given words.

Results for both Hard Assignment (Hard) and Segmental Assignment (Seg) are shown in Table 1. We transpose the attention matrix to run Seg so that a segment (word) can consist of multiple phones. As expected, both methods work well, with the more principled Seg performing slightly better—though it has a slight advantage, since (due to teacher forcing) it always generates the right number of word segments (yielding OS = 0).

However, these methods can only be applied to the annotated data, which is a significant weakness.

#### 4.2 Words as Targets

Next, we consider the more typical direction of decoding, predicting words given phones. We do not use hard assignment in this setting for the reasons described in Section 3.2.

Results (Table 2) show that, when using phones, flipping the model direction from  $w \rightarrow p$  to  $p \rightarrow w$  makes the results much worse. The Thresholding method is especially bad, so we focus on Seg for

Table 2: Word boundary scores on Mboshi for models with words as targets, using phones, phone frames, or acoustic feature frames as input ( $p \rightarrow w$ ,  $f \rightarrow w$ , and  $a \rightarrow w$ , respectively), with Thresholding or Segmental assignment. The first row is copied from Table 1. Unsupervised baselines (acoustic input only) are also shown: R15 (Räsänen et al., 2015); K17 (Kamper et al., 2017).

		P	R	F	OS
$w \rightarrow p$	Seg	93.5	93.5	93.5	0.0
$p \rightarrow w$	Thr	19.2	20.4	19.8	6.3
$p \rightarrow w$	Seg	58.0	58.0	58.0	0.0
$f \rightarrow w$	Seg	57.6	57.6	57.6	0.0
$a \rightarrow w$	Seg	13.4	13.4	13.4	0.0
a	R15	21.5	19.8	20.6	-8.1
a	K17	32.4	7.4	12.0	-77.1

the rest of the paper. The deterioration by simply flipping the model is unsatisfying, because the setting is simple enough that the model should be able to achieve near-perfect results by acting like a lexicon, mapping canonical pronunciations to words.

This decoding direction allows us to build models that take acoustic features as input and produce words.

However, once we replace phone transcriptions as input with acoustic features, the result (denoted  $a \rightarrow w$  in Table 2) is drastically worse—even underperforming one of the unsupervised baselines (note the change on attention structure between Figure 1b and Figure 1a).

To understand what causes the dramatic drop in performance, we explore an intermediate input representation where we replace each acoustic frame with its phone label. This input format, which we refer to as phone frames, has the same length as the acoustic input sequence and reflects the duration of each phone, while abstracting away from acoustic variability. Results of this experiment ( $f \rightarrow w$  in Table 2) show that most of the performance gap is recovered. This suggests that the model learns little about the phonetic variability in this experiment. This prompts us to work on a larger dataset, with the hope that the model is able to capture the phonetic variability given more data.

#### 4.3 Scaling to Larger Data

We investigate if by exposing a model to more speech data it would learn to normalize phonetic variance and close the gap between  $f \rightarrow w$  and  $a \rightarrow w$ . For this experiment we use English data (speech, lexicon phone sequences, and word transcriptions) from the MuST-C dataset (Di Gangi



Table 3: Word boundary scores on MuST-C using Segmental assignment with a variety of models. The first three models have 1 hidden layer, while Lg has 5. Unsupervised baselines are also shown: R15 (Räsänen et al., 2015); K17 (Kamper et al., 2017)

		P	R	F	OS
p → w	Seg	44.0	44.0	44.0	0.0
f → w	Seg	42.6	42.6	42.6	0.0
a → w	Seg	21.2	21.2	21.2	0.0
a → w (Lg)	Seg	19.9	19.9	19.9	0.0
a	R15	20.0	28.6	23.5	43.7
a	K17	23.7	26.1	24.8	10.2

et al., 2019). MuST-C provides translations to other languages; we don’t use these here but we do limit our data to the 145k English utterances (257h of speech) for which translations are available in all the languages<sup>7</sup>. Utterances have an average length of 6.5 seconds/18 words. We use the same speech feature extraction configuration as in the Mboshi experiments. Because we are using a larger dataset, we also try a deeper (5-layer) model for the speech input. Results are shown in Table 3. The results on p → w are lower than for Mboshi, suggesting that the longer utterances in MuST-C make the task more challenging. The speech in MuST-C is probably also harder than in Mboshi (TED talks vs. read speech); nevertheless, the performance on a → w is better on MuST-C than Mboshi, closer to the MuST-C f → w results. This suggests that adding more data does allow the model to learn more about acoustic variability. However, given the large size of this data set, all the results are underwhelming, and the results with speech still do not beat the unsupervised models.

## 5 Conclusion

Previous researchers had suggested a connection between attention weights and word alignments in both speech recognition and speech translation. However, we have experimented with several attention-based segmentation methods and demonstrated that these only succeed in the scenario where words are used as the input to the model—a scenario with limited application. Performance drops considerably for models with phones as input, and is no better than unsupervised segmentation for models using speech as input, even when the amount of training data is increased by two orders

<sup>7</sup>Find the list of files in Kaldi format in the following link [https://homepages.inf.ed.ac.uk/s1945848/must\\_c\\_insights.zip](https://homepages.inf.ed.ac.uk/s1945848/must_c_insights.zip).

of magnitude. Although in principle the transcriptions provide an additional source of information, using this to help segment words from speech will likely require a completely different approach.

## References

- Alexei Baeovski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *arXiv preprint arXiv:2105.11084*.
- Marcely Zanon Boito, Alexandre Bérard, Aline Villavicencio, and Laurent Besacier. 2017. Unwritten languages demand attention too! Word discovery with encoder-decoder models. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Marcely Zanon Boito, Aline Villavicencio, and Laurent Besacier. 2019. Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. In *Interspeech*.
- Marcely Zanon Boito, Aline Villavicencio, and Laurent Besacier. 2020. Investigating alignment interpretability for low-resource nmt. In *Machine Translation*. Springer.
- Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. NMTPY: a flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*.
- Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Interspeech*.
- Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems (NIPS)*.
- Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a multilingual speech translation corpus. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Pierre Godard, Marcely Zanon-Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. In *Interspeech*.

- David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *European Conference on Computer Vision (ECCV)*.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- B-H Juang and Lawrence R Rabiner. 1990. The segmental K-means algorithm for estimating parameters of hidden markov models. *Transactions on Acoustics, Speech, and Signal Processing*.
- Herman Kamper, Aren Jansen, and Sharon Goldwater. 2017. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*.
- Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Shruti Palaskar and Florian Metze. 2018. Acoustic-to-word recognition with sequence-to-sequence models. In *Spoken Language Technology Workshop (SLT)*.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *Workshop on Automatic Speech Recognition and Understanding*. IEEE.
- Okko Räsänen, Gabriel Doyle, and Michael C Frank. 2015. Unsupervised Word discovery from speech using automatic segmentation into syllable-like units. In *Interspeech*.
- Hao Tang, Liang Lu, Lingpeng Kong, Kevin Gimpel, Karen Livescu, Chris Dyer, Noah A Smith, and Steve Renals. 2017. End-to-end neural segmental models for speech recognition. *Journal of Selected Topics in Signal Processing*.
- Sarah Wiegrefe and Yuval Pinter. 2019. Attention is not not explanation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

## A Appendix

### A.1 Segmental Assignment

Given an attention map, the goal of segmental assignment is to find a segmentation that maximizes the amount of attention weights each word covers while making sure that the word segments are connected. To achieve this, we turn this into a problem of finding a maximum weighted path on a graph, where the edges of the graph are segments, the weights on the edges correspond to the amount attention weights covered, and the graph encodes the connectedness constraints.

Suppose the attention map is of dimension  $T \times K$ . Recall that  $T$  is the number of input tokens (such as speech frames) and  $K$  is the number of output tokens (such as words). We first create a vertex set  $V = \{(t, k) : \text{for } t = 0, \dots, T \text{ and } k = 0, \dots, K\}$ , a grid marking every element in the attention map. An edge is a pair of vertices  $(t_1, k_1)$  and  $(t_2, k_2)$  while satisfying  $t_1 < t_2$  and  $k_2 = k_1 + 1$ . That edge represents a segment of the  $k_2$ -th output token that aligns to  $t_1$  to  $t_2$  on the input side. This can be realized by defining the incoming edges

$$\text{in}((t_2, k_2)) = \left\{ \left( (t_1, k_2 - 1), (t_2, k_2) \right) : \right. \\ \left. \text{for } t_1 = 0, \dots, t_2 \text{ if } k_2 > 1 \right\}. \quad (3)$$

We assign the sum of attention weights from  $t_1$  to  $t_2$ , i.e.,  $\sum_{t=t_1+1}^{t_2} \alpha_{t, k_2}$  to the edge  $((t_1, k_2 - 1), (t_2, k_2))$ .

Once the graph is constructed, we find the maximum weighted path that starts at  $(0, 0)$  and ends at  $(T, K)$ . An example is shown in Figure 2. Note that due to the imposed constraints and in turn due to how the graph is constructed, segmental assignment only considers monotonic alignments.

### A.2 Hyperparameters and Implementation

We use the sequence-to-sequence implementation of nmtpytorch (Caglayan et al., 2017)<sup>8</sup>. The model comprises one-layer encoder and one-layer decoder with 0.5 dropout, except in the experiment of Section 4.3 where we use a five-layer encoder in the Large (Lg) model. We set a size of 256 to all hidden dimensions (i.e., source and target embedding, encoder, and decoder). We use the Adam optimizer,

<sup>8</sup><https://github.com/lium-1st/nmtpytorch>

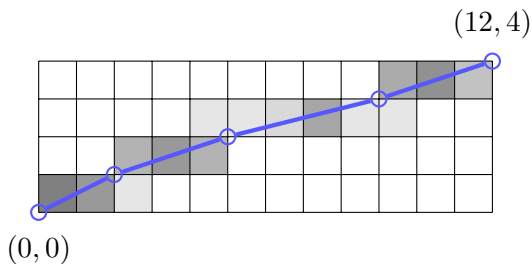


Figure 2: An example segmental assignment with 12 input tokens and 4 output tokens. Each box represents an element in the attention map. The darker the shade, the higher the attention weight. Edges of the maximum scoring path and the corresponding vertices are shown in blue. The shade being crossed by an edge is the amount of attention weights covered by the edge. The goal of segmental assignment is to find the maximum weighted path from the bottom-left corner to the top-right corner.

and the scheduler applies a decay factor of 0.5 after two consecutive epochs where loss does not decrease. All models are trained until cross-entropy loss on training reaches 0. The implementation of each model has around 3M and 19M of learnable parameters for the 1 and 5 layers encoder model, respectively. They are trained with one Nvidia GEFORCE GTX 1080 Ti. To reduce computation on Segmental Assignment, we set the maximum duration of a word to 4 seconds (400 frames for speech or phone frames representation) for Mboshi and 10 seconds for MuST-C. We set them by analyzing their performance on the development set.

Regarding the unsupervised speech baseline models, we use the unigram public implementation of Kamper et al. (2017)<sup>9</sup> with a minimum word segment duration of 250 ms. Because its performance is linked to the syllable segmentation method, we select the best configuration by fine-tuning Räsänen et al. (2015)<sup>10</sup>’s hyperparameter values on the development set.

### A.3 Results without Transcriptions

Finally, we consider a more traditional scenario where the model is exposed to unseen data. For this setting our model does not have access to transcriptions and therefore we do not use Teacher-Forcing. We evaluate the development set from Mboshi (used in Section 4.1), which has 1147 token types (where only 710 are observed during

Table 4: Results from 1 layer, and 5 layers (Lg) models for word segmentation of the development (unseen) set of the Mboshi and MuST-C dataset.

DS	Model	P	R	F	OS (%)
Mb	p → w	51.9	54.2	53.0	4.4
Mb	f → w	48.5	47.3	47.9	-2.3
Mb	a → w	16.2	14.7	15.4	-8.9
Mb	a → w (Lg)	14.1	14.9	14.5	6.0
MC	p → w	43.8	44.0	43.9	0.5
MC	f → w	30.1	30.1	30.1	0.2
MC	a → w (Lg)	17.9	20.7	19.2	15.8

training). In terms of absolute tokens, it has 2993, and only 516 out-of-vocabulary. We experiment with the MuST-C by using 1162 utterances from MuST-C’s superset unseen during training. In this case, the set has 3273 token types.

Surprisingly, for p and f → w, attention still produces meaningful segments although the model has not seen or early stopped with a development set. In that case, we observe a degradation in performance but not dramatic. The small number of unobserved absolute tokens do not have a critical effect on the segmentation performance of the model.

Finally, the small difference in performance between the train and development sets on a → w shows an already present weak segmentation signal not correlated with word in speech models.

<sup>9</sup>[link to Kamper et al., 2017 Github implementation](#)

<sup>10</sup>[link to Rasanen et al., 2015 Github implementation](#)