# LEARNING ACOUSTIC FRAME LABELING FOR PHONEME SEGMENTATION WITH REGULARIZED ATTENTION MECHANISM

*Binghuai Lin*[1] *, Liyuan Wang*[1]

[1]Smart Platform Product Department, Tencent Technology Co., Ltd, China

## ABSTRACT

Phoneme segmentation plays an important role in various speech processing applications such as keyword spotting, automatic pronunciation assessment, and automatic speech recognition. In this paper, we propose a method for phoneme segmentation based on a regularized attention mechanism. Specifically, the representations of speech utterance for each frame are extracted from a pre-trained acoustic encoder and combined with presumed phoneme sequences based on the attention mechanism. By fusing acoustic representations with these aligned phoneme representations, we learn phoneme labeling for each frame to obtain final segmentation. For better alignment between the pronounced phoneme sequence and utterance, we regularize the attention matrix utilizing an extra attention loss. The whole network is optimized by a multi-task learning framework (MTL). Experimental results based on the TIMIT and Buckeye corpora show the proposed method is superior to the previous baselines and reaches the state-of-the-art (SOTA) performance in F1 score and R-value.

***Index Terms***— Phoneme segmentation, attention mechanism, extra supervision, acoustic representations, phoneme labeling

## 1. INTRODUCTION

Phoneme segmentation, also referred to as phoneme alignment, is the task of positioning phonemes in a corresponding speech signal. It plays an important role in various speech processing applications, such as keyword spotting [1], automatic pronunciation assessment [2], automatic speech recognition (ASR) [3], etc.

Commonly, phoneme segmentation can be implemented in either supervised or unsupervised ways. Unsupervised methods conduct phoneme segmentation based on the speech signal without any target boundaries. Previous work obtained statistics on the change of the speech signal and predicted potential boundaries frame by frame [4, 5]. Supervised methods rely on the annotations of phoneme boundaries and other additional information such as the pronounced phonemes. The supervised setting can be further categorized into two types. One is the text-independent phoneme segmentation, where no phonemes are provided. Many previous studies have been conducted on text-independent phoneme segmentation. People utilized the information of vocal fold vibration contained in electroglottograph (EGG) to detect the voiced section in the speech data and then divided each voiced section into several candidate phonemes using the Viterbi algorithm [6]. A deep neural network such as recurrent neural network (RNN) was utilized as feature functions for the binary classification of phoneme boundaries [7, 8]. The other is the text-dependent phoneme segmentation, where the pronounced phonemes are provided in advance. In this paper, we focus on the latter. We take presumed phonemes as additional input for the segmentation.

The text-dependent phoneme segmentation, also called forced alignment, takes acoustic signals as well as transcriptions at the phone or word level as input. The common approach for forced alignment is based on a generative model of the speech signal using Hidden Markov Models (HMM). HMM for each acoustic phone is trained, and the sequence of frames in an utterance is aligned to the phone sequence by finding the sequence of hidden states that maximizes the utterance likelihood [9, 10]. Previous work further improves the segmentation accuracy by using more powerful statistical models for boundary correction that are conditioned on phonetic context and duration features based on HMM forced alignments [11]. However, as phoneme boundaries are not represented in the model, the training procedure does not explicitly optimize the boundary accuracy [11]. It also has other drawbacks such as convergence of the expectation-maximization (EM) procedure to local maxima and overfitting effects due to a large number of parameters [12]. Some work proposed a discriminative method for phoneme alignment. It devised the alignment function by mapping the Mel-frequency cepstral coefficients (MFCCs) of the speech utterance along with the target alignment into an abstract vector space and adopted a support vector machine (SVM) to separate correct alignments from incorrect ones [12]. However, it needs to design alignment functions in a handcrafted way. In this paper, instead of manually designed functions, we propose a discriminative method of learning phoneme alignment automatically.

Attention mechanism has been widely used in computer vision [13], machine translation [14], speech recognition [15],

These authors annotated with [1] contributed equally to this work.

etc. Though attention has been treated as a soft alignment between two sequences, whose score quantifies how well output at one position is aligned to input at another position, it has been found that the attention score differs significantly from alignments in the traditional sense [16]. It also has limitations when it comes to the requirements of specific alignments between input and output, as in the case of ASR which needs monotonic alignments between speech input and text input [17]. In this paper, the purpose of utilizing an attention mechanism is two folded: (1) acting as a fusion mechanism for acoustic and phoneme representations; and (2) acting as a soft alignment mechanism that can be regularized with an extra supervised attention loss.

In this paper, we propose a text-dependent method for phoneme segmentation. Speech feature representations are learned based on a pre-trained acoustic encoder, and phoneme representations of the pronounced phoneme sequences are encoded by a phoneme encoder. We utilize the attention mechanism to fuse the speech and phoneme sequences. Based on the fused speech and phoneme representations, we learn phoneme labeling for each frame to obtain the final segmentation. To regularize the attention scores for better alignment, we treat them as soft alignments and incorporate additional supervision over the attention probability distribution. Experimental results based on the TIMIT [18] and Buckeye [19] corpora show the proposed method outperforms the baselines and achieves the SOTA performance with F1-score and R-value.

## 2. PROPOSED METHOD

The network for phoneme segmentation is composed of an acoustic encoder as well as a phoneme encoder as shown in Figure 1. The acoustic encoder takes speech signals as input and acoustic representations as output. The phoneme encoder generates independent numerical representations for each phoneme. These two encoders are combined based on the attention mechanism, with varied attention scores indicating relevance between the input frame and phoneme sequences of different positions. Specifically, highest scores normally indicate the most relevant frames and phonemes. With the soft alignment, the weighted combination of the acoustic and phoneme representations is then utilized as input for phoneme classification for each frame. To achieve more appropriate alignment between these two sequences, the phoneme attention scores are regularized with an additional supervised target.

### 2.1. Feature representations

We obtain acoustic representations based on a pre-trained acoustic encoder called wav2vec 2.0 [20], which has been widely used in phoneme recognition, ASR, etc. Wav2vec 2.0 is a framework for self-supervised learning of representations from raw audio data. It consists of a multi-layer convolutional feature encoder trained via a contrastive task based on a large amount of unlabeled data, where the correct quantized latent audio representation is to be distinguished from the distractors. Based on the pre-trained acoustic encoder, we extract acoustic representations for $i$th frame of the utterance, denoted as $h^i_{\text{speech}}$.

As each phone type in the phone set shows distinct characteristics [21], we employ different numerical representations for each phoneme type. We also take into account the influence of the phoneme positioning in a word on the pronunciation by denoting phoneme positions with 'B', 'I', 'E', 'S' representing the beginning, middle, ending positions in a word as well as single-phoneme words. These phoneme properties are encoded as positional and phoneme embeddings, and by summing them up we obtain the final feature representations for $j$th phoneme, denoted as $h^j_{\text{phone}}$.

### 2.2. Attention-based fusion

Utilizing the attention mechanism, the soft alignment between aforementioned acoustic and phoneme representations is implied by the attention scores. We employ multi-head attention (MHA) mechanism following the previous work [14]. MHA models the relationship between queries, keys, and values. The attention score can be defined as Eq. (1) and Eq. (2):

$$\text{AttentionScore}(Q, K) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \quad (1)$$

$$\text{Attention}(Q, K, V) = \text{AttentionScore}(Q, K) * V \quad (2)$$

where $Q$ denotes the queries and $K$ denotes the keys with the dimension of $d_k$. $V$ represents the values with the dimension of $d_v$.

In this paper, we utilize representations of speech $\text{H}_{\text{speech}}$ as queries and representations of phonemes $\text{H}_{\text{phone}}$ as keys and values. Here, $\text{H}_{\text{speech}}$ is a combined representation of an utterance with the $i$th column $h^i_{\text{speech}}$. $\text{H}_{\text{phone}}$ is a combined representation of the phone sequence for an utterance with the $j$th element $h^j_{\text{speech}}$. We obtained weighted sum of phoneme representations based on Eq. (2). Then, the representation for $i$th frame is obtained by concatenated fusion of acoustic and summed phoneme representations as shown in Eq. (3).

$$h^i_{\text{fusion}} = [\text{Attention}(h^i_{\text{speech}}, \text{H}_{\text{phone}}), h^i_{\text{speech}}] \quad (3)$$

### 2.3. Network training and attention regularization

The final phoneme segmentation is obtained by classification of each frame. Classes include all possible phonemes and an additional silence label. Based on Eq. (3), we add

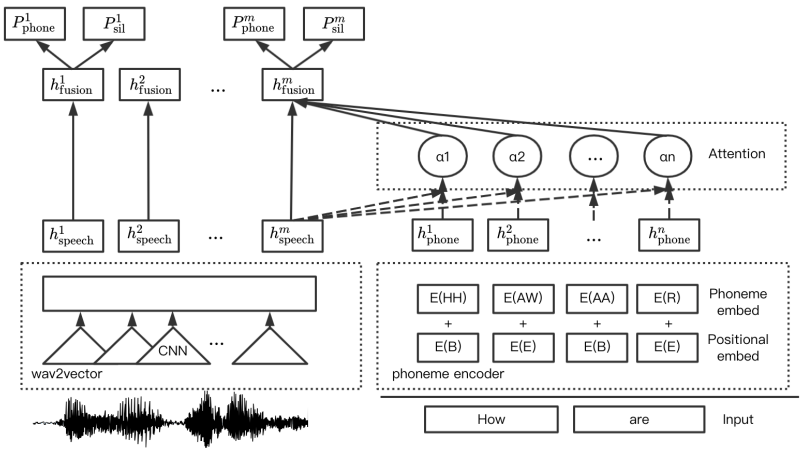


**Fig. 1**. Network structure of attention-based phoneme segmentation

a fully-connected (FC) layer followed by the SoftMax activation to calculate phoneme class probabilities. The cross-entropy classification loss between the predicted results and true phoneme sequence labels is defined in Eq. (4), where $m$ is the total number of frames, and $c$ is the number of classes.

$$L_{\text{phone}} = -\sum_{i=1}^{m}\sum_{j=1}^{c} y_{\text{phone}}{}_i^j \times \log(p_{\text{phone}}{}_i^j) \tag{4}$$

To further differentiate silence frames from non-silence frames, we utilize an extra loss for binary silence classification defined in Eq. (5), where $y_{\text{sil}}^i$ and $p_{\text{sil}}^i$ are the $i$th label and predicted result, respectively.

$$L_{\text{sil}} = -\sum_{i=1}^{m} y_{\text{sil}}^i \times \log(p_{\text{sil}}^i) + (1 - y_{\text{sil}}^i) \times \log(1 - p_{\text{sil}}^i) \tag{5}$$

We can treat the attention scores defined in Eq. (1) as the soft alignment probability distribution over different positions of the phoneme sequences for every frame. In this way we can regularize the attention scores utilizing the alignment labels such that more appropriate alignments can be obtained. Here, the alignment labels are denoted as zeros or ones in the alignment matrix, with one at position $(i, j)$ indicating the $i$th frame is aligned to the $j$th phoneme. Based on the predicted attention scores and the alignment labels, we can define an additional attention loss as in Eq. (6) to regularize the attention scores, where $N_P$ is the total number of pronounced phonemes in the sequence.

$$L_{\text{align}} = -\sum_{i=1}^{m}\sum_{j=1}^{N_P} y_{\text{align}}{}_i^j \times \log(p_{\text{attention}}{}_i^j) \tag{6}$$

The network is optimized based on an MTL framework. The total loss is defined as Eq. (7):

$$L_{\text{total}} = \lambda L_{\text{phone}} + \beta L_{\text{sil}} + \gamma L_{\text{align}} \tag{7}$$

where $\lambda$, $\beta$ and $\gamma$ are hyper-parameters to balance the loss among tasks of phoneme, silence classification and attention regularization.

## 3. EXPERIMENTS

### 3.1. Experimental setup

We conducted experiments based on the TIMIT and Buckeye corpora. For the TIMIT corpus, We used the standard train and test split, where we randomly sampled 20% of the training data for validation. For the Buckeye corpus, we split the corpus into training, validation, and test sets at the speaker level with a ratio of 80/10/10 and split long sequences into smaller ones following previous work [8].

Acoustic representations with a dimensionality of 384 from Wav2vec 2.0 are transformed utilizing $384 \times 32$ parameters to achieve same dimensionality with phoneme embeddings. The 61 phonemes in TIMIT and Buckeye are mapped to 39 different phonemes according to Carnegie Mellon University (CMU) Pronouncing Dictionary [22]. The positional and phoneme embeddings with the dimensionality of 32 are summed to obtain final phoneme representations. The MHA for the acoustic and phoneme representations consists of one single head. Two separate FC layers following the fused acoustic and phoneme representations based on attention mechanism have the dimension of $64 \times 40$ for phoneme classification and $64 \times 2$ for silence classification, respectively.

We evaluate the proposed method following previous work [5, 7, 8] using precision (P), recall (R), and F1-score with a tolerance level of 20ms. To tackle the drawback of the F1-score that a high recall and a low precision may yield relatively high F1-score, we utilize a complementary metric denoted as R-value as previous work did [8].

### 3.2. Comparative study

We compare the proposed method with several previous baselines with the tolerance of 20ms: the discriminative method [12] and Montreal [10]. Results on TIMIT and Buckeye are shown in Table 1.

**Table 1**. Comparison of the proposed method and previous forced alignment algorithms

| Corpora | Model | P | R | F1 | R-val |
|---|---|---|---|---|---|
| TIMIT | Ours | 93.42 | 95.96 | **94.67** | **95.18** |
| | Discrim [12] | 90 | 82.2 | 85.9 | 79.51 |
| | Montreal [10] | 83.9 | 81.6 | 82.7 | 85.16 |
| | SEGFEAT [8] | 92.67 | 93.03 | 92.85 | 93.91 |
| Buckeye | Ours | 88.49 | 90.33 | **89.40** | **90.90** |
| | SEGFEAT [8] | 85.40 | 89.12 | 87.23 | 88.76 |

From the results, we can see our proposed method outperforms the previous baselines by $1.8\%$ in F1 and $1.3\%$ in R-value on TIMIT, and $2.2\%$ in F1 and $2.1\%$ in R-value on Buckeye, which has reached the SOTA performance.

### 3.3. Ablation studies

To demonstrate the effectiveness of the proposed method, we make some ablation studies based on the TIMIT corpus. First, we remove the attention mechanism and conduct tasks such as phoneme classification based on the extracted acoustic representations only to investigate the effect of attention with extra phoneme information. Second, we remove the silence loss or attention loss in turn to demonstrate the effect of silence loss and attention regularization.

Results with attention (W Attention) and without attention (W/O Attention) are shown in Table 2. From the results, we can see that the proposed method with attention mechanism performs superior to that without attention mechanism in F1-score and R-value by a margin of $1.3\%$, indicating the benefit of exploring pronounced phoneme information to facilitate the phoneme classification.

**Table 2**. Comparison of the proposed method with and without attention

| Model | P | R | F1 | R-val |
|---|---|---|---|---|
| W Attention | 93.42 | 95.96 | **94.67** | **95.18** |
| W/O Attention | 91.69 | 94.99 | 93.31 | 93.87 |

We investigate the impact of the silence and attention loss by removing both or one of them. From results in Table 3, we can see different loss contributes differently to model performance. Adding the silence loss only to the phoneme classification loss has little gain in the performance. However, when combined with attention regularization, it improves a lot in F1 and R-value. We can further illustrate the effect of attention regularization by visualizing the output soft alignment from the the attention matrix. From Figure 2 and 3, we can see that with attention regularization, the attention matrix distribution can be closer to the true alignment, which can facilitate the phoneme classification.

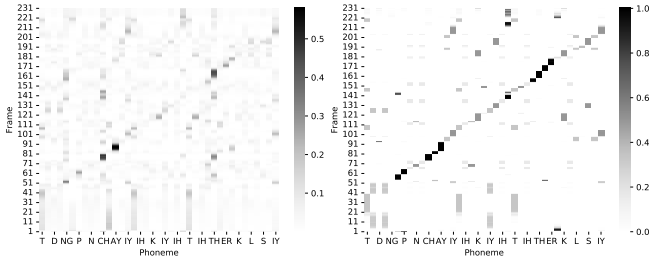


**Fig. 2**. Attention matrix W/O and W attention regularization

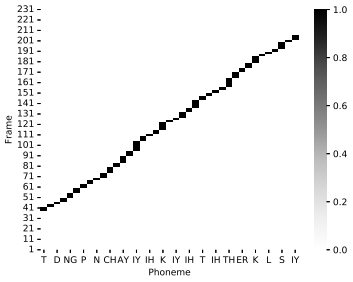


**Fig. 3**. Golden phoneme alignment

**Table 3**. Comparison results with different losses

| Model | P(%) | R(%) | F1(%) | R-val(%) |
|---|---|---|---|---|
| Loss(phone) | 92.96 | 95.0 | 93.97 | 94.73 |
| Loss(phone+sil) | 93.36 | 94.19 | 93.77 | 94.69 |
| Loss(phone+atten) | 93.1 | 95.9 | 94.28 | 94.9 |
| Loss(all) | 93.42 | 95.96 | **94.67** | **95.18** |

## 4. CONCLUSION

In this paper, we propose a method for phoneme segmentation. The speech feature representations are extracted from a pre-trained Wave2vec 2.0 acoustic encoder, and phoneme representations are derived from the combination of positional and phoneme embeddings. The attention mechanism is utilized to fuse the speech and phoneme representations as well as output the soft alignments. Then, the fused representations are fed into classification of each frame to obtain the final segmentation. The soft alignments indicated by the attention matrix can be further utilized to regularize the attention matrix based on the attention loss. Experimental results based on the TIMIT and Buckeye corpora show the proposed method achieves the SOTA performance with F1-score and R-value. In the future, we will apply the proposed method to other tasks such as phoneme mispronunciation detection.

## 5. REFERENCES


[1] Joseph Keshet, David Grangier, and Samy Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[2] Binghuai Lin and Liyuan Wang, "Deep feature transfer learning for automatic pronunciation assessment," *Proc. Interspeech 2021*, pp. 4438–4442, 2021.

[3] David Rybach, Christian Gollan, Ralf Schluter, and Hermann Ney, "Audio segmentation for speech recognition using segment features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4197–4200.

[4] Paul Michel, Okko Räsänen, Roland Thiolliere, and Emmanuel Dupoux, "Blind phoneme segmentation with temporal prediction errors," *arXiv preprint arXiv:1608.00508*, 2016.

[5] Felix Kreuk, Joseph Keshet, and Yossi Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," *arXiv preprint arXiv:2007.13465*, 2020.

[6] Lijiang Chen, Xia Mao, and Hong Yan, "Text-independent phoneme segmentation combining egg and speech data," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 6, pp. 1029–1037, 2016.

[7] Joerg Franke, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker, and Alex Waibel, "Phoneme boundary detection using deep bidirectional lstms," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[8] Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi, "Phoneme boundary detection using learnable segmental features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8089–8093.

[9] John-Paul Hosom, "Automatic phoneme alignment based on acoustic-phonetic modeling," in *INTERSPEECH*. Citeseer, 2002.

[10] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[11] Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Jiahong Yuan, Wen Wang, and Mark Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5552–5556.

[12] Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Dan Chazan, "Phoneme alignment based on discriminative learning," *Ninth European Conference on Speech Communication and Technology, 2005*, 2005.

[13] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[16] Philipp Koehn and Rebecca Knowles, "Six challenges for neural machine translation," *arXiv preprint arXiv:1706.03872*, 2017.

[17] Yingzhu Zhao, Chongjia Ni, Cheung-Chi Leung, Shafiq R Joty, Eng Siong Chng, and Bin Ma, "Cross attention with monotonic alignment for speech transformer.," in *INTERSPEECH*, 2020, pp. 5031–5035.

[18] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[19] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[21] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[22] K-F Lee and H-W Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.