*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation

*Felix Kreuk*[1], *Joseph Keshet*[1], *Yossi Adi*[2]

[1]Bar-Ilan University
[2]Facebook AI Research
felix.kreuk@gmail.com

## Abstract

We propose a self-supervised representation learning model for the task of unsupervised phoneme boundary detection. The model is a convolutional neural network that operates directly on the raw waveform. It is optimized to identify spectral changes in the signal using the Noise-Contrastive Estimation principle. At test time, a peak detection algorithm is applied over the model outputs to produce the final boundaries. As such, the proposed model is trained in a fully unsupervised manner with no manual annotations in the form of target boundaries nor phonetic transcriptions. We compare the proposed approach to several unsupervised baselines using both TIMIT and Buckeye corpora. Results suggest that our approach surpasses the baseline models and reaches state-of-the-art performance on both data sets. Furthermore, we experimented with expanding the training set with additional examples from the Librispeech corpus. We evaluated the resulting model on distributions and languages that were not seen during the training phase (English, Hebrew and German) and showed that utilizing additional untranscribed data is beneficial for model performance. Our implementation is available at: https://github.com/felixkreuk/UnsupSeg.

**Index Terms**: Unsupervised Phoneme Segmentation, Self-Supervised Learning, Contrastive Noise Estimation

## 1. Introduction

*Phoneme Segmentation* or *Phoneme Boundary Detection* is an important precursor task for many speech and audio applications such as Automatic Speech Recognition (ASR) [1, 2, 3], speaker diarization [4], keyword spotting [5], and speech science [6, 7].

The task of phoneme boundary detection has been explored under both supervised and unsupervised settings [8, 9, 10, 11]. Under the supervised setting two schemes have been considered: *text-independent* speech segmentation and *phoneme-to-speech alignment* also known as *forced alignment*, which is a text-dependent task. In the former setup, the model is provided with target boundaries, while in the latter setup, the model is provided with additional information in the form of a set of pronounced or presumed phonemes. In both schemes, the goal is to learn a function that maps the speech utterance to the target boundaries as accurately as possible.

However, creating annotated data of phoneme boundaries is a strenuous process, often requiring domain expertise, especially in low-resource languages [12]. As a consequence, unsupervised methods and Self-Supervised Learning (SSL) methods, in particular, are highly desirable and even essential.

In unsupervised phoneme boundary detection, also called *blind-segmentation* [13, 10], the model is trained to find phoneme boundaries using the audio signal only. In the self-supervised setting, the unlabeled input is used to define an auxiliary task that can generate labeled pseudo training data. This can then be used to train the model using supervised techniques. SSL has been proven to be effective in natural language processing [14, 15], vision [16], and recently has been shown to generate a useful representation for speech processing [17, 18].

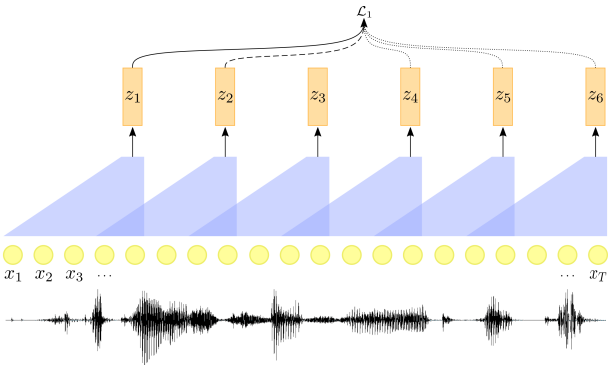



Figure 1: *An illustration of our model and SSL training scheme. The solid line represents a reference frame* $\mathbf{z}_1$*, the dashed line represents its positive pair* $\mathbf{z}_2$*, and the dotted lines represent negative distractor frames randomly sampled from the signal.*

Most of the SSL work in the domain of speech processing and recognition has been focused on extracting acoustic representations for the task of ASR [17, 18]. However, it remains unclear how effective SSL methods are when applied to other speech processing applications.

In this work, we explore the use of SSL for phoneme boundary detection. Specifically, we suggest learning a feature representation from the raw waveform to identify spectral changes and detect phoneme boundaries accurately. We optimize a Convolutional Neural Network (CNN) using the Noise Contrastive Estimation principle [19] to distinguish between pairs of *adjacent frames* and pairs of random distractor pairs. The proposed model is depicted in Figure 1. During inference, a peak-detection algorithm is applied over the model outputs to produce the final segment boundaries.

We evaluate our method on the TIMIT [20] and Buckeye [21] datasets. Results suggest that the proposed approach is more accurate than other state-of-the-art unsupervised segmentation methods. We conducted further experiments with larger amount of untranscribed data that was taken from the Librispeech corpus. Such an approach proved to be beneficial for better overall performance on unseen languages.



http://dx.doi.org/10.21437/Interspeech.2020-2398

**Our contributions:**

- We demonstrated the efficiency of SSL, in terms of model performance, for learning effective representations for unsupervised phoneme boundary detection.
- We provide SOTA results in the task of unsupervised phoneme segmentation on several datasets.
- We provide empirical evidence that leveraging more unlabeled data leads to better overall performance on unseen languages.

The paper is organized as follows: In Section 3 we formally set the notation and definitions used throughout the paper as well as the proposed model. Section 3 provides empirical results and analysis. In Section 2 we refer to the relevant prior work. We conclude the paper with a discussion in Section 5.

## 2. Related work

The task of phoneme boundary detection was explored in various settings. Under the supervised setting, the most common approach is the forced alignment. In this setup, previous work mainly involved hidden Markov models (HMMs) or structured prediction algorithms on handcrafted input features [22, 23]. In the text independent setup, most previous work reduced the task of phoneme segmentation to a binary classification at each time-step [24, 9]. More recently, [8] suggested using an RNN-coupled with structured loss parameters.

Under the unsupervised setting, the speech utterance is provided by itself with no boundaries as supervision. Traditionally, signal processing methods were used to detect spectral changes over time [25, 26, 27, 13], such areas of change were presumed to be the boundary of a speech unit. Recently, Michel *et al.* [10] suggested training a next-frame prediction model using HMM or RNN. Regions of high prediction error were identified using peak detection and flagged as phoneme boundaries. More recently, Wang *et al.* [28] suggested training an RNN autoencoder and tracking the norm of various intermediate gate values (forget-gate for LSTM and update-gate for GRU). To find phoneme boundaries, similar peak detection techniques were used on the gate norm over time.

In the field of self-supervised learning, Van Den Oord *et al.* [17] and Schneider *et al.* [18] suggested to train a Convolutional neural network to distinguish true future samples from random distractor samples using a probabilistic contrastive loss. Also called Noise Contrastive Estimation, this approach exploits unlabeled data to learn a representation in an unsupervised manner. The resulting representation proved to be useful for a variety of downstream supervised speech tasks such as ASR and speaker identification.

## 3. Model

Following the recent success of contrastive self-supervised learning [16, 17, 18], we propose a training scheme for learning useful representations for unsupervised phoneme boundary detection. We denote the domain of audio samples by $\mathcal{X} \subset \mathbb{R}$. The representation for a raw speech signal is therefore a sequence of samples $\mathbf{x} = (x_1, \ldots, x_T)$, where $x_t \in \mathcal{X}$ for all $1 \leq t \leq T$. The length of the input signal varies for different inputs, thus the number of input samples in the sequence, $T$, is not fixed. We denote by $\mathcal{X}^*$ the set of all finite-length sequences over $\mathcal{X}$.

Denote by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_L)$ a sequence of spectral representations sampled at a low frequency. Each element in the sequence is an $N$-dimensional real vector, $\mathbf{z}_i \in \mathcal{Z} \subseteq \mathbb{R}^N$ for $1 \leq i \leq L$. Every element $\mathbf{z}_i$ corresponds to a 10 ms frame of audio with a processing window of 30 ms. Let $\mathcal{Z}^*$ denote all finite-length sequences over $\mathcal{Z}$.

We learn an encoding function $f : \mathcal{X}^* \to \mathcal{Z}^*$, from the domain of audio sequences to the domain of spectral representations. The function $f$ is optimized to distinguish between pairs of *adjacent* frames in the sequence $\mathbf{z}$ and pairs of randomly sampled distractor frames from $\mathbf{z}$. Denote by $D(\mathbf{z}_i)$ the set non-adjacent frames to $\mathbf{z}_i$,

$$D(\mathbf{z}_i) = \{\mathbf{z}_j : |i - j| > 1 \ , \ \mathbf{z}_j \in \mathbf{z}\}. \tag{1}$$

Practically we use $K$ randomly selected frames from $D(\mathbf{z}_i)$, and denote it by $D_K(\mathbf{z}_i) \subset D(\mathbf{z}_i)$. The loss for frame $\mathbf{z}_i$ is defined as,

$$\hat{\mathcal{L}}(\mathbf{z}_i, D_K(\mathbf{z}_i)) = -\log \frac{e^{\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_{i+1})}}{\sum_{\mathbf{z}_j \in \{\mathbf{z}_{i+1}\} \cup D_K(\mathbf{z}_i)} e^{\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)}}, \tag{2}$$

where $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / ||\boldsymbol{u}|| \, ||\boldsymbol{v}||$ denotes the cosine similarly between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Overall, given a training set of $m$ examples $S = \{\mathbf{x}_i\}_{i=1}^m$, we would like to minimize the following objective function,

$$\mathcal{L} = \sum_{\mathbf{x} \in S} \sum_{\mathbf{z}_i \in f(\mathbf{x})} \hat{\mathcal{L}}(\mathbf{z}_i, D_K(\mathbf{z}_i)) \tag{3}$$

During inference, we receive a new utterance $\mathbf{x}$. We then apply the encoding function to get $\mathbf{z} = f(\mathbf{x})$. We set the score for a boundary at time $i$ to be the dissimilarity between the $i$-th frame and the $i+1$-th frame for $i = 1, \ldots, L-1$. That is

$$\mathrm{score}(\mathbf{z}_i) = -\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_{i+1}) \ . \tag{4}$$

Intuitively, $\mathrm{score}(\mathbf{z}_i)$ can be interpreted as the model's confidence that the next frame $\mathbf{z}_{i+1}$ belongs to a different segment than that of the current frame $\mathbf{z}_i$. Thus, times with high dissimilarity values are associated with segment changes, and are considered as candidates for segment boundaries. We apply a peak detection algorithm over the dissimilarity values, $\mathrm{score}(\mathbf{z})$ to get the final segmentation. The frames for which the score exceeds a peak prominence of $\delta$ are predicted as boundaries. The optimal value of $\delta$ is tuned in a cross-validation procedure.

Figure 2 presents an example utterance from TIMIT. The power spectrum of the utterance is presented in (a), the score function is presented in (b) and the corresponding learned representation $\mathbf{z}$ in (c).

## 4. Experiments

In this section, we provide a detailed description of the experiments. We start by presenting the experimental setup. Then we outline the evaluation method. We conclude this section with experimental results and analysis.

### 4.1. Experimental setup

The function $f$ was implemented as a convolutional neural network, constructed of 5 blocks of 1-D strided convolution, followed by Batch-Normalization and a Leaky ReLU [30] nonlinear activation function. The network $f$ has kernel sizes of $(10, 8, 4, 4, 4)$, strides of $(5, 4, 2, 2, 2)$ and 256 channels per layer. Finally, the output was linearly projected by a fully connected-layer. Overall the model was similar to the one proposed by [17, 18]. However, unlike the aforementioned prior work, the proposed model does not utilize a *context network*.

Table 1: *Comparison of phoneme segmentation models using TIMIT and Buckeye data sets. Precision and recall are calculated with tolerance value of 20 ms. Results marked with * are reported using our own optimization.*

| | | TIMIT | | | | Buckeye | | | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | Model | Precision | Recall | F1 | R-val | Precision | Recall | F1 | R-val |
| Unsupervised | Hoang *et al.* [29] | - | - | 78.20 | 81.10 | - | - | - | - |
| | Michel *et al.* [10] | 74.80 | 81.90 | 78.20 | 80.10 | 69.34* | 65.14* | 67.18* | 72.13* |
| | Wang *et al.* [28] | - | - | - | 83.16 | 69.61* | 72.55* | 71.03* | 74.83* |
| | **Ours** | **83.89** | **83.55** | **83.71** | **86.02** | **75.78** | **76.86** | **76.31** | **79.69** |
| Supervised | King *et al.*[24] | 87.00 | 84.80 | 85.90 | 87.80 | - | - | - | - |
| | Franke *et al.*[9] | 91.10 | 88.10 | 89.6 | 90.80 | 87.80 | 83.30 | 85.50 | 87.17 |
| | Kreuk *et al.*[8] | 94.03 | 90.46 | 92.22 | 92.79 | 85.40 | 89.12 | 87.23 | 88.76 |

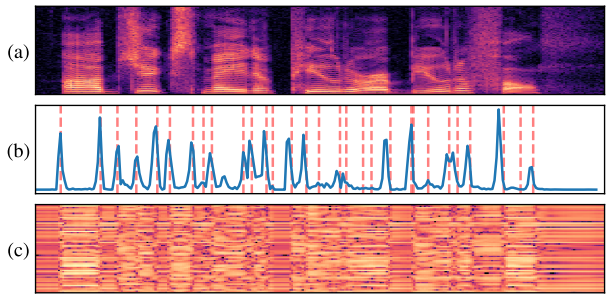



Figure 2: *An illustration of the prediction produced by our model: (a) the original spectrogram; (b) our model's output at each time step, red dashed lines represent the ground truth segmentation; (c) the learned representation* **z**.

Our experiments with such a network led to inferior performance, and therefore this component was omitted from the final model architecture.

We optimized the model using a batch size of 8 examples and a learning-rate of 1e-4 for 50 epochs. We follow an early-stopping criterion computed over the validation set. All reported results are averaged over a set of 3 runs using cross-validation with different random seed values. To get $D_K$ we experimented $K \in \{1, 3, 5, 7, 10\}$, but did not observe significant differences in performance.

We evaluated our model on both TIMIT and Buckeye corpora. For the TIMIT corpus, we used the standard train/test split, where we randomly sampled 10% of the training set for validation. For Buckeye, we split the corpus at the speaker level into training, validation, and test sets with a ratio of 80/10/10. Similarly to [8], we split long sequences into smaller ones by cutting during noises, silences, and un-transcribed segments. Overall, each sequence started and ended with a maximum of 20 ms of non-speech[1].

### 4.2. Evaluation method

Following previous work on phoneme boundary detection [10, 28], we evaluated the performance of the proposed models and baseline models using precision ($P$), recall ($R$) and F1-score with a tolerance level of 20 ms.

A drawback of the F1-score for boundary detection is its sensitivity to over-segmentation. A naive segmentation model that outputs a boundary every 40 ms may yield a high F1-score by achieving high recall at the cost of low precision. The authors in [31] proposed a more robust complementary metric denoted as *R-value*:

$$\begin{aligned} &\text{R-value} = 1 - \frac{|r_1| + |r_2|}{2} \\ &r_1 = \sqrt{(1-R)^2 + (OS)^2}, \;\; r_2 = \frac{-OS + R - 1}{\sqrt{2}} \end{aligned} \tag{5}$$

where $OS$ is an over-segmentation measure, defined as $OS = R/P - 1$. Overall the performance is presented in terms of Precision, Recall, F1-score and R-value.

### 4.3. Results

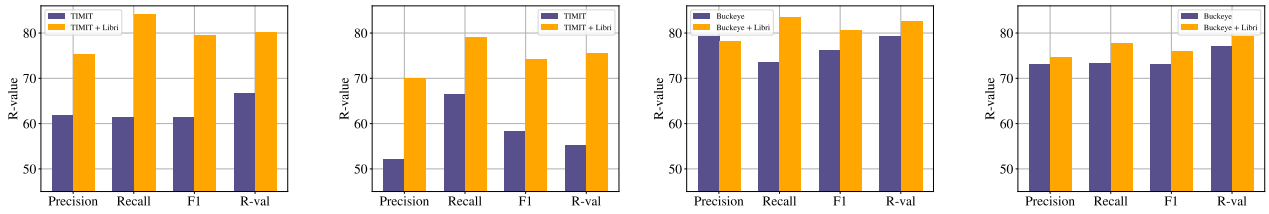
In Table 1 we compared the proposed model against several unsupervised phoneme segmentation baselines: Hoang *et al.* [29], Michel *et al.* [10], and Wang *et al.* [28]. We also report results for SOTA supervised algorithms in order to gauge the gap between the unsupervised and supervised methods. As the unsupervised baselines did not report results for the Buckeye data set, and there are no pre-trained models available, we optimized these models locally. For a fair comparison we verified that the performance of the reproduced models is comparable to the one originally reported on TIMIT. These results are marked with *.

Results suggest that the proposed model is superior to the baseline models over all metrics on both corpora. Notice, for the TIMIT benchmark, the proposed model achieves comparable results to a supervised method based on a Kernel-SVM [24]. Additionally, as opposed to the reported unsupervised baselines which are built using Recurrent Neural Networks, our model is mainly composed of convolutional operations, hence can be parallelized over the temporal axis.

### 4.4. The effect of more training data

By not relying on manual annotations, SSL methods allow leveraging large unlabeled corpora for additional training data. In this sub-section we explored the effect of expanding the training set with additional examples from the Librispeech corpus [32]. We evaluated the model under the following schemes: (i) training distribution and test distribution match; (ii) test distribution is different from the training set distribution, but both are from the same language; and (iii) test and training distributions are from different languages. In the following experiments, we denote by TIMIT+ and Buckeye+ the augmented

[1] All experiments were conducted at Bar-Ilan university.




(a) *Hebrew* (b) *German* (c) *Hebrew* (d) *German*

Figure 3: *Precision, Recall, F1, and R-value as a function of data added from Librispeech. All models were trained on English training data (sub figures (a) and (b) on TIMIT while sub figures (c) and (d) on Buckeye) and evaluated on both Hebrew and German data sets.*

Table 2: *Analysis of model performance on the TIMIT and Buckeye test sets before and after augmenting them with examples from Librispeech.*

| Training set | Test set | P | R | F1 | R-val |
|---|---|---|---|---|---|
| TIMIT | TIMIT | 83.89 | 83.55 | 83.71 | 86.02 |
| TIMIT+ | TIMIT | **84.11** | **84.17** | **84.13** | **86.40** |
| Buckeye | Buckeye | **75.78** | 76.86 | 76.31 | 79.69 |
| Buckeye+ | Buckeye | 74.92 | **79.41** | **77.09** | **79.82** |

Table 3: *Analysis of approach when evaluating the model on a test set that originates from a different distribution than that of the training set.*

| Training set | Test set | P | R | F1 | R-val |
|---|---|---|---|---|---|
| TIMIT | Buckeye | 67.48 | 73.71 | 70.41 | 73.10 |
| TIMIT+ | Buckeye | **71.17** | **81.66** | **76.05** | **76.53** |
| Buckeye | TIMIT | **86.26** | 79.63 | 82.80 | 84.61 |
| Buckeye+ | TIMIT | 86.19 | **80.10** | **83.03** | **84.90** |

versions of TIMIT and Buckeye, respectively. To better match recording conditions we chose different partition from Librispeech to augment TIMIT and Buckeye. For TIMIT+ we used the "train-clean-100" partition from Librispeech, while for Buckeye+ we used the "train-other-500" partition from Librispeech.

**In-domain test set** Results are summarized in Table 2. Surprisingly, the models trained on the augmented training sets showed minor improvements over the original models trained on the TIMIT and Buckeye data sets. In order to better understand the effect of more training data on model performance, we explore the use of out-of-domain test sets in the following paragraphs.

**Out-of-domain test set** We repeated the experiment from the previous paragraph, however this time with a cross dataset evaluation. In other words, we optimized a model on TIMIT and tested it on Buckeye and vice-versa. Results are summarized in Table 3. It can be seen that in cases where the training set and the test set originate from the same distribution (Table 2), adding more data leads to minor improvements in model performance. However, when these are coming from mismatched distributions as seen in Table 3, adding more data leads to an improvement in performance. For the TIMIT data set, the R-value for the model trained on TIMIT+ was improved by 3.43 points. For the Buckeye data set, we observed a smaller increase in performance.

**Multi-lingual evaluation** Finally, we analyzed the effect of more training data in the multi-lingual setup. To that end, we evaluated the proposed models, trained on TIMIT, TIMIT+, Buckeye, and Buckeye+ (English data), using two data sets from unseen languages. Specifically, we used a Hebrew data set [33] and the PHONDAT German data set [34] as test sets. Figure 3 presents the Precision, Recall, F1, and R-value for both data sets with and without additional training data from Librispeech.

Results suggest that utilizing additional unlabeled data yields an increase in performance on unseen languages. For example, when evaluated on the German data set PHONDAT, the TIMIT+ model improved from an R-value of 55.34 to an R-value of 75.58, while on the Hebrew data set the Buckeye+ model improved from an R-value of 79.25 to an R-value of 82.63. Notice, the improvement using TIMIT+ is larger by one order of magnitude comparing to the Buckeye+ improvement. One possible explanation for that is TIMIT being significantly smaller comparing to Buckeye, hence benefiting more from additional data. These results highlight the importance of additional diverse data sets in cases where there is a mismatch between training set and test set languages. Moreover, this suggests that the representations obtained by the suggested model are not tightly coupled with language-specific features.

## 5. Discussion and future work

In this work we empirically demonstrated the efficiency of self-supervised methods in terms of model performance for the task of unsupervised phoneme boundary detection. Our model reached SOTA results on both TIMIT and Buckeye data sets under the unsupervised setting, as well as showed promising results in terms of closing the gap between unsupervised and supervised methods. Moreover, we empirically demonstrated that using diverse datasets and leveraging more training data produced models with better overall performance on out-of-domain data coming from Hebrew and German.

For future work, we will explore the semi-supervised setting, where we are provided with a limited amount of manually annotated data. Additionally, we will explore the use of the proposed method on low-resource languages and under "in-the-wild" conditions. Lastly, we would like to explore the viability of such unsupervised segmentation methods in an unsupervised ASR pipeline.

# 6. References


[1] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz, "Transcribing radio news," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE, 1996, pp. 598–601.

[2] D. Rybach, C. Gollan, R. Schluter, and H. 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.

[3] C.-K. Yeh, J. Chen, C. Yu, and D. Yu, "Unsupervised speech recognition via segmental empirical output distribution matching," *arXiv preprint arXiv:1812.09323*, 2018.

[4] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.

[5] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[6] Y. Adi, J. Keshet, E. Cibelli, E. Gustafson, C. Clopper, and M. Goldrick, "Automatic measurement of vowel duration via structured prediction," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4517–4527, 2016.

[7] Y. Adi, J. Keshet, and M. Goldrick, "Vowel duration measurement using deep neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[8] F. Kreuk, Y. Sheena, J. Keshet, and Y. Adi, "Phoneme boundary detection using learnable segmental features," *arXiv preprint arXiv:2002.04992*, 2020.

[9] J. Franke, M. Mueller, F. Hamlaoui, S. Stueker, and A. Waibel, "Phoneme boundary detection using deep bidirectional lstms," in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[10] P. Michel, O. Räsänen, R. Thiolliere, and E. Dupoux, "Blind phoneme segmentation with temporal prediction errors," *arXiv preprint arXiv:1608.00508*, 2016.

[11] O. Rasanen, "Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, no. 36, 2014.

[12] M. Goldrick, J. Keshet, E. Gustafson, J. Heller, and J. Needle, "Automatic analysis of slips of the tongue: Insights into the cognitive architecture of speech production," *Cognition*, vol. 149, pp. 31–39, 2016.

[13] O. Räsänen, U. K. Laine, and T. Altosaar, "Blind segmentation of speech using non-linear filtering methods," *Speech Technologies*, pp. 105–124, 2011.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[17] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[19] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.

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

[21] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," *Columbus, OH: Department of Psychology, Ohio State University*, 2007.

[22] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, "Phoneme alignment based on discriminative learning," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[24] S. King and M. Hasegawa-Johnson, "Accurate speech segmentation by mimicking human auditory processing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8096–8100.

[25] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Ninth International Conference on Spoken Language Processing*, 2006.

[26] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–937.

[27] G. Almpanidis and C. Kotropoulos, "Phonemic segmentation using the generalised gamma distribution and small sample bayesian information criterion," *Speech Communication*, vol. 50, no. 1, pp. 38–55, 2008.

[28] Y.-H. Wang, C.-T. Chung, and H.-y. Lee, "Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries," *arXiv preprint arXiv:1703.07588*, 2017.

[29] D.-T. Hoang and H.-C. Wang, "Blind phone segmentation based on spectral change detection using legendre polynomial approximation," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 797–805, 2015.

[30] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[31] O. J. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the r-value," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] A. Ben-Shalom, J. Keshet, D. Modan, and A. Laufer, "Automatic tools for analyzing spoken hebrew."

[34] H. G. Tillmann and B. Pompino-Marschall, "Theoretical principles concerning segmentation, labelling strategies and levels of categorical annotation for spoken language database systems," in *Third European Conference on Speech Communication and Technology*, 1993.