*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Gate Activation Signal Analysis for Gated Recurrent Neural Networks and Its Correlation with Phoneme Boundaries

*Yu-Hsuan Wang, Cheng-Tao Chung, Hung-yi Lee*

College of Electrical Engineering and Computer Science, National Taiwan University

r04922167@ntu.edu.tw, f01921031@ntu.edu.tw, hungyilee@ntu.edu.tw

## Abstract

In this paper we analyze the gate activation signals inside the gated recurrent neural networks, and find the temporal structure of such signals is highly correlated with the phoneme boundaries. This correlation is further verified by a set of experiments for phoneme segmentation, in which better results compared to standard approaches were obtained.

**Index Terms**: autoencoder, recurrent neural network

## 1. Introduction

Deep learning has achieved great success in many areas [1][2][3]. For problems related to sequential data such as audio, video and text, significant improvements have been achieved by Gated Recurrent Neural Networks (GRNN), in which the hidden neurons form a directed cycle suitable for processing sequential data [4][5][6][7]. In addition to taking the neural network outputs to be used in the target applications, internal signals within the neural networks were also found useful. A good example is the bottleneck features [8][9].

On the other hand, in the era of big data, huge quantities of unlabeled speech data are available but difficult to annotate, and unsupervised approaches to effectively extract useful information out of such unlabeled data are highly attractive [10][11]. Autoencoder structures have been used for extracting bottleneck features [12], while GRNN with various structures can be learned very well without labeled data. As one example, the outputs of GRNN learned in an unsupervised fashion have been shown to carry phoneme boundary information and used in phoneme segmentation [13][14].

In this paper, we try to analyze the gate activation signals (GAS) in GRNN, which are internal signals within such networks. We found such signals have temporal structures highly related to the phoneme boundaries, which was further verified by phoneme segmentation experiments.

## 2. Approaches

### 2.1. Gate Activation Signals (GAS) for LSTM and GRU

Recurrent neural networks (RNN) are neural networks whose hidden neurons form a directed cycle. Given a sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$, RNN updates its hidden state $\mathbf{h}_t$ at time index $t$ according to the current input $x_t$ and the previous hidden state $\mathbf{h}_{t-1}$. Gated recurrent neural networks (GRNN) achieved better performance than RNN by introducing **gates** in the units to control the information flow. Two popularly used gated units are LSTM and GRU [15][16].

The signals within an LSTM recurrent unit are formulated as:

$$f_t = \sigma(\mathbf{W}_f x_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (1)$$

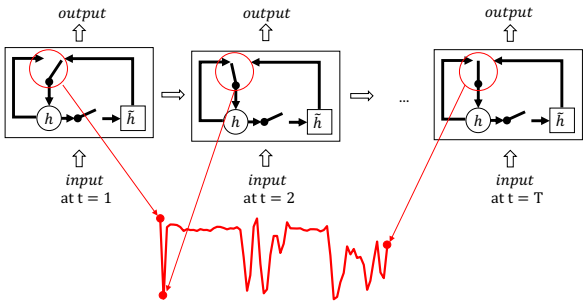


Figure 1: *The schematic plot regarding how the gate activation signals may imply for the update gate of a GRU.*

$$i_t = \sigma(\mathbf{W}_i x_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (2)$$

$$\widetilde{c_t} = tanh(\mathbf{W}_c x_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \widetilde{c_t} \quad (4)$$

$$o_t = \sigma(\mathbf{W}_o x_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (5)$$

$$h_t = o_t tanh(c_t) \quad (6)$$

where $f_t$, $i_t$, $o_t$, $c_t$, $\widetilde{c_t}$ and $h_t$ are the signals over the forget gate, input gate, output gate, cell state, candidate cell state and hidden state at time $t$, respectively; $\sigma(\cdot)$ and $tanh(\cdot)$ are the sigmoid and hyperbolic tangent activation functions respectively; $\mathbf{W}_\star$ and $\mathbf{U}_\star$ are the weight matrices and $\mathbf{b}_\star$ are the biases.

The GRU modulates the information flow inside the unit without a memory cell,

$$h_t = (1 - z_t) h_{t-1} + z_t \widetilde{h}_t \quad (7)$$

$$z_t = \sigma(\mathbf{W}_z x_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (8)$$

$$\widetilde{h}_t = tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (9)$$

$$r_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (10)$$

where $z_t$, $r_t$, $h_t$ and $\widetilde{h}_t$ are the signals over the update gate, reset gate, hidden state and candidate hidden state at time $t$, respectively; $\odot$ means element-wise product [17].

Here we wish to analyze the gate activations computed in equations (1), (2), (5), (8), (10) [18] and consider their temporal structures. For a GRNN layer consisting of $J$ gated units, we view the activations for a specific gate at time step $t$ as a vector $\mathbf{g}_t$ with dimensionality $J$, called *gate activation signals* (GAS). Here $\mathbf{g}_t$ can be $\mathbf{h}_t$, $\mathbf{i}_t$, $\mathbf{o}_t$, $\mathbf{z}_t$ or $\mathbf{r}_t$ above. Figure 1 is the schematic plot showing how GAS may imply for a gate in an gated unit.



http://dx.doi.org/10.21437/Interspeech.2017-877

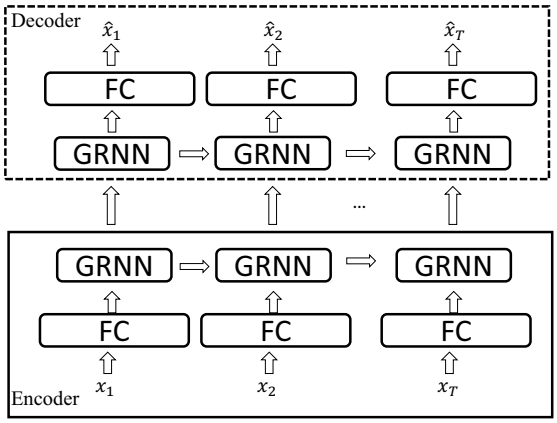


Figure 2: *AE-GRNN structure consisting of an encoder and a decoder. The encoder consists of gated recurrent layers (GRNN) stacking on feed-forward fully-connected layers (FC). Only one recurrent layer and one feed-forward layer are shown here for clarity. The structure of decoder is the mirrored structure of encoder.*

### 2.2. Autoencoder GRNN

Autoencoders can be trained in an unsupervised way, and have been shown to be very useful for many applications [19]. We can have an autoencoder with GRNN as in Figure 2 called AE-GRNN here. Given an input utterance represented by its acoustic feature sequence $\{x_1, x_2, ..., x_T\}$, at each time step $t$, AE-GRNN takes the input vector $x_t$, and produces the output $\hat{x}_t$, the reconstruction of $x_t$. Due to the recurrent structure, to generate $\hat{x}_t$, AE-GRNN actually considers all information up to $x_t$ in the sequence, $x_1, x_2, ..., x_t$, or $\hat{x}_t = \textit{AE-GRNN}(x_1, x_2, ..., x_t)$. The loss function $\mathcal{L}$ of AE-GRNN in (11) is the averaged squared $\ell$-2 norm for the reconstruction error of $x_t$.

$$\mathcal{L} = \sum_{n}^{N} \sum_{t}^{T_n} \frac{1}{d} \left\| x_t^n - \textit{AE-GRNN}(x_1^n, x_2^n, ..., x_t^n) \right\|^2 \quad (11)$$

where the superscript $n$ indicates the $n$-th training utterance with length $T_n$, and $N$ is the number of utterances used in training. $d$ indicates the number of dimensions of $x_t^n$.

## 3. Initial Experiments and Analysis

### 3.1. Experimental Setup

We conducted our initial experiments on TIMIT, including 4620 training utterances and 1680 testing utterances. The ground truth phoneme boundaries provided in TIMIT are used here. We train models on the training utterances, and perform analysis on the testing ones.

In the AE-GRNN tested, both the encoder and the decoder consisted of a recurrent layer and a feed-forward layer. The recurrent layers consisted of 32 recurrent units, while the feed-forward layers consisted of 64 neurons. We used ReLU as the activation function for the feed-forward layers [20]. The dropout rate was set to be 0.3 [21]. The networks were trained with Adam [22]. The acoustic features used were the MFCCs of 39-dim with utterance-wise cepstral mean and variance normalization (CMVN) applied.

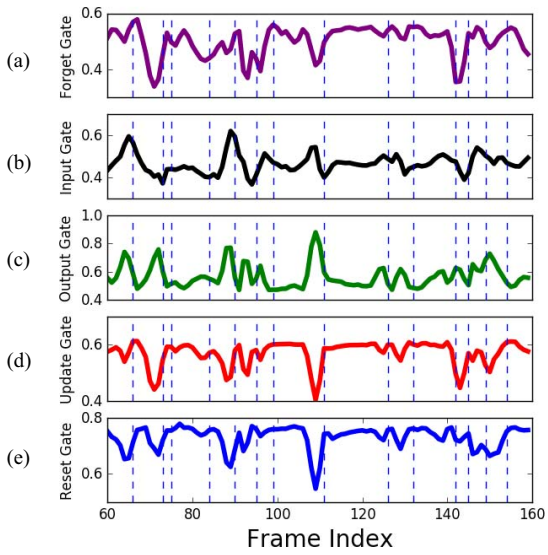


Figure 3: *The means of the gate activation signals over all gated units for different gates with respect to the frame index. The blue dashed lines indicate the phoneme boundaries.*

### 3.2. Initial Results and Observations

Figure 3 shows the means of the gate activation signals over all gated units in the encoder of AE-GRNN with respect to the frame index, taken from an example utterance. The upper three subfigures (a)(b)(c) are for LSTM gates, while the lower two (d)(e) for GRU gates. The temporal variations of GRU gates are similar to the forget gate of LSTM to some degree, and looks like the negative of the input gate of LSTM except for a level shift. More importantly, when compared with the phoneme boundaries shown as the blue dashed lines, a very strong correlation can be found. In other words, whenever the signal switches from a phoneme to the next across the phoneme boundary, the changes in signal characteristics are reflected in the gate activation signals. This is consistent to the previous finding that the sudden bursts of gate activations indicated that there were boundaries of phonemes in a speech synthesis task [23].

### 3.3. Difference GAS

With the above observations, we define difference GAS as follows. For a GAS vector at time index $t$, $\mathbf{g}_t$, we compute its mean over all units to get a real value $\bar{g}_t$. We can then compute the difference GAS as the following:

$$\Delta \bar{g}_t = \bar{g}_{t+1} - \bar{g}_t \quad (12)$$

The difference GAS can also be evaluated for each individual gated unit for each dimension of the vector $\mathbf{g}_t$,

$$\Delta \bar{g}_t^j = \bar{g}_{t+1}^j - \bar{g}_t^j \quad (13)$$

where $g_t^j$ is the $j$-th component of the vector $\mathbf{g}_t$. We plotted the difference GAS and the individual difference GAS for the first 8 units in a GRNN over the frame index for an example utterance as in Figure 4. We see those differences bear even stronger correlation with phoneme boundaries shown by vertical dashed

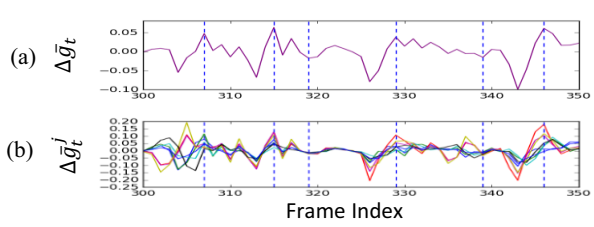


Figure 4: *Subfigure (a) shows the plot of $\Delta\bar{g}_t$ for forget gate of LSTM for an example utterance over frame index. Subfigure (b) shows the plots of $\Delta\bar{g}_t^j$ with different colors (only shows $j = 1$ to 8 for clarity). The blue dashed lines indicate phone boundaries.*

lines. All these results are consistent with the finding that the gate activations of forget gate over recurrent LSTM units in the same layer have close correlation with phoneme boundaries in speech synthesis [23], although here the experiments were performed with AE-GRNN.

# 4. Phoneme Segmentation

Because the GAS was found to be closely related to phoneme boundaries, we tried to use these signals to perform phoneme segmentation. The segmentation accuracy can be a good indicator to show the degree of the correlation between GAS and phoneme boundaries. In this section, we will first describe recurrent predictor model (RPM), an unsupervised phoneme segmentation approach, which servers as the baseline. Then we will describe how to use GAS in phoneme segmentation.

## 4.1. Baseline: Recurrent Predictor Model

RPM was proposed earlier to train GRNN without labeled data, and it was found the discontinuity on model outputs have to do with phoneme boundaries [13][14]. An RPM has only the lower half of Figure 2. The model output at time $t$, $\hat{x_t} = RPM(x_1, x_2, ..., x_t)$, is to predict the next input $x_{t+1}$. The loss function $\mathcal{L}$ used in training RPM is the averaged squared $\ell$-2 norm of prediction error,

$$\mathcal{L} = \sum_{n}^{N} \sum_{t}^{T_n - 1} \frac{1}{d} \left\| x_{t+1}^n - RPM(x_1^n, x_2^n, ..., x_t^n) \right\|^2 \qquad (14)$$

which is actually parallel to (11). The superscript $n$ indicates the $n$-th training utterance and $d$ indicates the number of dimensions of $x_t^n$. Because frames which are difficult to predict or with significantly larger errors are likely to be phoneme boundaries, the error signals $E_t$ of RPM,

$$E_t = \frac{1}{d} \left\| x_{t+1}^n - RPM(x_1^n, x_2^n, ..., x_t^n) \right\|^2 \qquad (15)$$

can be used to predict the phoneme boundary similar to GAS here. A time index is taken as a phoneme boundary if $E_t$ is a local maximum, that is $E_t > E_{t-1}$ and $E_t > E_{t+1}$, and $E_t$ exceeds a selected threshold.

## 4.2. GAS for Phoneme Segmentation

From Figure 4, a direct approach to use GAS for phoneme segmentation is to take a time index as a phoneme boundary if $\Delta\bar{g}_t$ is a local maximum, that is $\Delta\bar{g}_t > \Delta\bar{g}_{t-1}$ and $\Delta\bar{g}_t > \Delta\bar{g}_{t+1}$, and $\Delta\bar{g}_t$ exceeds a selected threshold.

GAS can also be integrated with RPM. Since RPM also includes GRNN within its structure, GAS can also be obtained and interpolated with the error signals obtained in (15) as in (16), where $w$ is the weight. A time index is taken as a phoneme boundary if $I_t$ is a local maximum and exceeds a selected threshold.

$$I_t = (1 - w)E_t + w\Delta\bar{g}_t \qquad (16)$$

# 5. Experiments Results for Phoneme Segmentation

Here we take phoneme segmentation accuracy as an indicator to show the correlation between GAS and phoneme boundaries. The setup is the same as in Section 3.1. In the segmentation experiments, a 20-ms tolerance window is used for evaluation. All GAS were obtained from the first recurrent layer. Different segmentation results were obtained according to different thresholds, we report the best results in the following tables.

## 5.1. R-value Evaluation

It is well-known that the F1-score is not suitable for segmentation, because over segmentation may give very high recall leading to high F1-score, even with a relatively low precision[14]. In our preliminary experiments, a periodic predictor which predicted a boundary for every 40 ms gave F1-score 71.07 with precision 55.13 and recall 99.99, which didn't look reasonable. It has been shown that a better evaluation metric is the R-value [24], which properly penalized the over segmentation phenomenon. The approach proposed in a previous work [25] achieved an R-value 76.0, while the 40-ms periodic predictor only achieved 30.53. Therefore, we chose to use R-value on the performance measure.

## 5.2. Comparison among different gates

The R-values using different gates of LSTM and GRU are shown in Table 1. The results for LSTM gates are consistent with the findings in the previous works [23][26]. In LSTM, the forget gate clearly captures the temporal structure most related to phoneme boundaries. GRU outperformed LSTM which is also consistent with earlier results [17][26]. The highest R-value is obtained with the update gate of GRU. The update gate in GRU is similar to the forget gate in LSTM. Both of them control whether the memory units should be overwritten. Interestingly, the reset gate of GRU achieved an R-value significantly higher than the corresponding input gate in LSTM. The reason is probably the location of reset gate. In GRU, the reset gate does not control the amount of the candidate hidden state independently, but shares some information of the update gate, thus has better access to more temporal information for phoneme segmentation [17]. The update gate in GRU was used for extracting GAS in the following experiments.

## 5.3. Comparison among different approaches

In Table 2, we compared the R-value obtained from the temporal information provided by RPM, without or with its GAS (rows (a)(b)(c)(d)). The best result in Table 1 is in row(e), which was obtained with update gates of GRU in AE-GRNN. We considered two structures of RPM: the same as AE-GRNN (4 layers) or only use the encoder part (2 layers, a feed-forward layer plus a recurrent layer). The latter used the same number

Table 1: *The comparison between different gates in gated recurrent neural networks.*

| Models | R-value |
|---|---|
| LSTM Forget Gate | 79.15 |
| LSTM Input Gate | 70.75 |
| LSTM Output Gate | 61.97 |
| GRU Update Gate | **82.54** |
| GRU Reset Gate | 78.94 |

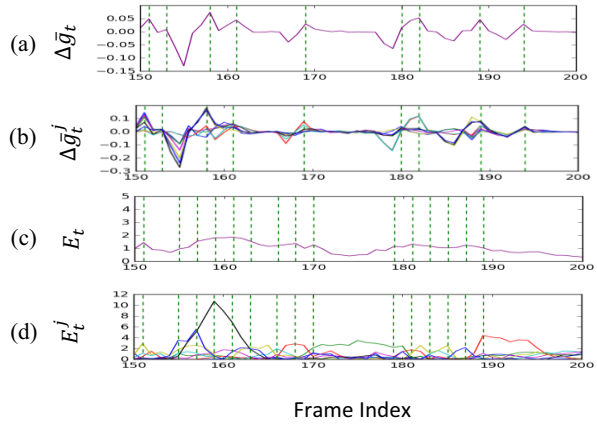


Figure 5: *Subfigures (a) and (c) show the plots of $\Delta\bar{g}_t$ and $E_t$, respectively, for an example utterance over frame index. Subfigures (b) and (d) show the plots of $\Delta\bar{g}_t^j$ and $E_t^j$, respectively, with different colors (only show $j = 1$ to 8 for clarity). The green dashed lines indicate segmentation results.*

of parameters as AE-GRNN. We also tested the conventional approach of using hierarchical agglomerative clustering (HAC) as shown in row (f) [27][28]. We further added a white noise with 6dB SNR to the original TIMIT corpus (the right column). The last row (g) is for the periodic predictor predicted a boundary every 80 ms, serving as a naive baseline. Precision-Recall curves in Figure 6 illustrate the overall performance on clean TIMIT corpus with different thresholds.

We found that the RPM performance was improved by the interpolation with GAS (RPM+GAS v.s. RPM). Larger improvements were gained when data became noisy. We further analyzed the segmentation results on clean corpus with the highest R-values of 2-layered RPM and AE-GRNN. We analyzed the results of AE-GRNN by observing $\Delta\bar{g}_t$ and $\Delta\bar{g}_t^j$. Likewise, we analyzed the results of RPM by observing $E_t$ and $E_t^j$, where $E_t^j$ indicates the squared prediction error in the $j^{th}$ dimension computed in (15). We showed their relations in Figure 5. We see that the curve of $E_t$ is smooth and significantly different from the sharp curve of $\Delta\bar{g}_t$. The smoothness led RPM to suffer from over segmentation. The smoothness was caused by the fact that there were always a subset of $E_t^j$ which were significantly large. On the other hand, the curves of $\Delta\bar{g}_t^j$ are more consistent. The consistency enables curve of $\Delta\bar{g}_t$ to be sharp and thus AE-GRNN would not suffer from over segmentation. This explains why GAS are helpful here.

Also, we can see RPM alone didn't benefit much from adding more layers (rows (c) v.s. (a)). Providing if RPM is powerful enough to predict next frames better, there will be no

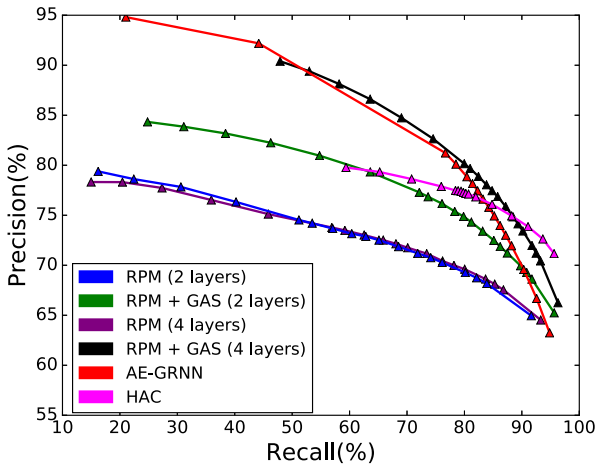


Figure 6: *The Precision-Recall curves of different approaches in Table 2. Different markers on the curves stand for the results of different thresholds.*

error signals and thus no temporal information. This side effect balanced the advantage of increased model size. Interestingly, not only GAS offered improvements (rows (b) v.s. (a) and rows (d) v.s. (c)), but with more layers, the interpolation with GAS achieved larger improvements (rows (d) v,s, (b)). The best performances in both clean and noisy corpora are achieved by 4-layered RPM with GAS. Last but not least, performance of approaches using GAS and HAC are more robust to noise.

Table 2: *The comparison of R-values for recurrent predictor model (RPM) without and with its internal GAS and GAS of AE-GRNN on clean and noisy TIMIT corpus. The performance of hierarchical agglomerative clustering (HAC) and a periodic predictor are also included.*

| Models | Clean | SNR-6dB |
|---|---|---|
| (a) RPM (2 layers) | 76.02 | 73.7 |
| (b) RPM + GAS (2 layers) | 79.94 | 79.16 |
| (c) RPM (4 layers) | 76.10 | 73.65 |
| (d) RPM + GAS (4 layers) | **83.16** | **81.54** |
| (e) AE-GRNN | 82.54 | 81.22 |
| (f) HAC | 81.61 | 80.41 |
| (g) Periodic Predictor | 62.17 | 62.17 |

## 6. Conclusions

We show that the gate activation signals (GAS) obtained in an unsupervised fashion have temporal structures highly correlated with the phoneme changes in the signals, and this correlation was verified in the experiments for phoneme segmentation. Also, our experiments showed that GAS bring improvements to RPM without additional parameters. Like bottleneck features, GAS are obtained from the element of neural networks, instead of networks' outputs, and both of them are shown to bring improvements. With the promising results of GAS shown in the paper, we hope GAS can brought the same improvements as the ones brought by bottleneck features.

# 7. References


[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1635–1638.

[3] B.-P. Network, "Handwritten digit recognition with," 1989.

[4] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," *CoRR*, vol. abs/1511.06984, 2015. [Online]. Available: http://arxiv.org/abs/1511.06984

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *CoRR*, vol. abs/1603.01360, 2016. [Online]. Available: http://arxiv.org/abs/1603.01360

[6] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[7] Y. Adi, J. Keshet, E. Cibelli, and M. Goldrick, "Sequence segmentation using joint rnn and structured prediction models," *arXiv preprint arXiv:1610.07918*, 2016.

[8] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–757.

[9] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Interspeech*, vol. 237, 2011, p. 240.

[10] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[11] C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8081–8085.

[12] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.

[13] J. F. Drexler, "Deep unsupervised learning from speech," Master's thesis, Massachusetts Institute of Technology, 2016.

[14] P. Michel, O. Räsänen, R. Thiollière, and E. Dupoux, "Improving phoneme segmentation with recurrent neural networks," *CoRR*, vol. abs/1608.00508, 2016. [Online]. Available: http://arxiv.org/abs/1608.00508

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[18] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[19] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder." in *Interspeech*. Citeseer, 2010, pp. 1692–1695.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5140–5144.

[24] O. J. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the r-value." in *Interspeech*, 2009, pp. 1851–1854.

[25] O. Räsänen, "Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level." in *CogSci*, 2014.

[26] W. Zaremba, "An empirical exploration of recurrent network architectures," 2015.

[27] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3989–3992.

[28] C.-a. Chan, "Unsupervised spoken term detection with spoken queries," Ph.D. dissertation, National Taiwan University, 2012.