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

Large scale evaluation of importance maps in automatic speech recognition

Viet Anh Trinh[1] *and Michael I Mandel*[1,2]

[1] The Graduate Center, CUNY, New York, USA
[2] Brooklyn College, CUNY, New York, USA
vtrinh@gradcenter.cuny.edu, mim@sci.brooklyn.cuny.edu

Abstract

This paper proposes a metric that we call the structured saliency benchmark (SSBM) to evaluate importance maps computed for automatic speech recognizers on individual utterances. These maps indicate time-frequency points of the utterance that are most important for correct recognition of a target word. Our evaluation technique is not only suitable for standard classification tasks, but is also appropriate for structured prediction tasks like sequence-to-sequence models. Additionally, we use this approach to perform a comparison of the importance maps created by our previously introduced technique using "bubble noise" to identify important points through correlation with a baseline approach based on smoothed speech energy and forced alignment. Our results show that the bubble analysis approach is better at identifying important speech regions than this baseline on 100 sentences from the AMI corpus.

Index Terms: importance map, saliency map, speech recognition, information bottleneck.

1. Introduction

Finding relevant information in input features X that is necessary for an output/task y has seen a surge of interest in the computer vision [1–5] and reinforcement learning communities [6–8]. [9] proposed the information bottleneck approach to address the problem and [10, 11] used the idea to improve model generalization. Our previous work proposed a correlational method to find regions of speech spectrograms that are important to a listener's correctly identifying the words it contains, and we applied it to both humans and automatic speech recognition (ASR) systems [12–14]. These "importance maps" or "saliency maps" reveal how the ASR uses speech features to derive the recognition. In this paper, we propose a method to evaluate the quality of predicted importance maps and apply them to saliency maps estimated for an ASR "listener."

The saliency map in speech has a similar meaning to the saliency map in computer vision. However, unlike in vision, where ground truth can be obtained from eye-tracking systems, in speech, we do not have a corresponding "ear-tracking" system. We thus propose a method to assess the quality of a predicted speech saliency map. The main idea of our approach is that the better the predicted saliency map, the higher the accuracy when the ASR uses only information from the important regions of the spectrogram. Similarly, if the important regions are removed from an observation, the ASR should have low accuracy.

To the best of our knowledge, we are among the first, if not the first, to propose a method to evaluate the saliency map of running sentences, a structured prediction problem. In computer vision, there is related work on evaluation methods for saliency maps in simple classification problems without ground truth. [15] proposed the MoRF method (Most Relevant First) to evaluate saliency maps by measuring model performance degradation when the n most relevant pixels are replaced by random values. [5] introduced the complementary LeRF method (Least Relevant First), where the least relevant features are removed. [16] recommended evaluating with a score measuring the area between the MoRF and LeRF curves created when the number of pixels n is varied.

Inspired by [5, 15, 16], we propose here an evaluation metric, the structured saliency benchmark (SSBM), that measures accuracy degradation when the most or least important time-frequency points are replaced with white noise in a structured prediction setting. A fundamental difference between our approach and these others is that they evaluate the accuracy of a single simple classifier, such as an image classifier, so they only consider how a saliency map affects the classification of a single object, not how it might affect other objects in the scene.

2. Method

The main idea of our method is to evaluate the quality of the predicted time-frequency importance regions for an utterance. Denote the predicted importance maps in the speech spectrogram from method M for word w as $I_M^w \in \{0,1\}^{F\times T}$, a binary matrix indicating whether time-frequency point $I_M^w(f,t)$ is important for the recognition of w (1) or not (0). If the ASR can correctly recognize word w and only word w using only the regions where $I_M^w = 1$ instead of using all the spectrogram points, and if it cannot recognize word w but can recognize all other words when presented with only the regions where $I_M^w = 0$, then we can conclude that method M has successfully identified the important regions for recognizing w. To measure this, we perform two tests. In the first case, we add noise everywhere in a sentence except the predicted important regions of w, which is equivalent to dropping the least relevant features (LeRF). In the second case, we add noise to the predicted important regions for w, equivalent to dropping the most relevant features (MoRF). To encourage specificity, saliency maps that select less signal energy as important are preferred to those that select more.

We define a new metric that we call the structured saliency benchmark (SSBM) to evaluate the accuracy of the analyzed words with respect to the accuracy of other words in the sentence and the predicted important speech energies.

$$\Delta_{\text{LeRF}} = \frac{a_w - a_o}{1 - e_{\text{LeRF}}} \qquad \Delta_{\text{MoRF}} = \frac{a_o - a_w}{e_{\text{MoRF}}} \tag{1}$$

$$SSBM = \Delta_{\text{LeRF}} + \Delta_{\text{MoRF}} \tag{2}$$

where a_w is the accuracy of analyzed word w, a_o is the averaged accuracy of the other words, e_{LeRF} is the proportion of energy that is dropped by the LeRF mask (dropped energy divided by utterance energy), and e_{MoRF} is the proportion of energy that is dropped by the MoRF mask. Thus, Δ_{LeRF} represents the accuracy of the analyzed word per unit (proportion) of energy,

http://dx.doi.org/10.21437/Interspeech.2020-2883

with the accuracy of other words as a penalty. We can see that if the importance maps of w are correct, then when the least important energy for w is removed, the accuracy of w, a_w, should be high while the accuracy of other words, a_o, should be low. Additionally, for two different importance maps with the same a_w and a_o, the map corresponding to higher e_{LeRF} (more unimportant energy dropped) should be better as should the one with the lower e_{MoRF} (less important energy preserved).

2.1. Saliency maps

We analyze the importance maps of two different approaches. The first is a bubble analysis method where a time-frequency point is predicted to be important when its audibility in noise is significantly correlated with speech intelligibility [12, 13]. The second is an energy-based baseline, where a time-frequency point in the spectrogram is predicted to be important when its energy is larger than a certain threshold. Future work will investigate methods based on feature gradients [2, 17–19], pertubation [4], etc., which are not straightforward to apply to structured prediction problem in speech.

The bubble analysis method [12, 13] identifies important regions by adding many instances of random noise to clean speech, then finding the spectrogram points that are revealed when the ASR recognizes the noisy speech correctly and hidden by noise when the ASR fails to recognize the utterance. Specifically, the noisy utterances are generated by adding many instances of random white noise to the clean speech to make these utterances inaudible. However, the noise level is decreased inside randomly placed oval-shaped bubbles to reveal the speech information inside. Denote as y_{ijk} the intelligibility, which has value one or zero (binary) when the ASR correctly or incorrectly recognizes the kth word in the jth noisy mixture of the ith clean utterance. In addition, the audibility $D_{ij}(f,t)$ is defined as a continuous variable that represents the inverse of the amount of noise added to a time-frequency point in a spectrogram, varying between zero (maximum noise) and one (no noise). A point-biserial correlation $c_{ik}(f,t)$ is computed between $D_{ij}(f,t)$ and y_{ijk}. The significance (p-value) of this correlation is examined under a two-sided t-test for every time-frequency point in the spectrogram [13]. The importance map is defined as the set of time-frequency points that have positive correlation and p-values smaller than a specific threshold.

We compare the bubble method with an energy-based baseline in which a time-frequency point in the spectrogram is considered important when its energy in a smoothed version of the spectrogram is greater than a certain threshold. Specifically, the linear frequency spectrogram has pre-emphasis applied, is converted to a mel spectrogram with 30 bins, and then is converted back to a linear frequency axis. The importance map of a word is then the set of high energy spectrogram points that are between the start and end frame of the target word in the forced alignment of the clean utterance produced by Kaldi.

2.2. LeRF and MoRF noise masks

The LeRF mask is created by adding maximum noise to unimportant regions while adding minimum noise to important regions. There is a transition between the two as shown in the top plot of Figure 1. The intention is that when maximum noise is added outside the important regions of a specific word, then the ASR should still be able to recognize this word, but should not be able to recognize the other words in the sentence. The procedure is slightly different for the two mask prediction algorithms, so each is described separately below.

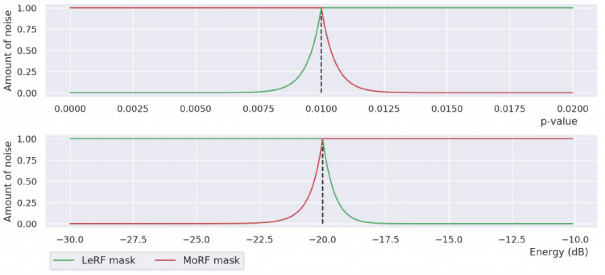

Figure 1: *Example mask transition functions for an arbitrary threshold. Top: Bubble analysis. Bottom: Energy-based*

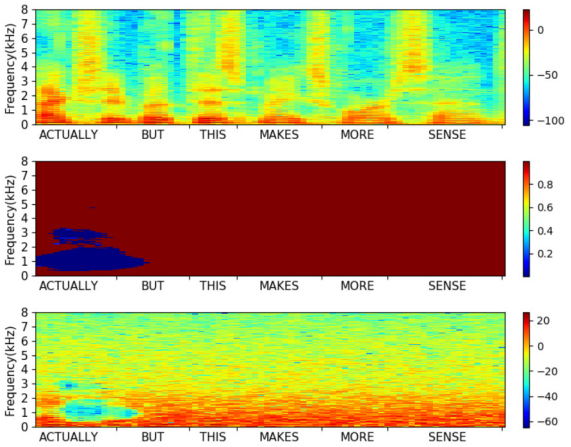

Figure 2: *Bubble analysis approach. From top to bottom: (a) Clean speech (b) LeRF mask created by dropping the least relevant features for the word "actually" with threshold* 4.64×10^{-7} *(time-frequency points that have p-value* $\geq 4.64 \times 10^{-7}$ *have a maximum amount of noise added to them). (c) Noisy mixture created by adding the mask in (b) to the clean speech in (a).*

The bubble analysis LeRF mask m^b_{LeRF}, at a single point is

$$q^b_{\text{LeRF}}(p) = -(d_1 - d_0)\frac{p-t}{\alpha t - t} \tag{3}$$

$$m^b_{\text{LeRF}}(p) = 10^{0.05\,\text{clip}(q^b_{\text{LeRF}}(p), d_0, d_1)} \tag{4}$$

where t is the threshold, p is the p-value of time-frequency points in the spectrogram, $\alpha < 1$ is a parameter controlling the size of the transition region while d_0 and d_1 control the minimum and maximum value of the mask, respectively.

The green line in the top plot in Figure 1 illustrates mask values for $t = 0.01$, $\alpha = 0.5$. In addition, $d_0 = -80$, $d_1 = 0$ leading to a minimum mask value of 0.0001 and maximum value of 1. As shown in this figure, a time-frequency point with a p-value larger than 0.01 has noise level 1 (maximum noise), while a point with a p-value smaller than 0.0075 has noise level 0.0001. Additionally, a visualization of a complete mask with threshold $t = 4.64 \times 10^{-7}$ is shown in the second row of Figure 2.

The bubble analysis MoRF mask is derived in a similar way as equations (3) and (4), however with $q^b_{\text{MoRF}}(p) = -q^b_{\text{LeRF}}(p)$. The red line in the top plot of Figure 1 shows the MoRF mask with the same parameters as the green line. In addition, a visualization of the mask is shown in the top plot of Figure 3. The MoRF and LeRF masks are almost complementary to each other, but are not exactly because the masks always decay smoothly towards 0 to mirror the logarithmic nature of intensity perception.

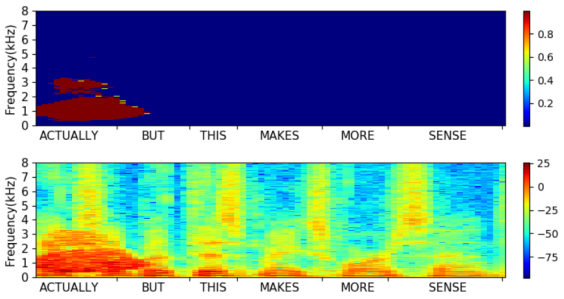

Figure 3: *Bubble analysis. Top: MoRF mask created by dropping the most important features of the word "actually" with threshold* 4.64×10^{-7}. *Bottom: Noisy mixture*

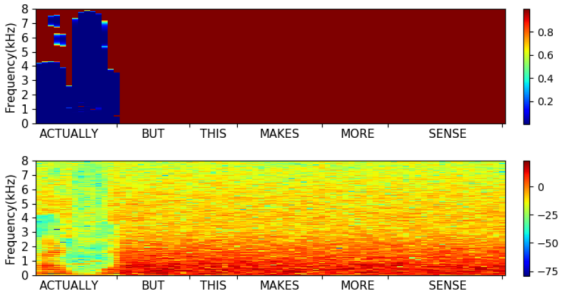

Figure 4: *Energy-based approach. Top: LeRF mask with* $t_{dB} = -20$. *Bottom: Noisy mixture*

Similarly, the LeRF mask for the energy-based approach is created by adding maximum noise to the time-frequency region with energy lower than a specific threshold t_{dB} in decibels (unimportant regions). The important regions have minimum noise added, except the transition area. The mask m is defined analogously to (4), but using

$$q^{e}_{\text{LeRF}}(a) = (d_1 - d_0)\frac{a - t}{\alpha t - t} \tag{5}$$

where a is the absolute magnitude of the time-frequency point in the spectrogram and $t = 10^{0.05 t_{\text{dB}}}$ is the threshold in magnitude. An example of the mask with a specific threshold $t_{\text{dB}} = -20$ dB is illustrated in the bottom plot of Figure 1 and Figure 4. The energy-based MoRF mask is formed by adding maximum noise to the time-frequency region with energy bigger than a specific threshold. The mask is derived the same as equation (5) except with $q^{e}_{\text{MoRF}}(a) = -q^{e}_{\text{LeRF}}(a)$.

To create the noisy speech, we multiply the spectrogram of a white noise signal by the mask and add the masked noise to the clean speech. Examples of the mask and the masked noisy speech are shown in the second and third rows of Figure 2.

3. Experimental setup

We utilize the AMI dataset [20], which includes 100 hours of English meeting recordings. We selected the Individual Headset Microphone (IHM) channels for our experiment. We followed the standard train/test split and chose 100 sentences (9 minutes) from the test set where the recognizer achieved 100% accuracy without additional noise added to be our set of clean speech. We created 1000 noisy mixtures for every clean utterance, leading to a dataset of 100,000 mixtures for the bubble analysis method.

We use Kaldi [21] as the ASR to perform the experiments. We choose the standard model in AMI recipe s5b with a time-delay neural network (TDNN) acoustic model and an n-gram language model from the SRI Language Modeling Toolkit (SRILM) [22]. The TDNN is a modification of a feed-forward neural network, where the hidden vector representation at a layer is derived from several vectors (window of size n) from the preceding layer. The time-domain utterances are sampled at 16 kHz and are transformed into the frequency domain using an STFT with window length 64 ms, and hop length 16 ms.

For the bubble analysis technique, we choose $d_0 = -80$, $d_1 = 0$ and $\alpha = 0.5$. We perform experiments with 25 different values of threshold t that are spaced evenly on a log scale from 10^{-8} to 10^{0}. For the energy-based technique, we use the same values of d_0, d_1, α, however we use 21 different values of thresholds t_{dB}, spaced evenly on a linear scale from -80 to 20 with a step size of 5.

4. Results

Here, we compare the bubble analysis and the energy-based approaches according to LeRF and MoRF curves and SSBM scores. Figure 6 allows a direct comparison between the two mask methods by characterizing each masked signal by the proportion of speech energy in the entire utterance that it obscures. This proportion could vary for different words at the same threshold, so this plot averages over masks that have the same proportion when rounded to the nearest percent.

The top plot of Figure 6 shows the accuracy of analyzed words when the least important features are dropped, averaged over the entire dataset. Perfect performance in this case would be in the top right corner, obscuring almost all of the speech while preserving recognition accuracy. In general, we can see that the bubble analysis method (blue line) achieves approximately the same accuracy as the energy-based method (orange line).

The bottom plot of Figure 6 shows the accuracy of analyzed words when the most important features are dropped on all 100 sentences. A perfect MoRF mask would be in the bottom left corner of the bottom plot, obscuring almost none of the speech while destroying recognition accuracy. This plot demonstrates that the bubble analysis method is better at reducing recognition accuracy than the energy-based method when both drop the same amount of important speech energy. In both plots, the orange lines are shorter than the blue lines because the important regions of a word are restricted to be between the start frame and end frame in the energy-based approach.

Figure 7 shows the SSMB scores (green lines) at various thresholds for both methods. For the bubble analysis method in the top row, we can see that the threshold of 4.64×10^{-8} obtains the best SSBM score of 6.5. This means that the increase in LeRF accuracy at higher thresholds is not worth the decrease in MoRF accuracy. For the energy-based method in the second row, the threshold of -65 dB achieves the highest SSMB score of 4.7, which is worse than that of the bubble analysis method. The spike in the corresponding LeRF curve at 15dB is caused by a small denominator (0.001). Note that MoRF and LeRF are not symmetric, and the SSMB considers both directions from the least/most important features; thus, it is more robust to artifacts and noises. Thus, the bubble analysis method produces better importance maps than the energy-based approach according to the LeRF and MoRF curves and the SSBM score.

First, we can see that the ASR does not need to observe all of the speech energy of a word to correctly identify it. For illustration, the ASR can recognize the word "actually" with a

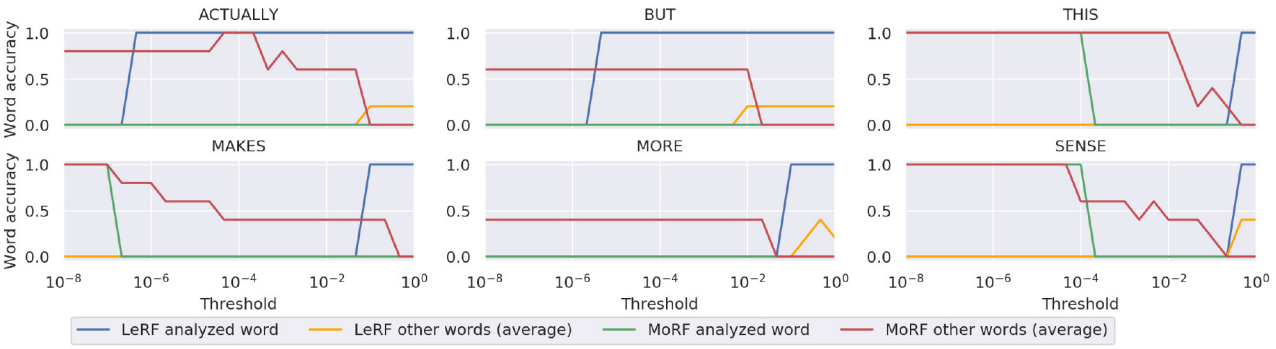

Figure 5: *Bubble analysis: Word accuracy on the sentence "actually but this makes more sense." with LeRF and MoRF masks*

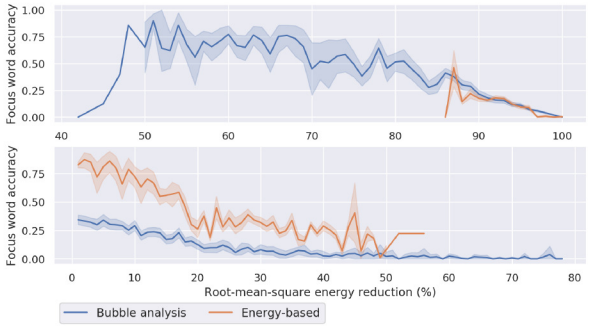

Figure 6: *Average accuracy of analyzed word with LeRF mask (top) and MoRF (bottom).*

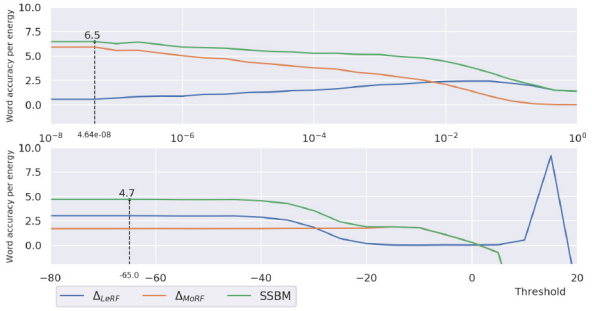

Figure 7: Δ_{LeRF} *and* Δ_{MoRF} *along with their combination into the SSBM score. Higher is better for all three. Top: bubble analysis, achieving SSBM of 6.5. Bottom: energy-based, achieving SSBM of 4.7 (accuracy per unit (percentage) of energy).*

threshold as low as 4.64×10^{-7} on the bubble analysis LeRF mask as in Figure 5 (blue line). This mask and its corresponding noisy speech are illustrated in the second and third row of Figure 2. As we can see, the mask only spans 400 Hz to 3200 Hz. Surprisingly, the clean speech lacks energy at those frequencies, but this does not prevent the ASR from correctly identifying the word. This saliency map achieves the same accuracy as the energy based alternative, while requiring less of the spectrogram to be audible, an efficiency reflected in its higher SSBM score.

Second, the threshold identifying which time-frequency points are important is varied across word. For example, in Figure 5 (blue line), the ASR needs to use all time-frequency points with p-value $< 4.64 \times 10^{-6}$ to correctly identify the word "but," however, the ASR must use all spectrogram points with

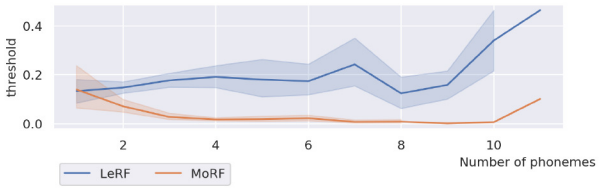

Figure 8: *Relationship between number of phonemes and threshold for the bubble analysis mask.*

p-value < 0.1 to recognize the words "more."

Figure 8 shows that word length may explain the variation in threshold for correct recognition. It shows the threshold at which a target words transitions from correct to incorrect recognition as a function of word length in phonemes. We can see that longer words typically require a higher LeRF threshold, meaning more speech is revealed, while they typically require a lower MoRF threshold, meaning less speech is obscured. Similar trends were observed with word length measured in syllables and characters.

5. Conclusion and future work

In this paper, we proposed an evaluation metric for structured saliency maps, where we measure the word accuracy when either keeping or dropping the most important regions. A gap in this accuracy is measured between the analyzed word and other words in the sentence with respect to the predicted important speech energies. Additionally, we compare saliency maps from a bubble analysis method and an energy-based baseline on sentences from the AMI meeting corpus. According to several metrics, the bubble analysis approach achieves a better importance map than its alternative. In the future, we will extend this evaluation to measure generalization of these predictions across ASR systems and use these importance maps to enhance speech recognition robustness in noisy conditions. We also hope that this speech saliency evaluation metric can facilitate a community evaluation on the topic of speech saliency, similar to those that have been organized around visual saliency [23].

6. Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. IIS-1750383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7. References

[1] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[2] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[3] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[5] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "A unified view of gradient-based attribution methods for deep neural networks," in *NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning-Now What?* ETH Zurich, 2017.

[6] S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2015, pp. 2125–2133.

[7] A. Goyal, Y. Bengio, M. Botvinick, and S. Levine, "The variational bandwidth bottleneck: Stochastic evaluation on an information budget," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=Hye1kTVFDS

[8] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, H. Larochelle, M. Botvinick, Y. Bengio, and S. Levine, "Infobot: Transfer and exploration via the information bottleneck," in *International Conference on Learning Representations*, 2018.

[9] N. Tishby, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, 1999, pp. 368–377.

[10] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2897–2905, 2018.

[11] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck int," in *Conf. on Learning Representations*, 2017.

[12] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Measuring time-frequency importance functions of speech with bubble noise," *Journal of the Acoustical Society of America*, vol. 140, pp. 2542–2553, 2016.

[13] V. A. Trinh and M. I. Mandel, "Directly comparing the listening strategies of humans and machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, under review.

[14] V. A. Trinh, B. McFee, and M. I. Mandel, "Bubble cooperative networks for identifying important speech cues," *Proc. Interspeech 2018*, pp. 1616–1620, 2018.

[15] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[16] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *International Conference on Learning Representations*, 2019.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[19] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[20] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[22] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[23] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit/tübingen saliency benchmark," https://saliency.tuebingen.ai/.