CAN WE TRUST EXPLAINABLE AI METHODS ON ASR? AN EVALUATION ON PHONEME RECOGNITION

Xiaoliang Wu, Peter Bell, Ajitha Rajan

School of Informatics, University of Edinburgh

ABSTRACT

Explainable AI (XAI) techniques have been widely used to help explain and understand the output of deep learning models in fields such as image classification and Natural Language Processing. Interest in using XAI techniques to explain deep learning-based Automatic Speech Recognition (ASR) is emerging. But there is not enough evidence on whether these explanations can be trusted. To address this, we adapt a state-of-the-art XAI technique from the image classification domain, Local Interpretable Model-Agnostic Explanations (LIME), to a model trained for a TIMIT-based phoneme recognition task. This simple task provides a controlled setting for evaluation while also providing expert annotated ground truth to assess the quality of explanations. We find a variant of LIME based on time partitioned audio segments, that we propose in this paper, produces the most reliable explanations, containing the ground truth 96% of the time in its top three audio segments.

Index Terms— Explanation, Phoneme Recognition

1. INTRODUCTION

Explainable Artificial Intelligence (XAI) techniques have been proposed in recent years to help explain and understand the output of deep learning (DL) models used in image classification [1, 2, 3] and NLP classification tasks [4, 5, 6]. Our work focuses on post-hoc explanations, i.e., explain an existing model that has been previously trained, which we treat as a black box system. Such post-hoc explanations are widely applicable as they can be used over models whose internal structure is not known. The need for explanations for AI systems has risen recently with the advent of safety regulations such as the recently proposed EU AI act \(^1\) that requires explanations to help users better understand the decisions made by AI systems. Although current explanation techniques have significantly advanced our understanding of DL model predictions, the reliability of these explanations has been largely overlooked. Some recent studies [7, 8] have demonstrated the limitations of current XAI techniques. For instance, [7] applied three different XAI techniques on a CNN-based breast cancer classification model and found the techniques disagreed on the input features used for the predicted output and in some cases picked background regions that did not include the breast or the tumour as explanations.

Literature on evaluating the reliability of XAI techniques is still in its nascency and can be broadly divided into two branches - (1) Studies that assume the availability of expert annotated ground truth, maybe in the form of bounding boxes for images, to evaluate the accuracy of explanations [9, 10, 11, 12] and (2) research that uses the idea of removing relevant (or important) features detected by an XAI method and verifying the accuracy degradation of the retrained models [13, 14, 15, 16, 17]. The first category requires human-annotated ground truth for evaluation while the second category incurs very high computational cost to verify accuracy degradation from retraining the models.

Research on explanations for ASR is still in its early stages. Wu et al. [18] modified image-based explanations for speech input in ASR, but the validity of the explanations was not evaluated. A key challenge in ASR is the absence of a clear mapping from words output to segments of audio, owing to the fact that ASR outputs are generally influenced by surrounding words, not just the immediate speech input.

Why Timit PR task? In an effort to evaluate the reliability and trustworthiness of explanations in the ASR context, we use the TIMIT [19] Phoneme Recognition (PR) task using the standard recipe from the Kaldi toolkit [20], as a simple but basic controllable task that is predictable with a phoneme language model with ground truth in the form of manual labeling and segmentation at the phoneme level. At this early stage, we believe it is important to properly validate the technique, and such detailed ground truth labelling – not found in any other data set – is essential for evaluating both the quality and reliability of the explanations generated.

We generate explanations for the PR system via LIME [1], a renowned XAI method designed for images. It employs a linear regression model to locally approximate the black box DL model’s prediction. We chose LIME due to its local perturbation approach, aligning with our belief that PR models exhibit local effects: phonemes are primarily influenced by adjacent phonemes, not distant ones.

To adapt LIME to produce explanations for the TIMIT PR task, first, we classify every phoneme in a PR transcription as correct or incorrect based on comparison with the expected transcription. Second, we apply LIME to generate explanations for each phoneme in the transcription using input speech perturbations. Third, we improve performance of LIME by focusing perturbations of the input audio to be within a limited window around the phoneme of interest using two LIME variations, LIME Window Segment (LIME-WS) and LIME Time Segment (LIME-TS). A segment refers to a distinct section within an audio and each of LIME-WS and LIME-TS has its own definition of segment which is then used as the basic unit of an explanation. We evaluate reliability of the basic LIME explanations and the variants, LIME-WS and LIME-TS, for the TIMIT PR task on Kaldi using the

---

\(^1\)https://artificialintelligenceact.eu/the-act/
ground truth labelling of the TIMIT dataset. We found explanations with LIME-TS are the most reliable for the TIMIT Phoneme Recognition task using Kaldi, capturing the ground truth 96% of the time in the top three segments of the explanation. LIME-WS outperforms LIME and LIME-WS by up to 44.8% and 17% on All speakers, respectively. Source code for X-PR and examples are available at https://anonymous.4open.science/r/X-PR-4560.

2. METHODOLOGY

Fig. 1. An outline of generating an explanation for a phoneme appearing in the output transcription.

Fig. 2. Different segmentation used by (LIME-WS,LIME) and LIME-TS.

Motivation Phoneme recognition systems, especially those based on complex algorithms, often function as a black box, making their decision-making opaque. For example, if a system misidentifies the phonemes ‘b’ and ‘p’ in words like ‘bat’ and ‘pat’, the developers need to understand why. Explanation techniques are designed to clarify such confusions, offering insight into the system’s decisions. This not only helps in refining the system but also builds developer and user trust.

What is an Explanation for a Phoneme? Taking cues from existing explanation methods helps us understand how to design explanations in other domains. In image tasks, models often highlight the areas of eyes, to explain why this image is labeled as a face. Similarly, for PR, an explanation should pinpointing key frames of speech for recognizing a phoneme. For example, as shown in the Figure 1, we have an input audio (divided into segments available in TIMIT), alongside its original transcription. For the phoneme ‘d’ which appears in the transcription, the explanation is the importance ranking of segments in the audio. The higher the rank, the more crucial that segment is for recognizing ‘d’. As shown in the Figure 1, the second segment emerges as the most important or the highest ranked for the phoneme ‘d’.

A general Framework of generating an Explanation: Figure 1 presents a high level overview of generating an explanation for a phoneme in the transcription output. We start with an input audio and its segments. We then perturb the audio by masking out segments randomly. To clarify, masking is done by setting the sample points of chosen segments to zero, creating silent intervals in the audio. In Figure 1, segments shaded in black within each mutant are the ones that have been masked. For the phoneme of interest, ‘d’ in Figure 1, we compute the importance ranking of segments in the audio as an explanation for ‘d’ which is based on the effect of the perturbations on the phoneme output. It is worth noting that when the segment corresponding to phoneme ‘s’ is masked (see the first mutant), we find the adjacent phoneme ‘d’ is wrongly recognized as ‘dh’. This is because the phoneme ‘d’ is affected when neighboring segments are masked.

We delve deeper into the LIME explanation approach and its variants, LIME-WS and LIME-TS, later in this section. To apply the LIME technique, we first need to treat the PR task as a classification task. To do this, we attach 0 or 1 label to every phoneme in the output transcription by aligning and comparing it with the expected transcription that is available in the TIMIT dataset. We implement classification of each phoneme output with the NIST sclite scoring tool.

2.1. Explanations using LIME and its variants for PR

In this section, we start by describing the base case which is a straightforward adaptation of LIME to work on the PR task. We then describe our variants, LIME-WS and LIME-TS, that applies perturbations to segments within a fixed window.

Base LIME Explanations: LIME, proposed in [1], is a black-box XAI technique that can be applied to any model without requiring information on its structure. Given a complex neural network (NN) model \( f(x) \) that takes in an input \( x \) and produces an output \( y \), the goal of LIME is to mimic the behavior of a complex model \( f(x) \) with a weighted linear regression model \( g(x) \) in the local area of a specific instance of interest \( x \). The weighted linear regression model \( g(x) \) is defined as:

\[
g_P(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d
\]

The complex NN model \( f(x) \) is the Kaldi system. \( P \) refers the positional index of the phoneme to be explained in the original output. As shown in the Figure 1, the specific instance \( x \) is the input audio, ‘d’ is the phoneme output to be explained and \( P \) is 2. To fit \( g_P(x) \) and get \( w_1 \) to \( w_d \), LIME needs several perturbed instances of \( x \) and their outputs. As shown in the right side of the Figure 1, mutants of the original audio \( x \) are created by masking out segments randomly. Input features \( x \) are the segments. Values of features \( x_1 \) to \( x_d \) in the mutant are 1 or 0, where 1 means that the segment at this position has not been masked while 0 implies the segment is masked. For example, in the Figure 1, for the uppermost mutant, the third segment has been masked. Consequently, the value of \( x_3 \) is set to 0, with all other feature values being 1.

Footnote 2: https://github.com/usnistgov/SCTK
For each mutant $m_i$, we align its transcription against the original (unmasked) output transcription, $y$. All transcriptions are from the trained Kaldi. After alignment, we may find some phonemes match the original output transcription while some others are incorrect. For those correct phonemes, the output of $m_i$ - represented as $f^p(m_i)$ - will be 1 while for others, $f^p(m_i)$ will be 0. $j$ refers the index of the positional index of phonemes in the original transcription. Continuing with the example of the uppermost mutant, denoted as $m_1$, we observe the following: ‘d’ to be explained (at the second position) has been wrongly identified as ‘dh’. Therefore, we have $f^s(m_1) = 0$, which is also the $f^p(m_1)$. Similarly, the fourth element, ‘w’, remains unaltered, resulting in $f^t(m_1) = 1$.

Then, the LIME model will compute how the masked segment in each of the mutants affects the output phoneme ‘d’ (bounded with a green box in the Figure 1). If the masked segment changes the output phoneme ‘d’ at that position, then it will have a high ranking (aggregated over many mutants with masked segments).

Using the mutants and the associated binary labels $f^p(m_i)$ after alignment to original transcription, LIME will start to fit $g^f(x)$ using the locally weighted least squares objective function, which is defined as:

$$L(g) = \sum_{i=1}^{n} \epsilon_i (f^p(m_i) - g^f(m_i))^2$$

In this equation, $n$ is the number of mutants and $\epsilon_i$ is a weight assigned to each mutant $m_i$ that reflects its closeness to the audio of interest $x$. It is computed as the cosine similarity between the instance $x$ and the mutant $m_i$, which is $\epsilon_i = \text{CosineSimilarity}(x, m_i)$. The weights $w_1$ to $w_d$ in the fitted linear regression model, $g^f(x)$, indicate the importance score of different audio segments for the selected output phoneme. We treat the ranking of segments based on their importance score as the explanation for each phoneme, as shown in Figure 1.

**Segment-based LIME with a sliding window (LIME-WS):**

In Base LIME, mutants for the original audio $x$ are created by masking random segments. However, given the local nature of PR task, distant audio segments likely will not impact the phoneme in focus. Considering this, removing ineffective mutants can optimize computation.

To realize this idea, we use a fixed length sliding window during the generation of perturbations for LIME explanations. The sliding window slides from left to right one segment at a time. Within the range delimited by the sliding window, a pre-determined number of segments are randomly chosen for masking, while keeping the segments outside the sliding window unchanged. We hypothesize that focusing on perturbations within this sliding window will result in higher quality explanations. Other steps in LIME-WS for fitting the linear regression model remain the same as LIME.

**Time Segment-based LIME with a sliding window (LIME-TS):** LIME and LIME-WS employ audio segmentation from the TIMIT dataset, segmented by linguistic experts. However, manual segmentations might be absent in common datasets like Librispeech and Common voice[21]. To overcome this challenge and generalize the applicability of our segment-based explanation technique, we investigate a method to split audio into uniform, non-overlapping segments via timestamps. Figure 2 contrasts the audio segmented from TIMIT and by timestamps. We split the audio into 70ms segments (to remain comparable with average length of manual segments), but this can be changed based on user choice.

### 3. EXPERIMENTS

We evaluate the reliability of explanation techniques using the TIMIT PR model from Kaldi. For generating explanations, we choose the TIMIT dataset owing to the ground truth mapping of phonemes to input speech and details like speaker’s gender. We generated explanations for all 630 speakers using their shared ‘SA1’ sentence. This uniform sentence allows comparison of explanation techniques across different demographics, focusing on the impact of factors like gender.

**Validity Metric** The validity metric evaluates the reliability of the three explanation methods – LIME, LIME-WS, and LIME-TS. The metrics are defined as $validity_{1} = \frac{N_1}{N}$, $validity_{3} = \frac{N_3}{N}$, and $validity_{5} = \frac{N_5}{N}$, where $N$ is the total number of phonemes, $N_1$ is phonemes with the top-ranked segment matching the ground truth, while $N_3$ and $N_5$ represent phonemes where the ground truth is within the top 3 and 5 ranks, respectively. For all three metrics, higher is better.

### 4. RESULTS AND ANALYSIS

![Fig. 3. The top five most frequently occurring transcription mistakes and their corresponding frequencies on different groups. There are three substitution mistakes on the left of the dashed blue line and two deletion mistakes on the right. For example, er → uw means that er is replaced by uw and Xih means that ih is deleted.](image)

**Comparison of Explanation Techniques.** The three methods, LIME, LIME-WS, and LIME-TS, successfully generate explanations for all audio samples in our dataset. In Table 1, we display the metrics $validity_{1}$, $validity_{3}$, and $validity_{5}$ for each technique across All speakers, and then segregated for Male and Female speakers. To benchmark these results, we also present validity scores obtained from random rankings—these values are shown after the / symbol in the table. It is evident that all techniques are more trustworthy than random rankings across all metrics. For example, in Table 1, LIME-WS scores 0.49, 0.76, and 0.83 across the validity metrics for All speakers, greatly surpassing the scores of 0.03, 0.10, and 0.15 from random ranking. Both LIME and LIME-TS also exceed the random ranking significantly, with
the differences verified statistically using one-way Anova followed by post-hoc Tukey’s test [22].

From Table 1, we see that LIME-WS and LIME-TS consistently perform better than LIME across all metrics for each speaker group. For example, on validity$_2$, they surpass LIME by 24% and 44.8% respectively. This outcome is in line with our expectations, as the sliding window introduced in LIME-WS and LIME-TPS ensures that perturbations are restricted to a local range. This restriction enables the explanation technique to focus on relevant segments as the influence of a phoneme is typically confined to a small number of adjacent phonemes.

For LIME-TPS versus LIME-WS, we find LIME-TPS performs better than LIME-WS in all cases. We verified this difference is statistically significant using one-way Anova and a follow-up Tukey’s HSD test [22] at a 5% significance level. For example, LIME-TPS outperforms LIME-WS by 17% on validity$_5$ over All speakers. LIME-TPS utilizes fixed-length time segments as the fundamental unit for generating explanations and additionally, LIME-TPS segments are slightly smaller than LIME-WS segments – 70 versus 78 ms. The corresponding ground truth explanation in the original audio always overlaps with multiple LIME-WS segments which are considered equally significant. Conversely, LIME-WS uses manually labeled audio segments as the unit of explanation that has a one to one correspondence with the ground truth explanation. We believe the smaller LIME-WS segments for perturbation and the validity measurement (top 1, 2 or 5 ranked segments) that considers ground truth overlapping segments as equally important helps LIME-TPS look more attractive than LIME-WS.

Overall, we find LIME-TPS to be most reliable among the three explanation methods, capturing the ground truth in 96% of the cases when considering the top 3 segments in the explanations. Additionally, it is easily generalizable, owing to its use of time-based segments rather than an expert labelled audio segment, as in LIME or LIME-WS.

Male versus Female Speakers. Table 1 shows that the validity scores for all three explanation methods are higher for Males compared to Females. This is consistent with the fact that the TIMIT dataset, used for training, contains 70% male speakers and only 30% females. Thus, Kaldi better recognizes the nuances of male speech, even when both male and female speakers are uttering the same sentence.

We explored the most commonly occurring transcription errors in the different speaker groups. Figure 3 illustrates the five most commonly occurring transcription errors and their corresponding frequencies across all three speaker groups. Except for the $sh \rightarrow s$ substitution, we observed significant variances in error frequencies between Female and Male speakers. For example, the $er \rightarrow uw$ error is more prevalent among Male speakers (0.54) than Female speakers (0.28). Using a Wilcoxon Signed Rank Test with a 5% significance level, we verified that these error frequency differences between Males and Females are statistically significant.

Explanations can help investigate possible causes for difference in error frequencies between genders. For instance, we examine the $er \rightarrow uw$ error. When comparing the LIME-WS explanations for Male and Female speakers regarding this error, we notice distinct patterns. Table 2 shows the three most recurring segments for this error. For Female speakers, these segments frequently center around position 8. In contrast, Male speakers often have two of their top three segments near position 34. Further analysis reveals the $er \rightarrow uw$ error appears at two key locations in sentence SA1: position 6 (surrounded by phonemes ‘vel’ ‘d’ ‘aa’) and position 34 (adjacent to phoneme ‘y’). Among Female speakers, 64% of these errors align with position 6, compared to 47% for Male speakers. This insight suggests potential areas for model improvement, tailored to each gender. Overall, explanations serve as a valuable tool to examine errors and compare speaker groups.

### Table 1

<table>
<thead>
<tr>
<th>Speaker Group</th>
<th>LIME / Random ranking</th>
<th>LIME-WS / Random Ranking</th>
<th>LIME-TPS / Random Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>validity$_1$</td>
<td>0.40/0.0</td>
<td>0.35/0.0</td>
<td>0.42/0.0</td>
</tr>
<tr>
<td>validity$_2$</td>
<td>0.62/0.06</td>
<td>0.54/0.06</td>
<td>0.62/0.06</td>
</tr>
<tr>
<td>validity$_5$</td>
<td>0.67/0.13</td>
<td>0.59/0.12</td>
<td>0.66/0.13</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Segment Position</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Phoneme output)</td>
<td>(Phoneme output)</td>
<td>(Phoneme output)</td>
</tr>
<tr>
<td>8 (d)</td>
<td>8 (d)</td>
<td>34 (er)</td>
</tr>
<tr>
<td>9 (aa)</td>
<td>9 (aa)</td>
<td>8 (d)</td>
</tr>
<tr>
<td>34 (er)</td>
<td>7 (dcl)</td>
<td>32 (y)</td>
</tr>
</tbody>
</table>

### 5. CONCLUSION

The focus in this paper is to evaluate quality of explanation for the Phoneme Recognition(PR) task. We adapt LIME from the image domain to the characteristics of the TIMIT PR task and propose two techniques, LIME-WS and LIME-TPS, to generate explanations. To assess the quality of explanations, we conduct experiments under controlled conditions with ground truth available. Our findings indicate that LIME-TPS is the most reliable, containing the ground truth audio segment in phoneme output explanations in 96% of the cases.

While our evaluation of explanation technique reliability provides initial insights, it remains a significant challenge to assess these techniques on more complex speech tasks owing to the involvement of other components, such as a language model, and the presence of long-span dependencies that make it challenging to create a ground truth mapping from word outputs to audio segments. We aim to address evaluation of explanations over more complex speech tasks in the future like using diffusion models [23].
6. REFERENCES


