# SPES: Spectrogram Perturbation for Explainable Speech-to-Text Generation

**Dennis Fucci** and **Marco Gaido** and **Beatrice Savoldi**
**Matteo Negri** and **Mauro Cettolo** and **Luisa Bentivogli**
Fondazione Bruno Kessler, Trento, Italy
dennisfucci@gmail.com
{mgaido,bsavoldi,negri,cettolo,bentivo@fbk.eu}

## Abstract

Spurred by the demand for interpretable models, research on eXplainable AI for language technologies has experienced significant growth, with feature attribution methods emerging as a cornerstone of this progress. While prior work in NLP explored such methods for *classification* tasks and *textual* applications, explainability intersecting *generation* and *speech* is lagging, with existing techniques failing to account for the autoregressive nature of state-of-the-art models and to provide fine-grained, phonetically meaningful explanations. We address this gap by introducing ***S**pectrogram **P**erturbation for **E**xplainable **S**peech-to-text Generation* (**SPES**), a feature attribution technique applicable to sequence generation tasks with autoregressive models. SPES provides explanations for each predicted token based on both the input spectrogram and the previously generated tokens. Extensive evaluation on speech recognition and translation demonstrates that SPES generates explanations that are faithful and plausible to humans.

## 1 Introduction

The recent advances of Artificial Intelligence (AI) and the emergence of foundation models (Bommasani et al., 2021) have come together with concerns about their risks and impact (Weidinger et al., 2021), as well as calls for a better understanding of their inner workings (Barredo Arrieta et al., 2020; Eiras et al., 2024). This need has been reinforced by the demands for transparency by legal frameworks like the EU AI Act and the US National AI Initiative Act (Panigutti et al., 2023; Roberts et al., 2024). In response, the field of eXplainable AI (XAI) has emerged prominently, with the goal of elucidating the reasoning behind system predictions (Doshi-Velez and Kim, 2017; Carvalho et al., 2019; Vilone and Longo, 2021; Pradhan et al., 2022). Among the various active XAI approaches (Ferrando et al., 2024), this paper centers on *feature attribution* methods, which aim to identify and quantify the relevance of each input feature in determining a model's final output.

Originally developed for image and text classification (Danilevsky et al., 2020; Kamakshi and Krishnan, 2023), feature attribution methods have been extended to other modalities—like speech (Becker et al., 2024; Pastor et al., 2024)—and to text generation (Sarti et al., 2023; Zhao et al., 2024, et al.). Still, XAI in the domain of speech-to-text (S2T) generative tasks is limited, with only a few preliminary works focusing on automatic speech recognition (ASR) (Mandel, 2016; Trinh et al., 2018; Trinh and Mandel, 2021; Kavaki and Mandel, 2020; Markert et al., 2021; Wu et al., 2023, 2024). This contrasts with the pivotal role of spoken language—arguably the most natural form of human interaction (Munteanu et al., 2013)—and consequently with the importance of S2T technologies, which are now fostered by foundation models that transcribe and translate at an unprecedented scale (Latif et al., 2023).

Moreveor, most existing feature attribution methods in S2T (Mandel, 2016; Trinh and Mandel, 2021; Markert et al., 2021; Wu et al., 2023, 2024) are not applied to autoregressive models, which predict each token iteratively, relying on both speech input and previous output tokens. Even the only method applied to autoregressive S2T models (Kavaki and Mandel, 2020) assumes conditional independence between prediction steps, disregarding the influence of previously generated text on each new prediction. Finally, some feature attribution methods are incompatible with the spectrogram input format (Markert et al., 2021; Wu et al., 2023, 2024)—the format widely used in modern S2T models and enabling the visualization of patterns corresponding to well-



Transactions of the Association for Computational Linguistics, vol. 14, pp. 772–801, 2026. https://doi.org/10.1162/TACL.a.684
Action Editor: Georgila Kallirroi. Submission batch: 4/2025; Revision batch: 9/2025; Published 6/2026.

understood phonetic phenomena, a property not afforded by alternatives such as raw waveforms or mel-frequency cepstral coefficients (MFCCs) (Fucci et al., 2024). Other methods that can handle spectrograms produce explanations that fail to highlight fine-grained patterns related to the spectral characteristics of speech, such as fundamental frequency and formants (Mandel, 2016; Trinh et al., 2018; Kavaki and Mandel, 2020; Trinh and Mandel, 2021).

To overcome the above limitations, we propose ***S**pectrogram **P**erturbation for **E**xplainable **S**peech-to-text Generation* (**SPES**), the first feature attribution technique that provides token-level explanations for autoregressive S2T models by considering both the speech input and previously generated tokens.[1] SPES employs a perturbation-based approach that adapts image segmentation techniques to spectrograms, enabling the identification of fine-grained, meaningful patterns. It also perturbs previously generated tokens, estimating the impact of these perturbations on the output probability distribution. This process is repeated multiple times, with impact scores aggregated to produce final heatmaps. Through quantitative assessments and in-depth analyses across two S2T tasks—ASR and, for the first time, speech translation (ST)—we show that SPES produces explanations that are both *faithful* (i.e., reflecting the true reasoning of the model) and *plausible* (i.e., aligned with human understanding of meaningful phonetic patterns).[2]

The paper is structured as follows: §2 provides an overview of feature attribution concepts and existing approaches across different fields (e.g., vision and NLP), discussing their strengths and limitations to motivate our design choices. §3 reviews S2T-specific techniques and their drawbacks. §4 details the SPES methodology, comprising input perturbation, impact estimation, and score aggregation. §5 outlines our evaluation strategies, while §6 describes the experimental setup. §7 presents our findings, followed by analyses of explanation plausibility in §8. Finally, §9 offers concluding remarks and the potential applications of SPES.

[1]The code is available at https://github.com/hlt-mt/FBK-fairseq under the Apache License 2.0.

[2]Our definitions of faithfulness and plausibility follow Jacovi and Goldberg, 2020 and are further discussed in §5.

## 2 Background: Overview of Feature Attribution Methods

Feature attribution methods produce a *saliency* map $S$, where each element $s_i$ represents the contribution of input feature $x_i$ to the prediction $y$ (Samek and Müller, 2019; Samek et al., 2021; Madsen et al., 2022). $S$ is typically visualized as a heatmap that highlights influential inputs. These methods are being used across different domains, e.g. NLP and computer vision, and can be grouped into four categories: gradient-based, amortized, decomposition-based, and perturbation-based, each offering distinct ways to generate $S$.

*Gradient-based* methods (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2019) use the model's gradients with respect to input features to measure feature relevance, where the gradient reflects the sensitivity of the prediction to small changes in each feature. While these methods are efficient, as they involve a single or a few backpropagation steps, they can be unstable due to gradient instabilities and sensitivity to noise (Ancona et al., 2018; Samek et al., 2021; Dombrowski et al., 2022).

*Amortized explanation* methods (Dabkowski and Gal, 2017; Yoon et al., 2019; Schwarzenberg et al., 2021; Chuang et al., 2023b) train a global explainer to produce saliency maps in a single forward pass. Although efficient at inference time, this approach requires separate training for each model and depends on the explainer's learning capacity, which can be influenced by biased training patterns (Chuang et al., 2023a).

*Decomposition-based* methods (Bach et al., 2015; Shrikumar et al., 2017) propagate predictions backward through the network, layer by layer, decomposing the prediction score into contributions from individual neurons according to predefined rules. This process continues until the score is distributed across the input features. While these methods are as efficient as gradient-based ones, their effectiveness depends on the choice of propagation rules, which can vary across models and applications, potentially limiting their generalizability (Montavon et al., 2019).

*Perturbation-based* methods (Zeiler and Fergus, 2014; Lundberg and Lee, 2017; Fong and Vedaldi, 2017) determine feature saliency by systematically altering or masking parts of the input and observing the impact on the model's out-

put. Despite being computationally intensive, particularly for high-dimensional data, these methods offer stable and accurate explanations by directly linking output variations to feature relevance. Moreover, they are independent of model architecture (Covert et al., 2021).

## 3 Related Works in S2T

Moving to the ***S2T domain***, feature attribution has received relatively little attention compared to other fields, with efforts focused only on ASR, leaving other tasks like ST unexplored. Moreover, some of the proposed methods do not base their explanations on spectrograms. Specifically, Markert et al. (2021) generate explanations based on MFCCs, while Wu et al. (2023, 2024) generate them directly on raw waveforms. This represents a crucial limitation, as spectrograms are the input representation of many widely used state-of-the-art S2T models—such as Whisper (Radford et al., 2023), SeamlessM4T (Communication et al., 2023), and OWSM (Peng et al., 2023). Also, unlike waveforms or MFCC, spectrograms are directly interpretable by humans as they provide a visual representation of frequency content over time, making it possible to identify patterns associated with phonetic and prosodic features (Stevens, 2000).

Henceforth we focus on those feature attribution techniques that work with spectrograms. Spectrograms provide a 2D visualization of frequency distributions over time, where darker areas indicate higher energy at specific frequencies, capturing the spectral characteristics of speech (Stevens, 2000). These techniques adopt either perturbation-based or amortized approaches.

The perturbation-based Bubble Noise technique introduced by Mandel (2016) and Trinh and Mandel (2021) obscures time-frequency patterns in the spectrogram using white noise, except in randomly positioned ellipsoidal bubble regions where the noise is attenuated, allowing speech characteristics to remain visible. To assess the impact of perturbations, this technique assigns a binary label to each predicted word based on whether the word changes under the perturbed input. Final saliency scores are calculated by correlating these labels with the noise mask, which tracks the amount of noise added to the spectrogram. Bubble Noise involves high computational costs to generate explanations, which escalate with increasing bubble counts. Additionally, the use of uniformly sized and shaped bubbles inherently overlooks the structural patterns within the spectrogram, potentially limiting the method's effectiveness in capturing fine-grained spectral details.

To reduce explanation time and provide fine-grained insights, Trinh et al. (2018) and Kavaki and Mandel (2020) propose an amortized explanation method. A mask estimator is trained to regulate white noise addition to individual features in the spectrogram aiming to maximize noise while preserving ASR accuracy. As a result, regions of the spectrogram where noise is minimized or absent are identified as salient for word prediction. However, in addition to the high costs associated with training separate explainers for different models, the resulting explanations highlight time-frequency patterns that only loosely correspond to meaningful spectral characteristics (Kavaki and Mandel, 2020).

Besides the above limitations, both techniques lack adaptation to the autoregressive nature of modern S2T models, where each token generation depends on both input speech and prior output text. Building on the accuracy and architectural independence of perturbation-based methods, we address these gaps by introducing a feature attribution technique that *i)* accounts for the autoregressive nature of modern S2T models, making it applicable to any autoregressive speech encoder-decoder model regardless of its architecture, and *ii)* captures fine-grained time-frequency patterns in spectrograms. We compare our approach to other perturbation-based methods that share these theoretical advantages.

## 4 SPES

Given an autoregressive S2T model that generates one token at a time, conditioned on both the speech input $X$—represented as a spectrogram—and the previously generated tokens $Y^{(k-1)} = [y_0, y_1, \ldots, y_{k-1}]$,[3] SPES produces two saliency maps: $\boldsymbol{S^X_{y_k}}$, which reveals the relevance of input features across time and frequency, and $\boldsymbol{S^{Y^{(k-1)}}_{y_k}}$, which highlights the relevance of previously generated tokens. The overall explanation for the predicted sequence $Y$, comprising $K$ tokens, is then formulated as:

$$S_Y = \{(S^X_{y_1}, S^{Y^0}_{y_1}), \ldots, (S^X_{y_K}, S^{Y^{(K-1)}}_{y_K})\} \quad (1)$$

[3]Where $y_0$ is the special token start of sentence <s>.

Preliminary experiments (detailed in Appendix A.2.2) indicated that separately perturbing the spectrogram and the previous output tokens yields more accurate saliency maps compared to joint perturbation. Therefore, we outline how these two components are perturbed independently to generate the corresponding saliency maps. Specifically, the following sections describe SPES in reference to the key components of perturbation-based approaches (Covert et al., 2020), namely *i)* selecting and perturbing features (§4.1), *ii)* estimating the impact of these perturbations on model predictions (§4.2), and *iii)* aggregating these impact estimations to derive final saliency scores (§4.3).

### 4.1 Input Perturbation

Input perturbation is commonly implemented by masking subsets of input elements. The effect of these perturbations on the model's output probabilities is then evaluated to derive the saliency maps $S_{y_k}$ (see §4.2). However, enumerating all possible subsets is infeasible because the number of perturbation combinations grows exponentially with input size (Khakzar et al., 2020; Covert et al., 2021). To approximate the expectation over this combinatorial space, perturbation-based approaches typically draw a finite number of random perturbations using Monte Carlo sampling (Castro et al., 2017; Petsiuk et al., 2018; Markert et al., 2021; Yang et al., 2021; Fel et al., 2023). Formally, for token $y_k$ let $\mathcal{M}_{y_k}$ denote the set of all possible $M$ binary masks over the input used to predict $y_k$. Monte Carlo sampling approximates the expected value for the saliency maps $\mathbb{E}_{M\sim\mathcal{M}_{y_k}}\left[\, S_{y_k} \mid M \,\right]$ by randomly drawing $N$ binary masks $M_n$ (the higher $N$, the more accurate and computationally expensive the estimation), and aggregating the impact on the model output probabilities obtained with each of them (see §4.3). In our setting, we draw $N^X$ and $N^Y$ samples for spectrogram features and previous output tokens, respectively. These hyperparameters control the variance of the estimator and hence trade off between computational cost and saliency map accuracy.

#### 4.1.1 Perturbing Spectrograms

To achieve explanations that highlight fine-grained spectral characteristics in spectrograms, we design perturbations that target time-frequency patterns using techniques from image processing, where morphological clustering is often employed to define the perturbation units (Seo et al., 2020; Yang et al., 2021).[4] This approach produces explanations that align more naturally with object contours (Ivanovs et al., 2021), in contrast to perturbations with uniformly sized and shaped units. A particularly effective clustering-based technique is the *Morphological Fragmental Perturbation Pyramid* (MFPP; Yang et al. 2021), which generates detailed saliency maps (Ivanovs et al., 2021). MFPP employs Simple Linear Iterative Clustering (SLIC; Achanta et al. 2012), a $k$-means-based algorithm that groups pixels according to spatial and color characteristics—representing, in our case, proxies for spectral patterns. By iteratively adjusting the number of clusters ($k$), the method produces a hierarchy of patches across varying granularities, from fine to coarse.

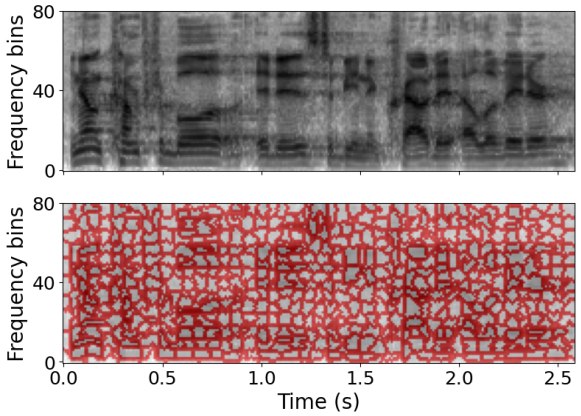



Figure 1: A mel-spectrogram (upper figure) and its segmentation obtained with SLIC (bottom figure).

As shown in Fig. 1, our SLIC-based segmentation of the spectrogram clusters nearby features with similar energy levels into distinct patches that respect the structure of spectral patterns, such as formants. Following MFPP, each spectrogram is segmented at three levels of granularity by varying $k$. This *multi-scale* approach creates patches at different resolutions, enabling the perturbation of neighboring features together or separately.

However, unlike images that have fixed dimensions, speech data vary in duration. Consequently, a fixed value of $k$ would result in larger patches for longer audio samples, leading to coarser explanations. To address this, SPES scales $k$ propor-

[4] In images, perturbations are typically applied either at the pixel level (Wagner et al., 2019) or to groups of pixels (Zeiler and Fergus, 2014; Ribeiro et al., 2016; Yang et al., 2021), referred to as *patches* or *superpixels*, which serve as the minimal processing units.

tionally to input length, ensuring consistent patch sizes across varying speech sample lengths. However, this approach has a limitation: as the number of patch combinations grows exponentially with $k$, exhaustive exploration of the search space requires a corresponding increase in the number of perturbations $N^X$, which becomes computationally intractable for large values of $k$.[5] Insufficient exploration, on the other hand, eventually lowers explanation quality (see Appendix A.2.3). To balance computational cost and effective exploration of patch combinations, SPES limits the increase of $k$ by setting a maximum value. Specifically, with $len(X)$ as the input length in seconds, $\phi$ as the number of patches per second, and $\tau$ as the threshold length in seconds, $k$ is determined as:

$$k = \min(len(X), \tau) \cdot \phi \tag{2}$$

Once the patches are defined, we randomly perturb them in combination across $N^X$ iterations. The number of patches perturbed at each iteration is controlled by a hyperparameter ($p_{spec}$), which specifies the probability of perturbing each patch—for example, with $p_{spec} = 0.5$, approximately half of the patches are perturbed.

Perturbation is implemented by setting the elements of the patch to 0 in a binary mask $M^X \in \{0,1\}^{T \times C}$, where $T$ represents the time dimension and $C$ represents the frequency dimension of $X$. The perturbed input $X'$ is then obtained by element-wise multiplication of the original input $X$ and $M^X$, which corresponds to setting the perturbed features to 0, the mean energy value in the utterance.[6] This perturbation method—the same used at training time by SpecAugment (Park et al., 2019), a common augmentation technique in most training settings, including ours—avoids introducing artifacts that fall outside the training distribution, helping the model remain robust to perturbations and aligning with strategies to prevent hallucinations during perturbations (Holzinger et al., 2022; Brocki and Chung, 2023). Although this advantage renders our approach particularly promising, alternative approaches, such as replacing masked elements with low-energy values, could be explored in future work.

#### 4.1.2 Perturbing Previous Output Tokens

Inspired by the effectiveness demonstrated in context mixing by the *value zeroing* (Mohebbi et al., 2023) perturbation strategy, which replaces selected token embeddings with zero vectors, SPES adopts the same approach for the perturbation of the previous output tokens. Specifically, for each of the $N^Y$ iterations, it builds a binary mask $M^{Y^{(k-1)}} \in \{0,1\}^{(k-1)\times 1}$, where each token is randomly selected for perturbation with probability $p_{tok}$, and the corresponding mask values are set to 0. The mask $M^{Y^{(k-1)}}$ is then multiplied with the embedding matrix of the previous output tokens to obtain their perturbed version $Y^{(k-1)'}$.

### 4.2 Perturbation Impact Estimation

The relevance of the features perturbed at each iteration is assessed by estimating how the perturbation impacts the model's ability to predict the output. In classification tasks, two common methods are used for this estimation, both based on the probability assigned by the model to the target class. One method uses the target class probability as the saliency score for non-perturbed features (Zeiler and Fergus, 2014; Petsiuk et al., 2018; Yang et al., 2021), where a high score indicates that the perturbed features do not include information relevant to predicting the target class. Alternatively, the score is computed as the difference between the original probability (without perturbation) and the probability obtained with the perturbed input (Ancona et al., 2018; Seo et al., 2020). In this case, the score is assigned to the perturbed features, where a high score suggests that the model's ability to recognize the target class is compromised in their absence.

In S2T tasks, assessing single-class probabilities corresponds to using the probability of the token being explained. We posit that this approach can be improved for two reasons. First, while classification tasks typically deal with a limited number of target classes, S2T tasks involve a BPE vocabulary (Sennrich et al., 2016) with thousands of tokens (Di Gangi et al., 2020; Fendji et al., 2022). This results in output probability distributions that are high-dimensional, dense vectors rather than sparse, peaked distributions (Zhao et al., 2023).

[5] As shown in Fig. 6 in Appendix B, the relationship between audio duration and computation time for generating explanations is quadratic, meaning that longer audio samples account for most of the computation time. Hence, increasing $N^X$ with the audio length would substantially increase computational costs.

[6] In line with standard practices in current S2T systems (Communication et al., 2023, among others), all input spectrograms undergo cepstral mean-variance normalization (Viikki and Laurila, 1998) before being fed into the model.

Second, the decoding process of S2T models often employs beam search (Sutskever et al., 2014), where the tokens with the highest probability are not always selected, and perturbations can significantly affect the predicted sequence by altering the probabilities of initially-neglected tokens.

Therefore, drawing on mechanistic interpretability research (Conmy et al., 2023; Zhang and Nanda, 2024), SPES estimates perturbation impact ($r^X$ for spectrograms and $r^{Y^{(k-1)}}$ for previous output tokens) by computing the KL divergence (Kullback and Leibler, 1951) between the probability distribution of an S2T model $\Psi$ fed with the original inputs ($X$ and $Y^{(k-1)}$) and that obtained with their perturbed counterparts ($X'$ and $Y^{(k-1)'}$). Specifically:

$$r^X = \text{KL}\left(\Psi(X, Y^{(k-1)}), \Psi(X', Y^{(k-1)})\right)$$
$$r^{Y^{(k-1)}} = \text{KL}\left(\Psi(X, Y^{(k-1)}), \Psi(X, Y^{(k-1)'})\right)$$

### 4.3 Scores Aggregation

According to the Monte Carlo strategy (see §4.1), the $r$ scores are aggregated across the $N$ perturbation iterations to produce the token-level saliency maps $S$:

$$S = \frac{\sum_{n=1}^{N} r_n(1 - M_n)}{\sum_{n=1}^{N}(1 - M_n)} \quad (3)$$

where $S$ is either $S^X$ or $S^{Y^{(k-1)}}$ and, accordingly, $M$ is either $M^X$ or $M^{Y^{(k-1)}}$, $N$ is either $N^X$ or $N^Y$, and $r$ is either $r^X$ or $r^{Y^{(k-1)}}$. Specifically, Eq. 3 sums the $r$ scores assigned to the perturbed features $(1 - M_n)$ in the numerator and weights them by the number of times each feature is perturbed in the denominator.

Lastly, SPES applies two normalization steps, introduced to improve the effectiveness and stability of explanations in practice. First, it applies z-score normalization to each token-level saliency map ($S^X_{y_k}$ and $S^{Y^{(k-1)}}_{y_k}$) independently. This step addresses the varying magnitudes of saliency scores across different tokens $y_k$ in the same output $Y$, preventing disproportionate emphasis on specific tokens. This is especially beneficial when aggregating token-level explanations at the sentence level, such as through averaging, to create a single saliency map that reflects the relevance of each feature to the prediction of the entire sentence. The effectiveness of this normalization has been validated, as detailed in Appendix A.2.1.

Second, SPES applies joint min-max normalization to $S^X_{y_k}$ and $S^{Y^{(k-1)}}_{y_k}$ for the same token $y_k$, scaling their scores to the $[0, 1]$ range. This paired normalization ensures interaction between the two maps: if one map has higher scores for a token, the other map will correspondingly receive lower scores, thereby reflecting the relative relevance of $S^X_{y_k}$ and $S^{Y^{(k-1)}}_{y_k}$ for predicting $y_k$.

## 5 Evaluating Explanations in S2T

Evaluating the quality of explanations is a complex task, ideally involving the assessment of several desirable properties across three dimensions: content, presentation, and user (Nauta et al., 2023). Within the **content** dimension, the foremost property is *correctness*, also known as *fidelity* (Tomsett et al., 2020) or *faithfulness* (Jacovi and Goldberg, 2020), which refers to "how well explanations match model reasoning" (Li et al., 2023). Moving to the **presentation** dimension, the quality of an explanation concerns not only its composition ("how something is explained") but also its *compactness*, adhering to the "less is more" principle. Accordingly, explanations should highlight the smallest possible set of features and avoid the inclusion of redundant elements (Sun et al., 2020). Lastly, concerning the **user** dimension, a desirable property is *plausibility*, defined as "how convincing the explanation is to humans" (Jacovi and Goldberg, 2020).

In this work, we consider all three dimensions to evaluate explanations. We inspect the content dimension by assessing faithfulness with the *deletion* metric, the most widespread method in classification tasks (Arras et al., 2017; Samek et al., 2017; Petsiuk et al., 2018; Tomsett et al., 2020; Samek et al., 2021; Gevaert et al., 2024) and also used in ASR (Trinh and Mandel, 2020). On the presentation axis, we measure compactness using the *size* metric. Finally, on the user axis, we assess plausibility through *dedicated analyses* in §8.

### 5.1 Deletion

This metric evaluates how quickly performance declines as input features, ranked by saliency scores, are gradually removed. Trinh and Mandel (2020) applied deletion to ASR explanations, adding noise to the most relevant features for each word in the output sentence. Thereby, they test whether the prediction of the target word changes under the noisy condition, while predictions for

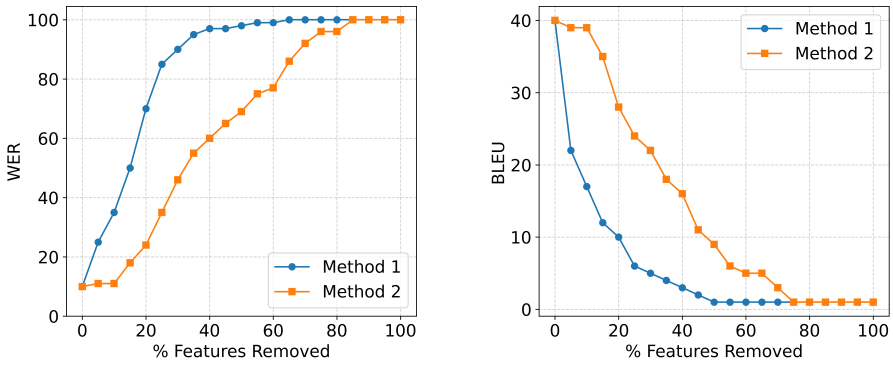



Figure 2: Illustrative deletion curves for ASR (left) and ST (right). The curves are fictitious and serve only to demonstrate how performance degrades as increasingly larger portions of salient features are removed. A steeper initial drop indicates that the attribution method identified critical features earlier, with the AUC summarizing this effect.

other words remain unaffected, since the noisy features are not expected to influence them. The process is repeated across various subsets of features with different saliency levels. However, this approach assumes conditional independence in predicting different words, an assumption that does not hold in state-of-the-art autoregressive S2T models, where early errors can impact subsequent predictions.

To address this limitation, we extend the deletion metric to the sentence level, allowing us to evaluate the impact of feature removal in the context of entire utterances rather than individual tokens. Specifically, we iteratively remove spectrogram features according to their saliency ranking and assess model performance on the complete sentence using standard ASR and ST metrics. At each step, the top $5\%$ of features are zeroed out, after which model performance is re-evaluated across the full test set. Performance is quantified using task-specific extrinsic metrics: for ASR, we compute WER[7] (↓), and for ST, we report BLEU (↑). These metrics are used to construct evaluation curves, which illustrate how performance degrades as features are removed in decreasing order of relevance.

To illustrate this process in detail, Figure 2 presents curves derived from fictitious explanations obtained with two methods (Method 1 and Method 2) on both ASR and ST. The x-axis shows the percentage of removed features, while the y-axis shows performance (WER for ASR, BLEU for ST). In the example, Method 1 causes a sharper degradation: even at $5\%$ feature removal, WER increases and BLEU decreases, whereas Method 2 initially remains stable. As feature removal progresses, both methods exhibit performance degradation, but Method 1 alters the curve more sharply than Method 2, indicating that it more effectively identifies the features critical to the prediction. To summarize this behavior with a single metric capturing the overall effect, we compute the area under the curve (AUC): for ASR, a higher AUC (higher WER), and for ST, a lower AUC (lower BLEU) correspond to faster performance degradation, reflecting more faithful explanations.

The main limitations of the deletion metric lie in *i)* its relatively high computational costs—as it involves multiple generations on the test set—and *ii)* its assumption that the majority of the input relevance is placed on only a few features, which is typically valid when using a softmax operation, but it may not be in other cases (Nauta et al., 2023). However, this assumption holds in our setting as the outputs come from a softmax operation.

### 5.2 Size

Size measures the compactness of an explanation as the percentage of input features identified as responsible for the model's prediction. In classification tasks, this involves incrementally adding the features with the highest saliency scores until the model's prediction matches the original (Sun et al., 2020). However, applying this approach in S2T tasks is complex since the output text changes continuously rather than in discrete steps as in classi-

[7]WER is computed with `editdistance 0.6.1` (https://github.com/roy-ht/editdistance). Scores are capped at 100 to mitigate instability from excessively high values due to long hallucinations.

fication. Wu et al. (2024) test two criteria to detect changes: *i)* WER greater than 0, which considers any modification—even irrelevant—as a prediction change, and *ii)* cosine similarity of SentenceBERT encodings lower than 0.5, which targets changes in output semantics. Since the choice of the similarity metric can lead to varying results and the optimal metric is debatable, we avoid isolating the subset of most salient features; instead, we assess explanation size across various subsets.

To this aim, we calculate the percentage of features with saliency scores above a threshold ranging from 0 to 1 in 5% increments and compute the AUC of the resulting curve as a comprehensive score. At the beginning of the curve, all features have saliency scores above 0. As the threshold increases, the percentage of included features decreases. A compact explanation will have most features below low thresholds, with only a few exceeding higher ones, while a more diffuse explanation will retain more features at higher thresholds.

By analyzing the curve's shape and its AUC, we can assess explanation compactness. Lower AUCs indicate more compact—hence clearer and more focused—explanations (Nauta et al., 2023). However, we caution that AUC evaluation for size may be vulnerable to adversarial artifacts, as an uninformative explanation with all saliency scores set to 0 would still yield a low AUC while providing no insights. Thus, size must be interpreted alongside faithfulness metrics for a comprehensive assessment.

## 6 Experimental Setting

**Tasks and Data** We test SPES on S2T tasks with English audio, focusing on ASR and ST into German (en→de) and Spanish (en→es) to assess its effectiveness across varying monotonicity in word alignment.[8] For evaluation, we use the MuST-C corpus (Cattoni et al., 2021): the en→de v2 subset for ASR and en→de ST, and the en→es v1 subset for en→es ST due to the absence of v2.

**Models** The ASR and two ST systems share the same architecture, consisting of a Conformer encoder and a Transformer decoder, and are fed with 80-feature mel-spectrograms (see Appendix A.1 for details on hyperparameters, training configurations, and data). We opted not to use widely adopted foundation models like Whisper (Radford et al., 2023) or SeamlessM4T (Communication et al., 2023) for several reasons: *i)* comparability across tasks, as Whisper only supports English translations; *ii)* performance, as our models outperform SeamlessM4T on these tasks (see comparison in Appendix A.1); and *iii)* computational efficiency, since our smaller models are more suited for extensive experiments while keeping computational costs manageable.

**SPES Hyperparameters** SPES hyperparameters were estimated through validation (see Appendix A.2). Spectrograms were clustered into patches using SLIC with default parameters from the scikit-image implementation,[9] except for sigma ($\sigma$), which was set to 0. We set $\phi$ to 400, 500, and 600 for the three-level multi-scale clustering, while $\tau$ was set to 7.5 for ASR and 5 for ST. The number of masking iterations, $N^X$, was set to 20,000 with $p_{spec}$ at 0.5. For previous output tokens, $N^Y$ was set to 2,000 with $p_{tok}$ at 0.4.

**Baselines** We compare SPES against two baselines. First, a ***feature-wise*** method that randomly perturbs previous output tokens and spectrograms at the level of single features without morphological clustering, and estimates the perturbation impact with KL divergence. For a fair comparison, we set $N^X$ to 20,000 and $N^Y$ to 2,000, as in SPES, with $p_{spec} = 0.7$ and $p_{tok} = 0.1$, optimized via validation (see Appendix A.3.1). Second, an enhanced version of the ***Bubble Noise*** technique that uses KL divergence as a perturbation impact estimator and is extended to also perturb the previous output tokens. Since Bubble Noise is significantly slower than SPES, especially at high bubble counts, we set $N^X$ to 1,000. This value corresponds to that used by Trinh and Mandel (2020) and results in a runtime comparable to SPES with $N^X$ set to 20,000. The number of bubbles per second is set to 10, with each bubble measuring approximately 0.43 seconds in width and 31 mel bins in height, as validated in Appendix A.3.2.

## 7 Results

### 7.1 Overall Results

The upper rows of Table 1 present Deletion and Size AUC scores for the two implemented baselines (*feature-wise* and *Bubble Noise*) and com-

[8] English and Spanish follow the SVO order, while German is an SOV language (Östling, 2015).

[9] https://scikit-image.org/docs/dev/api/skimage.segmentation.html

| | ASR (en) | | ST (en→de) | | ST (en→es) | |
|---|---|---|---|---|---|---|
| | **Deletion** (↑) | **Size** (↓) | **Deletion** (↓) | **Size** (↓) | **Deletion** (↓) | **Size** (↓) |
| feature-wise | 71.39 | 47.85 | 8.79 | 48.53 | 9.03 | 46.00 |
| Bubble Noise | 65.02 | 36.92 | 10.85 | 43.89 | 13.01 | 43.04 |
| SPES | **92.54** | **29.71** | **2.34** | **29.28** | **2.45** | **29.47** |
| – multi-scale | 92.22 | 29.94 | 2.49 | 29.62 | 2.56 | 29.78 |
| – clustering | 90.73 | 33.31 | 2.84 | 33.09 | 2.93 | 32.90 |
| – duration adaptation | 92.37 | 30.45 | 2.35 | 31.10 | 2.46 | 30.95 |
| – KL | 87.44 | 29.85 | 4.76 | 30.98 | 4.99 | 30.74 |

Table 1: Deletion and Size scores for the *feature-wise* and the *Bubble Noise* baselines as well as for SPES and the solutions tested in the ablation study.

pare them with SPES across ASR (en) and ST (en→de, en→es). Full deletion and size curves are provided in Fig. 7 and 8 in Appendix B.

Deletion and size scores reveal consistent trends across the three tasks. Looking at the baselines, we observe that the *feature-wise* method outperforms *Bubble Noise* in deletion scores, likely due to the differing number of perturbation iterations ($N^X$) assigned to the two baselines (see §6). However, despite the significantly lower $N^X$, *Bubble Noise* surpasses *feature-wise* in size scores.

When comparing SPES to the baselines, it clearly excels. In terms of deletion scores, SPES outperforms the *feature-wise* baseline by 21.15 points in ASR and up to 6.58 points in ST (en→es). The improvements are even more pronounced relative to the *Bubble Noise* technique (27.52 points in ASR and 10.56 in ST en→es). SPES also demonstrates its superiority in the size metric, with differences of up to 19.25 and 14.61 in ST (en→de) compared to *feature-wise* and *Bubble Noise*, respectively. Since the comparison between SPES and *Bubble Noise* involves different $N^X$ values, for the sake of fairness we carried out additional experiments using the same $N^X$ value (1,000) for both SPES and *Bubble Noise*. The results, presented in Table 13 in Appendix B, confirm SPES's superior performance under this condition as well. Overall, **SPES generates faithful and compact explanations**, accurately highlighting the input elements that are relevant to the S2T model's predictions while avoiding the inclusion of redundant information.

### 7.2 Ablation Study

To better understand how each design choice impacts the faithfulness and compactness of SPES-generated explanations, we conduct an ablation study. Firstly, we vary the granularity of spectrogram perturbations by testing: *i)* single-level segmentation instead of multi-level (– *multi-scale*), by using only 500 as $\phi$, *ii)* fixed grid segmentation[10] instead of clustering (– *clustering*), *iii)* removal of the duration-based patch adaptation of Eq. 2 (– *duration adaptation*), using a fixed number of patches (2,000, 2,500, and 3,000)[11] for all audio samples regardless of their duration. Lastly, we assess the contribution of KL divergence by replacing it with token-level probability difference—widely used in text applications (Sarti et al., 2023; Miglani et al., 2023)—as a method for estimating perturbation impact (– *KL*).

Table 1 (lower rows) presents the scores resulting from these variations. Looking at deletion, the largest performance drop across all tasks occurs when KL divergence is replaced with token-level probability difference (– *KL*). Ablating KL divergence reduces performance also in terms of size, although differences are less evident (ranging from 0.14 to 1.70 points). We can conclude that **using KL divergence to estimate perturbation impact is beneficial for producing explanations that are both more faithful and more compact**. Examining the ablations related to patch definition in the input spectrogram, we find that considering morphological aspects has the greatest impact. Indeed, fixed grid segmentation (– *clustering*) leads to decreased faithfulness for both ASR and ST, along with a 12-13% increase in explanation size across all tasks. This indicates that **accounting for morphology in spectrograms is crucial** for isolating patterns that models exploit for their predictions. **Adapting the number of patches to the in-**

[10] Obtained setting the sigma parameter in SLIC, which determines the smoothness of the segmentation contours, to 10.

[11] Validated on the dev set, as described in Appendix A.2.1.

| | **Alignment** (↑) |
|---|---|
| feature-wise | 51.10 |
| Bubble Noise | 72.00 |
| SPES | **82.97** |
| – multi-scale | 82.52 |
| – clustering | 81.09 |
| – duration adaptation | 82.96 |
| – KL | 82.59 |

Table 2: *Temporal alignment* scores between the top salient element and the segmentation in ASR.

**put speech duration also proves beneficial**: with a fixed number of patches (– *duration adaptation*), deletion scores are slightly worse but similar, and the size increases by 2-6%. Lastly, **using multiple segmentation levels improves both faithfulness and compactness**: ablating this option (– *multi-scale*) has a slight negative impact across all tasks, reflected in both deletion and size scores, with a degradation of approximately 1% in size for ST.

In summary, all the design choices of SPES contribute to providing faithful and compact explanations. Among them, estimating the impact of perturbations with KL divergence proves to be particularly crucial for achieving meaningful outcomes, followed by employing morphological clustering to define the units for perturbation.

## 8 Analyses of Plausibility

We assess the *plausibility* of SPES explanations (as defined in §5) by analyzing the saliency maps for both spectrograms ($\boldsymbol{S^X}$) and previous tokens ($\boldsymbol{S^{Y^{(k-1)}}}$). For $S^X$, we focus on ASR, where the monotonic audio-text alignment facilitates analysis, and compare SPES with the baselines and ablation variants. For $S^{Y^{(k-1)}}$, instead, we focus solely on SPES, as the introduction of $S^{Y^{(k-1)}}$ is one of its distinctive characteristics, and we examine both ASR and ST.

### 8.1 Explanations for Spectrograms

**Time Dimension** In ASR, manual inspection of $S^X$ examples suggests that saliency moves monotonically along the time axis as generation progresses from the first to the last token (see Appendix A.4 for an example of SPES explanations). This aligns with the expectation that token generation is guided by the time frames in which they are uttered, leading to time-localized explanations that shift from the beginning to the end of the audio.

We quantitatively verify this by comparing word-level saliency maps along the time axis[12] with word alignments from Gentle.[13] Specifically, we extract the top-1 feature's time position and check whether it falls within the corresponding word's time span. This is computed for each $S^X$ of every correctly predicted word, yielding a *temporal alignment* score—the percentage of cases where the top-1 feature falls within the expected time span. The scores in Table 2 show that SPES achieves the highest agreement at 82.97%, which is notable given the strict evaluation setting that considers only the top-1 feature. Similar to deletion and size metrics, ablation settings perform closely to SPES, ranging from 81.09% (SPES – clustering) to 82.59% (SPES – KL). Baselines show a larger gap, with Bubble Noise achieving 72.00% and the feature-wise setting only 51.10%. Additionally, for SPES, we compare the mean saliency scores for time frames inside and outside Gentle's boundaries. A Student's t-test confirms that the mean score is significantly higher inside the word's time span.

These analyses confirm that **SPES provides plausible insights into the time dimension**.

**Frequency Dimension** A manual inspection of SPES' explanations reveals that they are also well-localized in the frequency domain, often aligning with relevant spectral patterns, as shown in Fig. 10 and 11 (Appendix B). To evaluate whether SPES provides plausible insights along the frequency dimension, we extract saliency scores from the midpoint frame of the time span of individual phones.[14] The time spans are obtained using phonetic segmentation from Gentle. We then check whether the top-1 score in that frame falls within the typical frequency range for the phoneme under investigation, computing a *frequency alignment* score. Specifically, we analyze three vowels (/ɑ/, /ɔ/, /ɛ/) and two consonants (/s/, /n/), focusing on the first and second formants (F1-F2) for vowels, the frication range for the voiceless alveolar sibilant /s/, and the murmur range for the voiced alveolar nasal /n/, as these are crucial for their identification (Hillenbrand et al., 1995; Jongman

[12] We aggregate explanations at the word level by averaging token saliency maps and then take the maximum value over the frequency axis for each time frame.

[13] https://github.com/lowerquality/gentle.

[14] At midpoint, phone pronunciation is typically sustained.

| | ɑ | | ɔ | | ɛ | | s | | n | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | M | F | |
| feature-wise | 59.10 | 54.56 | 56.94 | 48.43 | 53.04 | 39.10 | 31.11 | 44.43 | 41.79 | 44.78 | 47.33 |
| Bubble Noise | 28.93 | 16.03 | 39.70 | 17.66 | 49.82 | 33.24 | 3.07 | 1.31 | 48.60 | 46.25 | 28.46 |
| SPES | **85.61** | **84.85** | **87.16** | 85.17 | 74.12 | 61.08 | 77.97 | 85.58 | 56.24 | 58.28 | **75.61** |
| – multi-scale | 81.25 | 82.50 | 84.18 | 84.02 | 70.69 | 59.66 | 75.07 | 83.79 | 53.53 | 56.98 | 73.17 |
| – clustering | 80.69 | 81.32 | 85.67 | 79.57 | **76.94** | **65.87** | **78.12** | 79.53 | **58.59** | **59.69** | 74.60 |
| – duration ad. | 85.26 | 83.97 | 86.12 | **86.00** | 73.27 | 61.44 | 74.18 | 86.25 | 52.76 | 53.33 | 74.26 |
| – KL | 84.98 | 83.97 | 86.42 | 84.32 | 72.56 | 59.66 | 77.77 | **86.88** | 56.42 | 59.66 | 75.26 |

Table 3: *Frequency alignment* scores between the top salient element and the F1-F2 ranges for the vowels /ɑ/, /ɔ/, and /ɛ/, as well as the noise ranges for the consonants /s/ and /n/, across genders (M, F) in ASR.

et al., 2000; Pruthi and Espy-Wilson, 2004).[15]

The scores in Table 3 show that SPES, including its ablation conditions, generates the most aligned explanations, with variations across phonemes. The lowest alignment occurs for /n/ (best score at 58.59% for males and 59.69% for females in the SPES – clustering), while the highest is for /ɔ/ (best score at 87.16% for males in SPES and 86.00% for females in SPES – duration adaptation). The feature-wise and Bubble Noise baselines perform significantly worse, and on average, SPES provides the highest scores.

Saliency peaks in frequency regions distinctive for the analyzed phones are also visible in the average explanation frame values, plotted in Fig. 9 (Appendix B). For vowels, notable peaks occur below 2,000 Hz for /ɑ/ and /ɔ/, where F1 and F2 are closer, and around 1,000–2,000 Hz for /ɛ/, where F1 and F2 are more distant. Similarly, saliency peaks are visible above 4,000 Hz for /s/ and below 500 Hz for /n/. Interestingly, when examining saliency curves by gender, female speakers' scores shift toward higher frequencies for vowels and /s/, reflecting known gender-based spectral differences due to shorter vocal tracts and smaller larynx sizes (Stevens, 2000).

These analyses indicate that **SPES can offer valuable insights into the frequency dimension**, capturing phonetic differences between sounds and sociophonetic characteristics such as gender.

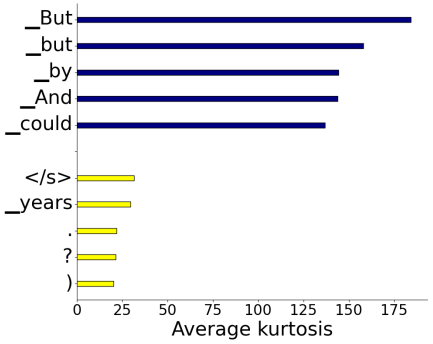



Figure 3: Top and bottom 5 tokens by kurtosis.

## 8.2 Spectrograms vs Previous Output Tokens

Our previous analyses show that SPES-generated saliency maps generally localize salient areas effectively in both time and frequency domains. However, manual inspections revealed that some spectrogram saliency maps appear "noisy", with scattered explanations that lack concentrated groups of features with higher saliency scores (for an example, see Appendix A.4). To investigate this behavior further, we explore whether this scattering is associated with specific token classes. Specifically, we hypothesize that, for some tokens, the less compact explanations in terms of speech features may stem from a higher contribution of previous output tokens to their prediction.

To quantify scattering, we use kurtosis, a measure of the "tailedness" of score distributions. Higher kurtosis indicates more peaked distributions, reflecting more focused explanations. For each token occurring over 100 times in the test set, we average the kurtosis values of all the spectrogram saliency maps related to its occurrences. In-

[15] We define the phoneme frequency ranges (in Hz) based on prior studies of American English production (Hillenbrand et al., 1995; Jongman et al., 2000; Maniwa et al., 2009; Pruthi and Espy-Wilson, 2004). For /ɑ/, $F1 \in [550, 950]$, $F2 \in [950, 1400]$ for males and $F1 \in [650, 1100]$, $F2 \in [1100, 1600]$ for females; for /ɔ/, $F1 \in [350, 750]$, $F2 \in [750, 1300]$ for males and $F1 \in [450, 900]$, $F2 \in [900, 1500]$ for females; for /ɛ/, $F1 \in [400, 700]$, $F2 \in [1300, 1900]$ for males and $F1 \in [450, 700]$, $F2 \in [1600, 2300]$ for females. For /s/, the frication range is $[4000, 7500]$ for males and $[4500, 8000]$ for females, while for /n/, the murmur range is $[100, 600]$ for both genders, as cross-gender differences are inconsistent (Kurowski, 2005).

| Token | ASR (en) | ST (en→de) | ST (en→es) |
|---|---|---|---|
| <s> | 4.75 | 11.49 | 4.82 |
| IT | 2.82 | 9.76 | 6.00 |
| LT | 92.43 | 78.75 | 89.18 |

Table 4: Percentage of instances where the initial token <s>, any intermediate token IT, or the latest token LT is the most relevant.

| Task | Token | IT | % |
|---|---|---|---|
| *ASR en* | ) | ( [47] | 40.5 |
| | **?** | **What** [11], **what** [7] | 29.1 |
| ST en→es | **?** | **¿** [35], **"¿** [7], | 82.2 |
| | ) | ( [148] | 33.1 |
| ST en→de | ) | ( [105] | 87.6 |
| | **zu** | **um** [23], **zu** [11], | 29.3 |
| | **an** | **sich** [6], **fängt** [4] | 25.1 |

Table 5: Tokens for which the highest saliency score is more often held by an IT. For each token, we provide the overall percentage of times it is primarily explained by *any* IT (%), and the frequency of each *individual* IT in holding the highest saliency in brackets.

terestingly, a trend emerges: tokens with higher kurtosis generally correspond to full words, while those with lower kurtosis are punctuation marks or special tokens. Fig. 3 shows the top and bottom 5 tokens by kurtosis. The bottom tokens include punctuation marks, the special end-of-sentence token </s>, and _years. The scattered explanations for these tokens suggests a lack of crucial audio information for their prediction. This is expected for punctuation, which often depends on grammatical rules rather than audio content. Similarly, </s> is always predicted based on a preceding strong punctuation mark, independent of the audio. In the case of _years, the model often relies on a preceding number in the output tokens, indicating a pattern-based prediction.

As an additional quantitative validation, we extract the 100 tokens with the highest and the 100 with lowest kurtosis. Two annotators, both authors of this paper, then judged whether it was plausible that the model relied more on the spectrogram for high-kurtosis tokens (top-100) and less for low-kurtosis tokens (bottom-100) when making predictions. Our results indicate that model's behavior was found plausible in 83% of instances.[16]

Overall, these findings reveal that **explanations obtained with SPES can be used to determine whether a token is predicted based on the input speech or on the textual context**, as it highlights the dependency of a predicted token on previous ones when plausible. In the next section, we further explore the plausibility of the textual relationships unveiled by SPES.

### 8.3 Explanations for Previous Output Tokens

**Positional Saliency** A manual inspection of $S^{Y^{(k-1)}}$ in both ASR and ST revealed that two types of tokens consistently receive higher saliency scores: the special start-of-sentence token <s> and the latest token (LT), i.e., the token immediately preceding the one being explained (for an example, see Appendix A.4). The distribution of saliency scores across token types aligns with our qualitative observations: LTs have the highest mean (see Fig. 12 in Appendix B), followed by <s>, whose distribution shifts towards higher scores compared to intermediate tokens (ITs). Student's t-tests confirm significant differences in mean saliency scores across groups ($p < 0.01$). Table 4 reports the percentage of $S^{Y^{(k-1)}}$ instances where <s>, an IT, or the LT receives the highest score. The LT is consistently the most important across tasks and language directions. Notably, this percentage is higher in ASR, where monotonic alignment reduces the importance of attending to tokens beyond the immediately preceding ones. In contrast, it decreases in ST, particularly in the en→de direction, which exhibits less monotonicity than en→es and where interdependencies across different positions in the target context may play a more significant role. The relevance of LTs underscores the critical role of the closest linguistic context. This finding is line with prior work that, by analyzing model component importance, demonstrated the significant focus of attention heads on LTs (Ferrando and Voita, 2024) and of final token states in terminal layers for predicting the subsequent tokens (Meng et al., 2022). Additionally, the significance of <s> echoes findings from research on "attention sinks" in autoregressive language models, which suggest that initial tokens act as anchors due to their positional

[16] Inter-annotator percentage agreement was very high (93.5%). Most disagreements were oversights and thus reconciled, except for three tokens. Given the subjectivity of judging plausibility, we did not enforce agreement, and these tokens were excluded from the analysis.

values (Xiao et al., 2024). In summary, concerning $S^{Y^{(k-1)}}$, **SPES highlights mechanisms that align with findings from previous research**.

**Intermediate Tokens** Despite the aforementioned trends in positional saliency, we observed instances where an `IT` holds the highest saliency score for the explanations of certain tokens (see Appendix A.4). To investigate this phenomenon, we analyze tokens occurring over 100 times in the test data for each task. For each unique token, we calculate the frequency with which an `IT` received the highest score in the explanation.

Table 5 presents examples from each language and task, showing the tokens where an `IT` held the highest score in at least 25% of their occurrences, along with their top two most frequent `IT`s, if available. Overall, these explanations appear plausible and linguistically motivated. First, punctuation marks are present across all languages. The token `)` is explained by its complementary `(`. Similarly, `?` is explained by `What`/`what` in English, and by `¿` in Spanish, i.e. by `IT`s that typically introduce questions. This signals that the preceding context plays a pivotal role in explaining token pairs that frequently co-occur. Also, it aligns with our findings on punctuation marks (§8.2): being characteristic of written text, punctuation does not have a direct correspondence in speech, reducing the saliency of the audio signal in the explanation.

By specifically looking at en→de, the token `zu` is preceded by the `IT` `um` to create infinitive sentences. This German grammatical construction corresponds to the single token "to" in English (e.g., "***um*** *Menschen aus ihren Stühlen* ***zu*** *holen*" – "**to** get people out of their chairs."). Along the same line, the token `an` acts as a separable prefix for the verb `fängt` (e.g. "*Es* ***fängt*** *später* ***an***" – "*it* ***starts*** *later*"). Also, `an` is part of fixed expressions like "*es fühlt* ***sich*** *so* ***an***" ("*it feels* ***like***"). These examples further demonstrate how **SPES relies on previous context to accurately interpret and explain tokens, particularly when there is no one-to-one mapping between source sentence and target translation.**

## 9 Conclusion and Final Remarks

In response to the need for feature attribution methods specifically designed for S2T autoregressive models, this work introduces **SPES** (***S****pectrogram* ***P****erturbation for* ***E****xplainable* ***S****peech-to-text Generation*). By employing a perturbation-based approach that ensures reliable explanations while remaining compatible with any model architecture, SPES produces token-level explanations across the predicted sentence. Its saliency maps are the first to effectively highlight contributions from fine-grained, acoustically meaningful patterns, such as formants, while also accounting for the influence of previous output tokens. We evaluate SPES on ASR and—for the first time—ST, supported by in-depth analyses. Results show that SPES provides explanations that are faithful, compact, and plausible. Future work could evaluate our approach on state-of-the-art S2T systems, such as Whisper, to verify whether our insights into model behavior generalize to larger models.

Overall, our work advances transparency and interpretability in S2T technology, with SPES offering diverse applications. It reveals which spectral features speech models rely on, potentially addressing long-standing questions about their alignment with human perception (e.g., Mandel, 2016; Trinh and Mandel, 2021). It can also provide insights into model errors, including hallucinations—a challenge in both language and speech models (Frieske and Shi, 2024; Sahoo et al., 2024; Atwany et al., 2025). For example, it can help identify whether errors stem from noisy and misleading input or reliance on incorrect spectral features. Additionally, it aids in bias detection, such as gender bias, a crucial factor in ASR (e.g., Attanasio et al., 2024) and ST (e.g., Bentivogli et al., 2020). Feature attribution methods for gender bias are well-documented in NLP (Jeyaraj and Delany, 2024; Chuan et al., 2024; Muntasir and Noor, 2025), and applying them in speech research could help determine whether acoustic cues or specific tokens influence gender assignment. By shedding light on these factors, SPES can be used for the development of reliable, robust, and fair speech models.

## Acknowledgments

This work was funded by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU, and by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People). We would also like to thank

the anonymous reviewers and the Action Editor, whose insightful comments helped us improve the current version of the paper.

## References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23181–23203, Vienna, Austria. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):1–46.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2024. AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, et al. 2021. On the Opportunities and Risks of Foundation Models.

Lennart Brocki and Neo Christopher Chung. 2023. Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. *Pattern Recognition Letters*, 176:131–139.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8).

Javier Castro, Daniel Gómez, Elisenda Molina, and Juan Tejada. 2017. Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research*, 82:180–188.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Ching-Hua Chuan, Ruoyu Sun, Shiyun Tian, and Wan-Hsiu Sunny Tsai. 2024. EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives. *Telematics and Informatics*, 91:102135.

Yu-Neng Chuang, Guanchu Wang, Fan Yang, Zirui Liu, Xuanting Cai, Mengnan Du, and Xia Hu. 2023a. Efficient XAI Techniques: A Taxonomic Survey.

Yu-Neng Chuang, Guanchu Wang, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanting Cai, and Xia Hu. 2023b. CoRTX: Contrastive Framework for Real-time Explanation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, et al. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Ian C. Covert, Scott Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. In *Advances on Neural Information Processing Systems*, volume 33 of *NIPS '20*. Curran Associates, Inc.

Ian C. Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: a unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566.

Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Mattia A. Di Gangi, Marco Gaido, Matteo Negri, and Marco Turchi. 2020. On Target Segmentation for Direct Speech Translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 137–150, Virtual. Association for Machine Translation in the Americas.

Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. 2022. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121(C).

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning.

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Imperial, Juan Arturo Nolazco, Lori Landay, Matthew Jackson, Phillip H. S. Torr, Trevor Darrell, Yong Lee, and Jakob Foerster. 2024. Risks and Opportunities of Open-Source Generative AI.

Thomas Fel, Melanie Ducoffe, David Vigouroux, Rémi Cadène, Mikael Capelle, Claire Nicodème, and Thomas Serre. 2023. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16153–16163.

Jean Louis K. E Fendji, Diane C. M. Tala, Blaise O. Yenke, and Marcellin Atemkeng. 2022. Automatic Speech Recognition Using Limited Vocabulary: A Survey. *Applied Artificial Intelligence*, 36(1):2095039.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A Primer on the Inner Workings of Transformer-based Language Models.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, Venice, Italy. IEEE Computer Society.

Rita Frieske and Bertram E. Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models.

Dennis Fucci, Beatrice Savoldi, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2024. Explainability for Speech Models: On the Challenges of Acoustic Feature Selection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 373–381, Pisa, Italy. CEUR Workshop Proceedings.

Arne Gevaert, Axel-Jan Rousseau, Thijs Becker, Dirk Valkenborg, Tijl De Bie, and Yvan Saeys. 2024. Evaluating feature attribution methods in the image domain. *Machine Learning*, 113(9):6019–6064.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, page 369–376, Pittsburgh, Pennsylvania, USA. Association for Computing Machinery.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040, Shanghai, China. International Speech Communication Association.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In *Speech and Computer*, pages 198–208, Leipzig, Germany. Springer International Publishing.

James Hillenbrand, Laura Getty, Michael Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97:3099–111.

Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. Explainable AI Methods - A Brief Overview. In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*, pages 13–38, Vienna, Austria. Springer International Publishing.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, Barcelona, Spain. The Institute of Electrical and Electronics Engineers, Inc.

Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguis-*

*tics*, pages 4198–4205, Online. Association for Computational Linguistics.

Manuela Jeyaraj and Sarah Delany. 2024. An Explainable Approach to Understanding Gender Stereotype Text. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 45–59, Bangkok, Thailand. Association for Computational Linguistics.

A. Jongman, R. Wayland, and S. Wong. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3 Pt 1):1252–1263.

Vidhya Kamakshi and Narayanan C. Krishnan. 2023. Explainable Image Classification: The Journey So Far and the Road Ahead. *AI*, 4(3):620–651.

Hassan Salami Kavaki and Michael I. Mandel. 2020. Identifying important time-frequency locations in continuous speech utterances. In *Interspeech 2020*, pages 1639–1643.

Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, and Nassir Navab. 2020. Improving Feature Attribution through Input-specific Network Pruning.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Kathleen Mary Kurowski. 2005. Acoustic properties for place of articulation in nasal consonants: A cross-linguistic study. *The Journal of the Acoustical Society of America*, 88(S1):S80–S80.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and Björn W. Schuller. 2023. Sparks of Large Audio Models: A Survey and Outlook.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. 2023. $M^4$: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. In *Advances on Neural Information Processing Systems*, volume 36, pages 1630–1643, New Orleans, LA, USA. Curran Associates, Inc.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances on Neural Information Processing Systems*, volume 30, page 4768–4777, Long Beach, CA, USA. Curran Associates, Inc.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 55(8).

Michael I. Mandel. 2013. Learning an intelligibility map of individual utterances. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4.

Michael I. Mandel. 2016. Directly Comparing the Listening Strategies of Humans and Machines. In *Interspeech 2016*, pages 660–664.

Kazumi Maniwa, Allard Jongman, and Travis Wade. 2009. Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, 125(6):3962–3973.

Karla Markert, Romain Parracone, Mykhailo Kulakov, Philip Sperl, Ching-Yu Kao, and Konstantin Böttinger. 2021. Visualizing Automatic Speech Recognition – Means for a Better Understanding? In *Proceedings of the 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pages 14–20, Virtual. International Speech Communication Association.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances*

*on Neural Information Processing Systems*, volume 35, New Orleans, LA, USA. Curran Associates, Inc.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using Captum to Explain Generative Language Models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying Context Mixing in Transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-Wise Relevance Propagation: An Overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, chapter 10, pages 193–209. Springer International Publishing, Cham.

Fahim Muntasir and Jannatun Noor. 2025. Explainable AI Discloses Gender Bias in Sexism Detection Algorithm. In *Proceedings of the 11th International Conference on Networking, Systems, and Security*, NSysS '24, page 120–127, New York, NY, USA. Association for Computing Machinery.

Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, page 2459–2464, Paris, France. Association for Computing Machinery.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s).

Robert Östling. 2015. Word Order Typology through Multilingual Word Alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, Queensland, Australia. The Institute of Electrical and Electronics Engineers, Inc.

Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. 2023. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1139–1150, New York, NY, USA. Association for Computing Machinery.

Daniel S. Park, William Chan, et al. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech 2019*, Graz, Austria. International Speech Communication Association.

Eliana Pastor, Alkis Koudounas, Giuseppe Attanasio, Dirk Hovy, and Elena Baralis. 2024. Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2221–2238, St. Julian's, Malta. Association for Computational Linguistics.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan

Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, {Jee Weon} Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing Whisper-Style Training Using An Open-Source Toolkit And Publicly Available Data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023*, 2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023. Institute of Electrical and Electronics Engineers Inc.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*, Northumbira University, UK. The British Machine Vision Association and Society for Pattern Recognition.

Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable Data-Based Explanations for Fairness Debugging. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD)*, pages 247–261, Philadelphia, PA, USA. Association for Computing Machinery.

Tarun Pruthi and Carol Y. Espy-Wilson. 2004. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43(3):225–239.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28492–28518, Honolulu, Hawaii, USA. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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 (KDD)*, page 1135–1144, San Francisco, California, USA. Association for Computing Machinery.

Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 2024. Global AI governance: barriers and pathways forward. *International Affairs*, 100(3):1275–1286.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3):247–278.

Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, chapter 1, pages 5–22. Springer International Publishing, Cham.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An Interpretability Toolkit for Sequence Generation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Robert Schwarzenberg, Nils Feldhus, and Sebastian Möller. 2021. Efficient Explanations from Empirical Explainers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 240–249, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM:

Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, Seoul, South Korea. IEEE Computer Society.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Dasom Seo, Kanghan Oh, and Il-Seok Oh. 2020. Regional Multi-Scale Approach for Visually Pleasing Explanations of Deep Neural Networks. *IEEE Access*, 8:8572–8582.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3145–3153, Sydney, NSW, Australia. PMLR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proceedings of the 2nd International Conference on Learning Representations*, Banff, AB, Canada.

Kenneth N. Stevens. 2000. *Acoustic Phonetics*. The MIT Press.

Youcheng Sun, Hana Chockler, Xiaowei Huang, and Daniel Kroening. 2020. Explaining Image Classifiers Using Statistical Fault Localization. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings*, pages 391–406, Glasgow, UK. Springer International Publishing.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, Sydney, NSW, Australia. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances on Neural Information Processing Systems*, volume 27, Montreal, QC, Canada. Curran Associates, Inc.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA. IEEE Computer Society.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2020. Sanity Checks for Saliency Metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6021–6029.

Viet Anh Trinh and Michael Mandel. 2021. Directly Comparing the Listening Strategies of Humans and Machines. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:312–323.

Viet Anh Trinh and Michael I. Mandel. 2020. Large Scale Evaluation of Importance Maps in Automatic Speech Recognition. In *Interspeech 2020*, pages 1166–1170.

Viet Anh Trinh, Brian McFee, and Michael I Mandel. 2018. Bubble Cooperative Networks for Identifying Important Speech Cues. In *Interspeech 2018*, pages 1616–1620.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Olli Viikki and Kari Laurila. 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Jörg Wagner, Jan Mathias Köhler, Tobias Gindele, Leon Hetzel, Jakob Thaddäus Wiedemer, and Sven Behnke. 2019. Interpretable and

Fine-Grained Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9089–9099, Long Beach, California, USA. IEEE Computer Society.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2023. Explanations for Automatic Speech Recognition. In *Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece. The Institute of Electrical and Electronics Engineers, Inc.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2024. Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10296–10300.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks.

Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You, and Yuan Zhu. 2021. MFPP: Morphological Fragmental Perturbation Pyramid for Black-Box Model Explanations. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pages 1376–1383, Los Alamitos, CA, USA. IEEE Computer Society.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2019. INVASE: Instance-wise Variable Selection using Neural Networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014 13th European Conference, Proceedings*, pages 818–833, Zurich, Switzerland. Springer International Publishing.

Fred Zhang and Neel Nanda. 2024. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2).

Shuai Zhao, Zhuoqian Liang, Jinming Wen, and Jie Chen. 2023. Sparsing and Smoothing for

the seq2seq Models. *IEEE Transactions on Artificial Intelligence*, 4(3):464–472.

# A Experimental Setting

## A.1 Model Implementation

In this section, we provide further details on the implementation of the ASR and ST models (see §6) used in our experiments.

**Models** Our models were implemented using the fairseq-S2T framework (Wang et al., 2020b). As audio features, we use log-compressed mel-filterbanks (80 channels), computed over windows of 25ms with a stride of 10ms using pykaldi.[17] We encode text into BPE (Sennrich et al., 2016) using SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 8,000 for English and 16,000 for Spanish and German. The length of audio features is reduced by a factor of 4 using two 1D convolutional layers with stride 2. The output of the convolutions is fed to a 12-layer Conformer (Gulati et al., 2020) encoder, and a 6-layer Transformer (Vaswani et al., 2017) decoder. We used 512 embedding features, 2,048 hidden features in the FFN, and a kernel size of 31 for Conformer convolutions. In total, the models have 113M parameters. As an objective, we minimize label-smoothed cross entropy (Szegedy et al., 2016) on the output of the decoder (using as a reference either the textual translation, for ST, or the transcripts, for ASR) with an auxiliary CTC (Graves et al., 2006) loss with the transcripts as reference, regardless of the target task, computed on the output of the 8th encoder layer. We optimize the model weights with the Adam optimizer (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$) and Noam learning rate scheduler (Vaswani et al., 2017) (inverse square-root) starting from 0 and reaching the 0.002 peak in 25,000 warm-up steps. The input features are normalized with Utterance-level Cepstral Mean and Variance Normalization and, at training time, we apply SpecAugment (Park et al., 2019). We set dropout probability to 0.1. The ASR model was trained for 250,000 steps while the ST models, one for each language direction, were trained for 200,000 steps, as their encoder was initialized with that of the ASR model. All trainings were carried out on four NVIDIA A100 GPUs (40GB of RAM) with 40,000 tokens per mini-batch and 2 as update frequency. All models were obtained by averaging the last 7 checkpoints.

[17] https://github.com/pykaldi/pykaldi.

**Data** Our ASR model is trained on the following datasets: CommonVoice (Ardila et al., 2020), LibriSpeech (Panayotov et al., 2015), TEDLIUM v3 (Hernandez et al., 2018), and VoxPopuli (Wang et al., 2021). For ST, instead, as training corpora we used MuST-C (Cattoni et al., 2021), EuroParl-ST (Iranzo-Sánchez et al., 2020), and CoVoST v2 (Wang et al., 2020a). Additionally, we translated the transcripts of the ASR datasets into the two target languages with the NeMo MT models.[18]

Table 6 reports the results of our models and compares them with SeamlessM4T (Communication et al., 2023).

| | ASR WER (↓) | ST en→de BLEU (↑) | ST en→es BLEU (↑) |
|---|---|---|---|
| Ours | **10.8** | **30.7** | **34.9** |
| Seamless | 15.8 | 23.0 | 31.8 |

Table 6: ASR (WER) and ST (BLEU) results of our models compared to SeamlessM4T on the MuST-C tst-COMMON set.

## A.2 Validation of SPES Hyperparameters

In this section, we describe the SPES hyperparameter validation process (see §6). Due to the high number and wide range of possible hyperparameter values, a full systematic evaluation was computationally prohibitive. Thus, we began with a manual inspection of explanations on a subset of MuST-C dev set samples, narrowing down value ranges for each hyperparameter and setting the perturbation iterations $N^X$ and $N^Y$ at 20,000 and 2,000, respectively, to balance explanation quality and computational cost.

Next, we conducted a systematic validation to select the best configuration using the *deletion* metric (see §5). We validated spectrogram perturbation hyperparameters, except for the threshold length $\tau$, in A.2.1 and output token perturbation hyperparameters in A.2.2. Given the reduced, yet still numerous, hyperparameter combinations after initial filtering, this validation used a subsample of 130 utterances, randomly chosen between 5 and 10 seconds to represent typical lengths and avoid rare length biases. Finally, we validated $\tau$ on the full validation set for both ASR and ST tasks to

[18] https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/nlp/machine_translation/machine_translation.html.

examine behavior across diverse lengths.

### A.2.1 Perturbing Spectrograms

For the perturbation of spectrograms (see §4.1.1), we validate three parameters: *i)* the number of patches per second ($\phi$), which determines the granularity of the perturbation; *ii)* the sigma value ($\sigma$) in SLIC, which controls the smoothness of segmentation contours; *iii)* the perturbation probability ($p_{spec}$), which defines the probability of each patch being perturbed at each iteration. At this stage, we also validated the effectiveness of applying standard normalization to the token-level saliency maps before aggregating them into a sentence-level explanation. Table 7 presents the deletion scores for all these aspects.

**Number of Patches per Second ($\phi$)** As described in §4, our segmentation of the spectrogram adopts a multi-scale approach, using three levels of granularity achieved using three $\phi$ values. As our validation set is quite homogeneous in terms of duration (in the $[5, 10]$ seconds range) and the validation procedure is computationally expensive, we do not adjust the number of patches based on duration and optimize three values for the numbers of clusters *k*. Then, we derive $\phi$ by dividing the best *k* values by 5 seconds. This choice aims to maximize the performance of the approach with fixed number of patches in our ablation study (- *duration adaptation*, see §7.2), ensuring that our method is not advantaged by more favourable hyperparameters. Among the various configuration tested, the combination $k = [2000, 2500, 3000]$ consistently yields better results with the only exception of the results obtained with $p_{spec} = 0.7$. Accordingly, our experiments with SPES use 400, 500, and 600 patches per second as the three clustering scales.

| $k$ | $\sigma$ | $\mathbf{p_{spec}}$ | Deletion (↑) | |
|---|---|---|---|---|
| | | | w/o norm | w/ norm |
| 500 1000 2000 | 0 | 0.3 | 84.90 | 91.94 |
| | | 0.5 | 91.95 | 92.87 |
| | | 0.7 | 91.75 | 91.90 |
| | 1 | 0.3 | 85.26 | 91.94 |
| | | 0.5 | 92.29 | 93.32 |
| | | 0.7 | 91.92 | 91.88 |
| 1000 1500 2000 | 0 | 0.3 | 84.49 | 92.45 |
| | | 0.5 | 93.41 | 94.56 |
| | | 0.7 | 93.38 | 93.09 |
| | 1 | 0.3 | 84.82 | 92.65 |
| | | 0.5 | 93.71 | 94.71 |
| | | 0.7 | 92.93 | 92.62 |
| **2000 2500 3000** | **0** | 0.3 | 83.02 | 91.88 |
| | | **0.5** | 94.12 | **95.37** |
| | | 0.7 | 93.73 | 93.28 |
| | 1 | 0.3 | 82.45 | 91.93 |
| | | 0.5 | 93.78 | 95.05 |
| | | 0.7 | 93.62 | 93.18 |
| 3000 3500 4000 | 0 | 0.3 | 80.01 | 90.25 |
| | | 0.5 | 92.39 | 94.24 |
| | | 0.7 | 92.99 | 92.51 |
| | 1 | 0.3 | 78.55 | 88.70 |
| | | 0.5 | 91.35 | 93.94 |
| | | 0.7 | 92.37 | 91.75 |
| 1000 2500 5000 | 0 | 0.3 | 83.00 | 91.06 |
| | | 0.5 | 92.46 | 93.94 |
| | | 0.7 | 92.56 | 92.07 |
| | 1 | 0.3 | 82.87 | 91.06 |
| | | 0.5 | 92.29 | 93.87 |
| | | 0.7 | 91.61 | 91.17 |
| Avg. | | | 89.47 | **92.04** |

Table 7: Deletion scores to validate $k$, $\sigma$, $p_{spec}$.

**Sigma ($\sigma$)** The preliminary manual inspection revealed that increasing $\sigma$ above 1 results in patch shapes similar to grids. Therefore, we considered two values for $\sigma$: 0 and 1. The former, the default value in the SLIC implementation in scikit-image, provides sharper patch contours, while the latter smooths them. The best value for $\sigma$ varies across the different configurations, with no clear trend emerging. This indicates that both values are effective in terms of perturbation. However, for the selected number of patches ($k = [2000, 2500, 3000]$), $\sigma = 0$ yields better results. For this reason, we set $\sigma$ to 0 in our experiments.

**Perturbation Probability ($p_{spec}$)** The initial manual inspection revealed that high and low $p_{spec}$ values do not produce good explanations. For this reason, we tested three $p_{spec}$ values: 0.3, 0.5, and 0.7. In all configurations of the other hyperparameters, the intermediate value provided better results. This suggests that a $p_{spec}$ of 0.3 may be too low, as the model remains robust with only a few perturbed features, while a value of 0.7 may result in excessive information loss. In light of this, a $p_{spec}$ of 0.5 proves to be an effective compromise.

**Normalization** For each parameter combination, we computed the scores using token-wise standard normalization. As discussed in §4.2, normalization levels off the differences between the score distributions of different tokens, which depend on the original probability distributions. In most combinations (specifically 20 out of the 30

| $p_{spec}$ | $p_{tok}$ | Deletion (↑) | |
|---|---|---|---|
| | | Joint | Separate |
| 0.500 | 0.001 | 92.40 | 94.25 |
| | 0.005 | 91.29 | 94.24 |
| | 0.010 | 90.76 | 94.25 |
| | 0.050 | 87.80 | 94.25 |
| | 0.100 | 83.72 | 94.49 |
| | 0.200 | 79.77 | 94.48 |
| | 0.300 | 77.48 | 94.45 |
| | **0.400** | 74.36 | **94.50** |
| | 0.500 | 71.02 | 94.49 |
| | 0.600 | 69.79 | 94.48 |
| | 0.700 | 66.57 | 94.47 |
| | 0.800 | 63.31 | 94.42 |
| | 0.900 | 62.71 | 94.42 |
| Avg. | | 77.77 | **94.40** |

Table 8: Deletion scores to validate $p_{tok}$ in SPES.

considered), including that with the highest score ($k = [2000, 2500, 3000]$, $\sigma = 0$, $p_{spec} = 0.5$), normalization turns out to be beneficial, and, on average, the scores with normalization were 2.57 points higher than those without normalization. Therefore, when aggregating token-level explanations at the sentence level, we use token-level normalization by default.

In summary, the selected parameters for spectrogram perturbation are: $\phi = [400, 500, 600]$, $\sigma = 0$, and $p_{spec} = 0.5$. Our validation also proved that normalizing the token-level saliency maps leads to better sentence-level explanations.

### A.2.2 Perturbing Previous Output Tokens

After determining an effective combination of parameters for the spectrogram perturbation, we moved to validate the perturbation of the previous output tokens (see §4.1.2). Specifically, we analyzed two aspects: *i)* whether it is better to jointly perturb the speech input and previous output tokens or not, and *ii)* determining the optimal value of the perturbation probability $p_{tok}$.

**Joint vs Separate Perturbation** Previous output tokens can be perturbed in two ways: jointly with the spectrogram or separately. In the joint approach, both the spectrogram features and the previous output tokens are perturbed simultaneously, potentially revealing interdependencies between them for next-token predictions, and saliency scores are accordingly assigned to the perturbed elements of both sides. However, Table 8 consistently shows higher scores for separate perturbations, indicating that joint perturbation degrades performance. This suggests that the perturbation of previous output tokens has a huge impact on the model's predictions, causing an indirect "blurring" effect on the explanations of the spectrograms—even if the $p_{tok}$ is set to values lower than 0.1, up to two orders of magnitude less than the probability of the spectrogram (e.g., $p_{tok} = 0.001$, $p_{tok} = 0.005$). Consequently, in our experiments, we compute saliency scores separately for the speech input and the prior tokens in SPES.

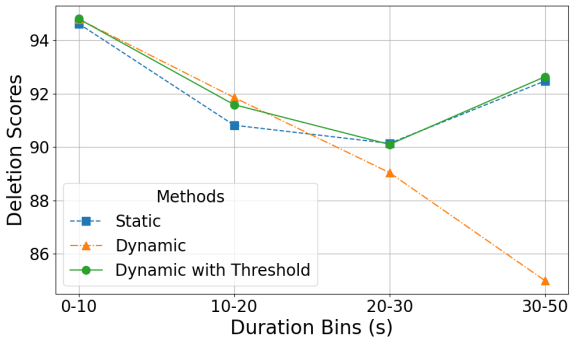



Figure 4: Variation of Deletion scores across samples of different lengths.

**Perturbation Probability ($p_{tok}$)** We considered $p_{tok}$ values in the range $[0.001, 0.900]$. In the *separate* setting, score differences were slight, following a similar trend to that observed for $p_{spec}$—specifically, higher values when approximately half of the tokens were removed (see Table 8). The best score was achieved with $p_{tok}$ set to 0.4.

In summary, in all our experiments the previous output tokens are perturbed separately from the spectrogram, with $p_{tok}$ equal to 0.4.

### A.2.3 Validation of the Dynamic Increase of Number of Patches and $\tau$

As discussed in §4.1.1, dynamically increasing the number of patches for long audio samples significantly raises the number of patch combinations, potentially reducing explanation quality. Fig. 4 shows deletion scores in ASR across input durations on the validation set, comparing a dynamic patch increase based on duration (*Dynamic*) versus a fixed number of patches (*Static*) with $k$ set to $[2{,}000, 2{,}500, 3{,}000]$. While *Dynamic* performs better for samples up to 20 seconds, its deletion score declines linearly with duration. In contrast, *Static* remains stable scoring 7.49 points higher than *Dynamic* for samples over 30 seconds.

| $\tau$ | Deletion | | |
|---|---|---|---|
| | **en (↑)** | **en→de (↓)** | **en→es (↓)** |
| 2.5 | 93.01 | 2.47 | 2.91 |
| **5.0** | 93.91 | **2.34** | **2.58** |
| **7.5** | **94.03** | 2.41 | 2.62 |
| 10.0 | 93.91 | 2.49 | 2.77 |

Table 9: Deletion scores to validate the duration threshold $\tau$ (in seconds) for stopping the linear increase in the number of patches.

| $\mathbf{p_{spec}}$ | **Deletion (↑)** |
|---|---|
| 0.5 | 74.79 |
| 0.6 | 76.63 |
| **0.7** | **79.29** |
| 0.8 | 78.49 |
| 0.9 | 58.67 |

Table 10: Deletion scores to validate $p_{spec}$ in the *feature-wise* configuration.

| $\mathbf{p_{tok}}$ | **Deletion (↑)** |
|---|---|
| **0.1** | **77.38** |
| 0.2 | 77.31 |
| 0.3 | 77.29 |
| 0.4 | 77.21 |
| 0.5 | 77.17 |
| 0.6 | 77.18 |
| 0.7 | 77.21 |
| 0.8 | 77.10 |
| 0.9 | 77.06 |

Table 11: Deletion scores to validate $p_{tok}$ in the *feature-wise* configuration.

To mitigate this issue, we introduced the threshold hyperparameter $\tau$ (see §4.1.1), which limits dynamic increases in the number of patches beyond certain audio lengths. Table 9 presents the validation results for $\tau$ on the full ASR and ST validation sets. For ASR, the optimal threshold is set at 7.5 seconds, while for ST, a lower threshold of 5.0 seconds is established. In Fig. 4, this *Dynamic with Threshold* approach effectively avoids the decline in faithfulness observed with *Dynamic*, consistently outperforming *Static* across all sample lengths, including those exceeding 30 seconds.

In summary, ASR uses $k = [3,000, 3,750, 4,500]$ for audio longer than 7.5 seconds, while ST uses $k = [2,000, 2,500, 3,000]$ for audio over 5.0 seconds.

### A.3 Validation of Baselines Hyperparameters

To ensure a fair comparison with SPES, we conducted a validation procedure to determine the optimal hyperparameters for the *feature-wise* and *Bubble Noise* baselines discussed in §6.

#### A.3.1 Feature-wise Perturbation

Since no segmentation of the spectrogram is involved in *feature-wise*, the only parameters to be validated are $p_{spec}$ and $p_{tok}$. The preliminary manual inspection revealed that low values for $p_{spec}$ were not effective in providing high deletion scores. Since the perturbation involves only sparse single features rather than groups of contiguous features (i.e., the patches), it is reasonable that the model is robust in recovering information even when a large percentage of features is perturbed. Therefore, we focused on values from 0.5 onward, whose deletion scores are shown in Table 10. The best result was achieved with 0.7.

Table 11 shows the deletion scores for $p_{tok}$, with the perturbation of the previous output tokens carried out separately, as in SPES. Similar to $p_{tok}$ in SPES, the differences among the scores are low, with lower values generally providing higher scores. However, in this case, $p_{tok}$ set to 0.1 yielded the highest result.

In summary, for the experiments with the feature-wise baseline reported in §7, we set $p_{spec}$ to 0.7 and $p_{tok}$ to 0.1.

#### A.3.2 Bubble Noise Perturbation

We implement the Bubble Noise perturbation method (Mandel, 2013; Trinh and Mandel, 2020, 2021), enhancing it to also produce explanations for previous output tokens (see §6). To perturb the spectrogram, we generate random noise that matches the intensity range of the speech input.[19] This noise replaces the spectrogram everywhere except within randomly placed ellipsoidal bubbles, where it is attenuated and then added back to the spectrogram.

Regarding bubble size and count, previous works use various numbers of bubbles per second, ranging from 18 to 80, all with a fixed size of 0.35 seconds and a height of approximately 180 Hz. We explore different configurations within a range that encompasses the values tested in earlier studies. Specifically, we aim to cover about

[19] We also tested white noise at various levels of SNR relative to the input spectrogram. However, noise matching the intensity range of the input spectrogram performed better.

| Bub/s | Width (s) | Height (mel) | Deletion (↑) |
|---|---|---|---|
| **10** | ∼0.43 | ∼31 | **69.40** |
| 35 | ∼0.25 | ∼15 | 59.07 |
| 115 | ∼0.13 | ∼9 | 56.97 |

Table 12: Deletion scores to validate bubble size and count.

half of the spectrogram with noise, aligning with SPES's perturbation level ($p_{spec} = 0.5$). Table 12 presents deletion scores for three levels of bubble granularity tested with $N^X$ = 1,000. The configuration with 10 bubbles per second yielded the best results, achieving deletion scores that were 10.33 and 12.43 points higher than the configurations with 35 and 115 bubbles per second, respectively.

In summary, Bubble Noise perturbation was set to 10 bubbles per second, each covering ∼0.43 seconds in width and ∼31 mel bins in height.

### A.4 Examples of Explanations

Fig. 5 shows $S^X$ and $S^{Y^{(k-1)}}$ for the transcription and the translation into German of the utterance "So what was the next step?". These visualizations highlight the following points, also detailed in §8.

1. Comparing ASR and ST $S^X$ maps reveals similar patterns; for example, the German tokens `war`, `nächste`, and `Schritt` align with the English counterparts `was`, `next`, and `step`, indicating a reliance on similar phonetic sequences.

2. In ASR $S^X$, salient areas move forward with each token, showing a monotonic progression in time (see §8.1). This effect is less pronounced in ST due to differences in word order, such as German `Was` and `also` (en. *what* and *so*) reversing the English order.

3. Most $S^X$ maps are frequency-localized. For instance, `So` (ASR) shows high saliency in the high-frequency range typical of alveolar sibilants and mid-frequency ranges for vowel formants (see §8.1 and Fig. 10).

4. Some $S^X$ maps lack clear salient areas, such as those for `?` in both ASR and ST (see §8.2).

5. In $S^{Y^{(k-1)}}$, the first and last tokens generally exhibit high saliency, with some intermediate tokens showing similar effects (see §8.3). For example, in ASR, `?` is influenced by `what`, and in ST, `der` (en. *the*) by `war` (en. *was*).

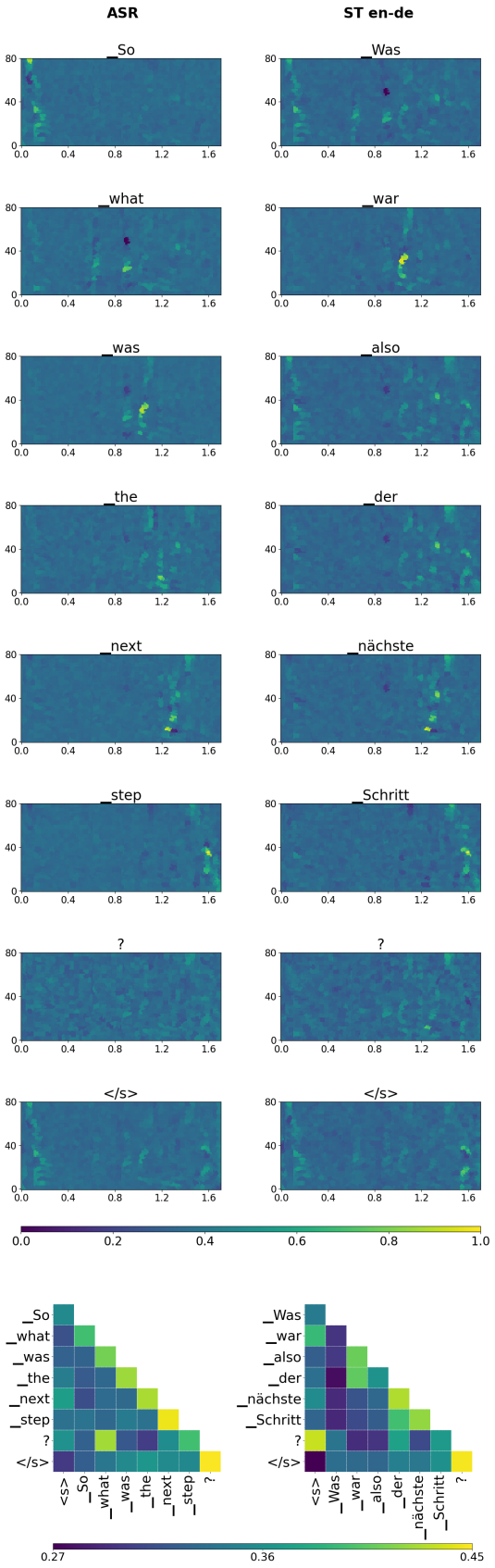



Figure 5: $S^X$ (top) and $S^{Y^{(k-1)}}$ (bottom) for ASR and ST on the utterance "So what was the next step?". In $S^{Y^{(k-1)}}$, rows show the tokens being explained, and columns the previous output tokens.

## B Complementary Results

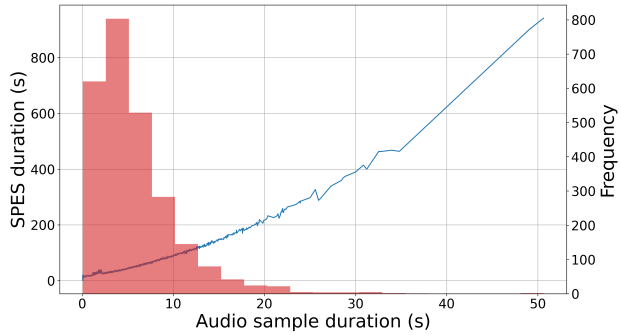
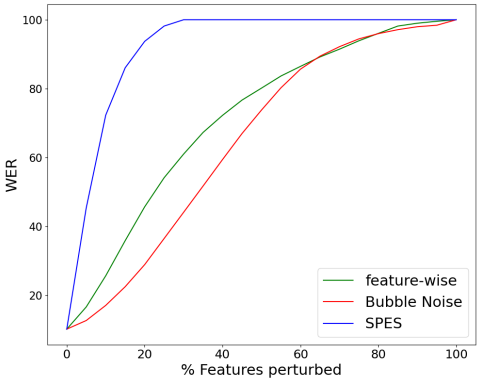


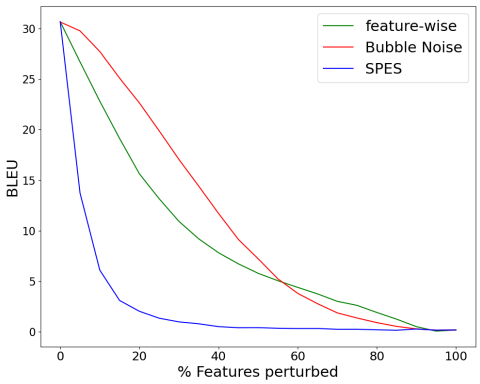

Figure 6: SPES explanation generation time across varying sample lengths and distribution of samples by length.

| | ASR (en) | |
|---|---|---|
| | **Deletion** (↑) | **Size** (↓) |
| Bubble Noise | 65.02 | 36.92 |
| SPES fine-grained | 75.22 | 39.54 |
| SPES coarse-grained | **79.71** | **26.91** |
| | **ST (en→de)** | |
| | **Deletion** (↓) | **Size** (↓) |
| Bubble Noise | 10.85 | 43.89 |
| SPES fine-grained | 6.84 | 40.06 |
| SPES coarse-grained | **5.14** | **28.58** |
| | **ST (en→es)** | |
| | **Deletion** (↓) | **Size** (↓) |
| Bubble Noise | 13.01 | 43.04 |
| SPES fine-grained | 7.41 | 39.66 |
| SPES coarse-grained | **6.10** | **27.17** |

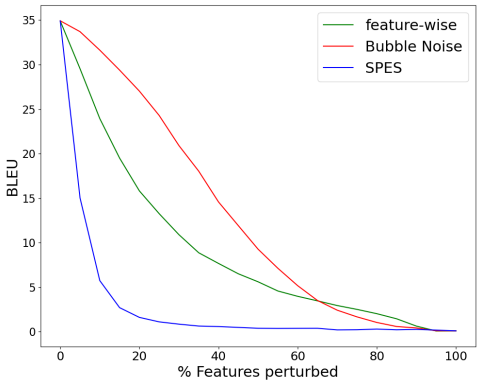
Table 13: Results comparing Bubble Noise and SPES with $N^X = 1,000$. SPES is evaluated in both a *fine-grained* setting ($\phi$ validated in Appendix A.2.1) and a *coarse-grained* setting ($\phi = [20, 40, 60]$), maintaining approximately 10 unperturbed patches per second—similar to Bubble Noise, which attenuates noise in 10 bubbles per second.




(a) ASR (en)

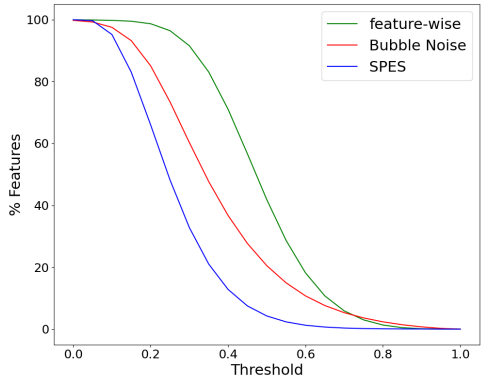



(b) ST (en→de)

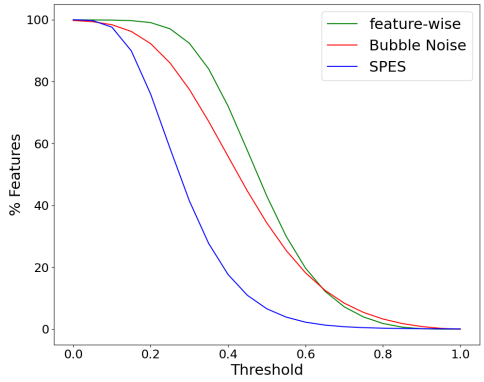



(c) ST (en→es)

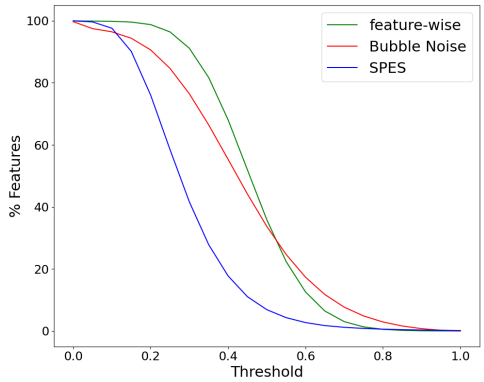
Figure 7: Deletion curves for the *feature-wise* strategy, the *Bubble Noise* technique, and SPES in the ASR and ST tasks.




(a) ASR (en)




(b) ST (en→de)




(c) ST (en→es)

Figure 8: Size curves for the *feature-wise* strategy, the *Bubble Noise* technique, and SPES in the ASR and ST tasks.

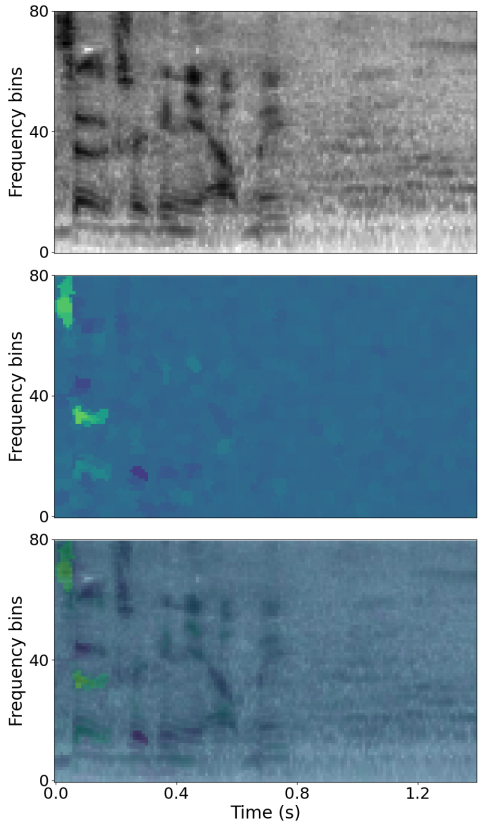



Figure 10: Example of saliency map (middle) for the token `so` (ASR), along with the corresponding spectrogram (top) and the map overlaid on the spectrogram (bottom).

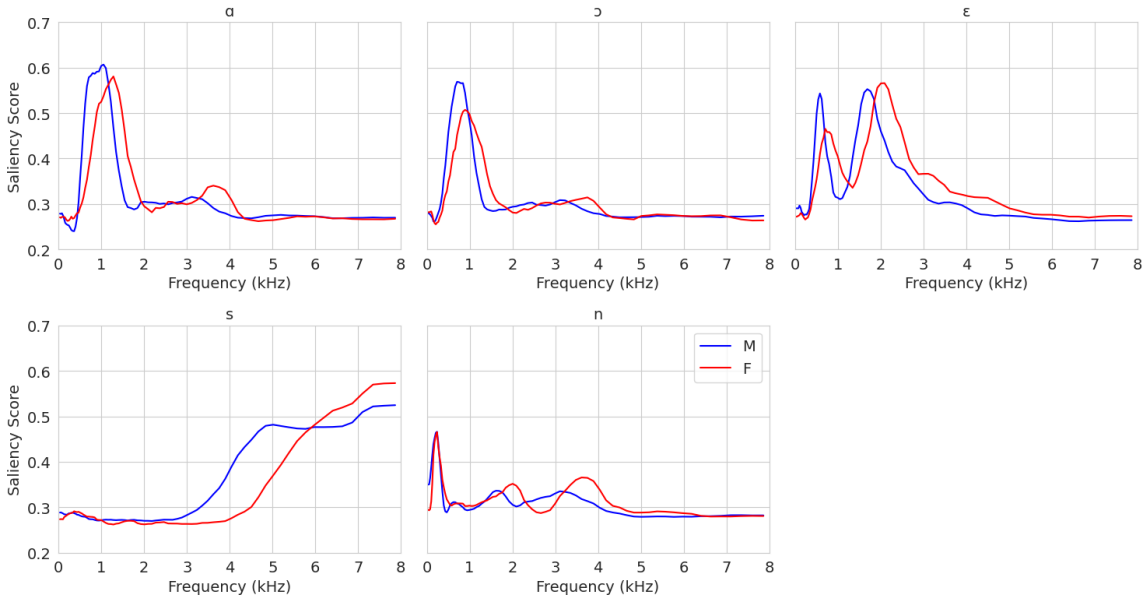



Figure 9: Average saliency scores over frequency for the phones... uttered by men and women.

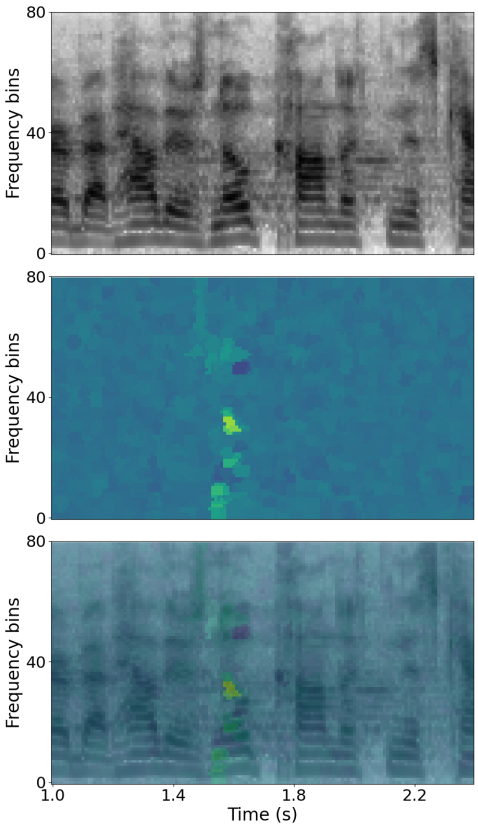



Figure 11: Example of saliency map (middle) for the token `no` (ASR), along with the corresponding spectrogram (top) and the map overlaid on the spectrogram (bottom).

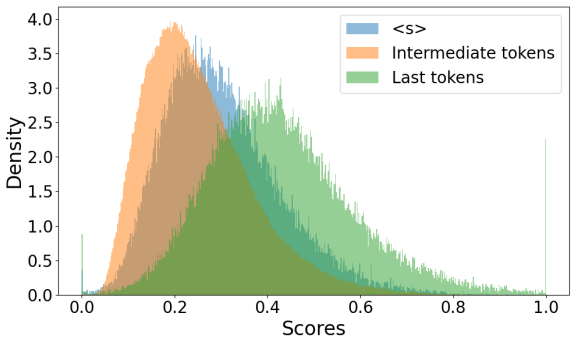



Figure 12: Distributions of saliency scores obtained in ST (en→de) for three types of tokens: `<s>`, *intermediate tokens* and tokens preceding the token to be explained (*last tokens*).