

How does a deep neural network look at lexical stress in English words?

Itai Allouche,¹ Itay Asael,¹ Rotem Rousso,¹ Vered Dassa,¹ Ann Bradlow,²  Seung-Eun Kim,² 
 Matthew Goldrick,²  and Joseph Keshet^{1,a)} 

¹Faculty of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel

²Department of Linguistics, Northwestern University, Evanston, Illinois 60208, USA

ABSTRACT:

Despite their success in speech processing, neural networks often operate as black boxes, prompting the following questions: What informs their decisions, and how can we interpret them? This work examines this issue in the context of lexical stress. A dataset of English disyllabic words was automatically constructed from read and spontaneous speech. Several convolutional neural network (CNN) architectures were trained to predict stress position from a spectrographic representation of disyllabic words lacking minimal stress pairs (e.g., initial stress *WAllet*, final stress *exTEND*), achieving up to 92% accuracy on held-out test data. Layerwise relevance propagation, a technique for neural network interpretability analysis, revealed that predictions for held-out minimal pairs (*PROtest* vs *proTEST*) were most strongly influenced by information in stressed versus unstressed syllables, particularly the spectral properties of stressed vowels. However, the classifiers also attended to information throughout the word. A feature-specific relevance analysis is proposed, and its results suggest that the best-performing classifier is strongly influenced by the stressed vowel's first and second formants, with some evidence that its pitch and third formant also contribute. These results reveal deep learning's ability to acquire distributed cues to stress from naturally occurring data, extending traditional phonetic work based around highly controlled stimuli.

© 2026 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/10.0042429>

(Received 8 July 2025; revised 15 January 2026; accepted 24 January 2026; published online 11 February 2026)

[Editor: Benjamin V Tucker]

Pages: 1348–1358

I. INTRODUCTION

While deep learning-based architectures have transformed many aspects of speech processing, those models often exhibit a nontransparent prediction process, obscuring how specific decisions are made. This work explores the application of one technique for shedding light on such architectures through the lens of a well-studied phenomenon in speech science and speech processing: lexical stress. Lexical stress refers to the prominence, accentuation, or emphasis of a particular syllable within a word. In many languages, lexical stress is crucial for communication as it can serve as the primary basis for a contrast in meaning. For example, in English, the words “*PROtest*” (noun; stress on initial syllable) and “*proTEST*” (verb; stress on final) have contrasting stress patterns but the same phonemes (with predictable changes to the quality of the initial vowel depending on stress).

A long tradition of phonetics research has examined the (relative) temporal, spectral, and energetic properties of stressed and unstressed syllables [Fry (1955, 1958); for reviews, see Cutler (2005), Gay (1978), Gordon and Roettger (2017), and Van Heuven (2018)]. Building on this work, automatic stress detection methods have used acoustic features such as fundamental frequency, amplitude, and energy

(Tepperman and Narayanan, 2005; Lieberman, 1960). Bayesian classifiers with multivariate Gaussian distributions have been applied to enhance stress detection accuracy, utilizing acoustic features such as peak-to-peak amplitude, normalized energy, and vowel duration (Waibel, 1986; Ying *et al.*, 1996). Similar to other areas of speech processing, recent work using deep learning architectures has significantly improved over these more traditional approaches. Convolutional neural networks (CNNs) and multi-distribution deep neural networks (MD-DNNs) outperformed traditional models in both English and Arabic stress detection in second language speech (Li *et al.*, 2018; Shahin *et al.*, 2016). An attention-based neural network combined with data augmentation through a neural network-based text-to-speech system has been introduced to detect lexical stress errors in second-language English learners, achieving significant improvements in precision and recalling Korzekwa *et al.* (2020). Classifiers applied to self-supervised learning representations from wav2vec 2.0 rather than traditional acoustic features also show high performance (Bentum *et al.*, 2024).

Although these advancements in deep learning have led to significant improvements in speech processing, they also introduce challenges, particularly in understanding classifier decisions and gaining insight into how these models operate. This issue of interpretability is crucial. Better understanding the acoustic features that drive model predictions can help

^{a)}Email: jkeshet@technion.ac.il

us verify that the model will be reliable when applied to novel data (which may have different distributions for critical features). Neural network interpretability has broadly become an active research area, with a number of techniques being explored across different domains. Here, we focus on layerwise relevance propagation [LRP; Bach *et al.* (2015)], a technique that was developed to provide insights into the decision-making processes of these complex models by understanding what specific portions of the input influence the model's output. We use this technique to investigate how CNNs—which have demonstrated superior performance in lexical stress detection—derive their predictions from the spectral features in their input.

This paper offers several novel contributions. First, a new speech dataset was automatically gathered without human annotation, using a pipeline built around deep learning systems. Second, we present a state-of-the-art model for lexical stress classification in disyllabic words and show its ability to generalize to novel disyllabic words. Last, we employ LRP to analyze the spectral features these classifiers use to achieve high accuracy.

This paper is organized as follows. In Sec. II, we describe how the dataset was generated. In Sec. III, we present how we built the classifier and how it was trained. In Sec. IV, we present the LRP technique, its application in our study, and our techniques for analyzing LRP results. In Sec. V, we empirically evaluate the classifiers. The LRP analyses are presented in Sec. VI, and the paper concludes with a discussion in Sec. VII.

II. DATASET CONSTRUCTION

Previous studies on lexical stress often depended on stress labels assigned by human annotators (Tepperman and Narayanan, 2005; Korzekwa *et al.*, 2020; Shahin *et al.*, 2016) or were assigned based on the canonical stress for the orthographic word taken from pronunciation dictionaries, such as the CMU Pronouncing Dictionary (Li *et al.*, 2018). However, manual annotation is costly and can introduce inter-annotator inconsistency, and the use of a single citation form can yield errors for orthographically ambiguous minimal pair words (e.g., *protest*). Prior work also failed to explicitly document measures to prevent lexical overlap between training and evaluation splits, leaving open the possibility that identical word types were seen during both stages (Tepperman and Narayanan, 2005; Li *et al.*, 2018; Korzekwa *et al.*, 2020; Shahin *et al.*, 2016). Reported accuracy may therefore overestimate true generalization. To avoid these concerns, we created a novel dataset, using a fully automatic pipeline, without utilizing human annotations (MLSpeech, 2025).

A. Selection of English disyllabic words

We prompted ChatGPT 4o (Hurst *et al.*, 2024) to generate 30 pairs of English disyllabic words that form stress minimal pairs: pairs of words with the same spelling but different meaning, depending on the primary stress

placement. Five of these items could not be incorporated into other parts of our pipeline. (The part-of-speech tagger, discussed below, failed to correctly annotate these forms.) The remaining 25 pairs are *address, conduct, content, contrast, convert, decrease, digest, export, extract, import, increase, insult, object, perfect, present, progress, project, protest, rebel, record, refuse, reject, subject, suspect, and transfer*; the initial and final syllable stress variants of these words will be referred to as “minimal pairs.” We also asked ChatGPT 4o to generate a larger set of 250 English disyllabic words that do not have stress minimal pairs (e.g., initial stress *WALlet*, final stress *exTEND*). After excluding one error, we retained 249 words (124 initial stress, 125 final stress). We refer to this as the set of “no minimal pair words.”

B. Extraction of words from existing corpora

These words were then extracted from three existing English speech corpora. The first dataset is LibriSpeech (Panayotov *et al.*, 2015), a widely used corpus consisting of ~1000 h of read speech derived from audiobooks. The second dataset is The Supreme Court corpus (Spaeth *et al.*, 2014), which includes recordings of oral arguments made before the United States Supreme Court. This dataset provides speech data from a highly formal and legal context, offering variation in speaker formality and legal terminology that can influence acoustics associated with lexical stress. The third dataset is TED-LIUM (Hernandez *et al.*, 2018), derived from TED Talks. It includes over 400 h of speech by speakers from various language backgrounds and nationalities, discussing a wide range of topics. The diversity in speaking styles and contexts in TED-LIUM further enriches the phonetic variation in our dataset.

The processing pipeline is shown in Fig. 1. First, we used a forced alignment algorithm to align the transcript with the audio. Specifically we used Montreal Forced Aligner [MFA; McAuliffe *et al.* (2017)] to align LibriSpeech and Supreme Court. The MFA works at time-resolution of 10 ms and was shown by Rouso *et al.* (2024) to outperform recent end-to-end automatic speech recognition (ASR)-based aligners such as WhisperX (Bain *et al.*, 2023) and MMS (Pratap *et al.*, 2024). For TED-LIUM, which already includes word-level timestamps, we used the provided alignments. Importantly, unlike MFA-aligned corpora (LibriSpeech and Supreme Court), where word boundaries are adjacent (e.g., “hello 0–0.3 second,” “my 0.3–0.4”), the TED-LIUM timestamps may be non-adjacent (e.g., “hello 0–0.3,” “my 0.35–0.4”), leaving short pauses between words.

Once the start time of each word is obtained, a 0.5-s window is opened. This window size was chosen to ensure that both syllables of every disyllabic word were fully included and to maintain a uniform input length compatible with our CNN architecture. We retained only those whose total duration was less than 0.5 s and zero-padded them as needed to reach exactly 0.5 s. Each minimal pair word was further segmented into syllables using a syllabification

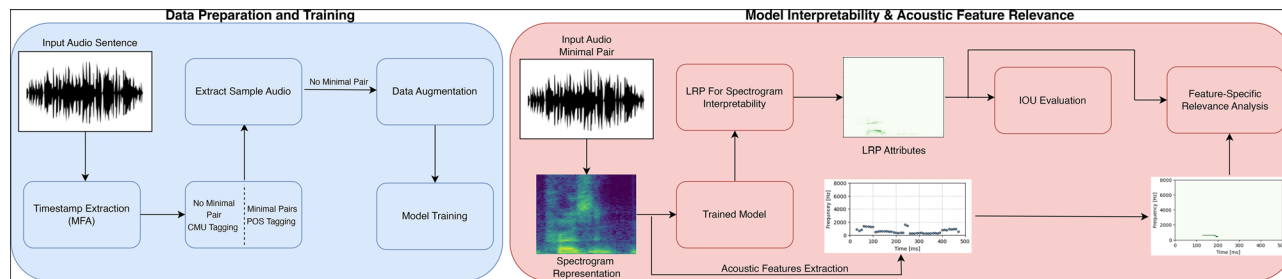


FIG. 1. End-to-end workflow for dataset construction, model training, and interpretability analysis. The blue-colored section illustrates the dataset creation and training process. Audio is collected from various datasets and extracted based on word- and phoneme-level timestamps. Words are tagged for stress using either part-of-speech tagging (minimal pairs) or dictionary entries (no minimal pair words). Once the audio is extracted, the training data (no minimal pair words) are expanded through data augmentation techniques to make CNN-based model training more robust. The red-colored section illustrates the processes of model interpretability and acoustic feature relevance measurement, applied to the test set of minimal pairs. Audio is converted into a spectrogram. This is processed by the trained model; LRP methods are applied to generate heatmaps, revealing the regions of the spectrograms that contributed most to the model’s predictions. The contribution of different (sub)syllabic regions to the LRP heatmap is quantified with the IOU metric. For the stressed vowel, acoustic features are extracted and used to generate feature-specific heatmaps; these are compared to the full heatmap to determine which acoustic features are most dominant in the model’s decision-making process.

algorithm (P2TK Developers, 2014). In addition, phoneme-level timestamps derived from the forced aligner were incorporated.

After identifying the location of target minimal pair words within the audio, we applied part-of-speech tagging system [SpaCy; Annex et al. (2020)] to the entire audio in order to determine whether each word was used as a noun or verb. The tagging follows linguistic conventions for English disyllabic words, where nouns typically carry primary stress on the initial syllable, referred to as initial stress (IS), and verbs carry primary stress on the final syllable, referred to as final stress (FS) (Lieberman, 1960). For the no minimal pair words, we used annotations from the CMU Pronouncing Dictionary (Li et al., 2018), as the stress placement is unambiguous in this case. During this process, we observed that certain words, specifically *address*, *content*, *export*, *extract*, and *decrease*, were particularly affected by inaccuracies in the part-of-speech tagging system. Many of these errors involved incorrectly labeling words with initial stress instead of final stress. To correct these inconsistencies, we manually reviewed and relabeled all instances of these words.

Overall, we had 7446 samples (124 types) and 3263 samples (125 types) for IS and FS, respectively, in the no minimal pairs set and 5475 samples and 1715 samples for the IS and FS, respectively, in the minimal pairs set, consisting of a total of 25 pairs.

C. Phonetic properties of extracted samples

To provide an initial verification that our procedures yielded sets of disyllabic words that contrasted in stress location, two well-known correlates of lexical stress were examined: amplitude (vowels in stressed syllables are louder than those in unstressed syllables) and duration (vowels in stressed syllables are longer than those in unstressed syllables) (Fry, 1955, 1958). For each sample, we calculated the ratio of the properties of the vowels of the initial versus final syllables. These relative measures allow us to control for potential differences in speech rate and amplitude across

samples. Our samples reliably showed the expected differences. Samples with initial stress had high ratios of amplitude [mean 0.6, standard deviation (s.d.) 0.18] and duration (mean 0.53, s.d. 0.15), reflecting relatively loud, long vowels in stressed initial versus unstressed final syllables. In contrast, samples with FS showed lower mean ratios for amplitude (mean 0.29, s.d. 0.14) and duration (mean 0.35, s.d. 0.12), reflecting relatively quiet, short vowels in unstressed initial versus stressed final syllables. To confirm that this is a reliable difference between sample types, we used bootstrap resampling (with 1000 replicates) to estimate the 95% confidence interval (CI) for differences in mean ratios between initial versus final stress samples. These differences were significant [amplitude mean difference, 0.305, 95% CI (0.302, 0.309), $p < 0.0001$; duration mean difference, 0.178, 95% CI (0.175, 0.18), $p < 0.0001$]. This provides independent confirmation that the samples in this dataset have contrasting stress patterns.

III. LEXICAL STRESS DETECTION

A. Model architectures

In our study, we investigated different CNN architectures to develop a binary classifier for identifying the IS or FS of stress-minimal word pairs. The input to all models consists of magnitude spectrograms obtained using the short-time Fourier transform (STFT), following z -score normalization (see Figs. 2 and 3 for examples). The STFT was computed with a sampling rate of 16 000 Hz, using a Hamming window of size 0.02 s and a window stride of 0.01 s. We experimented with several well-known CNN architectures. The LeNet-5 model consists of three connected convolutional layers with average pooling between the layers, followed by two dense layers in a 2×2 configuration (Lecun et al., 1998). The VGG11, VGG16, and VGG19 architectures consist of 8, 13, and 16 convolutional layers, respectively, with maximum pooling between the layers, followed by an additional 3 dense layers (Simonyan, 2014). Finally, the ResNet18 model consists of 18 convolutional layers with

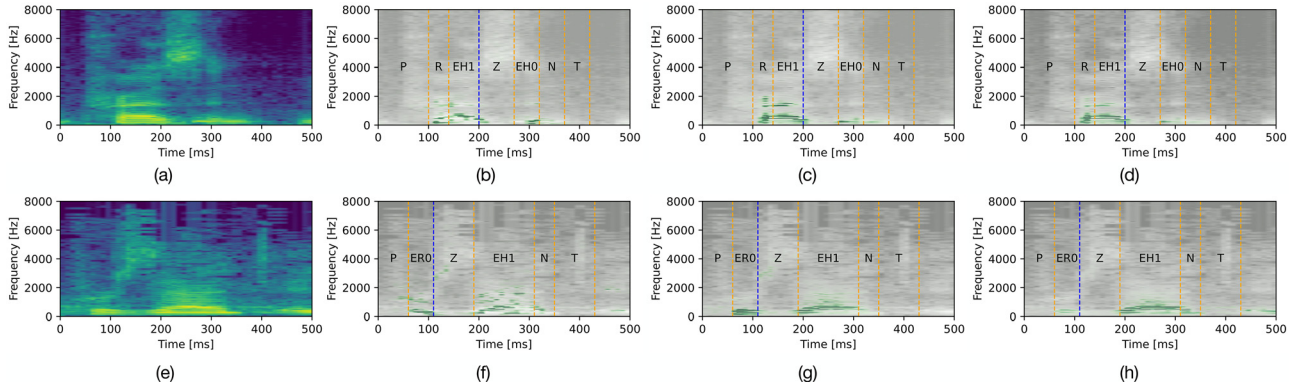


FIG. 2. Spectrograms and corresponding attributions for the words “*PRE*sent” and “*pre*SENT” using the VGG16 model. Orange vertical lines represent phoneme boundaries, with each label centered within its phoneme segment; 1 and 0/2 denote stressed and unstressed vowels, respectively. The blue vertical line represents the end of the initial syllable. Panel (a) shows the spectrogram of “*PRE*sent” (IS), while panel (e) shows the spectrogram of “*pre*SENT” (FS). Panels (b)–(d) and (f)–(h) display attributions using various methods— LRP_e , LRP_{21} , and LRP_{CMP} , respectively—shown over the spectrograms. These LRP variants effectively capture initial versus final stress contrasts, highlighting features at different scales. The LRP_{21} [panel (c) vs (g)] and LRP_{CMP} methods [panel (d) vs (h)] show clearly different patterns depending on the location of stress. Note that the phoneme labels differ slightly across the two rows; these reflect natural differences in pronunciation between initial and final stress members of this word pair (i.e., stress-related changes to vowel quality).

residual connections to address the vanishing gradient problem during training (He *et al.*, 2016).

B. Data splitting

The data were partitioned so that no word type appears in more than one split, thereby eliminating lexical overlap between training, validation, and test sets.

The training set was constructed using 201 out of the 249 word types from the no minimal pair words. Data augmentation techniques were employed to enlarge this set of examples. For each sample, we added a low-pass-filtered version of the sample (cutoff frequency at 3000 Hz) as well

as three versions of the sample mixed with noise. A sample of multi-talker babble from the VOICES corpus (Richey *et al.*, 2018) was combined with the original sample at three different levels [20, 10, and 3 dB signal-to-noise ratios (SNRs)]. These augmentations aimed to simulate real-world conditions, enhancing the robustness of our models. After the augmentation, we had 29 830 and 13 510 samples from the initial and final stress, respectively, corresponding to the 201 no minimal pair word types. The augmented dataset will be referred to as the “no minimal pairs train.”

The remaining 48 no minimal pair word types were divided into two lists of 24 words each, with one list used for the validation (i.e., hyper-parameter tuning; “no minimal

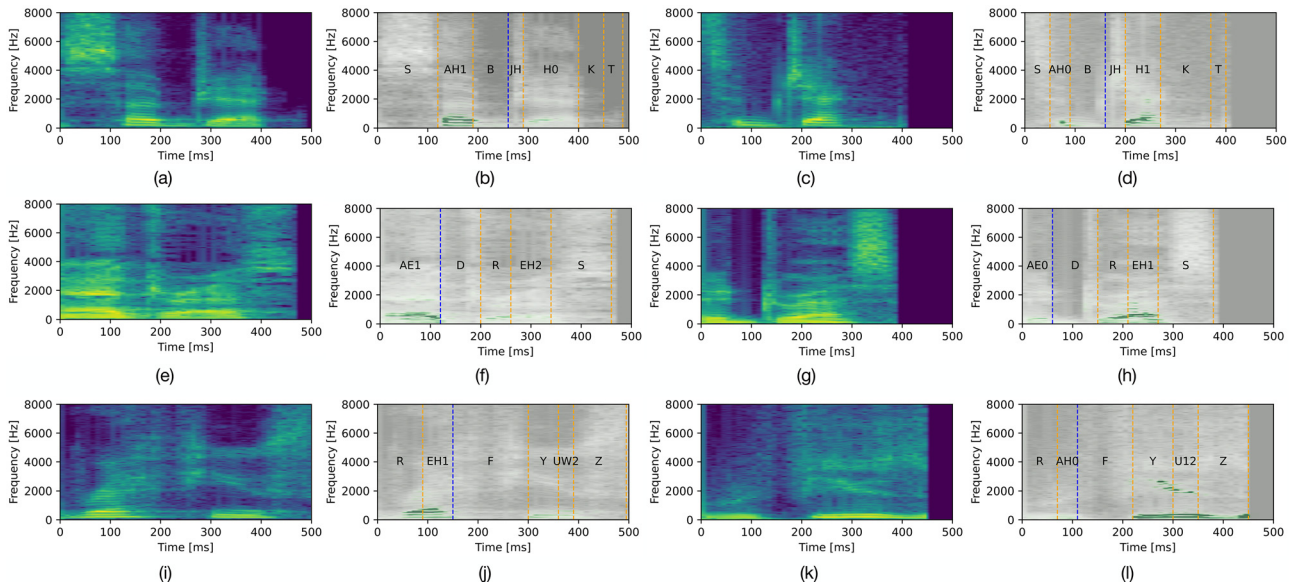


FIG. 3. Further illustrative examples of spectrograms and corresponding attributions for the words “*subject*,” “*address*,” and “*refuse*” with different primary stress location, using the VGG16 model. Vertical lines and annotation follow Fig. 2. Panels (a)–(d) show spectrograms and corresponding LRP_{CMP} heatmaps for the word “*subject*” with initial and final stress, respectively. Similarly, panels (e)–(l) display spectrograms and heatmaps of the words “*address*” and “*refuse*,” with different stress types. Note that the examples were drawn from datasets in which word segments were zero-padded to reach the fixed 0.5-s window length, as described in Sec. II B.

TABLE I. Composition of the train before and after data augmentations (w. aug.), validation, and test sets. Each set consists of disjoint word types (no overlap across splits). The table reports the number of unique word types and the total number of samples corresponding to initial stress (IS) and final stress (FS) words.

Set	No. of		
	Word types	IS samples (w. aug.)	FS samples (w. aug.)
No minimal pairs train	201	5966 (29 830)	2702 (13 510)
No minimal pairs validation	24	763	260
No minimal pairs test	24	717	301
Minimal pairs test	25	5475	1715

pairs validation”) and the other for the test set (“no minimal pairs test”). The 25 minimal pairs words will be referred to as “minimal pairs test.” Note that two test sets were used (“no minimal pairs test” and “minimal pairs test”) as the minimal pairs set may contain errors; stress was determined solely by part-of-speech tagging, which may be inaccurate. In contrast, the tagging for the no minimal pair words is more reliable; as the word form is unambiguous, the lexically determined stress pattern can be accurately identified. See Table I for dataset composition.

C. Training process

The models were trained using standard procedures for CNNs. To stabilize training, we employed an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5×10^{-4} and batch normalization. Training was performed for up to 50 epochs, with early stopping triggered after 15 epochs without validation improvement. All experiments were conducted on eight Nvidia A40 graphics processing units (GPUs) (Nvidia, Santa Clara, CA).

As the training data contain substantially more initial than final stress examples, a potential concern is an initial stress bias, whereby a classifier could achieve relatively high accuracy by disproportionately predicting the majority class, rather than relying on acoustically relevant cues to lexical stress. Such a bias could obscure the extent to which the model genuinely learns stress-related acoustic patterns. To mitigate this possibility, we employed the *focal loss* (FL) function of Lin *et al.* (2017), modifying the standard cross-entropy loss to focus more on hard-to-classify examples and to reduce the influence of the majority class during training. FL introduces a modulating factor, $(1 - p_t)^\gamma$, to the cross-entropy objective, reducing the relative loss contribution of well-classified examples and amplifying that of misclassified ones. Formally, it is defined as

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t), \quad (1)$$

where p_t denotes the predicted probability for the true class and $\gamma \geq 0$ is the focusing parameter that controls the rate at which easy examples are down-weighted. In our experiments, we set $\gamma = 2$ based on validation performance, which

provided a good balance between stability and discrimination.

IV. INTERPRETABILITY ANALYSIS METHODS

A. LRP

LRP has been extensively used in computer vision to provide insights into the contribution of individual input features to model predictions Kohlbrenner *et al.* (2020). Starting at the output, LRP attributes the classification prediction to units in the preceding layer: How *relevant* is each unit to the classifier producing one output versus another? This process is then repeated, back-propagating the relevance score of each unit to the classification prediction until the input layer is reached. This technique allows us to attribute relevance scores to each feature (e.g., how much does pixel X contribute to classifying an image as *dog* versus *cat*?).

In our study, we apply this technique to our lexical stress classifiers, back-propagating prediction to time-frequency bins in the input using the Captum interpretability library (Kokhlikyan *et al.*, 2020). This method identifies the regions of the spectrogram that have the most significant influence on the model’s classification of lexical stress. We specifically examine our best-performing VGG-based classifier, VGG16, and ResNet18. (For model performance comparisons see Sec. V and Table II.)

Denote by \mathbf{x}^l the input to the l -th layer. Denote by \mathbf{W}^l the weights associated with layer l . Denote by \mathbf{Z}^l is the elementwise preactivation of layer l and it is defined as $Z_{ij}^l = W_{ij}^l \odot x_i^l$. The relevance score of layer l is computed recursively from a higher layer, $l + 1$. In Kohlbrenner *et al.* (2020), the relevance score \mathbf{LRP}_Z is defined as

$$R_i^l = \sum_j \frac{Z_{ij}^l}{Z_j^l} R_j^{l+1}, \quad (2)$$

where $Z_j^l = \sum_i Z_{ij}^l$. The \mathbf{LRP}_Z method effectively traces the influence of each element of the input through the network,

TABLE II. Classification accuracy for each model architecture on each stress type and test set. (Chance-level performance is 50%.) The highest values for each test set are highlighted in boldface.

Test set	Classifier	Accuracy (%)		
		IS	FS	Overall
Minimal pair	LeNet	86.3	82.5	86
	VGG11	87.2	83.5	87
	VGG16	88.2	83.6	88
	VGG19	88.3	82.3	88
	ResNet18	87.74	84	87
No minimal pair	LeNet	91	85.6	90
	VGG11	91.8	88.3	91
	VGG16	92.6	90.6	92
	VGG19	92.2	91.3	92
	ResNet18	90.3	88.7	90

providing insight into how these impact the model’s predictions.

In addition to this method, we experimented with several types of LRP to determine the most effective method for identifying the regions of the spectrogram that are most influential for classifying words as stressed versus unstressed. The \mathbf{LRP}_ϵ rule extends the basic LRP method by addressing challenges such as division by zero and the influence of weak or noisy preactivations. It is defined as

$$R_i^l = \sum_j \frac{Z_{ij}^l}{Z_j^l + \epsilon \text{sign}(Z_j^l)} R_j^{l+1}. \tag{3}$$

In this rule, a small constant ϵ is added to the denominator along with the sign of Z_j^l . This modification prevents division by zero and mitigates the impact of small or noisy preactivations Z_{ij}^l on the relevance distribution. By incorporating ϵ , the \mathbf{LRP}_ϵ rule enhances the stability of relevance propagation, leading to more reliable and interpretable results. This adjustment is particularly beneficial when dealing with minimal or problematic contributions, as outlined by Bach *et al.* (2015).

The $\mathbf{LRP}_{\alpha\beta}$ rule, introduced by Bach *et al.* (2015), is an extension of the basic LRP method that differentiates the contributions of facilitatory and inhibitory activation flow. It is defined as

$$R_i^l = \sum_j \left(\frac{\alpha Z_{ij}^+}{Z_j^+} + \frac{\beta Z_{ij}^-}{Z_j^-} \right) R_j^{l+1}, \tag{4}$$

where $Z_{ij}^+ = \max\{Z_{ij}^l, 0\}$ and $Z_{ij}^- = \min\{Z_{ij}^l, 0\}$. In this rule, α and β are nonnegative parameters that weight the relevance distribution for facilitatory and inhibitory preactivations, respectively. The sum of α and β is constrained to 1 to ensure the conservation of relevance between layers. This approach allows for a more nuanced allocation of relevance based on the nature of the preactivations.

Earlier implementations of LRP, such as those proposed by Bach *et al.* (2015) and Lapuschkin *et al.* (2016), typically employed a single LRP rule applied uniformly throughout the network. This approach failed to account for the varying needs of different layers or components within the network, resulting in less accurate and insightful relevance attributions. To address these limitations, Kohlbrenner *et al.* (2020) introduced a composite strategy, which employs a combination of multiple LRP rules tailored to different parts of the network. In our analysis, an identity function was used to back-propagate activation for the first two convolutional layers closest to the input, $\mathbf{LRP}_{\alpha 1}$ ($\mathbf{LRP}_{\alpha\beta}$ with $\alpha = 1$) for the remaining convolutional layers, and \mathbf{LRP}_ϵ for the fully connected layers closest to the output. This combined approach, denoted as $\mathbf{LRP}_{\text{CMP}}$, enhances model interpretability and provides more accurate relevance attributions.

B. Intersection over union (IOU) in sub-syllabic regions of the LRP heatmap

We quantitatively assess the LRP analysis to assess how regions corresponding to different components of the initial and final syllables contribute to lexical stress classification using the IOU metric, adapted for analyzing relevance attribution in heatmaps. Specifically, we define the inside-total overlap ratio μ :

$$\mu = \frac{R_{in}}{R_{tot}}. \tag{5}$$

Here, R_{in} is the sum of all pixels of the spectrogram inside a bounding box derived from LRP heatmap and R_{tot} is the total sum of relevance values across the entire heatmap. This metric μ effectively measures the proportion of relevance attributed to the area of interest, providing insight into what portions of the signal the model has paid attention to. Higher values of μ indicate that a larger fraction of the relevance is concentrated within the target area.

The bounding boxes defining the regions of analysis were created using the phoneme-level timestamps and syllable boundaries described above, ensuring alignment with the relevant parts of the spectrograms. We set bounding boxes around the initial and final syllables and then, within each syllable, separate the vowel region from any other portion of the syllable (onset/coda).

C. Feature-specific relevance

In addition to the IOU metric, we introduce a measure to evaluate how the model focuses on specific acoustic features during decision-making. Focusing on the vowel within the stressed syllable, this analysis aims to isolate the contribution of each acoustic feature to the model’s decision-making process. We constructed *feature-specific heatmaps* that represent the relevance distribution as if the model were focusing exclusively on the given feature. This was done by extracting the fundamental frequency (F_0) and the first three formants (F_1, F_2 , and F_3), along with their formant bandwidth and sound intensity, using the default settings of the PRAAT analysis software (Boersma and Weenink, 2005). For each formant, heatmap values at each time point were generated by normalizing the intensity within the formant’s bandwidth and then scaling these normalized values by the sound intensity.

To quantify the similarity between these feature-specific heatmaps and the original relevance heatmap produced by LRP, we employed the Pearson correlation coefficient r . By comparing the strength of the correlation between feature-specific heatmaps, we gain insight into the relative importance of each acoustic feature in the model’s decision-making process. Additionally, we examine the ability of heatmaps created by all possible combinations of these features to determine how these different acoustic features interact to contribute to the model’s predictions.

Finally, we also examine the limitations of these specific features by considering the distribution of residual pixels left unexplained by any of these features (F_0 – F_3). We mask the LRP heatmap using the heatmap generated by combining all these features. We then examine the relative distribution of these unexplained pixels across the spectrum of the stressed vowels to consider what additional features, beyond pitch and formants, contribute to model performance.

V. MODEL PERFORMANCE

Classifier results on the test sets are summarized in Table II. The VGG16 architecture achieved classification accuracies of 88% and 92% in distinguishing between IS and FS in the disyllabic minimal pair and no minimal pair words test sets, respectively. The lower accuracy in the minimal pair test set may reflect errors in part-of-speech tagging. Note as well that human perception of lexical stress in disyllabic words based solely on acoustics (without access to syntactic or other contextual cues) is not at ceiling: Yu and Andruski (2010) reported first-language English listeners had a mean accuracy of 84% in this task.

Other architectures also showed strong results. VGG19 matched VGG16’s performance, despite utilizing a more complex architecture, suggesting that VGG16 is sufficient. VGG11 showed slightly lower performance on both test sets. While the other architectures also showed slightly lower performance, they still achieved a high level of performance (>86% on both test sets).

VI. INTERPRETABILITY RESULTS

We utilized LRP and our interpretability metrics to examine more closely the aspects of the input that influenced the strong performance of these classifiers.

A. LRP analysis

Inspection of the LRP heatmaps for both initial and final stress words suggested they highlighted distinct areas of relevance that align with the energy distribution patterns observed in the spectrograms. Figures 2 and 3 illustrate this for the best-performing VGG16 classifier. For initial stress words, the LRP heatmaps emphasized the initial syllable as the most relevant area for classification by VGG16. Conversely, for final stress words, the LRP heatmaps showed a concentration of relevance towards the end of the phoneme sequence. Figure 3 provides further illustrations of these differences, focusing on the $\mathbf{LRP}_{\text{CMP}}$ method.

B. Bounding box and IOU analysis

The IOU measurements provide a quantitative assessment of the overlap between the bounding boxes and the relevant pixels in the LRP heatmaps.

Table III shows IOU measurements (μ) for different LRP methods, examining VGG16 and ResNet18. The $\mathbf{LRP}_{\text{CMP}}$, $\mathbf{LRP}_{\alpha 1}$, and \mathbf{LRP}_{ϵ} methods suggest that both classifiers attend primarily to acoustic information in stressed

TABLE III. Average IOU values for different classifiers and LRP methods, separated by stressed versus unstressed syllables and syllable region (vowel versus all other components of syllable).

Classifier	LRP method	μ_{vowel}		μ_{notvowel}	
		Stress	No stress	Stress	No stress
VGG16	$\mathbf{LRP}_{\text{CMP}}$	0.432	0.135	0.276	0.156
	$\mathbf{LRP}_{\alpha 1}$	0.419	0.129	0.279	0.173
	\mathbf{LRP}_{ϵ}	0.428	0.137	0.27	0.164
	\mathbf{LRP}_z	0.301	0.154	0.318	0.226
ResNet18	$\mathbf{LRP}_{\text{CMP}}$	0.37	0.136	0.298	0.195
	$\mathbf{LRP}_{\alpha 1}$	0.361	0.128	0.255	0.255
	\mathbf{LRP}_{ϵ}	0.34	0.13	0.275	0.254
	\mathbf{LRP}_z	0.326	0.12	0.204	0.364

versus unstressed syllables and that the greatest concentration of pixels is specifically in the stressed vowel region. The basic \mathbf{LRP}_z method results are mostly consistent with these generalizations, but give some greater weight to unstressed syllables.

Overall, these findings suggest that the classifiers predict the location of stress by attending to features of the stressed syllable, in particular, to stressed vowels. However, it is important to note that under all methods, a considerable number of pixels lie outside of the stressed vowel—consistent with the distribution of cues to lexical stress throughout the syllable. Furthermore, many pixels lie outside of the stressed syllable—consistent with a view that sees lexical stress as a “relational” property of stressed as compared to unstressed syllables.

C. Feature-specific relevance analysis

We extended our IOU analysis by investigating how VGG16’s predictions are influenced by the phonetic features of the signal region that most strongly drives its responses, namely, the stressed vowels. We compared each of these features to the $\mathbf{LRP}_{\text{CMP}}$ -derived heatmap, focusing on the heatmap region corresponding to the stressed vowel. Figure 4 shows two feature-specific heatmaps corresponding to F_1 and F_2 in isolation. Additionally, the original relevance heatmap produced by VGG16 and $\mathbf{LRP}_{\text{CMP}}$ is presented for comparison.

As shown in Table IV, a heatmap generated solely using F_1 of the stressed vowel exhibits the strongest correlation with the observed heatmap. Unsurprisingly, this formant makes a significant contribution, as it correlates with tongue height, jaw height, and the overall opening of the oral cavity, which in turn are strongly associated with vowel sound intensity. However, the combination of F_1 and F_2 is a close second to F_1 alone, suggesting that the two acoustic features traditionally related to vowel quality in English work together to determine classifier predictions. Similarly, heatmaps combining these two formants with F_3 and, to a lesser extent, F_0 , are used to predict LRP responses. However, the outsize influence of F_1 is made clear by the low performance of the other formants and pitch in

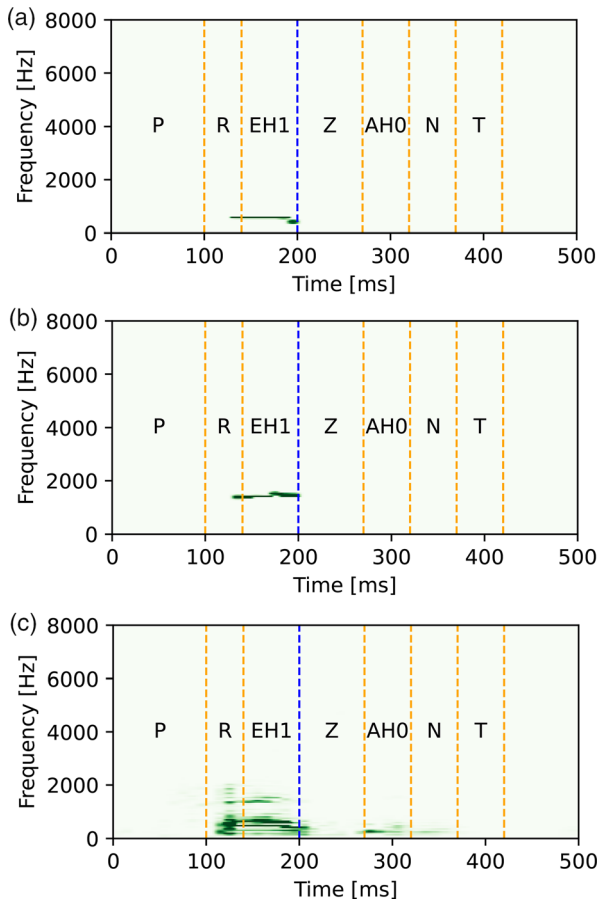


FIG. 4. Heatmaps of the first two formants (F_1 and F_2) and the original heatmap generated by LRP_{CMP} using VGG16 for the word “PREsent.” Panels (a) and (b) display the heatmaps generated for the two formants, considering bandwidth, spectrogram intensity, and voice intensity within the relevant time points. Finally, panel (c) illustrates the heatmap attributes generated by LRP_{CMP} utilizing VGG16. Orange vertical lines represent phoneme boundaries, and the blue vertical line represents the end of the initial syllable—the same as in Fig. 2.

isolation. When additional features that are less relevant to the model’s actual decision-making are combined (e.g., $F_1 + F_2 + F_3$), the overall similarity to the LRP-derived heatmap tends to decrease. This reduction arises because the inclusion of non-informative or weakly correlated regions introduces additional variance, thereby lowering the r value. In this sense, “more” does not necessarily mean “better”: Although these additional features may still play supporting roles in classification, their weaker correspondence to the model’s main decision cues can dilute the overall correlation with the relevance pattern.

While the r values are significant, they suggest that these features, even when considered in combination, fail to explain many aspects of the LRP heatmap for stressed vowels. To explore this, we examine the frequency distribution of residual (unexplained) pixels. As shown in Table V, the vast majority of unaccounted-for pixels lie in between F_0 and F_1 .

What other acoustic features might explain these pixels? Resolving questions like this is a key area for

TABLE IV. r and p values for heatmaps based on different features in the stressed vowel region (averaged across samples), using VGG16 and applying LRP_{CMP} method, sorted by r . (See text for details.)

Method	Acoustic feature	Mean $r \uparrow$	Mean $p \downarrow$
VGG16 + LRP_{CMP}	F_1	0.366	0.003
	F_1, F_2	0.359	0.003
	F_1, F_2, F_3	0.305	0.002
	F_0, F_1	0.3	0.002
	F_1, F_3	0.277	0.001
	F_0, F_1, F_2	0.245	0.001
	F_0, F_1, F_2, F_3	0.237	0.001
	F_0, F_1, F_3	0.205	0.001
	F_0, F_2	0.154	0.003
	F_2	0.154	0.008
	F_2, F_3	0.136	0.008
	F_0	0.116	0.004
	F_0, F_3	0.114	0.003
	F_3	0.048	0.02

development of this work. The phonetics literature suggests a number of possibilities (which are not mutually exclusive). Intensity has been proposed as a cue to stress (Cutler, 2005; Gay, 1978; Gordon and Roettger, 2017; Van Heuven, 2018). These pixels could reflect sensitivity to the concentration of energy in low frequencies during vowels. A number of prior studies have identified spectral tilt/spectral balance as an important cue to lexical stress (Gordon and Roettger, 2017; Van Heuven, 2018). Although spectral tilt measures often rely on information across a wider range of frequencies, spectral information in this range would provide a great deal of information about the steepness of spectral tilt. It is also possible that these pixels reflect sensitivity to harmonic structures. Inspection of Figs. 2 and 3 suggests that the classifier is attending to regularly spaced concentrations of energy in these lower frequencies—which could correspond to harmonics. The classifiers are using this information to track pitch, a contributor to lexical stress perception (Cutler, 2005). Our classifier relies on a standard spectrographic representation as input. The relatively poor resolution of this representation at low frequencies makes it a poor estimator of pitch. Tracking the spacing of harmonics (which have high amplitude in this region of the signal) would give the classifier access to this critical information. This could allow the classifier to estimate pitch with a higher degree of resolution than is available in the region of the spectrogram corresponding to F_0 , using information that can be readily

TABLE V. Relative distribution of pixels (averaged across samples) in the stressed vowel region of the residual LRP_{CMP} heatmap for VGG16. (See the text for details.)

Region	Initial stressed vowel	Final stressed vowel
Between F_0, F_1	0.93	0.886
Between F_1, F_2	0.07	0.1134
Between F_2, F_3	0	0
Above F_3	0	0

uncovered from the spectrographic input. It is also possible that the classifier is using the first and second harmonic amplitude difference, H1–H2, to measure voice quality (Garellek, 2019) differences related to stress (Gordon and Roettger, 2017).

VII. DISCUSSION

To explore what allows CNNs to outperform more traditional approaches, we built several novel classifiers for lexical stress detection in English disyllabic words, using LRP-based analyses to explore the spatiotemporal properties that drive model performance. A fully automatic pipeline was used to extract disyllabic words from read and spontaneous speech. CNNs trained on words without minimal pairs exhibited high performance on held-out no minimal pair words and words with minimal pairs. LRP analyses on the spectrogram input space for the CNNs revealed that the models were sensitive to acoustic properties in both stressed and unstressed syllables, but showed greater reliance on stressed syllables (particularly the stressed vowel). Analysis of our highest-performing model showed that F_1 of the stressed vowel exerted a stronger influence on predictions than other formants, but F_2 and, to a lesser extent, F_3 also contributed. Furthermore, these features failed to capture a substantial amount of variance in the LRP heatmap, suggesting additional features of the stressed vowel contribute to its predictions (e.g., spectral balance and pitch based on spacing of harmonics).

The classifiers’ greater reliance on the properties of stressed versus unstressed syllables and more specifically stressed vowels versus other syllable regions is similar to the focus of the phonetics literature on stressed syllables and stressed vowels (Cutler, 2005; Gay, 1978). Critically, differing from some early phonetics studies, the classifier does not focus *exclusively* on these regions, but also attends to information in both syllables, including non-vowel portions.¹ The use of information in the unstressed syllable resonates with the relatively widespread use of relativized or normalized cues to lexical stress (Cutler, 2005; Van Heuven, 2018). The use of information outside of the vowel is also consistent with phonetics research that examines duration and intensity of the whole syllable, the rime, or consonants (Cutler, 2005; Gordon and Roettger, 2017). More broadly, the use of consonantal information is consistent with the view that stress is not a property of vowels but of whole syllables, organized within more complex prosodic structures such as feet; these structures condition not only the properties of vowels but also surrounding consonants in stressed and unstressed syllables [see, e.g., Jensen (2000)]. Future work should continue to develop this general line of analysis by examining variation in stress driven by larger sentential and discourse contexts [see Van Heuven (2018) for a review and discussion].

While, in English, stressed versus unstressed syllables are typically distinguished by full versus reduced vowel quality, some disyllabic stress-pairs are pronounced with

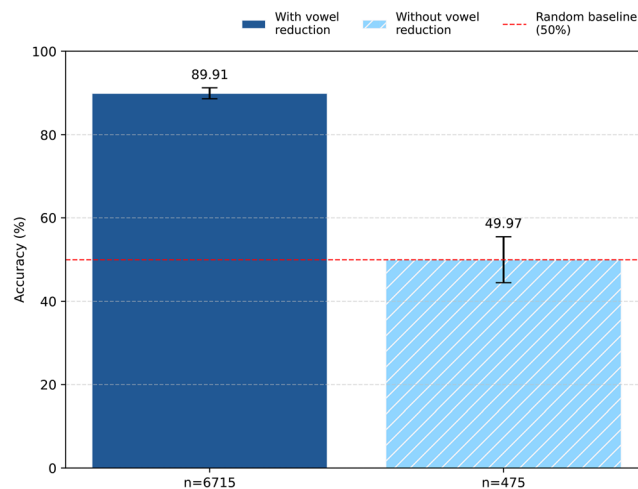


FIG. 5. Classification accuracy of the VGG16 model on the minimal pairs test set, aggregated by vowel reduction in the unstressed syllable. Bars show mean per-word accuracy, averaged across minimal pairs with vowel reduction (left) and without vowel reduction (right). Error bars indicate \pm standard error of the mean computed across words within each group. The number of test samples in each group is indicated by n . The dashed red horizontal line denotes chance-level (random baseline) accuracy at 50%.

full vowels in both syllables. Figure 5 summarizes classification accuracy on the minimal pairs test set after aggregating all minimal pairs words into two groups based on the presence or absence of vowel reduction in the unstressed syllable. Because the task is binary (initial vs final stress), chance-level performance corresponds to 50% accuracy. Tagging of minimal pairs words as “with” or “without” vowel reduction was based on a predefined list of English stress pairs (Bosker, 2024). Words treated as lacking vowel reduction include *import*, *insult*, *digest*, *increase*, and *transfer*. This classification was not based on token-level acoustic annotation; consequently, the actual degree of reduction in particular, productions may not precisely align with these assumptions. The figure reveals a clear performance gap: The models achieve substantially higher accuracy for words exhibiting vowel reduction than for those that do not. When vowel reduction is absent, mean accuracy falls to the chance level. This pattern is consistent with the nature of the training data, which consisted primarily of no minimal pairs disyllabic words in which vowel reduction serves as an important acoustic correlate of stress (Van Heuven, 2018). It also supports the interpretation that formant information, particularly cues related to vowel reduction, plays a central role in model performance, while also confirming that the models are sensitive to additional cues in cases where vowel quality differences are minimal.

Our feature-specific relevance analysis allowed us to connect the LRP results to formants of the stressed vowel. However, these analyses left a great deal of the LRP signal unexplained—not only in the stressed vowel, but also in other regions of the stressed syllable and the entirety of the unstressed syllable. This is consistent with prior work suggesting that formant information alone is insufficient for lexical stress classification (Bentum *et al.*, 2024). Future work

examining what acoustic features these pixels correspond to may provide new insights into the phonetics of English stress.

More broadly, this work suggests that deep learning may offer an important complement to traditional approaches to phonetics. Rather than starting with highly controlled materials like minimal pairs elicited under laboratory conditions, our classifier was trained on “messy” data—uncontrolled materials, varying along many dimensions beyond the contrast of interest, drawn from speech produced by a number of talkers in a variety of conditions. When exposed to this (naturally occurring) variability, the classifiers were able to extract regularities that fit with the previous phonetics literature. With increasingly powerful interpretability tools, deep learning may help phoneticians better understand the acoustic properties associated with linguistic contrasts.

ACKNOWLEDGMENTS

This work was supported by NSF DRL Grant No. 2219843 and BSF Grant No. 2022618.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The dataset used in this study is publicly available (MLSpeech, 2025). Code and a demo are also publicly available (Allouche, 2025).

¹It is possible that the distributed nature of stress cues reflects errors in our forced aligner. However, given the substantial temporal range of LRP pixels [see, e.g., Fig. 4(c)], it is likely that the results are robust to small changes to segment boundaries.

Allouche, I. (2025). “Minimal pairs lexical stress,” GitHub, <https://github.com/ItaiAllouche/minimalPairsLexicalStress>

Annex, A. M., Pearson, B., Seignovert, B., Carcich, B. T., Eichhorn, H., Mapel, J. A., von Forstner, J. L. F., McAuliffe, J., del Rio, J. D., Berry, K. L., Aye, K.-M., Stefko, M., de Val-Borro, M., Kulamani, S., and Murakami, S.-y. (2020). “SpicyPy: A pythonic wrapper for the SPICE toolkit,” *J. Open Source Softw.* 5(46), 2050.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS One* 10(7), e0130140.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). “WhisperX: Time-accurate speech transcription of long-form audio,” [arXiv:2303.00747](https://arxiv.org/abs/2303.00747).

Bentum, M., ten Bosch, L., and Lentz, T. (2024). “The processing of stress in end-to-end automatic speech recognition models,” in *Proceedings of Interspeech 2024*, Kos, Greece (ISCA, Stockholm, Sweden), pp. 2350–2354.

Boersma, P., and Weenink, D. (2005). “Praat: Doing phonetics by computer (version 6.2.09) [computer program],” <https://www.praat.org/> (Last viewed January 4, 2026).

Bosker, H. R. (2024). “List of minimal pairs differing only in lexical stress,” <https://osf.io/5d4ks/> (Last viewed January 4, 2026).

Cutler, A. (2005). “Lexical stress,” in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Oxford, UK), pp. 264–289.

Fry, D. B. (1955). “Duration and intensity as physical correlates of linguistic stress,” *J. Acoust. Soc. Am.* 27(4), 765–768.

Fry, D. B. (1958). “Experiments in the perception of stress,” *Lang. Speech* 1(2), 126–152.

Garellek, M. (2019). “The phonetics of voice 1,” in *The Routledge Handbook of Phonetics* (Routledge, London, UK), pp. 75–106.

Gay, T. (1978). “Physiological and acoustic correlates of perceived stress,” *Lang. Speech* 21(4), 347–353.

Gordon, M., and Roettger, T. (2017). “Acoustic correlates of word stress: A cross-linguistic survey,” *Linguist. Vanguard* 3(1), 20170007.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV (IEEE, New York), pp. 770–778.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Esteve, Y. (2018). “Ted-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation,” in *Speech and Computer: SPECOM 2018*, Lecture Notes in Computer Science, edited by A. Karpov, O. Jokisch, and R. Potapova (Springer, Cham, Switzerland), Vol. 11096, pp. 198–208.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A. et al. (2024). “GPT-4o system card,” [arXiv:2410.21276](https://arxiv.org/abs/2410.21276).

Jensen, J. T. (2000). “Against ambisyllabicity,” *Phonology* 17(2), 187–235.

Kingma, D. P., and J. Ba. (2014). “Adam: A method for stochastic optimization,” [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).

Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., and Lapuschkin, S. (2020). “Towards best practice in explaining neural network decisions with LRP,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, Scotland (IEEE, New York), pp. 1–7.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, T. (2020). “Captum: A unified and generic model interpretability library for PyTorch [computer program],” <https://captum.ai> (Last viewed January 4, 2026).

Korzekwa, D., Barra-Chicote, R., Zaporowski, S., Beringer, G., Lorenzo-Trueba, J., Serafinowicz, A., Droppo, J., Drugman, T., and Kostek, B. (2020). “Detection of lexical stress errors in non-native (L2) English with data augmentation and attention,” [arXiv:2012.14788](https://arxiv.org/abs/2012.14788).

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., and Samek, W. (2016). “Analyzing classifiers: Fisher vectors and deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV (IEEE, New York), pp. 2912–2920.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). “Gradient-based learning applied to document recognition,” *Proc. IEEE* 86(11), 2278–2324.

Li, K., Mao, S., Li, X., Wu, Z., and Meng, H. (2018). “Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks,” *Speech Commun.* 96, 28–36.

Lieberman, P. (1960). “Some acoustic correlates of word stress in American English,” *J. Acoust. Soc. Am.* 32(4), 451–454.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). “Focal loss for dense object detection,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy (IEEE, New York), pp. 2999–3007.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” in *Proceedings of Interspeech 2017*, Stockholm, Sweden (ISCA, Stockholm, Sweden), pp. 498–502.

MLSpeech (2025). “Lexical stress dataset,” https://huggingface.co/datasets/MLSpeech/lexical_stress_dataset (Last viewed January 4, 2026).

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia (IEEE, New York), pp. 5206–5210.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). “Scaling speech technology to 1,000+ languages,” *J. Mach. Learn. Res.* 25(97), 1–52.

P2TK Developers (2014). “P2tk: Phonetic and phonological toolkit, syllabified dictionary and syllabifier,” <https://sourceforge.net/p/p2tk/> (Last viewed January 4, 2026).

- Richey, C., Barrios, M. A., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., Lawson, A., Nandwana, M. K., Stauffer, A., van Hout, J., Gamble, P., Hetherly, J., Stephenson, C., and Ni, K. (2018). "Voices obscured in complex environmental settings (VOICES) corpus," in *Proceedings of Interspeech 2018*, Hyderabad, India (ISCA, Stockholm, Sweden), pp. 1566–1570.
- Rouso, R., Cohen, E., Keshet, J., and Chodroff, E. (2024). "Tradition or innovation: A comparison of modern ASR methods for forced alignment," [arXiv:2406.19363](https://arxiv.org/abs/2406.19363).
- Shahin, M. A., Epps, J., and Ahmed, B. (2016). "Automatic classification of lexical stress in English and Arabic languages using deep learning," in *Proceedings of Interspeech 2016*, San Francisco, CA (ISCA, Stockholm, Sweden), pp. 175–179.
- Simonyan, K. (2014). "Very deep convolutional networks for large-scale image recognition," [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Spaeth, H., Epstein, L., Ruger, T., Whittington, K., Segal, J., and Martin, A. D. (2014). "Supreme Court Database code book," <http://scdb.wustl.edu> (Last viewed January 4, 2026).
- Tepperman, J., and Narayanan, S. (2005). "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proceedings (ICASSP '05): IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, Philadelphia, PA (IEEE, New York), Vol. 1, pp. 1/937–1/940.
- Van Heuven, V. J. (2018). "Acoustic correlates and perceptual cues of word and sentence stress," in *The Study of Word Stress and Accent: Theories, Methods and Data* (Cambridge University Press, Cambridge, UK), pp. 15–59.
- Waibel, A. (1986). "Recognition of lexical stress in a continuous speech understanding system—A pattern recognition approach," in *ICASSP '86: IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan (IEEE, New York), Vol. 11, pp. 2287–2290.
- Ying, G., Jamieson, L., Chen, R., Michell, C., and Liu, H. (1996). "Lexical stress detection on stress-minimal word pairs," in *Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP '96*, Philadelphia, PA (IEEE, New York), Vol. 3, pp. 1612–1615.
- Yu, V. Y., and Andruski, J. E. (2010). "A cross-language study of perception of lexical stress in English," *J. Psycholinguist. Res.* **39**, 323–344.