Interspeech 2024
1-5 September 2024, Kos, Greece

The Processing of Stress in End-to-End Automatic Speech Recognition Models

Martijn Bentum[1], *Louis ten Bosch*[1], *Tom Lentz*[2]

[1]Center for Language Studies, Radboud University Nijmegen, The Netherlands
[2]Tilburg University, The Netherlands

martijn.bentum@ru.nl, louis.tenbosch@ru.nl, t.o.lentz@tilburguniversity.edu

Abstract

Listeners use stress to facilitate word recognition and speech segmentation. Classical ASR systems did not incorporate stress in their recognition process. In contrast, end-to-end ASR systems may use the information carried by stress. The present study shows that Wav2vec 2.0 is indeed sensitive to stress, and that this sensitivity is not a mere reflection of acoustic correlates of stress. Diagnostic classifiers of the CNN output reveal vowel-specific stress representations, that perform on par with acoustic features. Stress classifiers trained on transformer layers outperform classifiers based on acoustic correlates, but degrade when context is removed, showing that higher layers take the relative nature of stress into account. Results obtained by testing a stress classifier on a vowel it is not trained on, show that stress processing is to some extent abstract, i.e., the classifier does not simply detect a set of stressed vowel representations but rather, their common denominator.

Index Terms: word stress, explainable AI, ASR

1. Introduction

Many languages, including English, distinguish stressed and unstressed syllables. Stress can have a lexical function, distinguishing words that are otherwise (phonemically) identical, but can also be used by human listeners to segment continuous speech into words [1]. Classical ASR systems used to disregard stress as it added little to their performance. This paper investigates whether current End-to-End (E2E) models are sensitive to stress.

Stressed syllables are articulated with more effort [2], therefore tend to be longer, have a higher fundamental frequency, a higher intensity, a shallower spectral tilt, more peripheral formant frequencies, and their constituent phonemes are more resistant to co-articulation [2, 3, 4, 5, 6]. These acoustic correlates differ in their reliability as cues for stress [6]; in addition, they are hard to disentangle from other prosodic landmarks, most notably the intonation contour [7]. Which syllable is stressed can sometimes be predicted based on language-specific rules [2, 8] (e.g., heavier syllables are preferentially stressed). However, English stress placement can be lexical, e.g., *REcord* (noun) and *reCORD* (verb), the difference is carried by stress alone.

Stress is important for human listeners. Misunderstandings may occur when speakers stress unexpected syllables [9, 10, 11, 12]. Furthermore, human listeners utilize stress patterns for speech segmentation: They are faster to recognize a word such as *mint* in the pseudoword *mintesh* compared to the pseudoword *mintayve* [1, 13], because the latter contains the stressed syllable *tayve* and listeners segment this syllable out, splitting the target *mint*, making it harder to recognize.

Despite the importance for human listeners, stress recognition was not incorporated in classical automatic speech recognition (ASR) systems (e.g. [14]). Possibly, this is due to the effectiveness of a language model combined with bottom-up segmental information to identify items in the lexicon, which is also underspecified for distinctions such as the previous *record* example. In addition, stress realizations are heterogeneous and relative to the context [15] (e.g., syllable duration with respect to local speech rate). Furthermore, stress mostly surfaces as a supra-segmental feature with a temporal range that goes beyond the local spectro-temporal information encapsulated in the conventional 10ms spaced feature vectors. This asynchrony makes the integration of stress with the segmental stream harder.

The architecture of E2E-models is not predefined as classical ASR-models (e.g. acoustic model, language model, lexicon), but learned by a hierarchical network of neuronal layers. As a consequence, the E2E-model is free to represent and process aspects of speech not specified by design, such as stress. It is possible that E2E-models are like previous ASR-models, disregarding stress, but they might also have converged on a strategy capitalizing on the information provided by stress, like humans.

For the current study, we limit the scope to words spoken and recorded in isolation. This simplifies the labeling of stress considerably, though it means the presented results entangle word and sentence level stress [7]. We study one specific E2E-model, namely Wav2vec 2.0 (W2V) [16]. This model is ideally suited for the study of the emergence of stress representations, because the unsupervised pre-training phase of the model does not impose any labeling, affording the model freedom to identify salient representations for speech processing, potentially including stress. The W2V model is split in a convolutional neural network (CNN) block and a transformer block. The CNN block processes raw audio with a step size of 20 ms and an analysis window of 25 ms, which allows for the study of local effects of stress, while the transformer layers generate contextualized outputs (context window in the order of seconds), which allows for the study of contextualized representations of stress.

During pre-training, the output of the CNN block is quantized into codewords [16]. The quantized CNN outputs (normally only used during training) help the model to better generalize speech input. These discrete representations with a short 25 ms time span allow to probe possible abstract representation of stress in the model at the segmental level. For example, in Appendix D in [16], the authors show that sets of codewords relate to specific phonemes. The potential number of codewords is 102,400 [16]. With such a wealth of usable discrete units to describe the speech signal, it is possible that stress specific information is encoded as different codeword variants for the same phoneme.

10.21437/Interspeech.2024-44

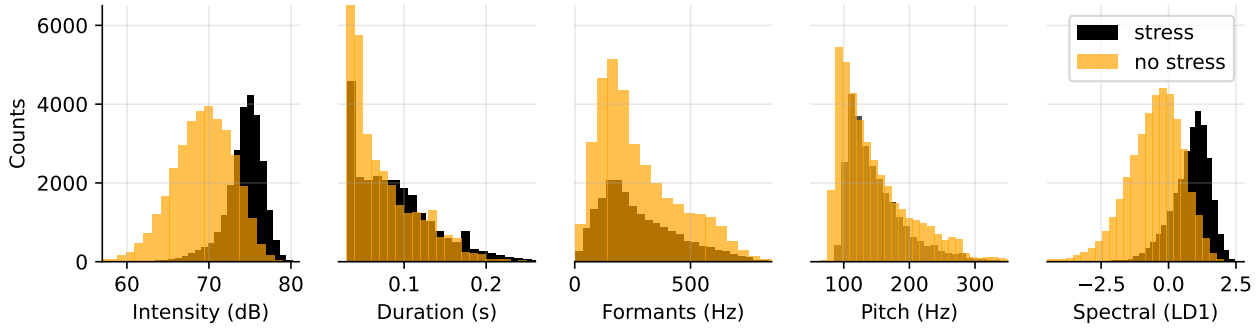

Figure 1: *Distributions of acoustic correlates of stress for stressed and unstressed vowels for the Complete Dataset.*

In the current study we aim to answer the following questions: To what extent does the information carried by acoustic correlates of stress percolate through the CNN layer? Do the Transformer layers process this information and take context into account, improving stress sensitivity? And if so, do the Transformer layers capture stress as one common feature that each syllable may have, or do they instead recognize stressed and unstressed versions of each phoneme separately? We hypothesize that if stress is as useful to E2E models as it is for human listeners, that the CNN layer captures at least the same stress information as the acoustic cues of stress do, and that higher (transformer) layers reflect syllable stress even more accurately.

2. Method

For all experiments we used a publicly available pre-trained W2V model[1] [17]. [To ensure author anonymity, the link to the source code will be added after the review process].

2.1. Materials

We used the MALD-corpus [18], containing 26,793 English words (we excluded the pseudowords). The words were recorded in isolation, spoken by the same 28 year old North American male speaker. There is one recording per word and all word recordings were down-sampled to 16 kHz.

2.2. Syllabification and stress pattern identification

We syllabified and identified the stress pattern for the words in the MALD corpus using Celex [19] and the Python module Prosodic[2]. We ignored secondary stress and only indicated primary stress or no stress for each syllable.

For all words present in Celex (24,458), we extracted the phonemic transcription of the MALD and Celex word. If the phonemic transcription of MALD and Celex for a given word mismatched, we used the Needlemann-Wunch algorithm [20] to optimally align both transcriptions. Subsequently we copied the syllable boundaries and stress patterns from Celex to the MALD transcription. For all MALD words not present in Celex (2,335), we applied the Prosodic module to generate syllable boundaries and stress patterns for 968 words (it failed for 1,367 words).

The resulting Complete Dataset contains 25,426 word types, 7,483 syllable types and 62,771 syllable tokens, with 25,426 stressed and 37,345 unstressed syllable tokens. There are 4,477 single syllable words and 3,711 hapax syllables. From the Complete Dataset, we selected 35 syllable types with at least 50 tokens and a balanced percentage of stressed realizations (i.e. between 40% - 60% stress). We will refer to this subset as the Balanced Dataset, which was created to rule out the potential confound of syllable type on stress classification.

2.3. Acoustic correlates of stress

For the vowels in stressed and unstressed syllables in the Complete Dataset, we computed values for the following acoustic correlates of stress: Intensity, Duration, Formants, Pitch and spectral balance. For each feature, the distribution of values for both stressed and unstressed vowels can be seen in Figure 1.

Intensity: We computed the mean Intensity (dB) of each vowel as follows $10\log_{10}(\overline{x^2}/4 * 10^{-10})$, whereby x denotes the audio samples corresponding to the vowel.

Duration: The duration of each vowel was based on the time-aligned phoneme transcriptions of the MALD-corpus [18].

Formants: We operationalized peripherality of formant frequencies as the Euclidean distance of the mean F1 and mean F2 of each monophtong vowel to the global mean F1 and F2 of all monophtong vowels. F1 and F2 values for all monophtong vowels in the Complete Dataset were computed with Praat [21].

Pitch: We computed the mean pitch for each vowel with the Librosa Python package [22] (version 0.10.1).

Spectral: We defined spectral balance according to [5] and computed the intensity (dB) in four consecutive frequency bands (0 - 500, 500 - 1000, 1000 - 2000, 2000 - 4000). Based on these intensity values, we applied a linear discriminant analysis (LDA) as implemented in the scikit-learn Python package [23] (version 1.3.0), to distinguish between stressed and unstressed vowels. The first linear discriminant (LD1) score for stressed and unstressed vowels is shown in Figure 1.

For each of the acoustic correlates of stress we trained 100 classifiers on random splits of the data into training (67%) and test (33%) sets. In addition to previously mentioned features, we added F1-F2 (mean F1 & mean F2) as a separate feature. Furthermore, we created the Combined feature, which consists of the features: Intensity, Duration, Pitch, F1-F2 and Spectral balance (consisting of the intensity of 4 frequency bands).

For the one dimensional features: Intensity, Duration, Formants and Pitch we trained classifiers based on the kernel density estimator (scipy.stats.gaussian_kde) from the Scipy Python package [24] (version 1.11.2), by estimating a probability den-

[1] https://huggingface.co/facebook/wav2vec2-large-xlsr-53
[2] https://github.com/quadrismegistus/prosodic

sity function (PDF) for both the stressed and unstressed vowels for a given feature. The stress label for a given vowel can be computed by applying the stressed and unstressed PDF to the feature value and select the label corresponding to the PDF with the higher value. For the multi-dimensional features: F1-F2 (mean F1, mean F2), Spectral balance (intensity values for the four frequency bands), we trained LDA-classifiers.

2.4. Diagnostic stress classifiers (probing)

We applied the W2V model to all words in the dataset and extracted the feature encoder outputs (CNN) and the output at the following Transformer layers: 1, 6, 12, 18, 21, 24. Based on these outputs we trained layer specific Multi-Layer Perceptron (MLP) classifiers with the scikit-learn Python module [23]; the procedure was repeated a 100 times on MLP classifiers on random splits of the data into training (67%) and test (33%) sets. Our use of probes to understand the information captured in intermediate layers of W2V is based on [25] (see for a linguistic context, e.g., [26]).

2.4.1. Occlusion

We define occlusion as setting a section of the word audio recording to silence. We created different occlusion conditions to investigate the effect of context on W2V processing of stress: *no occlusion* - the complete audio recording of a word and *occlusion* with only the audio of the vowel, and all other audio of the word recording set to silence. We used the materials in the Balanced Dataset to create the occlusion conditions and applied the process described in the preceding Section (2.4) on the whole word recordings and the occluded recordings with only the vowel.

2.4.2. Codewords

A codebook is used during pre-training of a W2V model. The codebook can also be used during inference, mapping the output of the CNN block to a specific codevector in the codebook. The resulting vector sequence is a discrete representation of the speech input. We created codevectors for all frames for each word in the dataset to test whether there are sets of codevectors that distinguish between stressed and unstressed version of a given vowel. Figure 2 shows the conditional probability for stressed and unstressed vowels given the codevector. The values on x-axis are ordered to show clusters of codevectors that are most related to a given vowel (stressed or unstressed).

In addition, we trained an LDA-classifier based on the codevectors to distinguish stressed and unstressed vowels; again, to ensure generalizability of the results this procedure was repeated 100 times on random splits of the data into training (67%) and test (33%) sets.

2.4.3. Leave-one-out and leave-one-in

To test whether the W2V model fractionates stress into phonemic specific representations or alternatively stress is processed as a more abstract supra-segmental feature, we trained stress classifiers with leave-one-out and leave-one-in datasets. In the leave-one-out dataset, a specific vowel (e.g. /i/) is removed from the training materials and used exclusively for testing. For the leave-one-in dataset this approach is reversed i.e., the classifiers is trained exclusively on a specific vowel (e.g. /i/) and tested on all other vowels.

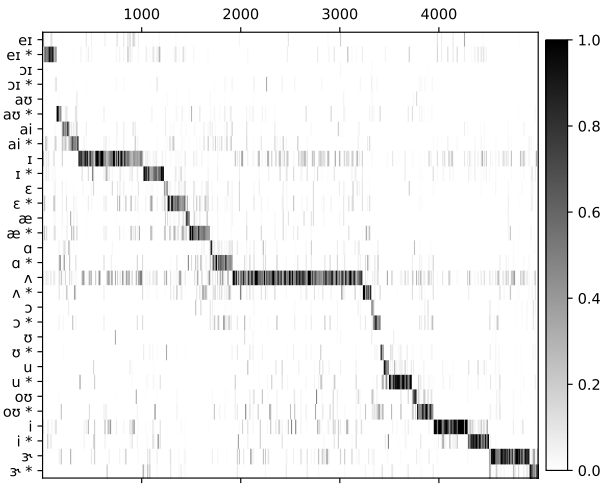

Figure 2: $P(phoneme|codevector)$ *Along the x-axis all codevectors are arranged that corresponded to a vowel at least once. Along the y-axis both the stressed (*) and unstressed version of each vowel are shown in pairs.*

2.5. Matthews correlation coefficient (MCC)

To compare the performance of the different stress classifiers, we use the MCC metric [27, 28]. MCC is unaffected by unbalanced datasets and only produces a high score when the classifier achieves good results for true positives, true negatives, false positives and false negatives [29]. The MCC score ranges in the interval [-1, 1], with a score of 0 corresponding to chance.

3. Results

3.1. Vowel stress classification performance

We computed feature values for different acoustic correlates of stress based on the word recordings in the Complete Dataset. Based on these features we trained stress classifiers (see Section 2.3). In addition, we applied the W2V model to the same materials and trained stress classifiers based on the outputs of different W2V layers (see Section 2.4). Furthermore, to control for the potential confound of syllable identity, we used the materials in the Balanced Dataset. Figure 3 shows the performance of all these stress classifiers. The Combined feature performs best of the acoustic correlates, closely followed by Intensity. The best classification overall is obtained with transformer layer 18, one of the W2V outputs, with a slight drop in performance for the Balanced Dataset based classifiers.

3.2. Context effects

To test investigate the influence of context we created an occlusion (see 2.4.1), with the word recordings set to silence except for the vowel (*occlusion*). The classifier performance results are shown in Figure 3. In the *no occlusion* (orange) condition, classifiers outperform the *occlusion* (grey) condition across the board, however, the difference between the occlusion conditions is most prominent in the transformer layers where the no occlusion classifiers perform better, indicating that the higher W2V layers stress processing is based more on context.

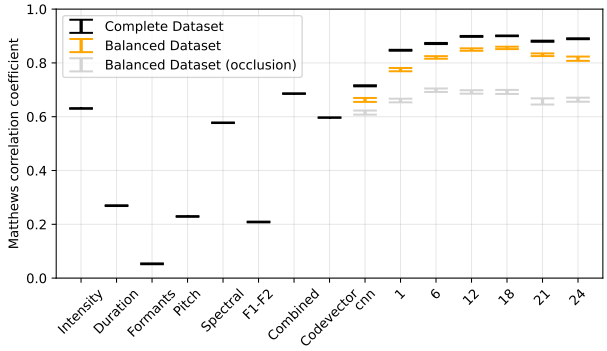

Figure 3: *The performance of classifiers trained on acoustic correlates of stress and different layers of the W2V model. For the Complete Dataset the results are shown in black. For the Balanced Dataset the results are shown in orange and grey, with orange for the whole word recordings (no occlusion) and grey (occlusion) based on only the the vowel. Error bars contain the 99% CI of the mean.*

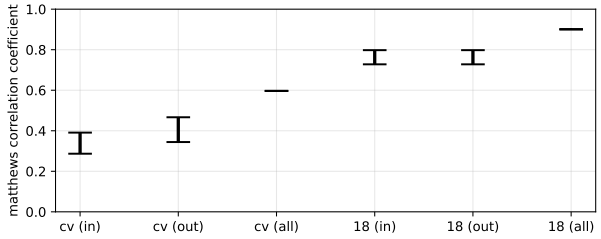

Figure 4: *For the Complete Dataset, the performance of stress classifiers based on leave-one-out (out) or leave-one-in (in), for either codevectors (cv) or W2V layer 18. Error bars contain the 99% CI of the mean.*

3.3. Vowel-specific or generalized stress

To test whether stress is processed in a vowel specific manner or alternatively in a more abstract supra-segmental manner in the W2V model, we computed codevectors for all materials in the Complete Dataset. Figure 2 shows that stressed and unstressed realizations of a given vowel map to distinct clusters of codevectors, indicating that stress is processed at the segmental level in the early W2V layers.

To further investigate whether higher layers in the W2V model abstract away from vowel specific stress processing we created leave-one-out and leave-one-in datasets (see Section 2.4.3) for both codevectors and Transformer layer 18. If stress processing is to some extent generalized across vowels, the classifiers trained on leave-one datasets should show reasonable performance. If however, stress processing is segment specific, the performance should drop. Figure 4 shows that for both the codevectors and layer 18 the classifiers based on the leave-one datasets perform worse compared to the classifiers trained on the whole dataset. For the codevectors this drop in performance is more pronounced, especially for the leave-one-in case. For transformer layer 18 the classifiers based leave-one-in perform the same as the leave-one-out, indicating that stress information is less vowel specific in the higher W2V layers.

4. Discussion and conclusion

We investigated stress processing in W2V, an E2E model. Stress representations were probed with stress classifiers, using layers of the W2V model as input. We compared the performance of these probes with classifiers trained on acoustic correlates of stress. Based on previous literature we identified Intensity, Duration, more peripheral formant frequencies (Formants), Pitch, and Spectral balance as acoustic features to correlate with stress realizations. In addition, we combined several acoustic features into the Combined feature.

For words in the MALD-corpus, Intensity and Spectral balance were the best performing single acoustic features, while the other features performed poorly. The robustness of Spectral balance has been reported before [5], however, Duration is typically a far more robust feature [6]. The current results might be due to speaker-specific idiosyncrasies (only one speaker in the dataset), with the high performance of Intensity probably due to the studio conditions of the recordings (i.e., always spoken towards the microphone at a similar distance). The Combined feature outperformed all single acoustic features and was almost on par with the CNN based classifier. This parity tentatively suggests that the CNN layer faithfully forwards acoustic information related to stress to higher layers. All W2V based classifiers (except for the codevectors) outperformed the acoustic feature based classifiers, with the best performance for Transformer layer 18. Hence, information about stress that is not a simple reflection of its acoustic properties seems to be extracted by higher W2V layers. To control for the confound of syllable identity, which to some extent predicts stress, we also trained classifiers on the Balanced Dataset (see Section 2.2). The result based on the Balanced Dataset (see Figure 3, orange vs. black) revealed that this only minimally influenced the performance of the W2V based stress classifiers.

We investigated to what extent W2V 'generates' abstract representations of stress. First, a test with codevectors (based on the CNN output), revealed that at the early processing stages, stress is 'fractionated' into vowel specific representations (see Figure 2). This finding was corroborated by a leave-one-out and leave-one-in experiment, revealing that codevector based stress representation do not generalize well over different vowels. In contrast, the result for leave-one-(in or out) with the classifier based on Transformer layer 18 showed that these representations generalize over different vowels in a much better way, which suggests that, at this stage in the model, stress representations are more abstract (see Figure 4). Furthermore, we tested the context-sensitivity of the different W2V layers by comparing results for occluded materials (with only the vowel present) to those obtained with intact recordings. We found a clear effect of occlusion and as expected, this was most pronounced for higher W2V layers (see Figure 3, orange vs. grey).

Future studies could improve the current results by incorporating spontaneous speech from multiple speakers, which would improve the ecological validity and might also better disentangle word and sentence level stress.

In summary, we found that E2E models process stress. The CNN block represents stress in a segment specific manner, with different codevectors for stressed and unstressed versions of a given vowel. Furthermore, stress representations at this stage are not much influenced by the wider context and the CNN based classifiers perform on par with simple acoustic features. In contrast, the Transformer layer based classifiers clearly outperform acoustic features, usefully incorporate context and seem to have a more generalized stress representation.

5. Acknowledgements

All authors participate in the Dutch NWO/NWA project InDeep (https://www.nwo.nl/en/projects/nwa129219399), led by J. Zuidema (Univ. of Amsterdam).

6. References

[1] A. Cutler and D. Norris, "The role of strong syllables in segmentation for lexical access." *Journal of Experimental Psychology: Human perception and performance*, vol. 14, no. 1, p. 113, 1988.

[2] V. J. Van Heuven, "Acoustic correlates and perceptual cues of word and sentence stress," *The study of word stress and accent: Theories, methods and data*, pp. 15–59, 2018.

[3] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[4] P. Lieberman, "Some acoustic correlates of word stress in American English," *The Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[5] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[6] M. Gordon and T. Roettger, "Acoustic correlates of word stress: A cross-linguistic survey," *Linguistics Vanguard*, vol. 3, no. 1, p. 20170007, 2017.

[7] D. R. Ladd and A. Arvaniti, "Prosodic prominence across languages," *Annual Review of Linguistics*, vol. 9, pp. 171–193, 2023.

[8] M. H. Kelly, "Word onset patterns and lexical stress in English," *Journal of Memory and Language*, vol. 50, no. 3, pp. 231–244, 2004.

[9] M. Benrabah, "Word-stress-a source of unintelligibility in English," *International Review of Applied Linguistics in Language Teaching*, vol. 35, no. 3, 1997.

[10] S. Shattuck-Hufnagel and A. E. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of psycholinguistic research*, vol. 25, pp. 193–247, 1996.

[11] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[12] A. Cutler, "Errors of stress and intonation," 1980.

[13] ——, "Segmentation problems, rhythmic solutions." *Lingua*, vol. 92, pp. 81–104, 1994.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[15] H. van den Heuvel, D. van Kuijk, and L. Boves, "Modeling lexical stress in continuous speech recognition for Dutch," *Speech Communication*, vol. 40, no. 3, pp. 335–350, 2003.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[18] B. V. Tucker, D. Brenner, D. K. Danielson, M. C. Kelley, F. Nenadić, and M. Sims, "The massive auditory lexical decision (MALD) database," *Behavior research methods*, vol. 51, pp. 1187–1204, 2019.

[19] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database (cd-rom)," 1996.

[20] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283670900574

[21] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[22] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[25] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2018, *arXiv 1610.01644*.

[26] T. Ashihara, T. Moriya, K. Matsuura, T. Tanaka, Y. Ijima, T. Asami, M. Delcroix, and Y. Honma, "SpeechGLUE: How well can self-supervised speech models capture linguistic knowledge?" *Interspeech*, 2023.

[27] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[28] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[29] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.