Automatic pitch accent classification through image classification

Na Hu[1], *Hugo Schnack*[2], *Amalia Arvaniti*[1]

[1]Radboud University, the Netherlands
[2]Utrecht University, the Netherlands
na.hu@ru.nl, h.g.schnack@uu.nl, amalia.arvaniti@ru.nl

Abstract

The classification of pitch accents has posed significant challenges in automatic intonation labelling. Previous research primarily adopted feature-based approaches, predicting pitch accents using a finite set of features including acoustic features (F0, duration, intensity) and lexical features. In this study, we explored a novel approach, classifying pitch accents as images represented in pixels. To evaluate this method's effectiveness, we used a relatively simple classification task involving only two types of pitch accents (H* and L+H*). The training of a basic neural network model for classifying images of these two types of accents (N= 2,025) yielded an average accuracy of 93.5% across 10 runs on the test set, showcasing the potential effectiveness of this new approach.

Index Terms: pitch accent classification, image classification, machine learning

1. Introduction

Intonation refers to utterance-level language-specific modulations of fundamental frequency (F0). F0 modulations can be modeled using various approaches, including the Fujisaki model [1], MOMEL [2], Tilt [3], polynomials [4], Functional Principal Components Analysis (FPCA) [5], and phonological models. The phonological models of intonation, in particular, generally follow the Autosegmental-Metrical approach (AM) [6], [7], which describes intonation in terms of pitch events, namely pitch accents and edge tones.

The AM approach to intonation has been widely adopted in intonation research and shows potential to enhance automatic speech applications [8], because it provides a framework for investigating the relationship between intonational forms and their meanings. For instance, AM delineates a clear distinction between H* and L+H* accents. Studies suggests that in American English, the former is realized as high pitch, encoding new information, while the latter is realized as rising pitch encoding contrastive information [9].

However, manually labeling intonation based on AM conventions requires extensive training, and is both time-consuming and susceptible to inter- and intra-annotator inconsistency [10]. Therefore, there is significant demand for automatic solutions. Considerable research effort has been dedicated to this pursuit, using different algorithms and features. The current state of automatic prosodic labeling achieves identification rates exceeding 90% in binary tasks, specifically in determining the presence or absence of an accent or a prosodic boundary. In contrast, pitch accent classification remains challenging and is highly dependent on the number of classes and speakers. When dealing with three pitch accent classes (high, downstepped, and low), [11] reported an accuracy rate of 80.2% using a multilevel hierarchical model based on a decision tree framework; [12] achieved an accuracy rate of 87.2% through the use of bagging and boosting with Classification and Regression Trees (CART); [13] employed neural networks and obtained an accuracy rate of 84%; [14] reported 81.3% accuracy using support vector machine (SVM) with a linear kernel. However, when dealing with a higher number of classes, the accuracy rates significantly decrease. For classification involving four classes (H*, L+H*, !H*, and L*), [8] reported an accuracy rate of 56.4% using a multi-layer perceptron classifier (MLP). For classification involving seven classes (H*, L+H*, !H*, H+!H*, L+!H*, L*, and L*+H), [15] obtained an accuracy rate of 63.99% using a confidence weighted combination of ensemble sampled SVMs, while [16] reported an accuracy of 70.8% using a fusion of neural networks and decision trees. Focusing on Dutch, [17] obtained an accuracy rate of 75.4% using SVM on seven types of pitch accents defined in the ToDI framework (H*L, H*, L*, L*H, H*LH, !H*L, !H*, L*HL) [18].

The aforementioned works are feature-based approaches, incorporating a finite set of features that include static acoustic features such as F0 point values, intensity, and duration [19], and dynamic ones such as slope approximations [20] or polynomial coefficients (horizontal line, sloped line, parabola and wave) [4], as well as lexical features. While the feature-based approach is widely adopted, it poses two potential limitations. First, the selection of features largely depends on researchers' expertise [21]. The second, and more important limitation, is that the features used in this approach only represent specific aspects of an F0 contour rather than the "full picture". An analogy illustrating this disparity is akin to describing an elephant with features like fan-shaped ears, a long nose, and chunky legs, rather than directly providing a depiction of an elephant. As an alternative to the feature-based approach, an instance-based approach to contour classification involves using entire contours for classification. This approach has been explored in a few studies. For example, [22] used hierarchical clustering to classify contours represented by sequences of F0 values extracted at different time points. Similarly, [23] assessed the effectiveness of k-means clustering for clustering F0 series and a bidirectional LSTM (long short-term memory) neural net classifier for contour labeling. Although these two studies did not explicitly focus on pitch accent classification, the methodologies they introduced can potentially be applied to address that issue as well.

The current study uses a similar approach to [22] and [23], classifying pitch accents based on their corresponding F0 contours rather than a set of features. In contrast to these two studies, our approach involves using image classification techniques to classify F0 contours presented as images, as opposed to a sequence of F0 values. The use of image classification techniques to classify pitch accents is theoretically sound, relying on computer vision to perceive and

interpret different patterns in F0 trajectories. This mechanism is essentially similar to manual annotations of pitch accents, where annotators follow labelling guidance, using human vision to identify phonetic evidence such as a "sharp rise in pitch" for L+H* and a "gradual" rise for H* as outlined in the MAE_ToBI training materials provided by MIT [24].

To our knowledge, this approach is novel for pitch accent classification, though widely used in various domains dealing with other types of time-series data, including medicine, entomology, astronomy, signal processing [25], micro- and macro- economics, finance, demographic data [26], and EEG signals [27]. The most widely adopted architecture for time-series classification is Convolutional Neural Networks (CNNs). Some studies modify the traditional CNN architecture, using 1D time-series signals as input, while others transform time series into images using image transformation methods such as recurrence plots (RP) [27], [28], Gramian Angular Fields (GAF) [29], Markov Transition Fields (MTF) [25], or a combination of these imaging methods [26]. The image classification approach is particularly intriguing for pitch accent classification because it preserves temporal information by transforming time series into images using image transformation methods such as GAF. This can be advantageous over time-series clustering, as introduced in [22], where F0 series for clustering share the same duration. However, in this initial exploration, we did not apply image transformation. Instead, we trained a basic neural network model without convolutional layers to classify images of F0 contours presented in pixels. To assess this method, we conducted a relatively straightforward classification task involving only two types of pitch accents: H* and L+H*, instead of dealing with multiple classes as in previous research. Although we only focused on these two accents, they are notoriously difficult to disentangle in manual annotations, often leading to disagreement among annotators [10].

2. Methods

2.1. Data set

The data set involved unscripted speech from 8 native speakers of Southern British English (5 females, 3 males, aged 18-54 years, mean age 29.25 years) elicited using three tasks. Each speaker completed a storytelling task and a map task [30]. Additionally, pairs of speakers engaged in informal discussions about unusual objects, such as a nose flute, a magnetic cable organizer, and a telescopic magnetic gripper. The recorded speech data were annotated for pitch accents by one expert annotator based solely on the F0 shape of the accented syllable and adjacent segments: accents were annotated as L+H* if they consisted of a deliberate F0 dip at the onset of the accented syllable, and as H* if not. The annotation yielded 2,025 accents: 1,764 H* and 261 L+H*. A second expert annotator independently annotated 12% of the accents using the same criteria, with unweighted Cohen's Kappa showing high inter-annotator agreement (0.85, C.I. = 0.81 – 0.89).

2.2. Preprocessing

The images of F0 curves were prepared in the following steps. First, F0 values (in Hz) were extracted from accented words at a time-step of 5 ms using PRAAT [31]. Second, the raw F0 curves were processed to remove F0 doubling and halving and interpolated to fill F0 gaps. Third, the resulting curves were normalized by speaker using z-scores to eliminate individual speaker characteristics. Fourth, to eliminate micro-phonetic variations at the morphological level, the speaker-normalized curves were smoothed using the B-spline method with 6 knots and a lambda of 10^4 using the `fda` package (see [5] for more information on smoothing), and time-registered based on a common landmark—the onset of accented vowels—to ensure that the curves are aligned around the same landmark and have the same duration, namely, the mean duration of all curves. Fifth, the resulting curves were saved as PNG images (480 by 480 pixels), resized to a lower resolution (28 by 28 pixels) in greyscale (pixel values 0 to 1). Figure 1 shows an example curve evolving through these steps. The corresponding pitch accent labels were transformed into arrays of integers: 0 for H* and 1 for L+H*.

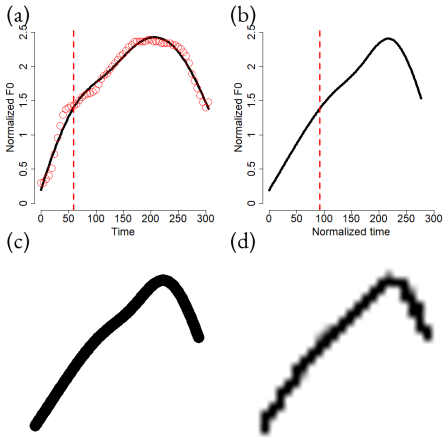

Figure 1: *(a) Raw F0 tracks (red circles) and smoothed curve (black line) with the accented vowel onset annotated as the red dashed line; (b) Landmark-registered curve; (c) Curve saved as a 480 by 480 pixels image without axes information; (d) Image rescaled to 28 by 28 pixels for training and testing.*

The entire set of images were randomly partitioned 10 times into training and testing sets at a ratio of 90% to 10%. The training set was used to train the network, while the testing set was used to evaluate the model's accuracy in classifying the images of curves. Figure 2 shows 10 randomly selected images from the training set.

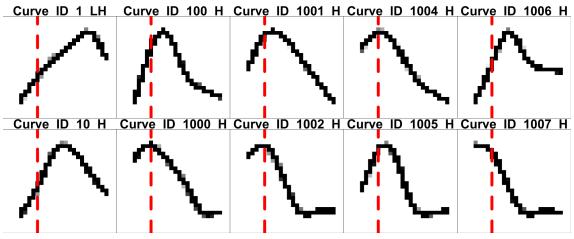

Figure 2: *10 randomly selected images from the training set with curve IDs and human-assigned labels shown on top of each subplot. The vertical red dashed line in each subplot denotes the accented vowel onset, depicted here for illustrative purposes and not present in the training and testing images.*

2.3. Model structure

A basic neural network without convolutional layers was trained to learn to associate the images of curves and their

corresponding labels using the `Keras/TensorFlow` package in `R` [32] (code available at https://osf.io/esb4g/?view_only=4409bab750b24dcd9be308bcb26a2d66). This network consisted of three layers. The first layer, "`layer_flatten`", transformed the image format from a 2D array (28 by 28 pixels) to a 1D array of 28*28=784 input features. The flattening layer was followed by a sequence of two dense, or fully-connected, layers. The first dense layer had 128 nodes (or neurons) with a `ReLU` activation function, and the second layer was a 2-node `softmax` layer. This output layer returned an array of 2 probability scores that sum to 1. Each node within this layer contains a score indicating the probability that the current image belongs to one of the two pitch accent classes. Although the `softmax` layer is usually used for multi-class problems, we opted to include it (instead of a single node with a `sigmoid` activation function) as it allows an easy increase in the number of output nodes when dealing with additional pitch accent classes. Using the `adam` optimizer and `sparse_categorical_crossentropy` as the loss function, the model was fit in 5 epochs.

Model training and testing was conducted 10 times on the 10 randomly partitioned sets of images, and the average accuracy score and the loss were calculated across the 10 runs.

3. Results

The average accuracy across 10 runs on the test dataset was 93.5% and the average loss 14.9%. Table 1 shows the test accuracy and loss for each run.

Table 1: *Model accuracy and loss on the test set in each run*

Run No.	Accuracy	Loss
1	0.921	0.196
2	0.956	0.133
3	0.960	0.097
4	0.921	0.171
5	0.926	0.155
6	0.966	0.106
7	0.887	0.257
8	0.931	0.126
9	0.941	0.130
10	0.941	0.122
Average	0.935	0.149

To gain a more comprehensive understanding of the model's predictability, we examined the prediction accuracy for each class. Out of the 203 randomly selected images for testing in the 10th run, 175 were annotated as H*, and 28 were annotated as L+H*. Among the 175 imaged labeled as H*, 170 were predicted as such, while 5 were predicted as L+H*. For the 28 images of L+H*, 23 were predicted as such, while the remaining 5 were predicted as H*. Thus, the model's accuracy in predicting H* was 97%, while the accuracy for predicting L+H* accents was 82%. Figure 3 presents 10 randomly selected correct predictions for each class made by the trained model in the 10th run. In this figure, human-assigned pitch accent labels are shown on top of each subplot before brackets, and model-predicted labels are shown within brackets. The vertical red dashed line in each subplot denotes the accented vowel onset, depicted here for illustrative purposes, and not present in the training and testing images. As shown in the figure, the model was quite successful in detecting patterns in F0 trajectory and assigning a correct pitch accent label to it. For example, curve 1199 (panel (a): row 1, column 1), which shows a clear rising pattern, is labeled as L+H* by the human expert and correctly predicted by the model as such. In contrast, curve 1019 (panel (b): row 2, column 1) exhibits a typical falling pattern, labeled as H* by the human expert and correctly predicted as such by the model. The model not only distinguishes typical rising and falling patterns, but also had successfully learned the association between F0 peak alignment and pitch accent labels. When F0 peaks occur noticeably after the accented vowel onset, as in the curves shown in Figure (3a), the model correctly identifies the curves as L+H*. Conversely, when F0 peaks occur before or align with the accented vowel onset, as in the curves shown in Figure (3b), the model correctly identifies them as H*.

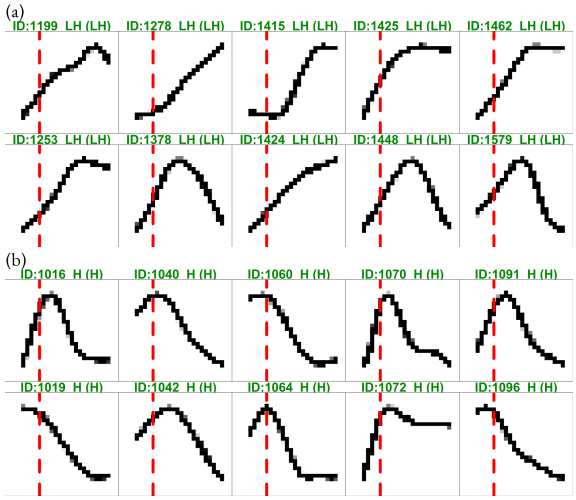

Figure 3: *10 randomly selected correct predictions for each class ((a): L+H*; (b) H*) made by the model in the 10th run. The vertical red dashed line in each subplot denotes the accented vowel onset.*

The output of the 10th run of the model was used to inspect the instances for which the predicted labels differed from those provided by humans, which was the case for ten test images (5%). Among these, the model predicted H* for five contours labeled as L+H* by humans, and L+H* for the remaining five contours labeled as H* by humans. These instances are presented in Figure 4, following the same visualization style as seen in Figure 3. Upon scrutinizing these discrepant cases, two plausible causes for the disparity between human-assigned labels and the model's predictions emerge. These cases are discussed in the next section.

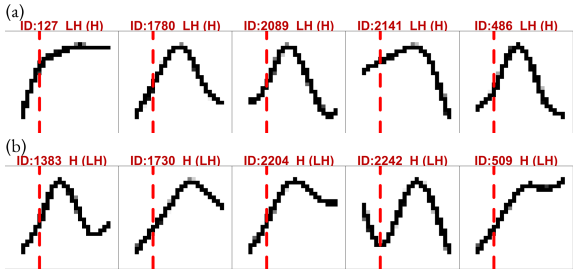

Figure 4: *All 10 cases with discrepancies between human-assigned labels and those predicted by the model in the 10th run: (a): L+H* predicted as H*; (b) H* predicted as L+H*. The vertical red dashed line in each subplot denotes the accented vowel onset.*

4. Discussion and conclusions

In this study, we trained a basic neural network model to learn the association between pitch accent labels provided by a human expert and F0 curves represented as pixel images. We used the trained model to predict the pitch accent type for new curves represented as images. The outcome from a relatively simple classification task involving two types of pitch accents (H* vs. L+H*) yielded an average accuracy rate of 93.5% across 10 runs on the test set, indicating the effectiveness of this new approach. This initial exploration demonstrates that image classification techniques can successfully discern different patterns in F0 series. This is in line with the success of this technique in handling other types of time-series data. The effectiveness of the image classification approach can be attributed to the fact that it takes into account the entire F0 information in a curve, rather than relying on a finite set of features describing the curve shape, as is the case for the existing feature-based models mentioned in the introduction.

While the overall accuracy is high, the accuracy rates for each class differ, with H* having higher accuracy than L+H* (97% vs. 82%). This disparity can be attributed to imbalance in the dataset, where there are a lot more instances of H* than L+H*. Apart from acquiring more L+H* instances for training, the accuracy for L+H* could be improved by employing techniques such as class weighting or data augmentation.

There were, however, some instances, in which the model's prediction did not agree with the classifications of the human annotator. There are two plausible explanations for these discrepancies. First, it seems that, in certain cases, these discrepancies might be better annotated with a different label than the one originally assigned. Specifically, instances originally labeled as H* by the human expert might be better annotated as L+H* given the presence of a clear F0 dip at the onset of the accented syllable. This applies to at least four out of the total five predicted H* cases, namely curves 1730, 2204, 2242, and 509 (in Figure 4(b)). Similarly, two instances originally labeled as L+H* by the human expert, namely curves 127 and 2141 (in Figure 4(a)), might be more appropriately labeled as H*, because the F0 dip at the beginning of the accented syllable is not as visible as it is in other cases. This potential concern involves approximately 3% of the instances in the test set (6 out of 203). While this fraction of data might contribute to the model's discrepant predictions, intra- and inter-annotator inconsistency is hard to avoid and has been frequently observed in the manual AM annotations [10], highlighting the need for alternative approaches that can generate labels more consistently. It is also possible that the F0 rises causing the discrepancy between human and machine are artifacts of interpolation, leading to a marked rise in the curve not accessible to the former but used by the latter. While the questionable instances constitute only a small fraction of the data, a sanity check of the interpolated curves, which was not conducted in the current experiment due to its exploratory nature, is necessary to identify the real cause of the issue and further enhance data quality. After resolving these questionable instances, we could retrain the model to see if the accuracy can be further improved. The other plausible cause for the disparity between the labels provided by the human expert and the model's prediction is related to the ambiguity in the pitch accent type, concerning curves 1780, 486, and 1383 (as shown in Figure 4). Although an F0 dip seems to be present at the beginning of the accented syllable, the F0 peak following the dip is quite close to the beginning of the accented syllable. Therefore, the pitch accent category to which these curves belong is ambiguous and thus they could potentially fit into either category, as indicated by the close probability scores for these curves belonging to the two classes (e.g., 0.57 vs. 0.43). Curves with these atypical shapes constitute the overlaps between these two types of pitch accents [33], [34], [35], [36], despite the fact that the accents are categorically different in terms of phonetic realization in standard British English [37]. Such curves pose challenges not only to manual annotations but also to automatic classification. One potentially useful solution that avoids relying on human-assigned labels is unsupervised learning, enabling algorithms to determine the clustering of contours without the guidance of pre-provided labels. Subsequently, the machine-formed clusters can be compared to the grouping of curves suggested by human-provided labels to assess if the algorithm picks the same traits as those considered relevant by human annotators. However, it is unpredictable which aspects of F0 variation algorithms would consider important when forming clusters.

Although this initial exploration has demonstrated the potential of using image classification to predict pitch accents, it is hard to directly compare our results with those of previous studies on automatic pitch accent classification, because those works generally involve more classes and trained their classifiers on different data sets, with the Boston University Radio News Corpus (BU-RNC) [38] being the most commonly used corpus. To compare the effectiveness of the current method with those used in previous studies, it is essential to involve more classes to classify and conduct experiments on the BU-RNC corpus.

While the image classification technique was applied to pitch accents in the current study, we do not intend to restrict its application to this specific task. Given its working mechanism as a pattern recognition technique, we believe that the approach showcased in the current study can be applied as a generic method for pitch contour classification, addressing various research questions. Presumably, after being trained on data containing the labels of interest, the models can automatically assign labels to new contours, expediting the data annotation process.

This preliminary exploration has shown a high accuracy of the image classification technique in classifying H* and L+H* in unscripted speech in standard British English, demonstrating its potential effectiveness in F0 contour classification.

5. Acknowledgements

This work has received financial support from the European Research Council (grant no. ERC-ADG-835263) awarded to Amalia Arvaniti.

6. References

[1] H. Fujisaki, "Modelling the process of fundamental frequency contour generation," in *Speech perception, production and linguistic structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisasaka, Eds., Amsterdam: IOS Press, 1992, pp. 313–328.

[2] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and experiment*, M. Horne, Ed., Dordrecht: Kluwer Academic Publishers, 2000, pp. 51–88. doi: 10.1007/978-94-015-9413-4_4.

[3] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *J Acoust Soc Am*, vol. 107, no. 3, pp. 1697–1714, 2000, doi: 10.1121/1.428453.

[4] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Lang Speech*, vol. 50, no. 3, pp. 281–310, 2007, doi: 10.1177/00238309070500030101.

[5] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *J Phon*, vol. 49, pp. 16–40, 2015, doi: 10.1016/j.wocn.2014.10.001.

[6] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," Massachusetts Institute of Technology. PhD Thesis., 1980.

[7] D. R. Ladd, *Intonational Phonology*. 2008. doi: 10.1017/cbo9780511808814.

[8] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric F0 features," in *Proceedings of ICASSP – IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008. doi: 10.1109/ICASSP.2008.4518667.

[9] J. B. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in communication*, P. R. Cohen, J. L. Morgen, and M. E. Pollack, Eds., MIT Press, Cambridge, MA; London. Lewis., 1990.

[10] A. K. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of ICSLP 2000 – the 6th International Conference on Spoken Language Processing*, 2000. doi: 10.21437/icslp.2000-521.

[11] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Comput Speech Lang*, vol. 10, no. 3, 1996, doi: 10.1006/csla.1996.0010.

[12] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of ICSLP*, 2002, pp. 16–20.

[13] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proceedings of ICASSP – IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004. doi: 10.1109/icassp.2004.1326034.

[14] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Proceedings of INTERSPEECH*, 2005.

[15] A. Rosenberg, "AuToBI - A tool for automatic ToBI annotation," in *Proceedings of INTERSPEECH*, 2010, pp. 146–149.

[16] C. González-ferreras, D. Escudero-mancebo, C. Vivaracho-pascual, and V. Cardeñoso-payo, "Improving automatic classification of prosodic events by pairwise coupling," in *Proceedings of IEEE Transactions on Audio, Speech, and Language Processing*, 2012, pp. 2045–2058.

[17] N. Hu, B. Janssen, J. Hanssen, C. Gussenhoven, and A. Chen, "Automatic analysis of speech prosody in Dutch," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020. doi: 10.21437/Interspeech.2020-2142.

[18] C. Gussenhoven, "Transcription of Dutch Intonation," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed., 2005. doi: 10.1093/acprof:oso/9780199249633.003.0005.

[19] G. Demenko and A. Wagner, "The stylization of intonation contours," in *Proceedings of the International Conference on Speech Prosody*, 2006. doi: 10.21437/speechprosody.2006-48.

[20] G. A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proceedings of HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006. doi: 10.3115/1220835.1220864.

[21] B. D. Fulcher, "Feature-Based Time-Series Analysis," in *Feature Engineering for Machine Learning and Data Analytics*, 2018. doi: 10.1201/9781315181080-4.

[22] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *J Int Phon Assoc*, vol. 53, no. 1, pp. 159–188, Apr. 2023, doi: 10.1017/S0025100321000049.

[23] J. Cole, J. Steffman, S. Shattuck-Hufnagel, and S. Tilsen, "Hierarchical distinctions in the production and perception of nuclear tunes in American English," *Lab Phonol*, vol. 14, no. 1, 2023, doi: 10.16995/labphon.9437.

[24] N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, "ToBI for Prosodic Transcription of American English," MIT Opencourseware 6.911, 2006.

[25] Z. Wang and T. Oates, "Spatially encoding temporal correlations to classify temporal data using convolutional neural networks," *arXiv:1509.07481*, 2015.

[26] W. Jiang, D. Zhang, L. Ling, and R. Lin, "Time Series Classification Based on Image Transformation Using Feature Fusion Strategy," *Neural Process Lett*, vol. 54, no. 5, 2022, doi: 10.1007/s11063-022-10783-z.

[27] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," in *Proceedings of International conference on machine vision*, 2017.

[28] R. K. Tripathy and U. Rajendra Acharya, "Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework," *Biocybern Biomed Eng*, vol. 38, no. 4, 2018, doi: 10.1016/j.bbe.2018.05.005.

[29] Z. Wang and T. Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *AAAI Workshop - Technical Report*, 2015.

[30] A. H. Anderson *et al.*, "The Hcrc Map Task Corpus," *Lang Speech*, vol. 34, no. 4, pp. 351–366, 1991, doi: 10.1177/002383099103400404.

[31] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.43," retrieved 8 September 2018.

[32] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.r-project.org/

[33] D. R. Ladd, *Simultaneous Structure in Phonology*. 2014. doi: 10.1093/acprof:oso/9780199670970.001.0001.

[34] J. M. Scobbie, "Interface and Overlap in Phonetics and Phonology," in *The Oxford Handbook of Linguistic Interfaces*, 2012. doi: 10.1093/oxfordhb/9780199247455.013.0002.

[35] J. B. Pierrehumbert, "Phonological Representation: Beyond Abstract Versus Episodic," *Annu Rev Linguist*, vol. 2, 2016, doi: 10.1146/annurev-linguistics-030514-125050.

[36] J. B. Pierrehumbert, "Word-specific phonetics," in *Laboratory Phonology 7*, 2008. doi: 10.1515/9783110197105.

[37] J. Kim, N. Hu, S. Gryllia, R. Orrico, and A. Arvaniti, "Delineating H* and L+H* in Southern British English," in *Proceedings of Speech Prosody 2024*, 2024.

[38] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus." 1995.