



# Wav2vec behind the Scenes: How end2end Models learn Phonetics

Teena tom Dieck<sup>1</sup>, Paula-Andrea Pérez-Toro<sup>1,2</sup>, Tomás Arias-Vergara<sup>1,2,3</sup>,  
Elmar Nöth<sup>1</sup>, Philipp Klumpp<sup>1</sup>

<sup>1</sup>Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

<sup>2</sup>GITA Lab, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia

<sup>3</sup>Department of Otorhinolaryngology, Head & Neck Surgery, University Hospital Erlangen

{teena.tom.dieck;philipp.klumpp}@fau.de

## Abstract

End2end models became extremely popular in recent years. Whilst they excel at tasks like acoustic modelling or full-fledged speech recognition, the decision making process can be quite complex to retrace due to their black-box character.

As end2end models learn high-level feature extraction on-the-fly, outputs from hidden layers from within the network had been used as feature vectors in various studies to perform transfer learning. It is therefore crucial to understand how extracted hidden activations transport information collected from the signal. Furthermore, is the traditional categorization into feature extractor and temporal analysis still applicable on the sub-parts of end2end models?

By the example of Wav2vec 2.0, we show how an acoustic model learns to perform a frequency analysis on a speech waveform. Our experiments also show that phonetic information about speech production is preserved in extracted feature vectors. Ultimately, our findings highlight how different parts of an end2end model encode information on an entirely different level. Whilst the influence of gender is quite large on early feature vectors, it vanished after temporal contextualization. At the same time, hidden activations which included context information were superimposed by language-related patterns.

**Index Terms:** Acoustic Modelling, Feature Extraction, Phonetic Analysis

## 1. Introduction

Over the course of the past years, end2end models for acoustic modelling and automatic speech recognition (ASR) experienced a significant increase in both popularity and power. Any sequence understanding task can be broken down into two major steps. At first, features are extracted, traditionally over a window of fixed size. This window can be shifted over the sequence of arbitrary length to compute feature vectors (FV) at different time steps. The second part of a sequence understanding problem then encompasses the actual analysis of temporal patterns. Instead of using only the FV of one particular time step, a model should be able to evaluate local information against the background of its temporal context. Modern end2end architectures solve the task of feature extraction and temporal contextualization in one framework, allowing each component to learn from the respective other.

Among the most popular end2end architectures were Wav2vec [1] as well as its successor Wav2vec 2.0 [2]. Un-supervised (prediction of next time step) and self-supervised (prediction of masked time steps) pretraining strategies were used to learn a feature extraction which was not biased towards the final task, for example phoneme recognition. This pretraining is computationally very expensive, but the

following adaptation to a particular domain, often referred to as fine-tuning, only requires the addition of a linear classification layer to train the model on a specific task. All preceding layers had already been ‘warmed up’, in other words, they already performed a high-quality feature extraction and temporal analysis, and during the process of fine-tuning, only little adaptations are required (compared to a randomly initialized ‘cold start’) to push the parameters in the right direction.

While learning both feature extraction as well as temporal analysis in one large model does have advantages, it also comes at the risk of having a black-box solution which lacks interpretability. Along with the growth of deep architectures, methods to understand what happens inside of a neural network during the process of decision-making became increasingly important. Initially designed for the domain of image-processing, Grad-CAM [3] provided a technique to visualize which parts of an input signal contributed strongly to the final decision by exploiting gradient information during back-propagation. Another relevant approach is activation maximization [4]. Here, after a successful training, the process of back-propagation is flipped to find the input that maximizes the output of a certain node (a node represents a class). Looking at the inputs or outputs is not the only way to collect information about a deep model, its latent dimensions also hold information which describe the input signal. This property is used in representation learning (RL) approaches, e.g. auto-encoders, to learn lower-dimensional representations of data that preserves as much information as possible. RL is generally referred to as an unsupervised learning method, but representations can also be extracted from any model trained in a supervised manner. The idea of analyzing hidden layer activations has been employed in the domains of image [5, 6] and speech processing [7, 8, 9, 10]. A study [11] similar to the one presented here evaluated how hidden layers of a multi-layer perceptron trained with mel frequency cepstral coefficients (MFCCs) as inputs encoded phonetic clusters and vowel characteristics. For end2end models, such an approach is particularly interesting. It could shed light on how information is transformed throughout different stages of the neural network. Is Wav2vec 2.0 simply a powerful acoustic model, or can it still be broken down into a feature extractor and a temporal context analysis? Furthermore, if Wav2vec 2.0 was trained as an acoustic model to predict phonetic symbols, is it possible to derive information related to fundamental phonetic concepts? This work aims to answer these questions by fine-tuning a Wav2vec 2.0 base architecture on a multi-lingual phone recognition task to visualize intermediate network activations. Our experiments show that end2end models are capable of encoding information about vowel phonation, place and manner of articulation of consonants, and that hidden activations from different parts of

the model can yield very different results.

## 2. Materials & Methods

### 2.1. The Common Phone Dataset

To fine-tune our Wav2vec 2.0 model, we used the multilingual *Common Phone* [12] (CP) corpus. It comprises 116 hours of speech samples collected from more than 11,000 speakers in English, French, German, Italian, Russian and Spanish. The dataset is gender-balanced and provides phonetic annotation with a total of 101 different symbols from the International Phonetic Alphabet [13] (IPA). Training of the acoustic model as well as all experiments described throughout this work made use of the .wav audio files which had been converted to the standard format of 16 kHz sampling rate, 16 bits per sample and single-channel configuration. We used the CP training split to train the acoustic model. The test split was used afterwards to evaluate the overall performance with respect to phone error rate (PER) and to compute all hidden layer activations used in the scope of presented experiments.

### 2.2. The Acoustic Model

We used the base Wav2vec 2.0 model [2] which had been pre-trained on LibriSpeech [14], a corpus comprising 960 hours of read English speech. A final linear layer was appended to the model to map the outputs of the transformer [15] to 102 target nodes. The additional output node besides the 101 phone targets represented the blank token required for connectionist temporal classification [16] (CTC). Parameters were tuned with Adam optimizer [17]. The learning rate was initially set to  $3 \cdot 10^{-6}$  and increased linearly to  $3 \cdot 10^{-5}$  in the first ten epochs. It was then kept constant for 30 epochs, and would ultimately decay exponentially by a factor of 0.96 for the final 120 epochs. Instead of showing the model the entire training data once in every epoch, it was instead only given a subset of 5,000 randomly selected samples from the training split.

When the final model was used for inference, the most probable phone sequence was estimated through a beam search (beam width = 10) and CTC decoding. As we did not want to induce any language-related bias in phone transition probabilities, no language model was utilized during decoding. The final model achieved a global PER of 17.8 % on the development and 18.1 % on the test set. Detailed results for the individual languages can be found in [12].

### 2.3. Hidden Layer Activations

All hidden layer activations that were used in presented experiments were extracted from predictions on the test set of CP. The end2end model was not trained in a frame-wise fashion but with CTC, thus many of the output frames were ultimately mapped to the blank token. To get activations at the correct time frame, we chose to always extract them if the final classification layer emitted a phone symbol. The motivation is quite straightforward: CTC can be understood as a state machine, which would output a blank token whenever the system should remain in the current (previously emitted non-blank) state. If the system was convinced that the state had actually changed, a new symbol would be emitted. At this exact point in time, where the model is strongly convinced of the new symbol being present, we extracted hidden layer activations.

Two layers had been considered for activation extraction. The first one was the 768-dimensional output of the final transformer layer. It was chosen not only because temporal analysis had

been completed at this stage, but also because it served as the input to the final classification layer, which from this vector had come to the decision to output a non-blank symbol. The second hidden activation output was collected from the last convolutional layer at the same time step. It is quite common to refer to the convolutional neural network (CNN) part of Wav2vec as the feature extraction unit. We therefore wanted to access this 512-dimensional vector to evaluate how it would be altered during temporal contextualization in the following transformer layers. Ultimately, for a single audio sample, the presented approach yielded 3 sequences of equal length. The first sequence described the recognized phone symbols, the second and third the respective hidden activations from the transformer and CNN for each of the emitted symbols.

### 2.4. Experiments

To visualize how FVs were distributed in the high-dimensional space, they were decomposed via a 2-dimensional principal component analysis (PCA). Depending on the individual experimental setup, different sets of FVs were chosen to estimate the decomposition. We chose PCA over other methods of dimensionality reduction because it A) does not take any ground truth information into account (e.g. like linear discriminant analysis) and B) only determines the directions of maximum variance in the high-dimensional space, hence resulting in a function that can be used to transform unseen data into the 2 principal components. Other techniques, particularly those from the domain of manifold learning, only project data to a lower dimension, without the ability to later transform unseen data points.

The goal of the first experiment was to find out if an end2end model would be capable to reproduce the findings from [11], stating that the arrangement of FVs closely resembled the vowel triangle. The triangle reflects the tongue configuration during phonation with respect to the position (from front to back) as well as the elevation towards the palate (from open to close). It is spanned by the open front vowel [a], the close front vowel [i] as well as the close back vowel [u] [13]. This pattern could be preserved by models that perform any kind of frequency analysis, because position and elevation are directly linked to formant frequencies  $F1$  (more front, higher  $F1$ ) and  $F2$  (more open, higher  $F2$ ), respectively [18]. To achieve more stable results, the elongated variants of the triangle vowels were used to compute the PCA. Previous work had already shown that formant patterns are more pronounced in elongated vowels [19], likely because the speaker had more time to precisely position the tongue. Furthermore, we used an equal number of samples per vowel to prevent any bias, and used the FVs from the transformer layer, because it would summarize the entire vowel, without onset or offset effects. The estimated PCA would then allow transformation of any vowel sample into the 2-dimensional plane. We did this for the triangle vowels, and additionally for the vowels [e] and [o] to visualize how they would arrange in the PCA that had never seen any of these vowels. Lastly, the plosive [p] was included to show what would happen to a phone that is unrelated to the vowel triangle.

To find out whether the first experiment would also support the assumption of short vowels being less accurately pronounced, the same PCA was used to visualize the mean values of all vowels from the first analysis in their long and short variations. The theory states that all short vowels should be located closer to the center of the triangle [19].

To investigate to what extent Wav2vec would be able to preserve information about place and manner of consonant articulation, transformer FVs were collected from all fricative and plosive

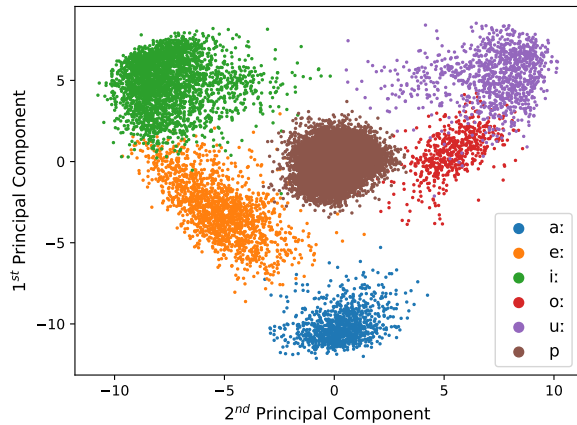


Figure 1: Scatter plot of vowel productions estimated over [a:], [i:] and [u:]. [e:] and [o:] are correctly reflected inside the vowel triangle. Productions of plosive [p] are closely scattered around the origin.

phones. Productions of elongated or palatalized variants were neglected, because they would heavily bias the PCA result. The decomposition was then estimated over an equal number of random samples drawn from each group. In the resulting plot, only the mean values of the PCA decomposition of every phone were visualized. For each plosive, an arrow was drawn from its mean value to the mean value of the fricative with equal or (if no equal could be found) very similar place and manner of articulation to highlight their relationship.

The aim of the last experiment was to demonstrate how the last CNN and the last transformer layer would encode information in very unique ways. In the first part, we selected the same three elongated vowels [a:], [i:] and [u:] as in the vowel triangle experiment. Two PCA decompositions were estimated, one for the CNN and one for the transformer FVs, by selecting an equal number of FVs per phone *and* per gender (to avoid a biased result). The resulting PCA mean values of both genders would shape a triangle, with one phone in each corner. The question at hand is whether there are differences between female and male speakers in FVs from the CNN or the transformer. Secondly, we performed a similar experiment only with the phone [a], the vowel that was found to be produced frequently among all six languages. Instead of balancing with respect to gender, FVs were balanced with respect to language, and afterwards visualized. The question remained the same: Are there differences between the results for the CNN and the transformer?

### 3. Results

Figure 1 illustrates realizations of vowels, projected into the 2-dimensional plane with a PCA estimated from [a:], [i:] and [u:]. To recreate the triangle often shown in the literature, the first principal component had to be plotted on the y-axis, the second on the x-axis. Additionally, both values had to be multiplied by  $-1$ , which effectively inverted the direction of the axis. While these operations may appear artificial, it was surprising to find that they exactly match with the properties of the original  $F1$ - $F2$  plane.  $F1$  is found on the y-axis of the triangle, and both  $F1$  and  $F2$  values are descending for increasing x and y values. The close-mid front vowel [e:] was correctly located below the fully closed [i:], the slight shift to the right indicated the lower  $F2$  frequency. The same was observed for the [o:],

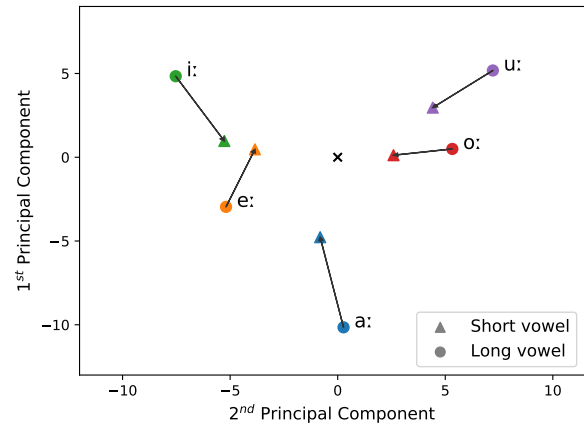


Figure 2: Mean PCA values of short and long vowel productions. The origin is marked with 'x'.

which was found below the [u:] and a little further towards the left (increasing  $F2$ ).

Figure 2 depicts the mean values of all elongated vowels and their short counterparts. As described in the literature, short variants tend to be located closer to the center of the triangle than the elongated ones.

Realizations of plosives and fricatives are shown in Figure 3. Several of the plosives, like [t], [k] or glottal stop [ʔ] were located very close to fricatives with equal or similar place and manner of articulation. For others, the distance was larger, but the path of the arrow depicting the correspondence was never interrupted by another phone class. Note that fricative [β] is also a valid correspondence of plosive [b], however, we did not include a second arrow because [β] and [v] are already very similar. Lastly, the fricatives that did not have any articulatory counterpart in the class of plosives were found mostly on the left side.

In the last experiment, FVs extracted from the CNN and transformer had been compared. In the first part, differences between female and male speakers were to be identified. The result is shown in Figure 4. Vowel productions from both groups ap-

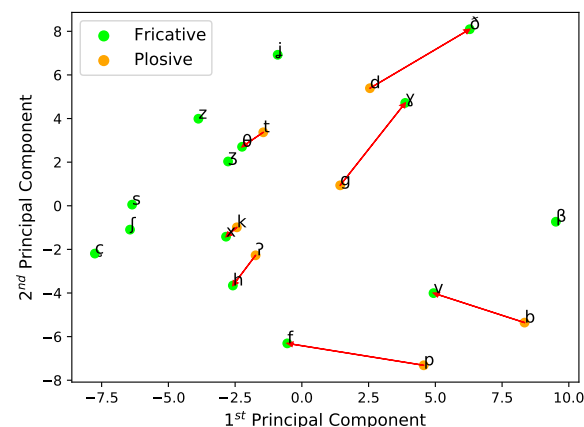


Figure 3: Mean PCA values of plosive and fricative realizations. The arrows from a plosive to a fricative phone indicate correspondences of equal or very similar place and manner of articulation.

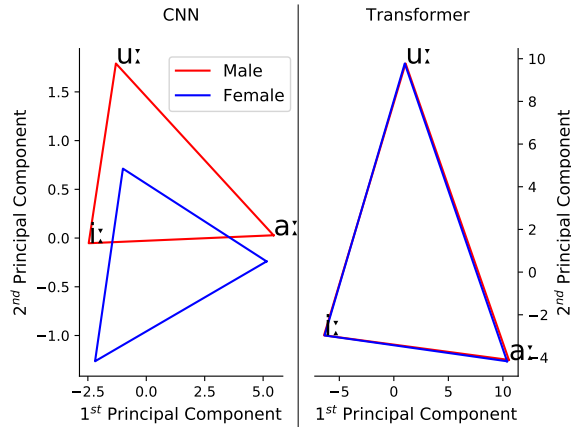


Figure 4: Mean PCA values of vowel realizations [a:], [i:] and [u:] from female and male speakers. Left side shows the results from CNN feature vectors, right side those from the transformer.

pear to show no difference with respect to the first two principal components of FVs from the transformer. At the same time, all FVs that were extracted at the same time steps, but from the last convolutional layer, seem to encode large amounts of gender information. This clearly shows that the dominance of a speaker's gender in the high-dimensional point cloud from the CNN FVs was reduced greatly by the following transformer layers.

If the transformer was able to reduce the importance of gender information, does it, at the same time, increase the impact of other parameters? Figure 5 shows the same arrangement of principal components computed for realizations of the open front vowel [a], clustered according to the respective language it was observed in. This time, the PCA result from CNN FVs of the different languages are very similar, and their mean values were all close to the origin. For the same realizations of [a], the transformer FVs encode plenty of language-related information, up to the extent that one could determine a linear decision boundary between certain languages.

#### 4. Discussion

Rediscovering the pattern of a vowel triangle from the first two principal components of vowel FVs was a clear indicator that the end2end model had learned to perform a frequency analysis on the raw waveform, and that formant information was found to be of significant help for successful vowel classification. The PCA even recovered the original (inverted) orientation of  $F1$  and  $F2$  on their respective axes. Furthermore, the transition of short vowels towards the center of the triangle, which had been covered in previous studies [19, 20], was preserved. Most consonants, as demonstrated with the plosive [p], would scatter very closely to the origin, as they did not contain much information along the directions of the first two principal components in the high-dimensional space. Small deviations from the origin were only observed for consonants which also included some voicing.

Along with formant information, knowledge about place and manner of consonant articulation was preserved decently in hidden layer activations from the transformer. The major difference between a plosive and a fricative that share place and manner of articulation lies in stream of air, which is constant in the latter, but interrupted and then released in the former. Such finding could be utilized in a medical setup, for example in dysarthria

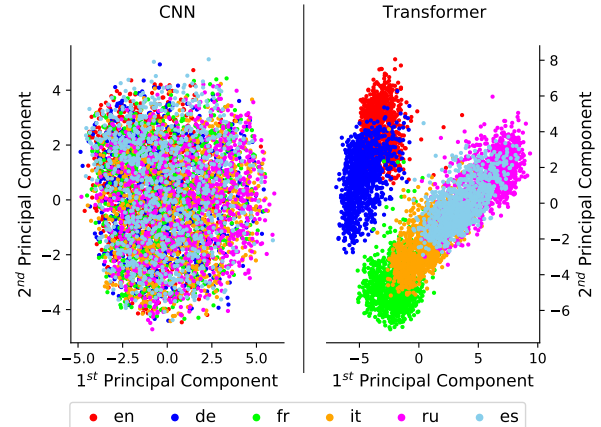


Figure 5: Realizations of vowel [a] in all of the six languages. Left side shows the results from CNN feature vectors, right side those from the transformer.

assessment, to evaluate articulation of patients. A model that learns the articulatory relationships between plosives and fricatives from healthy speech would not be restricted to the limited amount of pathological data.

The last experiment clearly showed that FVs extracted from the last CNN and the last transformer layer can encode information on an entirely different level. The recovery of gender information from the transformer output was basically impossible. This may be caused by the fact that gender information does not play any role in phone classification, hence, the transformer tries to diminish its influence as much as possible. The CNN result was strongly affected by gender information, which confirmed the assumption of the convolutional layers serving as a feature extractor. This theory received more support in the second part of the experiment. Putting it very simple, an [a] is an [a], no matter what language it was produced in. There may be slight differences, but, from a phonetic perspective, the realizations of the same phone (not phoneme!) in different languages should be very similar. This is exactly what was observed for the hidden outputs of the CNN feature extractor. According to this logic, temporal context was included in the transformer layers, and suddenly, the German [a] is very much different to the Spanish one. This is very likely caused not by the vowel [a] itself, but by the surrounding phones which ultimately shape phonotactic information. Thus, the PCA of hidden activations from the last transformer layer not only provide an acoustic model, but also comprise language model information.

#### 5. Conclusion

End2end models do not only yield state of the art results on acoustic modelling and ASR problems, but they also learn phonetic concepts such as a vowel triangle and the associated  $F1$ - $F2$  plane. Details about place and manner of articulation are also well preserved. These findings are particularly important because one does not have to explicitly search for the phonetic information, but it greatly affects the distribution of data in the high-dimensional space. If hidden activations are extracted from an end2end model, it is crucial to carefully select the correct layer. As our experiments have shown, feature extraction and temporal contextualization transform the entire information depending on the task that the model was optimized for.

## 6. References

- [1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *Proc. Interspeech 2019*, pp. 3465–3469, 2019.
- [2] 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. 12449–12460, 2020.
- [3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [4] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.
- [5] E. Saraee, M. Jalal, and M. Betke, “Visual complexity analysis using deep intermediate-layer features,” *Computer Vision and Image Understanding*, vol. 195, p. 102949, 2020.
- [6] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea, “Visualizing the hidden activity of artificial neural networks,” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 101–110, 2016.
- [7] P. Klumpp, T. Bocklet, T. Arias-Vergara, J. C. Vásquez-Correa, P. A. Pérez-Toro, S. Bayerl, J. R. Orozco-Arroyave, and E. Nöth, “The phonetic footprint of covid-19?” in *INTERSPEECH*, 08 2021, pp. 441–445.
- [8] P. Klumpp, T. Arias-Vergara, J. C. Vásquez-Correa, P. A. Pérez-Toro, J. R. Orozco-Arroyave, A. Batliner, and E. Nöth, “The phonetic footprint of parkinson’s disease,” *Computer Speech & Language*, vol. 72, p. 101321, 2022.
- [9] Y. Belinkov and J. Glass, “Analyzing hidden representations in end-to-end automatic speech recognition systems,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [10] T. Nagamine and N. Mesgarani, “Understanding the representation and computation of multilayer perceptrons: A case study in speech recognition,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 2564–2573.
- [11] N. T. Vu, J. Weiner, and T. Schultz, “Investigating the learning effect of multilingual bottle-neck features for asr,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [12] P. Klumpp, T. Arias-Vergara, P. A. Pérez-Toro, E. Nöth, and J. R. Orozco-Arroyave, “Common phone: A multilingual dataset for robust acoustic modelling,” *arXiv preprint arXiv:2201.05912*, 2022.
- [13] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.
- [14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [16] A. Graves, “Connectionist temporal classification,” in *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 61–93.
- [17] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [18] W. J. Hardcastle, J. Laver, and F. E. Gibbon, *The handbook of phonetic sciences*. John Wiley & Sons, 2012.
- [19] E. Nöth, *Prosodische Information in der automatischen Spracherkennung: Berechnung und Anwendung*. Walter de Gruyter, 2012, vol. 259.
- [20] M. Picard, “Vowel harmony, centralization, and peripherality: the case of pasiego.” *Linguistics*, vol. 39, no. 1, 2001.