# Measuring foreign accent strength using an acoustic distance measure

Martijn Bartelds[1], Wietse de Vries[1], Caitlin Richter[2], Mark Liberman[2], Martijn Wieling[1]

[1]*University of Groningen, the Netherlands*
[2]*University of Pennsylvania, United States of America*
m.bartelds@rug.nl, ricca@sas.upenn.edu, myl@cis.upenn.edu, m.b.wieling@rug.nl

## Abstract

*Pronunciations from different speakers are often compared using phonetic transcriptions, even though transcribing speech is time-consuming and error prone. To understand whether this process can be omitted when the goal is to quantify pronunciation differences, we investigate several acoustic-only methods for representing and comparing pronunciations. Specifically, we compute numerical feature representations based on Mel-frequency cepstral coefficients and a pre-trained Transformer-based neural network, and make word-level comparisons using Dynamic Time Warping. We use the speech of non-native and native speakers of English as input to these models, and evaluate the algorithms by comparing their output to human judgements of accent strength. Our results show that the Transformer-based approach outperforms the already well-performing transcription-based method, while being minimally affected by individual speaker differences. These results suggest that phonetically transcribing speech is not necessary to quantify pronunciation differences when a pre-trained Transformer-based neural model is available.*

**Keywords:** acoustic distance, mel-frequency cepstral coefficients, neural networks, pronunciation variation, speech.

## 1. Introduction

Phonetic transcriptions are frequently used to investigate and quantify pronunciation differences between speakers (Nerbonne and Heeringa 1997; Livescu and Glass 2000; Gooskens and Heeringa 2004; Heeringa 2004; Wieling, Bloem, et al. 2014; Chen et al. 2016; Jeszenszky et al. 2017). However, transcribing speech using a phonetic alphabet is time consuming, labor intensive, and variation might be the result of transcriber differences (Hakkani-Tür, Riccardi, and Gorin 2002; Bucholtz 2007; Novotney and Callison-Burch 2010). In addition, a set of discrete symbols may not be sufficient to include all phonetic details relevant for studying speech (Mermelstein 1976; Duckworth et al. 1990; Cucchiarini 1996). We therefore investigated the effectiveness of acoustic-only methods (not requiring phonetic transcriptions) to quantify pronunciation differences.

In this paper, we introduce two distinct acoustic-only methods to quantify pronunciation differences between non-native and native speakers of English.[1] Specifically, we transform unprocessed audio samples into numerical feature representations based on Mel-frequency cepstral coefficients (MFCCs) (Bartelds, Richter, et al. 2020) for the first method, and create self-supervised neural features from the pre-trained `wav2vec 2.0` model (Baevski et al. 2020) for the second method. Subsequently, we use dynamic time warping (DTW) (Müller 2007) to quantify the difference between the feature representations. We compare the resulting acoustic-only differences to phonetic transcription-based distances based on the same data, and human native-likeness judgments collected by Wieling, Bloem, et al. (2014) to assess whether our acoustic-only methods are a valid measurement technique. For reproducibility, our code is publicly available.[2]

Our work is related to that of Huckvale (2004) and Moustroufas and Digalakis (2007), who also investigated pronunciation differences without requiring phonetic transcriptions. However, these studies only considered a limited set of speech segments (i.e. vowels), or included speakers from a single language background, respectively. Moreover, for our second method, we take advantage of recent advances in neural (deep learning) approaches (Baevski et al. 2020; Kahn et al. 2020), which seem to improve over earlier neural techniques (e.g., Qian et al. 2017).

## 2. Materials

### 2.1. Datasets

We extracted audio samples from the Speech Accent Archive (Weinberger 2015) and from another study on Dutch accented English speech (Wieling, Blankevoort, et al. 2019). The Speech Accent Archive covers a wide variety of language backgrounds, where each speaker reads the same 69-word paragraph shown in Example (1).

(1) *Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*

In line with Wieling, Bloem, et al. (2014), we selected 280 speech samples of non-native English speakers from 99 different language backgrounds as our target speakers dataset, and 115 speech samples of U.S.-born L1 speakers as our reference native speakers dataset. The target speakers dataset was selected based on the perceptual evaluation data available (Wieling, Bloem, et al. 2014). We decided not to rely on a single reference native speaker, as the native English raters also had different regional backgrounds. For comparison, we also include the results of Wieling, Bloem, et al. (2014) regarding the

[1]This paper provides a short overview of two separate studies which discuss the MFCC approach, and the `wav2vec 2.0` approach, respectively: Bartelds, Richter, et al. 2020 and Bartelds, Vries, et al. 2020.

[2]MFCC-based model: `https://github.com/Bartelds/acoustic-distance-measure`. Neural model: `https://github.com/Bartelds/neural-acoustic-distance`

transcription-based difference.

The second dataset contains speech samples from native speakers of Dutch, and is therefore used to investigate whether our methods can also differentiate between smaller accent differences. These speakers were presented with the same paragraph used for the Speech Accent Archive, but they only read the first two sentences aloud. These recordings were collected at a science event held at the Dutch music festival Lowlands (Wieling, Blankevoort, et al. 2019). We included 62 speech samples of sober participants that exclusively had Dutch as their native language. For comparison, we phonetically transcribed the pronunciations and calculated the transcription-based differences (i.e. using the Levenshtein distance; Wieling, Bloem, et al. 2014) on the basis of these transcriptions.

### 2.2. Perceptual data

We evaluate our methods by using human ratings of native-likeness. We choose this evaluation procedure, because it is frequently used in previous work to evaluate accentedness in speech (Koster and Koet 1993; Munro 1995; Magen 1998; Munro and Derwing 2001).

The native-likeness ratings for the Speech Accent Archive speech samples were collected by Wieling, Bloem, et al. (2014). An online questionnaire was created where 1143 native U.S.-born speakers of English rated the accentedness of at most 50 speech samples on a 7-point Likert scale.

Native-likeness ratings for the speech recordings of the Dutch speakers data were provided by a different group of native English speakers (Offrede et al. 2020). In total, 115 participants rated the accent strength of target samples on a 5-point Likert scale using an online questionnaire similar to that of Wieling, Bloem, et al. (2014).

The native-likeness scores given to each target non-native speaker from both datasets are averaged. In this way, we obtain a single measure of similarity with native English speech for each non-native speaker.

## 3. Methods

### 3.1. Mel-frequency cepstral coefficients

For each target and reference audio sample from both datasets, we compute 39-dimensional MFCCs. These 39 dimensions include 12 cepstral coefficients and a single energy coefficient in each frame together with their first and second order derivatives. The coefficients are computed with a 25 ms sliding window and a stride of 10 ms. Speaker-based cepstral mean and variance normalization is used to reduce the influence of noise by applying a linear transformation to the coefficients of the MFCC feature representations.

### 3.2. Wav2vec 2.0

The `wav2vec 2.0` approach consists of a convolutional encoder, a Gumbel Softmax quantizer, and a 24 layer Transformer model. This model is trained as a single end-to-end model such that the encoder outputs are optimized for use in the Transformer. During pre-training on the large unlabeled Librispeech dataset (Panayotov et al. 2015), the objective was a contrastive task where the model had to predict spans of randomly masked frames with the full audio fragment as context. Representations can be extracted from the encoder (512 dimensions), the quantizer (768 dimensions), or the 24 Transformer layers (1024 dimensions). Previous work has found that Transformer layers can iteratively add information to the feature representations, while information may be lost in the final layers of the model (Tenney, Das, and Pavlick 2019). We therefore investigate the relation between the different Transformer layers, and subsequently select the best performing layer for extracting feature representations based on (a development set of) 25% of the Speech Accent Archive dataset.

### 3.3. Transcription-based approach

The phonetic transcriptions for the Speech Accent Archive dataset were obtained from Wieling, Bloem, et al. (2014). For the Dutch speakers dataset, the audio samples were phonetically transcribed by a single transcriber using the International Phonetic Alphabet. The phonetic transcription-based pronunciation differences are calculated using the adjusted Levenshtein distance algorithm of Wieling, Margaretha, and J. Nerbonne (2012), which is currently the best performing phonetic transcription-based method.

### 3.4. Quantifying pronunciation differences

We compute acoustic-only pronunciation differences by comparing the target non-native samples from both datasets to the 115 reference native samples. To include only comparable segments of speech, we automatically time-align the speech samples with a word-level orthographic transcription obtained from the Penn Phonetics Lab Forced Aligner (Yuan and Liberman 2008). After the extraction of the acoustic features, we use the time alignments to compare pronunciations of the same word (by two speakers) using DTW. The word-based pronunciation differences are then averaged to determine the pronunciation difference between a target non-native speaker and a reference native speaker. The difference between the pronunciation of a non-native speaker and native English speech is computed by averaging the pronunciation differences between the non-native speaker and set of native speakers from the reference dataset. The performance of our methods is evaluated by computing the Pearson correlation between the produced acoustic distances and the averaged human judgments of native-likeness for the target non-native samples. The transcription-based pronunciation difference from native English is measured similarly to the setup described for computing the acoustic-only pronunciation differences (but instead of DTW, the adjusted Levenshtein distance algorithm is used).

## 4. Results

### 4.1. Investigating transformer layers

We compute the Pearson correlation between the acoustic distances on the basis of the individual Transformer layers and the averaged human native-likeness ratings using the Speech Accent Archive development dataset. We evaluate the performance of the best performing Transformer layer on the full dataset, but excluding the development set (i.e. the test set) to prevent overfitting. These best performing layers are also used to compute the final results on the Dutch speakers dataset.

In Figure 1, we observe relatively stable correlations between layers 9 and 18 (ranging between $r = -0.85$ and $r = -0.87$), with the highest correlation for layer 10 ($r = -0.87, p < 0.001$). Evaluation of layer 10 on the test set (i.e. the full data, excluding the evaluation set) a shows a correlation of $r = -0.85$ ($p < 0.001$), which is highly similar to the correlations calculated on the development set.

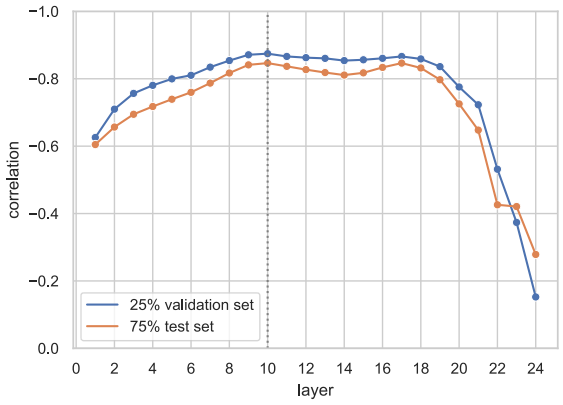


Figure 1: Pearson correlation coefficients between acoustic distances and human native-likeness judgments for different Transformer layers in the `wav2vec 2.0` model. The vertical dotted line marks the best performing layer chosen on the basis of the 25% development set of the Speech Accent Archive data.

### 4.2. Model performance

Table 1 shows the results of the different approaches. We observe that `wav2vec 2.0` achieves the best performance for the Speech Accent Archive dataset ($r = -0.85, p < 0.001$). For comparison with the other approaches, this correlation was computed on the basis of the complete dataset (i.e. including the evaluation set), as the correlation on the test set was not different from the results on the basis of the full dataset. The modified $z$-statistic of Steiger (1980) indicates that this correlation is a significant improvement over the results obtained using MFCCs ($z = 7.69, p < 0.001$) and the adjusted Levenshtein distance of Wieling, Bloem, et al. (2014) ($z = 4.31, p < 0.001$).

Results computed on the basis of the pronunciations of the native Dutch speakers also show a strong performance of `wav2vec 2.0` ($r = -0.70, p < 0.001$). For this dataset, we find a significant improvement over the use of MFCCs ($z = 3.78, p < 0.001$), but there is no (significant) difference compared to using the adjusted Levenshtein distance.

Table 1: Pearson correlation coefficients $r$ between human native-likeness ratings, and computed differences using MFCCs, `wav2vec 2.0`, and the adjusted Levenshtein distance, both for the Speech Accent Archive dataset (SAA) and native Dutch speakers dataset (DSD). All correlations are significant at the $p < 0.001$ level.

| Model | SAA | DSD |
|---|---|---|
| MFCCs | -0.71 | -0.34 |
| wav2vec 2.0 | **-0.85** | **-0.70** |
| Adjusted LD | -0.77 | -0.70 |

## 5. Discussion and conclusion

We investigated and compared two acoustic-only methods to compute pronunciation differences between non-native and native speakers of English. We evaluated the results by comparing the acoustic differences to phonetic transcription-based pronunciation differences and human judgments of native-likeness.

We found that a time-consuming and labor intensive phonetic transcriptions process can be omitted when a (language-specific) `wav2vec 2.0` model is available. Furthermore, we demonstrated that Transformer-based neural speech representations provide a more reliable and accurate representation of pronunciation variation compared to 39-dimensional MFCCs.

Even though `wav2vec 2.0` performed strongly on the Speech Accent Archive dataset, we found no significant improvement over a transcription-based difference method when using the native Dutch speakers dataset. While the pronunciations from this dataset were relatively similar, the human perceptual ratings were also provided using a less detailed (i.e. 5-point) Likert scale than the one that was used for the Speech Accent Archive accent ratings (i.e. 7-point Likert scale). This may have had an effect on distinguishing these two pronunciation difference measures regarding human performance.

While the architecture of `wav2vec 2.0` was designed for improving performance in the domain of transforming speech to text, we show that speech representations extracted from the hidden Transformer layers can also be successfully applied for investigating pronunciation variation. Despite our promising results, `wav2vec 2.0` models are currently only available for English. To create such models for other languages, large amounts of data and computing resources need to be available. To compute acoustic-only pronunciation differences when these are not available, the language-independent MFCC method can be used as an alternative. Future work should explore whether the existing `wav2vec 2.0` models could be adapted to other (low-resource) languages for modeling speaker variation.

## 6. Acknowledgments

The authors thank Hedwig Sekeres for creating the transcriptions of the Dutch speakers dataset.

## 7. References

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *arXiv preprint arXiv:2006.11477*.

Bartelds, Martijn, Caitlin Richter, Mark Liberman, and Martijn Wieling (2020). "A New Acoustic-Based Pronunciation Distance Measure". In: *Frontiers in Artificial Intelligence* 3, p. 39. DOI: 10.3389/frai.2020.00039.

Bartelds, Martijn, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling (2020). "Neural Representations for Modeling Variation in English Speech". In: *arXiv preprint arXiv:2011.12649*.

Bucholtz, Mary (2007). "Variation in transcription". In: *Discourse Studies* 9.6, pp. 784–808.

Chen, Nancy F, Darren Wee, Rong Tong, Bin Ma, and Haizhou Li (2016). "Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL". In: *Speech Communication* 84, pp. 46–56.

Cucchiarini, Catia (1996). "Assessing transcription agreement: methodological aspects". In: *Clinical Linguistics & Phonetics* 10.2, pp. 131–155.

Duckworth, Martin, George Allen, William Hardcastle, and Martin Ball (1990). "Extensions to the International Phonetic Alphabet for the transcription of atypical speech". In: *Clinical Linguistics & Phonetics* 4.4, pp. 273–280.

Gooskens and Heeringa (2004). "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data". In: *Language variation and change* 16.3, pp. 189–207.

Hakkani-Tür, Dilek, Giuseppe Riccardi, and Allen Gorin (2002). "Active learning for automatic speech recognition". In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. IEEE, pp. IV–3904.

Heeringa (2004). "Measuring dialect pronunciation differences using Levenshtein distance". PhD thesis. Citeseer.

Huckvale, Mark (2004). "ACCDIST: a metric for comparing speakers' accents". In: *Eighth International Conference on Spoken Language Processing*.

Jeszenszky, Péter, Philipp Stoeckle, Elvira Glaser, and Robert Weibel (2017). "Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German". In: *Journal of Linguistic Geography* 5.2, pp. 86–108.

Kahn, Jacob, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. (2020). "Libri-light: A benchmark for asr with limited or no supervision". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7669–7673.

Koster, Cor J and Ton Koet (1993). "The evaluation of accent in the English of Dutchmen". In: *Language learning* 43.1, pp. 69–92.

Livescu, Karen and James Glass (2000). "Lexical modeling of non-native speech for automatic speech recognition". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*. Vol. 3. IEEE, pp. 1683–1686.

Magen, Harriet S (1998). "The perception of foreign-accented speech". In: *Journal of phonetics* 26.4, pp. 381–400.

Mermelstein, Paul (1976). "Distance measures for speech recognition, psychological and instrumental". In: *Pattern recognition and artificial intelligence* 116, pp. 374–388.

Moustroufas, N and Vassilios Digalakis (2007). "Automatic pronunciation evaluation of foreign speakers using unknown text". In: *Computer Speech & Language* 21.1, pp. 219–230.

Müller, Meinard (2007). "Dynamic time warping". In: *Information retrieval for music and motion*, pp. 69–84.

Munro, Murray J (1995). "Nonsegmental factors in foreign accent: Ratings of filtered speech". In: *Studies in Second Language Acquisition* 17.1, pp. 17–34.

Munro, Murray J and Tracey M Derwing (2001). "Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate". In: *Studies in second language acquisition* 23.4, pp. 451–468.

Nerbonne and Heeringa (1997). "Measuring Dialect Distance Phonetically". In: *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonolby*. Association for Computational Linguistics (ACL), pp. 11–18.

Novotney, Scott and Chris Callison-Burch (2010). "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 207–215.

Offrede, Tom F., Jidde Jacobi, Teja Rebernik, Lisanne de Jong, Stefanie Keulen, Pauline Veenstra, Aude Noiray, and Martijn Wieling (2020). "The Impact of Alcohol on L1 versus L2". In: *Language and Speech*. DOI: 10.1177/0023830920953169.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (Apr. 2015). "Librispeech: An ASR Corpus Based on Public Domain Audio Books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Qian, Yao, Keelan Evanini, Xinhao Wang, Chong Min Lee, and Matthew Mulholland (2017). "Bidirectional LSTM-RNN for Improving Automated Assessment of Non-Native Children's Speech." In: *INTERSPEECH*, pp. 1417–1421.

Steiger, James H (1980). "Tests for comparing elements of a correlation matrix." In: *Psychological bulletin* 87.2, p. 245.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (July 2019). "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4593–4601. DOI: 10.18653/v1/P19-1452.

Weinberger, Steven (2015). *Speech accent archive*. URL: http://accent.gmu.edu/index.php.

Wieling, Martijn, Gerwin Blankevoort, Vera Hukker, Jidde Jacobi, Lisanne Jong, de, Stefanie Keulen, Masha Medvedeva, Mara Ploeg, van der, Anna Pot, Teja Rebernik, Pauline Veenstra, and Aude Noiray (Aug. 2019). "The influence of alcohol on L1 vs. L2 pronunciation". English. In: *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Australasian Speech Science and Technology Association Inc., pp. 3622–3626.

Wieling, Martijn, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne (2014). "Measuring foreign accent strength in English: Validating Levenshtein distance as a measure". In: *Language Dynamics and Change* 4.2, pp. 253–269.

Wieling, Martijn, Eliza Margaretha, and John Nerbonne (2012). "Inducing a measure of phonetic similarity from pronunciation variation". In: *Journal of Phonetics* 40.2, pp. 307–314.

Yuan, Jiahong and Mark Liberman (2008). "Speaker identification on the SCOTUS corpus". In: *Journal of the Acoustical Society of America* 123.5, p. 3878.