Low-resource Accent Classification in Geographically-proximate Settings: A Forensic and Sociophonetics Perspective

Qingcheng Zeng[1], *Dading Chong*[2], *Peilin Zhou*[1], *Jie Yang*[1]

[1]School of Public Health, Zhejiang University, Hangzhou, China
[2]Peking University, China

qingchengzeng@outlook.com, 1601213984@pku.edu.cn, zhoupalin@gmail.com, jieynlp@gmail.com

Abstract

Accented speech recognition and accent classification are relatively under-explored research areas in speech technology. Recently, deep learning-based methods and Transformer-based pretrained models have achieved superb performances in both areas. However, most accent classification tasks focused on classifying different kinds of English accents and little attention was paid to geographically-proximate accent classification, especially under a low-resource setting where forensic speech science tasks usually encounter. In this paper, we explored three main accent modelling methods combined with two different classifiers based on 105 speaker recordings retrieved from five urban varieties in Northern England. Although speech representations generated from pretrained models generally have better performances in downstream classification, traditional methods like Mel Frequency Cepstral Coefficients (MFCCs) and formant measurements are equipped with specific strengths. These results suggest that in forensic phonetics scenario where data are relatively scarce, a simple modelling method and classifier could be competitive with state-of-the-art pretrained speech models as feature extractors, which could enhance a sooner estimation for the accent information in practices. Besides, our findings also cross-validated a new methodology in quantifying sociophonetic changes.

Index Terms: Geographically-proximate accent classification, Accent modelling, Forensic phonetics, Sociophonetics

1. Introduction

Accents and dialects, with minor differences, are ubiquitous in human languages (we will use both terms interchangeably in this paper). Generally speaking, accent related research could be divided into accented speech recognition and accent classification.

Accented speech recognition usually refers to transcribing accented speech data into texts. [1] firstly adopted context independent Hidden Markov Model (HMM) units to classify five European accents by both accent group and gender group. In the machine learning era, various levels, ranging from acoustic, phonetic, phonotactic to prosodic features were used in machine learning models and managed to achieve lower Word Error Rate (WER). For example, [2] tested different levels of features on English, Arabic, and Portuguese comprehensively. The accent classification systems before ASR systems enhanced the performance by a relatively large degree. After [3] and [4], the transformer architecture and the pretraining became the new paradigm for natural language processing (NLP) tasks [5, 6] and the trend soon surged in speech technology research [7, 8, 9, 10, 11]. Recently, [12] introduced an unsupervised style embedding method and witnessed 14.8% WER reduction based on transformer architectures on AESRC2020 dataset [13]. Generally, accented speech recognition usually models the accent as an exterior variable and merges it into downstream neural networks afterwards.

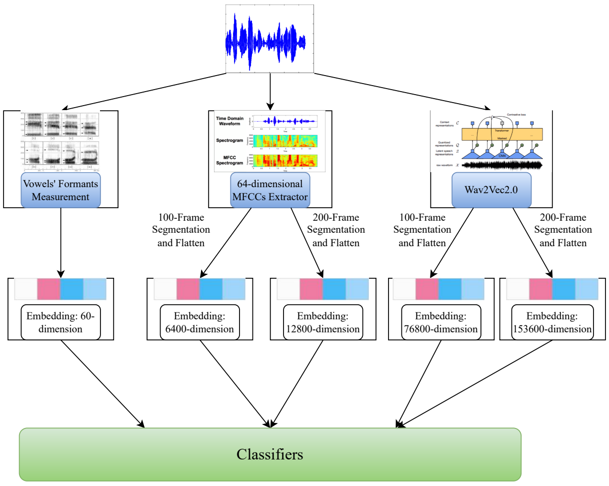
Time Domain Waveform
Spectrogram
MFCC Spectrogram
Vowels' Formants Measurement
64-dimensional MFCCs Extractor
Wav2Vec2.0
100-Frame Segmentation and Flatten
200-Frame Segmentation and Flatten
100-Frame Segmentation and Flatten
200-Frame Segmentation and Flatten
Embedding: 60-dimension
Embedding: 6400-dimension
Embedding: 12800-dimension
Embedding: 76800-dimension
Embedding: 153600-dimension
Classifiers

Figure 1: *The Overall Binary Classification Strategy*

Accent classification, which means classifying speech data into their corresponding accent categories, is relatively under-explored due to its indirect downstream applications. [14] explored foreign accented English classification with Mel-cepstrum coefficients, energy and first order differences. [15] tried to classify Arabic and Indian accents via Support Vector Machines (SVMs) based on MFCCs, which showed 75% to 97.5% accuracy with high precision and recall. [16] compared Gaussian Mixture Models (GMM), GMM-Universal Background Model (GMM-UBM), and i-vector in classifying Dravidian accented English. [17] used convolutional neural networks (CNNs) and Gated Recurrent Units (GRUs) to classify Ao's accents with approximately 6 hours of speech.

Forensic phonetics is a rapidly growing research field whose main research tasks consist of speaker identification and disputed utterance analysis [18]. Accent classification is one of the most important factors in speaker identification. [19] firstly proposed Y-ACCDIST modelling and used SVMs as the classifier to enhance accent classification in forensic applications. However, little attention was paid to forensic applications along with the development of speech technology.

Generally, speech technology in forensic applications encountered two main problems. Firstly, the scarcity of data. It is conceivable that in forensic scenarios, only short and fragmented raw speech data are available. Consequently, deep learning based methods are hard to be fully utilized. For example, [20] used 178 hours of raw speech data to improve bidirectional long short-term memory (BiLSTM) based recurrent neural networks' speech recognition performance. [19] could only utilize about 3 hours of data to classify geographically-proximate accents. Besides, almost all accent classification research focused on foreign accented English, such as Singapore accented English or Chinese accented English. Forensic applications tend to focus more on geographically-proximate accents on the contrary.

Sociophonetics explores language variations at a social scale [21]. The methods to quantify and compute language variations are relatively new and accent classification is a window to discover machine learning models' potential in sociophonetics research.

In this paper, we mainly combined three kinds of accent modelling methods, formant measurements, MFCCs, and wav2vec2.0 with two kinds of simple classifiers, logistic regression and multiple layer perceptrons (MLPs) to explore their accent classification performances at five urban varieties in Northern England. To summarize, our main contributions are:

1) To the best of our knowledge, we firstly utilized pretrained speech models as feature extractors in forensic-like applications.
2) We compared three accent modelling methods in accent classification and verified the competitiveness of traditional methods in low-resource settings.
3) Our results suggested the feasibility of traditional classification systems as a quick screening tool in forensic practices. Besides, our results also offered a potential methodology in detecting sociophonetic variations.

2. Data and Methods

2.1. Data

The data in this paper are from [22], which were extracted from the EDAC corpus in [23]. In total, 105 speakers were selected out from the five urban varieties, Leeds (N = 27), Liverpool (N = 17), Manchester (N = 23), Newcastle upon Tyne (N = 19), and Sheffield (N = 19), among which 59% are female and 67% had a higher education degree. The reading essay, "The Boy Who Cried Wolf", is a 216-word text including all English vowels. In total, 105 speakers have about 141 minutes' raw speech data. There are several advantages using this dataset. Its five adjacent urban varieties correspond well to forensic needs to classify fine-grained samples. Besides, the data were recorded via mobile phone applications which also aligns with forensic scenario characteristics.

2.2. Experimental Setup

The overall binary classification strategy is shown in Figure 1. Raw speech data are modelled by different methods and generated flattening embeddings. Accent embeddings are then processed by different classifiers to get the final results. We will discuss the experimental details in the following parts.

2.2.1. Feature Extraction and Segmentation

In this paper, we used three methods based on phonetic transcription, signal processing, and pretraining respectively, which basically covered the history of speech technology research.

Formant Measurements Formant modelling is a traditional phonetic method which only considers vowels. Its real-world applications has been verified extensively [24][25][26][27]. In this paper, firstly, speech data were forced-aligned via an HTK-based forced aligner [28]. Then the first two formants of each vowel were measured automatically, through phonetic analysis software, Praat [29]. Different strategies were adopted for monophthongs and diphthongs. For monophthongs, the formants were extracted at midpoints. For diphthongs, onglides and offglides were defined as 20% and 80% of the vowel duration. Formants were extracted at onglides and offglides. After automatic measurement and hand-correction, formants' data were z-scored to reduce the effects on models' coefficients because the second formant (F2 onwards) is bigger than the first formant (F1 onwards). In total, each speaker will be modelled by 60 features extracted from the dataset.

MFCCs MFCCs are one of the mostly used cepstral coefficients in signal processing, which are obtained as the inverse discrete cosine transform of the log energy in mel frequency bands. The concrete formula is shown below:

$$mfcc(t,c) = \sqrt{\frac{2}{M_{mfcc}}} \sum_{m=1}^{M_{mfcc}} \log\left(\widetilde{X}_m(t)\right) \cos\left(\frac{c(m-\frac{1}{2})\pi}{M_{mfcc}}\right) \quad (1)$$

where *t* represents the number of frames and *c* represents the dimension of MFCCs.

In this paper, MFCCs were extracted in 64 dimensions. Considering the high-dimension after modelling, we attempted to cut MFCCs into fixed number of frames, for example, 15, 100, and 200. On the one hand, data were augmented via dimensional reduction, from which we got about 45000 samples; on the other hand, this also helped to check the critical number of frames in the specific accent classification task.

Wav2vec2.0 Wav2vec2.0 [30] is a representative self-supervised model that outperformed semi-supervised methods in speech recognition on the Librispeech dataset [31]. It was trained in a similar fashion to BERT, masking a fixed proportion of time steps in the latent encoder space. Besides, its performances in speaker verification and language identification are also remarkable [32]. In this paper, we used *wav2vec-base* as our feature extractor to retrieve final layer output as speech representations for downstream classification tasks. Seemingly, we also did dimensional reduction via cutting representations into specific length of frames.

2.2.2. Classifiers

In this paper, we mainly used two classifiers, logistic regression and MLPs due to several reasons. Firstly, logistic regression has been tested to be one of the most robust linear classification methods. And MLPs is the simplest classifier which introduced non-linearity activation functions. Besides, both classifiers are relatively simple and can be trained in relatively quick speed, which aligns well with real-world forensic applications. At last, the results we tested on RNNs led to two intractable problems, easily overfitting and extremely slow training speed.

2.2.3. Classification Strategy and Evaluation Metrics

[22] and [17] both trained classifiers using one-versus-all classification strategy, which we are following along in this paper. Take the training of Liverpool as an example, 10% of the data will be selected out as the test objects. A training set with an equal number of Liverpool speakers and other cities' speakers will be generated for each sampling. Here, other cities' dataset

Table 1: *Results of Formant Measurements Modelling*

Model	Logistic Regression			MLPs		
Metric City	P	R	F1	P	R	F1
Leeds	68.54	69.37	68.95	69.23	69.11	69.17
Liverpool	89.56	91.88	90.70	89.33	89.29	89.31
Manchester	68.77	72.30	70.49	70.50	70.70	70.60
Newcastle	79.20	76.53	77.84	76.21	77.00	76.60
Sheffield	59.35	59.42	59.38	62.16	61.84	62.00

is under-sampled to create a relatively balanced training set. Finally, a classifier will be trained on this new dataset, and we tested the accuracy of the test objects. This training procedure was repeated one hundred times for each sampling with differently sampled training sets to guarantee our results' robustness.

We evaluated our models' performance based on three metrics, precision (P), recall (R) and F1 score (F1).

3. Results

Due to the limitation of space, the binary classification results under formant modelling, MFCCs + 200-frame segmentation, and wav2vec2.0 + 100-frame segmentation are listed in Table 1 to Table 3 respectively. In Table 4, we listed several experimental setups' average F1 scores which have relatively good performances.

The speech representations generated by wav2vec2.0-base, combined with 100-frame segmentation and logistic regression, performed best at 82.03 F1 score, followed by MFCCs and 200-frame segmentation at 80.88 and 81.02 F1 scores respectively. The traditional formant method have relatively mediocre performance at 73.47 and 73.54, slightly better than MFCCs and 100-frame segmentation settings. Several different numbers of the frames were tested in the experiments, 100-frame and 200-frame segmentation have best overall performance. The F1 scores' performance along with frame segmentation and the interaction between classifiers and frames will be further discussed in the next section.

The comparison across three different modelling methods showed that wav2vec2.0 performed best averagely, at 77.56 F1 score, followed by MFCCs' 75.05 and formant modelling's 73.50. Besides, wav2vec2.0 has a smaller standard deviation compared to MFCCs, which supported pretrained models' strong capacity in speech representations. However, given the fact that formant modelling did not undergo any "augmentation", its results are quite impressive as a whole.

Jumping into different cities' classification accuracy, on the one hand, there is a rather clear trend that MFCCs and wav2vec2.0 have more averagely distributed classification results than pure formant measurements. For example, Sheffield has the lowest classification accuracy in formant modelling. Among other setups, the highest accuracy for identifying Sheffield accent improved from 62.00 to 84.10, and even higher than Newcastle accent which has high accuracy previously. On the other hand, formant modelling could identify Liverpool accents incredibly well, whose accuracy is never outperformed in following experiments. These results reflect different foci in accent modelling methods.

Table 2: *Results of MFCCs Modelling + 200 Frames Segmentation*

Model	Logistic Regression			MLPs		
Metric City	P	R	F1	P	R	F1
Leeds	69.04	68.09	68.56	80.40	79.37	79.88
Liverpool	72.39	75.39	73.86	83.50	85.50	84.49
Manchester	67.38	63.94	65.61	80.26	83.14	81.67
Newcastle	69.10	71.51	70.28	78.82	79.27	79.04
Sheffield	71.34	74.33	72.80	77.45	82.72	80.00

Table 3: *Results of wav2vec2.0 Modelling + 100 Frames Segmentation*

Model	Logistic Regression			MLPs		
Metric City	P	R	F1	P	R	F1
Leeds	80.45	82.69	81.55	71.75	71.04	71.39
Liverpool	86.11	86.87	86.49	76.66	85.37	80.78
Manchester	76.61	78.62	77.60	68.53	72.48	70.45
Newcastle	80.20	80.58	80.39	70.52	71.34	70.93
Sheffield	82.99	85.25	84.10	77.21	79.58	78.38

4. Discussion

In this section, we will discuss accent modelling methods, machine learning classifiers, frame segmentation, and the methods' potential insights in language variation and sociophonetics.

4.1. Accent Modelling

The three accent modelling methods we mainly explore showed different performances in the order of wav2vec2.0, MFCCs, and formant measurements. Recalling the modelling methodology, we only measured vowels' F1s and F2s as our feature engineering output, without any consonants' information. On the contrary, wav2vec2.0 and MFCCs took the whole speech data into consideration and better results could be expected. The differences in F1 scores are mainly at cities with high accuracy and low accuracy. For example, MFCCs managed to improve Sheffield's F1 score by over 13 points. However, at the same time, the previous highest F1 score in recognizing Liverpool dropped at about 17 points. Generally, the results using wav2vec2.0 followed the similar trends to MFCCs.

[22] stated the general trend of dialect leveling at Northern England and the fact that Liverpool and Newcastle are the most distinct two accents compared to other accents. And random forests' classification results verified this finding well. In this paper, we partially verified the prior knowledge from MFCCs and wav2vec2.0's perspective. Among all experimental settings, models could classify Liverpool accents best. However, the accuracy in identifying Newcastle accents could not apparently outperform others, which indicates that Newcastle's consonantal information might hide it among other accents. Besides, formant modelling's extraordinary performance in identifying Liverpool indicated that formant modelling is quite useful in filtering accent outliers.

To sum up, our results successfully find out a general function for three accent modelling methods and suggest an elementary workflow in forensic accent classification. Firstly, we should delimit several candidate accents awaiting for classification. Then, we should consider using formant modelling to select out the most distinct accent according to requests. Afterwards, more comprehensive accents modelling methods should be adopted and classify the accents as a quick screening tool. At last, we should go into details, analyze its acoustic features accordingly to finalize our judgments.

Table 4: *Different Experimental Setups' Average F1 Score under Binary Classification*

Experimental Setup	Average F1 Score
Formant + LR	73.47
Formant + MLPs	73.54
MFCCs + 100 Frames + LR	68.07
MFCCs + 200 Frames + LR	80.88
MFCCs + 100 Frames + MLPs	70.22
MFCCs + 200 Frames + MLPs	81.02
wav2vec2.0 + 100 Frames + LR	**82.03**
wav2vec2.0 + 200 Frames + LR	76.58
wav2vec2.0 + 100 Frames + MLPs	74.37
wav2vec2.0 + 200 Frames + MLPs	77.26

4.2. Classifiers

Generally, MLPs have better performance than its counterparts, which aligns with our intuition. The only exception is the wav2vec2.0 and 100-frame segmentation setting with the highest overall performance.

As the dimensions increase, the logistic regression models' generalization ability decrease due to its linear quality. We tested logistic regression using both MFCCs and wav2vec2.0 without doing any segmentation, which leads to a over 620000-dimension vector. The average F1 scores are dropping to less than 40. Although logistic regression is quite robust in most cases, when it comes to high-dimensional computing, MLPs are a better choice with relatively stable generalization ability. The turning point between logistic regression and MLPs is basically a rule of thumb. In our settings, we can basically say 100-frame fits better to logistic regression and 200-frame fits better to MLPs. As single dimension is increasing, more advanced classifiers should be adopted.

In this paper, we only classified accents via flattening the feature vectors into linear arrays rather than high-dimensional arrays. However, sequential information and interaction is crucial in speech information. Future work should consider using CNNs and RNNs as classifiers with advanced data augmentation techniques. Besides, a comprehensive database including more geographically-proximate accents will definitely benefit training neural networks and fine-tuning pretrained speech models.

4.3. Frame Segmentation

In the results part, we only showed 100-frame and 200-frame segmentation because of their relatively high performance. Figure 2 witnessed the average F1 score changes along with different lengths' frame segmentation. Averagely, 200 frames, which approximately equals to 10 phonemes, could classify the accents best in simple classifiers' settings.

Segmenting raw speech data into fixed length of frames is a disguised way of data augmentation. Compared with formant modelling which can not make full use of data and have to encounter multicollinearity problem, segmenting according to frames both improved accuracy and generalization ability. In practices, we should consider using it as a regular way to enhance classification models.

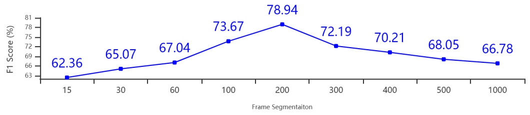

Figure 2: *Average F1 Score Changes along with Different Frames of Segmentation*

4.4. Sociophonetic Changes

Quantifying sociophonetic changes automatically and discovering potential changes are rarely explored topics in computational sociolinguistics. The core difficulty lies at the trade-off between accuracy and machine learning models' interpretability. [22] firstly tried to quantify sociophonetic changes via random forests. Although they argued that machine learners and human learners were following different patterns in classification, it was acceptable to use machine learning models as an exploratory tool in discovering changes. However, when it comes to logistic regression, the strong multicollinearity between formant variables made the interpretability highly fluctuant. We tried to extract the top ten most important variables from logistic regression models and compared with those from random forests. There is just one common variable. Although this variable was not seen in previous literature, further investigation is still needed to verify its real-world authenticity.

[22] stated another hypothesis that the classification accuracy could reflect the degree of dialect leveling. For example, the results in Table 1 implied that Liverpool and Newcastle accents are distinct ones and the rest three accents are undergoing dialect leveling. Our accent modelling methods which take overall information improved the Sheffield accent by a lot, implying that besides vowels' formants, Sheffield accents' overall information made it stand out among five accents. Previously, most variationist sociolinguistics research focused on vowels' features due to its relative scarcity and easy-to-measure acoustic features. MFCCs and pretrained models should make a comprehensive detection possible. In this paper, our experiments successfully verified the robustness of speech representations extracted from pretrained speech models, which corresponds to [33]. And our results further verified the feasibility of frame segmentation. This suggests that sociophonetic research should consider diving into fine-grained segments and pretrained models. Furthermore, we could compare across speakers in specific dimensions and even locate changes more accurately.

5. Conclusions and Future Work

In this paper, we explored geographically-proximate accent classification tasks, firstly utilized pretrained speech models in forensic-like settings and verified their robustness, compared across different accent modelling methods and classifiers, and proposed a general forensic accent classification workflow. We hope our work could encourage the utilization of advanced speech technology in forensic phonetics research.

As for the future work, previous computational linguistics research [34] showed BERT encoded different levels of features at various layers. [33] initially showed that Transformer-based models could capture intonational and durational differences better but further investigation towards layers is still needed. Besides, pretrained models' potentials in fine-tuning have not been fully utilized. We are looking forward to seeing the enhancement of accent classification and speech representations after fine-tuning on a medium amount of speech data.

6. References

[1] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1784–1787.

[2] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*. Columbia University, 2011.

[3] 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.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] N. Chen, F. Liu, C. You, P. Zhou, and Y. Zou, "Adaptive bidirectional attention: Exploring multi-granularity representations for machine reading comprehension," in *ICASSP*, 2020.

[6] F. Liu, X. Wu, C. You, S. Ge, Y. Zou, and X. Sun, "Aligning source visual and target language domains for unpaired video captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] C. You, N. Chen, and Y. Zou, "Contextualized attention-based knowledge transfer for spoken conversational question answering," in *INTERSPEECH*, 2021.

[8] ——, "Self-supervised contrastive cross-modality representation learning for spoken question answering," in *Findings of the Association for Computational Linguistics: EMNLP*, 2021.

[9] W. Xu, P. Zhou, C. You, and Y. Zou, "Semantic transportation prototypical network for few-shot intent detection," in *INTERSPEECH*, 2021.

[10] N. Chen, C. You, and Y. Zou, "Self-supervised dialogue learning for spoken conversational question answering," in *INTERSPEECH*, 2021.

[11] C. You, N. Chen, F. Liu, S. Ge, X. Wu, and Y. Zou, "End-to-end spoken conversational question answering: Task, dataset and model," *arXiv preprint arXiv:2204.14272*, 2022.

[12] S. Li, B. Ouyang, D. Liao, S. Xia, L. Li, and Q. Hong, "End-to-end multi-accent speech recognition with unsupervised accent modelling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6418–6422.

[13] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," *CoRR*, vol. abs/2102.10233, 2021. [Online]. Available: https://arxiv.org/abs/2102.10233

[14] L. M. Arslan and J. H. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[15] C. Pedersen and J. Diederich, "Accent classification using support vector machines," in *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*. IEEE, 2007, pp. 444–449.

[16] R. K. Guntur, K. Ramakrishnan, and M. Vinay Kumar, "An automated classification system based on regional accent," *Circuits, Systems, and Signal Processing*, pp. 1–21, 2022.

[17] M. Tzudir, S. Baghel, P. Sarmah, and S. M. Prasanna, "Excitation Source Feature Based Dialect Identification in Ao — A Low Resource Language," in *Proc. Interspeech 2021*, 2021, pp. 1524–1528.

[18] M. Jessen, "Forensic phonetics," *Language and linguistics compass*, vol. 2, no. 4, pp. 671–711, 2008.

[19] G. Brown, "Y-accdist: An automatic accent recognition system for forensic applications," Ph.D. dissertation, University of York, 2014.

[20] W. Rao, J. Zhang, and J. Wu, "Improved blstm rnn based accent speech recognition using multi-task learning and accent embeddings," in *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, 2020, pp. 1–6.

[21] J. Hay and K. Drager, "Sociophonetics," *Annu. Rev. Anthropol.*, vol. 36, pp. 89–103, 2007.

[22] P. Strycharczuk, M. López-Ibáñez, G. Brown, and A. Leemann, "General northern english. exploring regional variation in the north of england with machine learning," *Frontiers in Artificial Intelligence*, p. 48, 2020.

[23] A. Leemann, M.-J. Kolly, and D. Britain, "The english dialects app: The creation of a crowdsourced dialect corpus," *Ampersand*, vol. 5, pp. 1–17, 2018.

[24] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol. 12, no. 2, pp. 143–173, 2005.

[25] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and gaussian mixture models," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[26] M. Jessen, "MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis," in *Proc. Interspeech 2021*, 2021, pp. 411–415.

[27] J. J. Lo, "Cross-Linguistic Speaker Individuality of Long-Term Formant Distributions: Phonetic and Forensic Perspectives," in *Proc. Interspeech 2021*, 2021, pp. 416–420.

[28] J. Yuan and M. Y. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, pp. 3878–3878, 2008.

[29] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.2.10)," 2022. [Online]. Available: http://www.praat.org

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[31] 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)*, 2015, pp. 5206–5210.

[32] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.

[33] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, p. 101137, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447022000122

[34] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. [Online]. Available: https://aclanthology.org/P19-1356