Exploring Speaker Voice Similarity Across British Accents Using WavLM

Bahne J. Thiel-Peters

January 14, 2025

A Introduction

Speaker identification is a fundamental task in speech technology, aiming to determine a speaker's identity based on their voice. This capability has broad applications in areas such as security, personalization, and automatic transcription. Modern advances in this field often rely on pre-trained speech models, such as WavLM, which uses self-supervised learning to extract high-quality speech embeddings. These embeddings encode speaker-specific and linguistic characteristics, enabling effective downstream tasks like speaker verification, diarization, and similarity analysis.

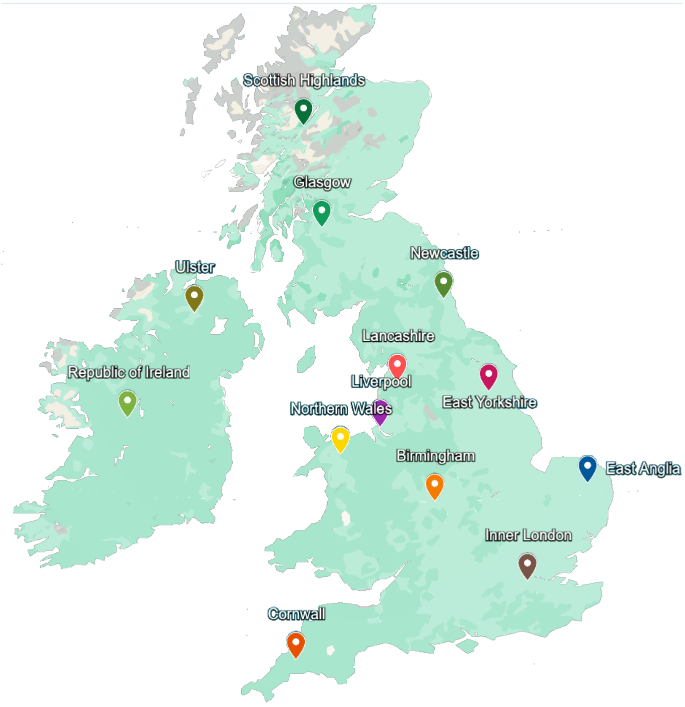

Figure 1: A geographical representation of the accents included in the ABI dataset.

The English language is spoken across a wide geographical area and exhibits significant variation in accents. These accents can be classified in many ways based on linguistic, cultural, or geographical criteria. One such classification is illustrated in the map shown in Figure 1, which highlights the regional diversity of accents across the British Isles. Accents are not just markers of geographical origin but also reflect intricate variations in pronunciation, rhythm, and intonation, posing unique challenges for speaker identification systems. The objective of this study is to evaluate the performance of the WavLM model in analyzing speaker similarity. By assessing the model's ability to capture subtle differences in speech characteristics across speakers and accents, this work aims to identify the strengths, limitations, and potential biases of pre-trained speech models in speaker identification tasks. The code and detailed implementations for this study are available on this GitHub repository.

B Dataset and Exploratory Data Analysis (EDA)

The dataset utilized in this study is **The Accents of the British Isles (ABI)** corpus, as described by S. M. D'Arcy et al. (2004). This corpus comprises approximately *95 hours* of audio data, categorized into *14 distinct British accents*. Each speaker within the dataset recorded the same texts under identical technological and environmental conditions, ensuring that the primary variations in the audio recordings arise from the accents themselves. The recordings include both short and long phrases. However, this paper focuses exclusively on the *short phrases*, which are particularly suitable for the analysis of the similarity of the intended speaker. The dataset features recordings from speakers aged between *18 and 50 years*, providing a relatively consistent demographic range. This structured setup makes the ABI corpus a robust resource for analyzing speaker voice similarity and evaluating accent-based differences.

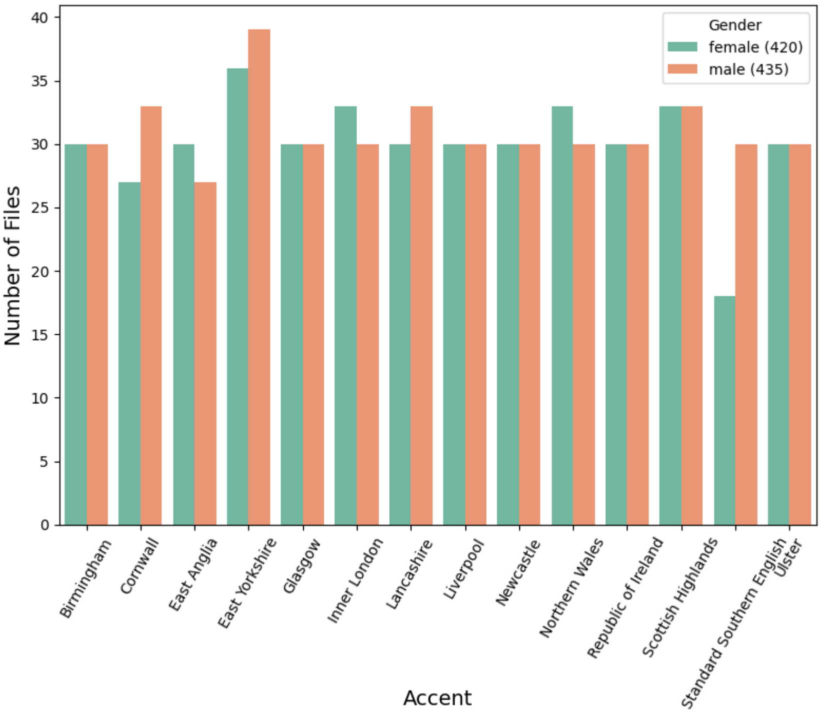

Figure 2: A column chart showing the number of audio files for each accent, categorized by gender.

Figure 2 visualizes the data distribution of audio files across different *accents* and *genders* in the dataset. The corpus includes a total of 855 audio files, with approximately *50.8772%* of the recordings coming from male narrators, reflecting a nearly even gender distribution overall. This uniformity supports balanced analyses across genders. On average, each accent contains *61 audio files*. For most accents, the difference in the number of recordings between female and male speakers does not exceed 5, ensuring a relatively equal representation. However, the Standard Southern English accent shows a notable imbalance, with only 18 female and 30 male recordings—a difference of *approximately 40%*. *East Yorkshire* stands out with the highest number of recordings, totaling *75 audio files*. This makes it particularly prominent in the dataset, offering more robust data for analysis within this accent group. The dataset's structured and balanced nature across accents and genders provides a strong foundation for analyzing speaker voice similarity, with slight deviations offering potential insights into specific groups.

Figure 3 illustrates the distribution of *audio durations*, in seconds, for each accent in the dataset. The median duration for most accents lies around *40–55 seconds*, indicating a consistent length across recordings. *East Yorkshire* exhibits the largest variability in audio duration, with several outliers extending beyond *140 seconds*, making it a notable exception compared to other accents. Conversely, accents such as *Cornwall and Lancashire* display more compact distributions, with minimal variability in duration. The presence of *outliers* in several accents (e.g., East Yorkshire, Northern Wales, and Republic of Ireland)

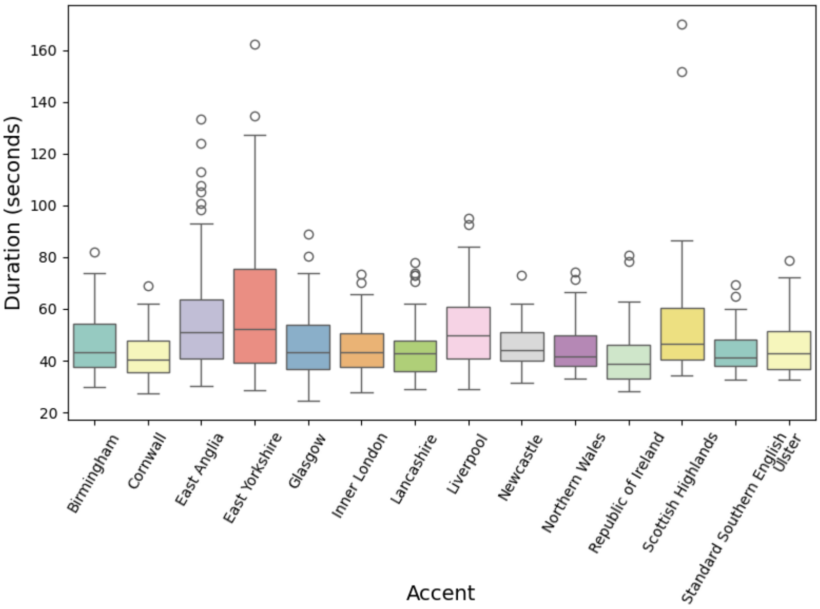

Figure 3: A boxplot showing the distribution of audio durations for each accent.

reflects occasional recordings with significantly longer durations. These variations could introduce challenges like ensuring uniformity in the analysis part. Overall, the dataset's duration distribution remains relatively balanced, supporting consistent evaluation across accents, while providing insights into unique cases with higher variability.
The *linguistic aspects* of the dataset have been explored in prior research, such as the study by S. D'Arcy and Russell (2008) and Ferragne and Pellegrino (2010). The linguistic insights from such studies offer valuable context for understanding the variability and distinctiveness of the accents within the dataset.

C Data Preprocessing

The preprocessing steps were minimal, as the dataset's audio recordings were already clean and consistent due to the standardized recording conditions. To ensure compatibility with the WavLM model, which was trained on *16kHz* data, all audio files were resampled to 16kHz. Additionally, *mono audio* was enforced for the same reason. Given the high quality and uniformity of the recordings, no further preprocessing was deemed necessary, as this setup is expected to yield optimal results for the speaker similarity evaluation.

D Methodology

This section outlines the computational methods and techniques used to analyze speaker voice similarity. It describes the processes involved in extracting and aggregating embeddings, calculating similarity scores, and the tools and metrics used for evaluating the model's performance.

D.1 Embeddings

In speech processing, an *embedding* is a high-dimensional vector representation that corresponds to a single audio input. This vector captures the essential acoustic and linguistic features of the audio signal, effectively encoding the unique characteristics of the speech within that specific recording. This process transforms the complex information contained in the raw audio into a structured and compact format, facilitating various speech analysis tasks by providing a quantifiable representation of each audio sample.

D.2 WavLM and Embedding Extraction

WavLM is an advanced *self-supervised* learning model designed to handle a wide range of speech processing tasks, from recognition to speaker identification. The model leverages a Transformer encoder architecture equipped with a gated relative position bias, enhancing its ability to capture sequential relationships in speech data. The core component of WavLM is its ability to process *masked speech inputs*, simulating noisy environments or overlapped speech, which trains the model to effectively extract clean speech representations from complex acoustic scenes.

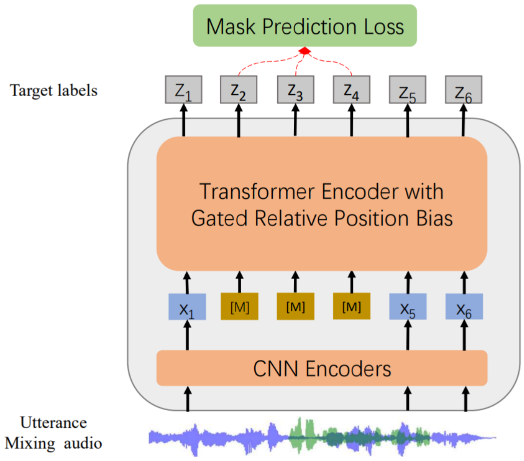

Figure 4: WavLM's Model Architecture (Chen et al., 2022)

As shown in Figure D.2, Embeddings are extracted by passing audio signals through a series of convolutional neural network (CNN) encoders that preprocess the signal into a form suitable for the Transformer encoder. The model processes masked segments of the audio signal, predicting the masked parts, which forces the model to develop a deep, contextual understanding of the speech content and speaker characteristics (Chen et al., 2022). Each audio file is processed individually to extract a *512-dimensional embedding*. These embeddings are directly used in subsequent analyses without aggregation. This approach ensures that each recording is treated as an independent sample.

D.3 Cosine Similarity

Once *embeddings* are extracted, similarity between speaker voices is computed using *cosine similarity*. This metric measures the cosine of *the angle between two vectors*, providing a value that indicates how similar the vectors are in terms of direction, regardless of their magnitude. The cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is calculated using the following formula:

$$CosineSimilarity(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$$

Where $\mathbf{u} \cdot \mathbf{v}$ is the dot product of the vectors $\mathbf{u}$ and $\mathbf{v}$, and $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$ are the Euclidean norms (or magnitudes) of the vectors (Miesle, 2023). This formula effectively captures the orientation similarity between embeddings, making it particularly useful for speaker identification tasks, as it highlights how phonetically or acoustically similar two voices are.

D.4 Dimensionality Reduction

To facilitate visualization and further analysis, *dimensionality reduction* is performed using *t-SNE with a perplexity setting of 20*. This technique reduces the high-dimensional space of embeddings down to two dimensions, preserving the relative distances between points as much as possible (Pajak, 2023).

D.5 Experimental Setup and Evaluation Pipeline

The *experimental pipeline* was developed using *Python 3.10.12*, the *HuggingFace Transformers* library *version 4.47.1*, and *PyTorch version 2.5.1*, and executed on an *NVIDIA T4 GPU*. This setup ensured robust computational performance necessary for processing large-scale speech data. The audio samples were pre-processed as detailed in Section C. The **evaluation pipeline** involved comparing embeddings to analyze the model's performance in maintaining consistent representations for individual speakers across different recordings and in differentiating among the vocal features of various speakers. The *cosine similarity* scores from these comparisons were used to determine an *optimal threshold* that maximized accuracy and minimized misclassifications. As described in Section D.4, *dimensionality reduction using t-SNE* was applied to the embeddings to facilitate visualization and further analysis, preserving relative distances between data points effectively.

To quantitatively evaluate model performance, the following metrics were computed:
Accuracy: The percentage of correct predictions (same/different speaker) across all comparisons.
Precision: This metric evaluates the proportion of true positive predictions out of all positive predictions made by the model. It indicates how many of the predicted "same speaker" pairs were actually correct, thus highlighting the model's ability to avoid false positives.
Recall: This metric measures the proportion of true positive predictions out of all actual positive instances. It reflects the model's ability to identify "same speaker" pairs correctly, emphasizing the avoidance of false negatives.
F1-Score: A harmonic mean of precision and recall, providing a single measure of performance.
Confusion Matrix: Visualized the classification performance, highlighting true positives, false positives, true negatives, and false negatives.

The setup was iteratively refined, with parameters and thresholds adjusted based on preliminary results to achieve optimal performance on the dataset.

E Evaluation and Results

This section presents the evaluation of the model's performance using threshold-based metrics and visualizations to demonstrate its efficacy in distinguishing speakers and analyzing voice similarities across accents.

E.1 Threshold-Based Evaluation Metrics

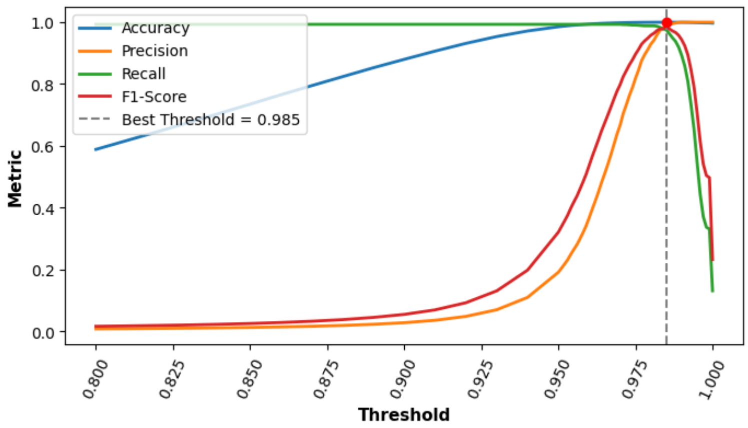

Figure 5: The relationship between the threshold and accuracy, precision, recall, and F1-score.

Figure 5 illustrates the relationship between the *threshold* and various evaluation metrics, including *accuracy, precision, recall, and F1-score.* The accuracy metric steadily increases as the *threshold* rises, peaking at a value of *0.999876* at a **threshold of 0.985**. At this same threshold, the *F1-score* reaches its optimal value of *0.982449*, driven by a balanced trade-off between precision and recall. In the early stages, precision is notably low because the model predicts ”same speaker” frequently, leading to a high number of *false positives.* The precision curve begins to improve significantly around a threshold of 0.94, which serves as its turning point. On the other hand, recall experiences a sharp decline at higher thresholds. This drop occurs because the model becomes more restrictive, predicting ”same speaker” less often, which increases the number of *false negatives.* The F1-score reflects this dynamic balance between precision and recall, highlighting the optimal performance point near the best threshold.

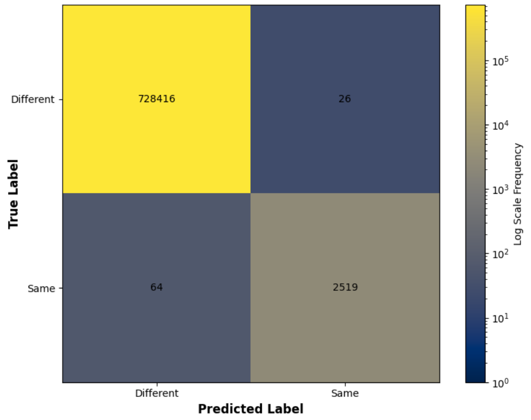

Figure 6: A confusion matrix depicting the model’s classification performance at the optimal threshold.

To better understand the model’s performance at the best threshold, Figure 6 presents a confusion matrix. At a threshold of 0.985, the matrix reveals an overwhelming majority of correct classifications for both ”different speakers” (728,416 true negatives) and ”same speakers” (2,519 true positives). However, some errors persist, including 64 false negatives and 26 false positives. These results demonstrate the model’s effectiveness in distinguishing speakers while emphasizing the trade-offs between precision and recall at varying thresholds. Overall, the combination of the evaluation metrics and confusion matrix underscores the robust performance of the model, particularly at the **optimal threshold of 0.985**.

E.2 Visualization of Speaker Similarity Across Accents

Figure 7 visualizes the *cosine similarity* between speaker *embeddings* across different accents and genders in a **heatmap**. The matrix is symmetric, reflecting the commutative property of cosine similarity. The diagonal of the matrix shows perfect similarity (value of 1.0) since each embedding is compared to itself. The plot is ordered by accents, with each accent divided into two consecutive blocks: the first block representing female speakers and the second block representing male speakers. For example, the upper line of squares corresponds to the similarity of female Glasgow speakers with all other accents, where odd-indexed blocks represent females and even-indexed blocks represent males. The heatmap highlights some key trends:

- **Inter-Accent Similarities**: Greenish regions indicate moderate similarities between accents. For instance, *Lancashire, East Yorkshire, and Birmingham* male audio recordings show relatively higher similarity to female audio recordings from other accents, visible in the second row of these accents’ blocks when compared to odd-indexed columns.

- **Notable Accent Trends**: Male speakers from East Anglia exhibit relatively high similarity across several accents, making this group stand out as more uniform in

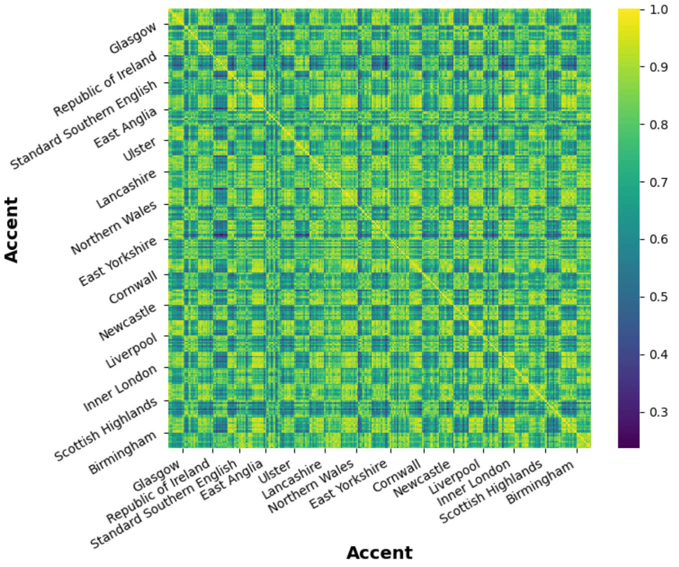

Figure 7: A heatmap visualizing the cosine similarity of speaker embeddings across accents and genders.

their embeddings.

- **Geographical Patterns**: *Ulster and Republic of Ireland* accents show a particularly high similarity for both genders, which may stem from their geographical proximity and shared linguistic traits.
- **Gender-Based Trends**: Overall, *female accents* tend to have higher similarity to other accents compared to males, suggesting a potential trend of more distinct male voice characteristics in the embeddings.
- **Unusual Similarity Patterns**: Certain samples exhibit consistently low similarity (blue areas) or unusually high similarity (green areas) to most other samples. These anomalies may be attributed to factors such as significant age differences among speakers or atypical voice pitches.

This heatmap provides valuable insights into accent-specific trends and gender influences, while also revealing potential outliers that may require further investigation.

E.3 Speaker Embedding Visualization

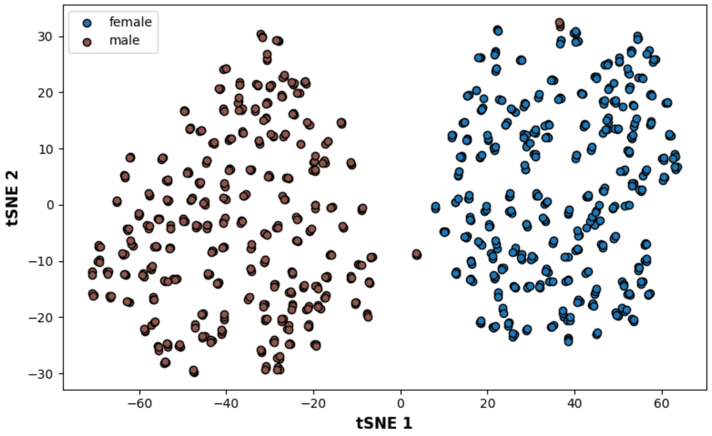

Figure 8: A t-SNE plot showing speaker embeddings grouped by gender.

A t-SNE visualization (D.4) is shown in figure 8 and demonstrates embeddings *grouped by gender*: male and female. The plot shows clear separability between the two genders, with distinct clusters for male and female speakers. Female embeddings appear slightly more compact than male embeddings, which may explain the **gender-based trends** observed in the heatmap (7). This indicates that the model captures gender-specific vocal characteristics, such as pitch and timbre, effectively. The lack of significant overlap between the two clusters highlights the model's robustness in differentiating between genders. However, since gender is a broad attribute, further analysis on speaker-level characteristics is needed to assess its precision for more granular tasks.

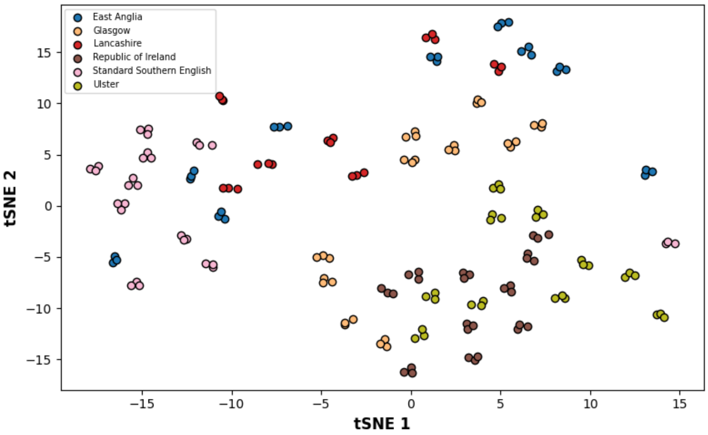

Figure 9: A t-SNE plot showing speaker embeddings grouped by accents.

To better differentiate between speaker embeddings, Figure 9 focuses exclusively on male speakers and visualizes embeddings for six different accents. The embeddings for *Ulster* and *Republic of Ireland* are highly interspersed, showing significant overlap. This may be attributed to their **geographical proximity** and shared linguistic traits, as previously noted in the heatmap analysis E.2. *East Anglia's* Embeddings are widely distributed across the entire plot, suggesting that this accent may be more **general** and less distinct compared to others. This could explain why *East Anglia* showed relatively high similarity to multiple accents in the heatmap. In contrast, the embeddings for *Standard Southern English* form a tightly clustered group, indicating a high degree of **internal consistency** and distinctiveness compared to the other accents.

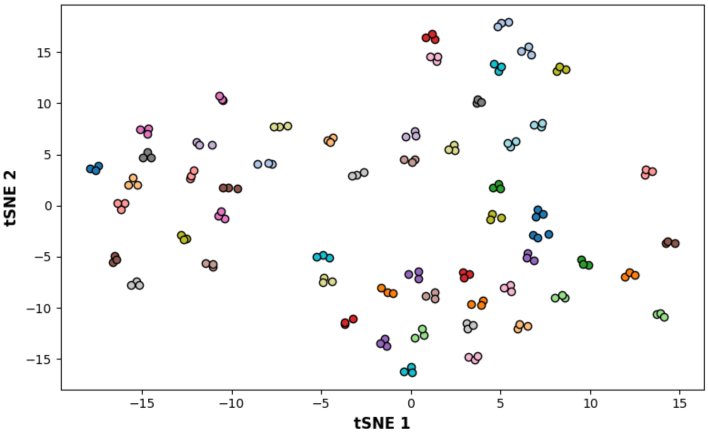

Figure 10: A t-SNE plot showing speaker embeddings grouped by speakers.

Figure 10 visualizes the embeddings for male speakers, with colors representing *individual speakers*. Each speaker's embeddings form distinct clusters with no overlap, highlighting the model's excellent ability to distinguish between speakers. This strong separation aligns with the metrics analysis (E.1), where high accuracy and F1-scores further confirmed the model's robustness in capturing speaker-specific characteristics. The proximity of each speaker's audio samples within their respective clusters underscores the model's effectiveness in this task.

F Analysis and Discussion

This section critically evaluates the results obtained, reflecting on the model's performance, its potential limitations, and underlying biases. The WavLM model exhibited strong performance in speaker similarity analysis, as demonstrated by its high accuracy and F1-scores in the evaluation metrics (E.1). The t-SNE visualizations (E.3) further highlighted the model's ability to group embeddings distinctly by gender and speaker identity, showcasing its robustness in capturing speaker-specific characteristics. This success can largely be attributed to the extensive pretraining of WavLM on over 200,000 hours of audio data. However, despite the strong performance of the WavLM model, several potential biases and limitations must be considered, as outlined in the following paragraphs.

Accent Neutrality in Pretraining. The pretraining dataset lacked a specific focus on British accents, with European Parliament recordings likely including British speakers but not emphasizing their accents. This limitation may account for the overlap observed in the heatmap (7) and figure 9 for accents such as *Ulster and Republic of Ireland*, suggesting that the model struggles to capture subtle distinctions in accents it was not explicitly trained on.

Gender-Based Trends. The heatmap (7) and figure 8 revealed that male embeddings tend to exhibit lower similarity compared to female embeddings. This trend could be due to males expressing a wider pitch range in shortpassages, such as questions or commands, which dominate this dataset. Alternatively, female speakers may adopt more uniform speech patterns in such scenarios, resulting in higher intra-gender similarity. This observation may not generalize to datasets with more diverse or longer speech samples.

Age Range of Speakers. The dataset's age range of 18 to 50 ensures diversity, but differences in vocal characteristics across this spectrum may influence the embeddings. For example, older speakers might produce lower or rougher tones, while younger speakers closer to 18 may have higher-pitched voices. Such variations could create subtle clustering biases tied to age.

Accent Neutrality and Dataset Outliers. Some speakers reportedly lacked strong accents, as noted in the dataset documentation. These neutral accents may have led to the outliers observed in the heatmap (7), where certain embeddings appeared less similar to other samples. This highlights a limitation in the dataset's accent consistency, which can affect the interpretation of accent-specific characteristics.

G Conclusion and Future Work

This study robustly demonstrates the *WavLM* model's capabilities in the analysis of speaker voice similarity, focusing on a variety of *British accents*. The model achieved impressive results, marked by high accuracy and F1-scores, which confirm its effectiveness in capturing subtle differences in speaker characteristics. These results are particularly insightful for understanding the influence of *regional and gender-based* variations in speech, revealing complex patterns of similarity and distinctiveness across accents. The findings enhance our understanding of the linguistic diversity within the British Isles and suggest directions for refining speech recognition technologies to better handle these variations. Specifically, the gender-related insights point to the need for speech systems to adapt more finely to different vocal characteristics to improve both accuracy and user experience. Moreover, the successful application of the WavLM model in this context underscores its potential for deployment in critical areas such as security and communication technologies, where precise speaker identification is paramount. Future initiatives could address the challenges identified in the analysis by exploring various approaches to refine the precision and applicability of speaker similarity assessments::

- **Fine-Tuning with Additional Accents Data:** To improve the model's ability to distinguish between different accents, it would be beneficial to fine-tune it on a more diverse accents dataset. This could include using the long passages from the ABI corpus or incorporating data or the *Open source Multi speaker Corpora of the English Accents in the British Isles* from Demirşahin et al. (2020). Fine-tuning has shown promising results in similar applications, such as emotion recognition using WavLM (Diatlova et al., 2024).

- **Alternative Similarity Metrics:** Exploring other similarity metrics or techniques, such as building a Siamese network, could provide new insights into accent differentiation. This approach has been successfully applied in tasks like *spoofing* speech detection, using the *wav2wav* architecture (Xie et al., 2021).

- **Utilization of an Alternative Dataset:** To address some of the *limitations* identified in the ABI corpus, particularly the inconsistencies in accent representation and the influence of neutral accents, future studies could benefit from using an alternative dataset. This would help to validate the findings from this study and enhance the generalizability of the model.

References

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.

D'Arcy, S., & Russell, M. J. (2008). Experiments with the ABI (Accents of the British Isles) Speech Corpus.

D'Arcy, S. M., Russell, M. J., Browning, S. R., & Tomlinson, M. J. (2004). The Accents of the British Isles (ABI), corpus.

Demirşahin, I., Kjartansson, O., Gutkin, A., & Rivera, C. (2020). Opensource Multispeaker Corpora of the English Accents in the British Isles.

Diatlova, D., Udalov, A., Shutov, V., & Spirin, E. (2024). Adapting WavLM for Speech Emotion Recognition.

Ferragne, E., & Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics.

Miesle, P. (2023). *What is cosine similarity: A comprehensive guide* [Accessed: 2025-01-13]. Retrieved January 13, 2025, from https://www.machinelearningplus.com/nlp/cosine-similarity/

OpenAI. (2024). *ChatGPT 3.5* [This tool was used to formalise ideas in the form of bullet points.]. chatgpt.com

Pajak, A. (2023, June). *T-sne: T-distributed stochastic neighbor embedding* [Published on Medium, June 27, 2023. Accessed: 2025-01-13]. Retrieved January 13, 2025, from https://medium.com/@pajakamy/dimensionality-reduction-t-sne-7865808b4e6a

Xie, Y., Zhang, Z., & Yang, Y. (2021). Siamese Network with Wav2vec Feature for Spoofing Speech Detection.