# Exploring Phonetic Features in Language Embeddings for Unseen Language Varieties of Austrian German

**Lorenz Gutscher** and **Michael Pucher**
Signal Processing and Speech Communication Laboratory, Graz University of Technology
Austrian Research Institute for Artificial Intelligence, Vienna, Austria
lorenz.gutscher@ofai.at michael.pucher@tugraz.at

## Abstract

Vectorized language embeddings of raw audio data improve tasks like language recognition, automatic speech recognition, and machine translation. Although embeddings exhibit high effectiveness in their respective tasks, unraveling explicit information or meaning encapsulated within the embeddings proves challenging. This study investigates a multilingual model's ability to capture features from phonetic, articulatory, variety, and speaker categories from brief audio segments comprising five consecutive phones spoken by Austrian speakers. Within the employed model for extraction, German serves as one of the pretrained languages used. However, the manner in which the model processes Austrian varieties presents an intriguing area for investigation. Using a k-nearest neighbor classifier, it is tested whether the encoded features are prominent in the embedding. While characteristics like variety are effectively classified, the accuracy of phone classification is particularly high for specific phones that are characteristic of the respective dialect/sociolect.

## 1 Introduction

Language embeddings are high-dimensional vectors in a continuous space that describe language-specific features like word order, prosody, speed, and accent. Embeddings can be obtained from either an orthographic perspective (such as word embeddings) or an acoustic perspective (representing spoken language). Utilizing these embeddings enhances precision in various domains including text classification, machine translation, automatic speech recognition, accent detection, and Language Identification (LID) (Hou et al., 2020). Transformer networks revolutionized these fields by using an attention mechanism that captures complex relationships within the words of a sentence or the audio features of utterances (Vaswani et al., 2017). This study focuses on the acoustic embeddings of spoken language, specifically those derived from the output of the final layer of a deep learning LID task. The objective of an LID system is to determine the language of a written text or an utterance.

Systems for LID can be adapted to identify accents and dialects within a language if labels are available. In cases of low-resource languages, the data itself is often not sufficient for training. Cross-lingual transfer can help to increase performance on tasks such as language modeling, translation, or language identification (Conneau et al., 2020). A pre-trained multilingual model can either be fine-tuned on the unseen data or just used as is. The amount of data enables generalization on unseen data and extraction of language-specific content from the embedding.

Standard Austrian German (SAG) is a special case in this context, as it belongs to the same language family as Standard German (SGG). The Austrian dialect landscape is very rich, with notable differences in vocabulary and pronunciation not only between SGG and SAG but also between SAG and other Austrian dialects (Elspaß and Kleiner, 2019; Kleene et al., 2016). The primary objective of this paper is to leverage a multilingual LID system, initially trained on 107 languages, without additional fine-tuning specifically for Austrian varieties. The aim is to assess the model's ability to generalize to unseen varieties and effectively map language features within the latent space.

The contribution of this paper is:

- It demonstrates the usability of multilingual models for low-resource languages without fine-tuning.
- It reveals that characteristic phones of a variety are distinctly represented within the embeddings.
- It shows the spatial mapping of unseen varieties of Austrian German and suggests that

quinphones are effective for classifying these varieties, contributing to better methods for handling and classifying dialectal variations.

The paper is structured as follows: Section 2 delves into previous research where acoustic embeddings are scrutinized for their potential to encapsulate language-specific features. Section 3 outlines the extraction of embeddings and further processing steps for experiments. Section 4 presents a dataset description and the classification results of four key feature groups: Phone classification, variety classification, classification of articulatory features and phone categories, and speaker classification. Each section of the respective feature group offers a presentation of the results, followed by an analysis.

## 2 Related Work

Language embeddings are investigated for properties of phonology, morphology, and syntax in (Bjerva and Augenstein, 2018) after fine-tuning language embeddings on specific Natural Language Processing tasks using text data. The method of feature probing through a k-Nearest Neighbor (kNN) classifier yields the conclusion that information pertaining to the investigated properties is encapsulated within the embeddings, exhibiting varying degrees of efficacy in accordance with the task-specific relevance of these properties. This concept is further pursued in (Östling and Kurfalı, 2023) with respect to typological features, asserting that multilingual language embeddings capture linguistic information when trained on the correct downstream tasks. The application of multilingual transfer learning to utilize acoustic embeddings derived from triphones, as described in (Kamper et al., 2021), demonstrates the capacity for extraction of phonetic content and language information for zero-resource languages. Using acoustic embeddings (Belinkov and Glass, 2017), an in-depth analysis of an Automatic Speech Recognition model at the frame level to incorporate phonetic features is conducted. The investigation aims to ascertain the layers within an end-to-end model where phones and sound classes are prominent. In (English et al., 2023) the wav2vec 2.0 model (Babu et al., 2022) is probed to contain three broad phonetic classes (voicing, frication, and nasals) within different layers of the model. (Linke et al., 2023) investigates read and spontaneous speech from Austrian and Hungarian varieties, showing evidence that parameters of speaking style are encoded in the pre-trained XLS-R model and that Austrian German is mapped separately from German German. In (Gutscher et al., 2023) the effectiveness of a pre-trained Language Identification (LID) model in mapping Austrian varieties within latent space is demonstrated. The model successfully distinguishes these varieties from SGG and other European languages. In (Zuluaga-Gomez et al., 2023) the internal categorization of the wav2vec 2.0 embeddings is analyzed through t-Distributed Stochastic Neighbor Embedding (t-SNE), and it is observed that there is a level of clustering based on phonological similarity.

## 3 Methods

In typical settings, acoustic language embeddings are extracted at the sentence or utterance level. In this work, it is hypothesized that valuable language information is not only present in sentence or utterance embeddings but also in smaller units, specifically in quinphones. Therefore, the dataset is divided into chunks of audio consisting of five consecutive phones. Quinphones find frequent application in Hidden Markov Models (HMMs) owing to their capacity to encapsulate contextual dependencies among phonetic units. HMMs of quinphones are capable of capturing the influence of adjacent phones, thereby contributing to the pronunciation of words. The language embedding is extracted with a multilingual LID system[1] for all quinphones, as depicted in Figure 1, representing each quinphone with a 2048-dimensional vector (no further classification based on the embeddings is done). The system employs the XLS-R model (Conneau et al., 2021) which builds on the wav2vec 2.0 architecture and underwent fine-tuning using the voxlingua107 dataset (Valk and Alumäe, 2021) (107 languages). Wav2vec 2.0 is initially trained on publicly available datasets encompassing 128 languages, providing substantial variability and encompassing a wide array of linguistic contexts. Utilizing quinphones is advantageous because the multilingual LID system mentioned above, with its default settings, requires a minimum sample length of 400 samples (25 ms) to extract embeddings due to the minimum size of the kernel filters. If single phones were used instead of quinphones, this minimum length requirement would pose problems for phones shorter than 25 ms.

[1] https://huggingface.co/TalTechNLP/voxlingua107-xls-r-300m-wav2vec

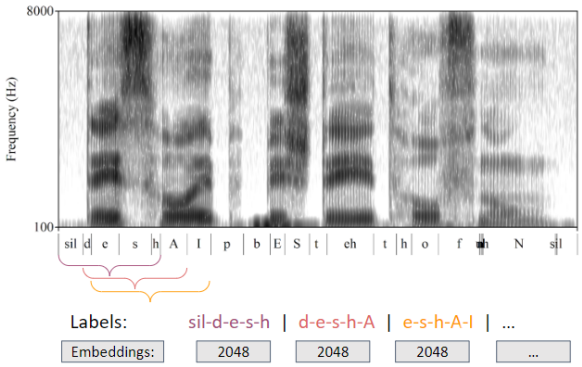


Figure 1: Process of extracting embeddings from quinphones

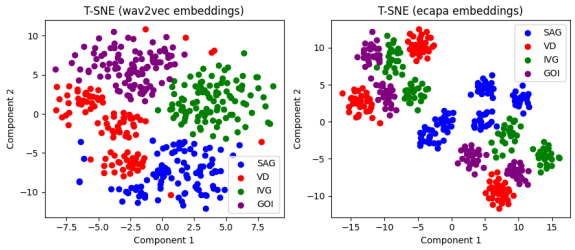


Figure 2: Visualization of four varieties of Austrian German using t-SNE with wav2vec 2.0 (left) and ECAPA-TDNN (right) LID models.

The goal of this paper is to test the model's ability to classify segmental and phonetic features for both seen and unseen Austrian varieties. Training on low-resource data can lead to speaker embeddings instead of language embeddings due to the limited number of speakers. To address this, a pre-trained model was utilized. The effectiveness of probing for features in language embeddings is shown in (Singla et al., 2022; Hewitt and Manning, 2019).

To investigate the clustering of language varieties, a sample set of 100 utterances per variety is employed, and t-SNE is used to visualize the potential clustering of the high-dimensional embedding vectors. Two models are compared in this analysis: the wav2vec 2.0 XLS-R and the Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Networks (ECAPA-TDNN) (Desplanques et al., 2020) model[2] (both fine-tuned on LID using the voxlingua107 dataset). The ideal visual output would exhibit clear spatial separation between the four language varieties. As illustrated in Figure 2, the wav2vec 2.0 model effectively disentangles speaker and variety information, resulting in more generalized clusters compared to the ECAPA-TDNN model. Conversely, the ECAPA-TDNN model reveals a bias towards encoding speaker-specific information, resulting in smaller clusters primarily representing individual speakers. Further analysis of the ECAPA-TDNN model reveals an additional layer of gender-based clustering. This model initially segregates the data into two primary clusters based on gender, aligned along a diagonal axis from the bottom left (Component 1: -10, Component 2: -15) to the top right corner (Component 1: 5, Component 2: 10) of the t-SNE plot. Within these primary gender clusters, further subdivision into smaller clusters occurs, each representing different speakers.

The process of data pre-processing involves the following: For each audio chunk (quinphone), a corresponding label file is created containing information about all five phone states. To avoid overlapping quinphones in the training and test sets, chunks of the same utterances are not split between those groups. For each probing feature in the datasets, binary targets are constructed, and an approximate nearest neighbor classifier is trained using the FAISS package (Douze et al., 2024). The parameter for determining the number of nearest neighbors is set to k=10, employing the Euclidean distance as the distance metric. This choice of k is designed to enhance the classification of infrequent instances, avoiding dependence solely on the clustering of instances associated with identical words. For each feature, the target is binary, which means there are only two possibilities for building the targets: (a) The feature is eminent in the current quinphone, or (b) the feature is not eminent in the current quinphone. The position of a feature within a quinphone is not taken into account. Infrequently observed features, occurring below the minimum threshold of 200 instances in the training set, are systematically excluded. The intrinsic operational principle of the kNN algorithm leads to a statistical bias concerning the classification accuracy score between frequent and non-frequent features, whereby the likelihood of accurate classification increases when there is a greater abundance of data points related to the feature in the training set. A dummy classifier is employed to rectify this effect. It randomly shuffles the binary target values, emulating random guessing, but taking into account the number of ones and zeros for these features. The

[2] https://huggingface.co/TalTechNLP/voxlingua107-epaca-tdnn

output of the classification metric from the dummy classifier is then subtracted from the metric of the actual classifier. The impact of the dummy classifier is particularly pronounced in categories where the majority of targets are predicted to be positive targets. The F1-score is employed for evaluation, representing the harmonic mean between precision and recall while considering both balanced and unbalanced target sets. It characterizes a trade-off between instances classified as false positives and false negatives. The focal point of interest does not reside solely in the absolute performance of the classification of individual features, but rather in discerning the degree to which certain features are encoded more effectively than others.

## 4 Experiments

Building upon the methods described above, the experiments aim to evaluate the performance of the proposed approach in classifying phonetic and articulatory features, along with variety and speaker groups, across diverse Austrian varieties.

### 4.1 Data overview

**Dataset.** The dataset consists of 16 kHz WAV recordings with corresponding labels in the format of HTS (Zen et al., 2007) label files containing detailed temporal-aligned phone annotations in addition to linguistic and prosodic information. The dataset comprises four distinct varieties, each contributing unique linguistic characteristics.

- The SAG variety utilizes data extracted from the Wiener Corpus of Austrian Varieties for Speech Synthesis (WASS) (Pucher et al., 2015; Toman and Pucher, 2015).
- The Viennese (VD) variety draws from the Viennese Sociolect and Dialect Synthesis (VSDS) corpus (Pucher et al., 2010).
- Additionally, the dataset includes Innervillgraten (IVG) and Bad Goisern (GOI) varieties, both sourced from the Goisern and Innervillgraten Dialect Speech (GIDS) corpus (Schabus et al., 2014).

The dataset is reduced to achieve balance among varieties, ensuring that each variety has an equal number of data points and approximately the same number of speakers (SAG: 5, VD: 3, IVG: 4, GOI: 4). Upon segmenting the utterances into labeled

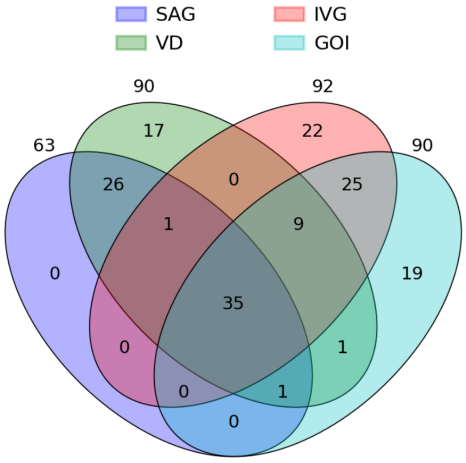


Figure 3: Phone set overlaps for SAG, VD, IVG, and GOI

units utilizing the provided time-codes from the annotations, the training set comprises 185,496 quinphones (90%), while the test set contains 20,610 quinphones (10%). To ensure a balanced evaluation, the test set was further refined for each feature to include an equal representation of 1-targets and 0-targets. The mean duration of a quinphone is 500 ms.

**Variety description.** All four varieties have shared phones and (except for SAG) between 17 and 22 unique phones. The numbers of overlapping phone sets are illustrated in Figure 3.

- SAG is the standard variety of German spoken in Austria. In SGG, for example, the high vowels [i] - [ɪ], [y] - [ʏ], and [u] - [ʊ] are clearly differentiated by quality (Davis and Mermelstein, 1990). This difference in quality is rather small to non-existent in SAG, though these phones still exist in SAG. A difference in vowel quality between SAG and SGG is the low vowel [ɑ] in SAG, which is [a] in SGG. In general, the transition between standard and dialect can be described by different processes that do not necessarily result in unique phones. For our analysis on the phonetic level, we are focusing on the different phones.
- The VD is an East Bavarian sociolect, nowadays mainly spoken by older, male, working class German speakers in Vienna, and has characteristic processes like monophthongization, which result in unique phonetic differences. The Viennese monophthongization is a form of assimilation, whereby one part of the diphthong is assimilated to the other (Moos-

müller, 2011).

1. <Haus> (Engl. “house"): [haʊ̯s] → [hɒːs]
2. <weit> (Engl. “wide"): [vaɐ̯t] → [væːd̥]

- The GOI dialect is a Central Bavarian dialect spoken in the region of Bad Goisern and has a significant number of diphthongs that arise through diphthongization of vowels, as shown in Example 1 below for the word <Schwester> (Engl. “sister"). Another source of new diphthongs is the vocalization of the lateral (/l/-Vocalization), as shown in Example 2 for the word <bald> (Engl. “soon"). This is a prominent feature in the Central Bavarian varieties and accours in word-medial and word-final positions. The vocalization of the lateral is perceived as a dialect feature and thus widely suppressed by standard variety speakers, including those who strive for a standard variety. Another characteristic phone of GOI is the uvular trill [ʀ].

  1. <Schwester> (Engl. “sister"): [ʃvesda] → [ʃvɛɛsda]
  2. <bald> (Engl. “soon"): [bɑld] → [bɔed]
  3. <recht> (Engl. “right"): [reɕd] → [ʀɛeɕd]

- The IVG dialect is a South Bavarian dialect spoken in East Tyrol and uses a fricativized trill [ʀ] or the uvular fricative [χ] as a characteristic phone, transcribed as [ʀχ] in our data. Another distinctive phone of IVG is the palatal approximant [ʎ].

  1. <warten> (Engl. “to wait"): [vɑːdn] → [vɔːʀχdn]
  2. <Zahl> (Engl. “number"): tsɑl] → [tsɔːʎ]

### 4.2 Phone classification

In the phone evaluation, positive classification targets in the dataset indicate the presence of at least one phone within a quinphone instance. Following the exclusion of exceedingly rare instances, the evaluation yields a total count of 132 phones from the initial pool (24 phones are excluded). The computed average F1-score stands at 0.42 with a standard deviation of ± 0.1. The phones with the highest F1-scores, after subtraction of associated dummy scores (denoted in brackets), are delineated in Table 1. The characteristic [æː] monophtongs described in Section 4.1 achieve an F1-score of 0.63 (0.0), while [ɒː] achieves 0.1 (0.0). For a full list of all phone classification results see Figure 6 in the appendix.

The phone category exhibits the second-best results, demonstrating significant variations among different phones. Notably, the distinct [ʀ] and several diphthongs from the GOI phone set attain a commendable score. Within the IVG phone set, the phones [ʀχ] and [ʎ] exemplify that phones incorporating language-specific features contribute to an elevation in the classification score and are especially well classified. This phenomenon is similarly observed for the VD monophthong [æː]. The specific vowel quality [ɑ] of SAG on the other hand, is not well classified. Given that the dummy classifier yields a score close to 0, these results demonstrate a classification performance significantly surpassing random chance for the dialect/sociolect varieties.

### 4.3 Classification of articulatory and phone categories

Within the 54 articulatory and phone classes, the average score is 0.23 with a standard deviation of ± 0.2. Only the categories retroflex (0.69), affricates (0.66), aspirated (0.64), voiced fricative (0.6), syllabic (0.55), and ʊ-vowel (0.52) achieve F1-scores over 0.5, indicating moderate classification (see Figure 4). Other categories, such as vowel types (low, high, closed, etc.), consonant types (front, fortis, short, etc.), and fricative types (central, back, front, unvoiced), consistently exhibit values below 0.5.

The group of articulatory and phone classes has the lowest score, indicating that this category is not well represented within the embedding in the case of quinphones. Only six out of 54 categories achieve F1-scores over 0.5. The other 48 categories lack sufficient information for a reliable classification in this quinphone setup. Compared to the variety, phone, and speaker categories, the articulatory and phone categories achieve the lowest average score over all features. It is proposed that the task of Language Identification (LID) does not effectively train models to represent these features in a compound manner in quinphones. Moreover, the setup of using quinphones is likely to contain

Table 1: Phones with highest F1-scores

| Phone | F1-score (dummy) | Phone | F1-score (dummy) |
|---|---|---|---|
| [ʁ] | 0.70 (0.0) | [ɛɑ] | 0.66 (0.0) |
| [ʎ] | 0.69 (0.0) | [ɕ] | 0.66 (0.0) |
| [f] | 0.67 (0.02) | [ʁχ] | 0.66 (0.0) |
| [ɔɛ] | 0.66 (0.0) | [ɛ] | 0.66 (0.01) |
| [ɑɛ] | 0.66 (0.0) | [ɔɐ] | 0.65 (0.0) |

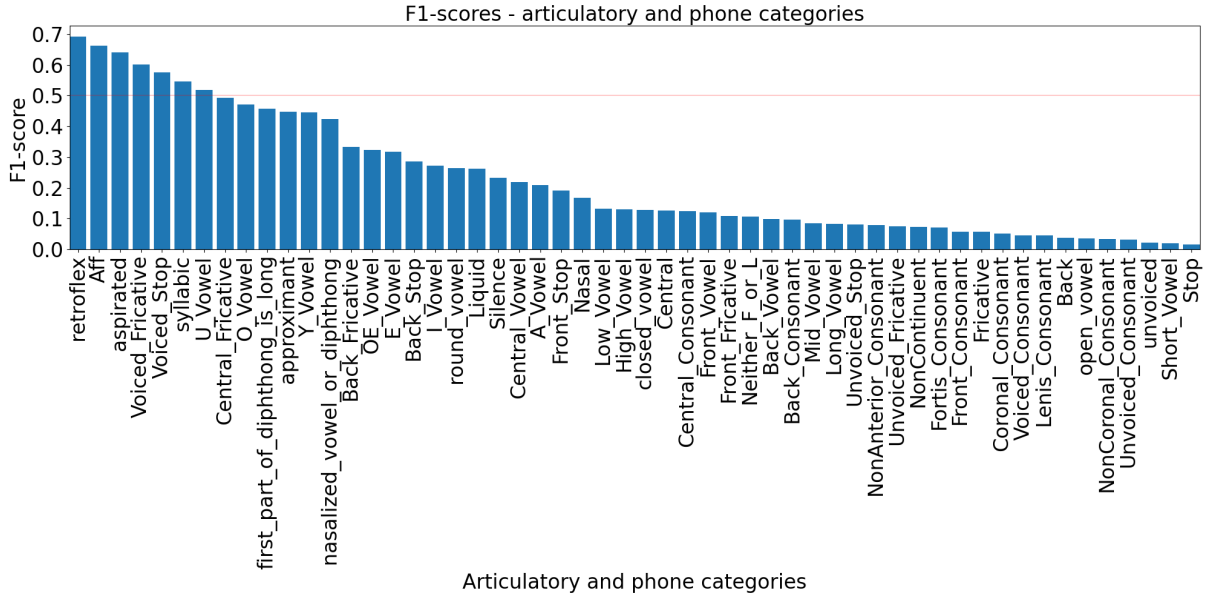


Figure 4: F1-scores for classification of articulatory and phone categories

Table 2: F1-scores for Austrian varieties

| Variety | F1-score (dummy) |
|---|---|
| IVG | 0.9 (0.04) |
| GOI | 0.89 (0.03) |
| SAG | 0.84 (0.05 |
| VD | 0.68 (0.04) |

more features within one quinphone (for example, consonants and vowels), making the training set very unbalanced.

### 4.4 Variety classification

The variety category comprises four distinct varieties, achieving an average score of 0.83 $\pm$ 0.1. As illustrated in Table 2, IVG attains the highest F1-score of 0.9 (0.03), followed by GOI with 0.89 (0.03), SAG with 0.84 (0.05), and VD with 0.68 (0.04).

The classification outcomes for quinphones demonstrate significant language-related cues within the variety category. While this phenomenon was demonstrated at the utterance level in Figure 2, it is noted that the representations of utterance embeddings from IVG and GOI show fewer outliers and less overlap compared to VD and SAG. This distinctiveness potentially contributes to improved classification results for IVG and GOI at the quinphone level. The authors suggest that the decreased performance of VD stems from the proximity of small speech units to SAG, leading to misclassifications in certain instances.

### 4.5 Speaker classification

The final group reflects speaker-related information embedded in the quinphone audio data. The average score is 0.36 with a standard deviation of $\pm$ 0.22. Notably, SPO (SAG) and HPO (VD) stand out with the highest scores of 0.9 and 0.73, respectively, while the remaining 14 speakers exhibit scores ranging from 0.51 to 0.07 (see Figure 5). The dummy classifier consistently yields a score of 0 in all cases.

The speaker classification achieves the second-lowest accuracy, suggesting that the embedding does not effectively capture information about the

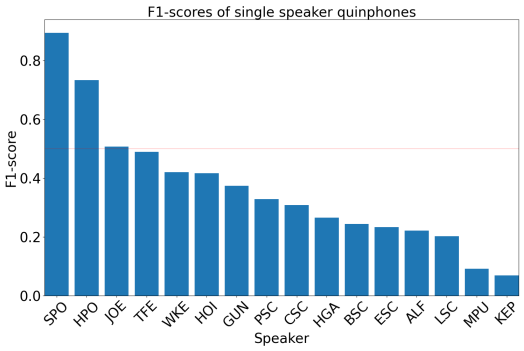


Figure 5: F1-scores for single speaker classification

speaker's voice, which is expected for a model trained on the task of LID. Speaker classification only exceeds F1 values of 0.5 for the SPO (SAG), HPO (VD), and JOE (VD) speakers. SPO and HPO, both professional radio and TV speakers, could potentially exhibit distinct speaking styles due to their professional backgrounds. This divergence in speaking style is likely manifested in the embeddings, leading to a more pronounced separation between these two speakers compared to others. JOE, as the singular youth voice in the corpus, could impart a unique linguistic imprint to the embeddings, potentially resulting in distinguishable language characteristics.

## 5 Conclusion

Understanding the intricacies of multilingual language embeddings in capturing phonetic features for unseen language varieties holds significant importance in advancing the capabilities of automated language processing systems. This study explores the efficiency of multilingual language embeddings derived from short audio segments (quinphones) in capturing phonetic features for Austrian German varieties. It shows that the multilingual wav2vec 2.0 model (fine-tuned on the task of LID) disentangles speaker and language information for unseen varieties of Austrian German. Furthermore, it indicates that individual phones within a quinphone are sufficient for the model to group or model specific varieties. This supports the utilization of comprehensive multilingual language identification embeddings in diverse applications, including automatic speech recognition, accent recognition, and language identification. It is particularly relevant for low-resource languages, where fine-tuning poses challenges.

## 6 Limitations

In this study, we opted to split utterances into non-overlapping segments to mitigate the potential issue of similar embeddings arising from overlapping segments. However, it is important to note that despite this precaution, instances of repeated single words between training and testing splits may still arise, albeit infrequently. Furthermore, a noteworthy limitation of our methodology pertains to its applicability to languages that lack closely related counterparts in the pre-trained model, unlike German and Austrian German. This discrepancy may hinder the extension of our findings to languages not adequately represented in the model's training data.

## 7 Ethical Considerations

No new data was recorded in this study. The datasets utilized are anonymized, employing pseudonyms and removing identifying information to ensure the privacy and confidentiality of the speakers. Explicit consent was obtained from each individual speaker for the use of recordings for research purposes. The findings do not marginalize any dialects or reinforce any power dynamics. Furthermore, the explainability of models that can be achieved through an analysis on the phonetic level contributes to making deep learning models more transparent to the potential user.

## References

A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Neural Information Processing Systems (NIPS 2017)*, volume 2017-December, pages 2442–2452.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proc. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 907–916.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for

speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Steven B. Davis and Paul Mermelstein. 1990. German. *Journal of the International Phonetic Association*, 20:48–50.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech 2020*, pages 3830–3834.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint 2401.08281*.

Stephan Elspaß and Stefan Kleiner. 2019. Forschungsergebnisse zur arealen Variation im Standarddeutschen. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Deutsch: Sprache und Raum. Ein internationales Handbuch der Sprachvariation*, pages 159–184. De Gruyter.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2023. Discovering phonetic feature event patterns in transformer embeddings. In *Proc. Interspeech 2023*, pages 4733–4737.

Lorenz Gutscher, Michael Pucher, and Víctor Garcia. 2023. Neural speech synthesis for austrian dialects with standard german grapheme-to-phoneme conversion and dialect embeddings. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 68–72.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *Proc. Interspeech 2020*, pages 1037–1041.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2021. Improved acoustic word embeddings for zero-resource languages using multilingual transfer. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 29, pages 1107–1118.

Andrea Kleene, Alexandra N. Lenz, Hans Bickel, Ulrich Ammon, Juliane Fink, Andrea Gellan, Lorenz Hofer, Karina Schneider-Wiejowski, Sandra Suter, Jakob Ebner, and Manfred Michael Glauninger. 2016. Variantenwörterbuch des Deutschen – die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. In *Variantenwörterbuch des Deutschen*. De Gruyter.

J. Linke, M.S. Kádár, G. Dobsinszki, P. Mihajlik, G. Kubin, and B. Schuppler. 2023. What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers. In *Proc. Interspeech 2023*, pages 5371–5375.

S. Moosmüller. 2011. Sound changes and variation in the Viennese dialect. In *On Words and Sounds: A selection of papers fromt the 40th PLM, 2009*, pages 134–147. Cambridge Scholars Publishing.

Michael Pucher, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann, and Dietmar Schabus. 2010. Resources for speech synthesis of viennese varieties. In *Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Michael Pucher, Markus Toman, Dietmar Schabus, Cassia Valentini-Botinhao, Junichi Yamagishi, Bettina Zillinger, and Erich Schmid. 2015. Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games. In *Proc. Interspeech 2015*, pages 1625–1629.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2014. Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347.

Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. 2022. What do audio transformers hear? Probing their representations for language delivery & structure. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 910–925.

Markus Toman and Michael Pucher. 2015. An Open Source Speech Synthesis Frontend for HTS. In *Proc. of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302*, pages 291–298.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda. 2007. The HMM-based speech synthesis system (HTS) version 2.0. In *6th ISCA Workshop on Speech Synthesis, SSW 2007*, pages 294–299. The International Society for Computers and Their Applications (ISCA).

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *Proc. Interspeech 2023*, pages 5291–5295.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations. *Computational Linguistics*, 49(4):1003–1051.

# A Appendix

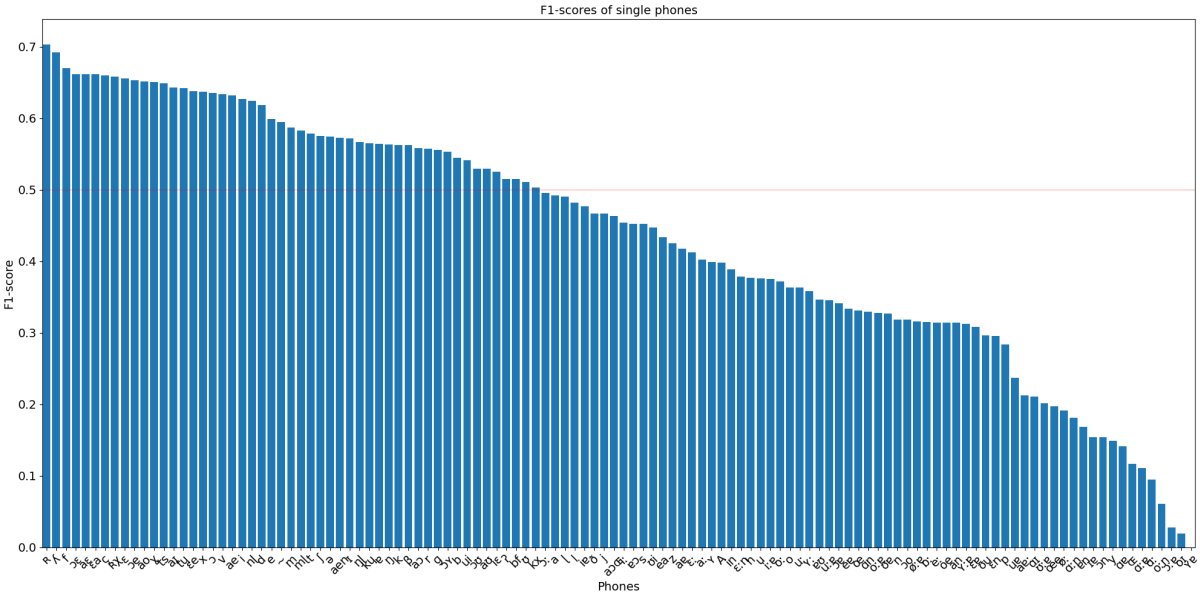


Figure 6: F1-scores for classification of all phones