Leveraging Linguistic Similarity for Low-Resource Speech Transcription

Valentina Fedchenko, Eric Jordan
ERTIM (INALCO), LACITO (CNRS)
65 rue des Grands Moulins, 7 rue Guy Môquet
{valentina.fedchenko, eric.jordan}@inalco.fr

Abstract

This study investigates how large-scale, self-supervised acoustic models (like XLSR and MMS) represent linguistic similarity and whether this can optimize Automatic Speech Recognition (ASR) for low-resource and dialectally diverse languages. While these models excel at cross-lingual transfer learning, their internal representations of fine-grained dialectal variation remain opaque. We focus on Yiddish, a language with a complex dialect continuum, to test if a model's internal acoustic similarity metric—Acoustic Token Distribution Similarity (ATDS)—predicts ASR performance. Our methodology involved fine-tuning models on Yiddish dialects and measuring ATDS between Yiddish and related languages. Results confirm that ATDS is a meaningful predictor: higher acoustic similarity in the model's latent space correlates with lower character error rates (CER) after fine-tuning. This relationship is strongest in mid-to-upper layers of the MMS model and for in-domain data. Crucially, ATDS captures model-dependent acoustic similarity, which does not always align with genealogical linguistic relationships but remains a practical indicator of transfer learning potential. We conclude that ATDS is a valuable tool for selecting donor languages to develop more efficient, dialect-sensitive ASR systems for language documentation, even if its absolute values require careful interpretation against linguistic knowledge.

Keywords: transfer-learning, linguistic similarity, dialectal diversity, low-resource context, automatic speech recognition, Yiddish dialects

1. Introduction

Over the past few years, automatic speech recognition (ASR) and speech processing have undergone a profound transformation, driven by large-scale self-supervised acoustic models such as wav2vec 2.0 (Baevski et al., 2020), XLSR (Conneau et al., 2021; Babu et al., 2021), and MMS (Pratap et al., 2024). Trained on thousands of hours of unlabelled audio across hundreds of languages, these models have demonstrated remarkable performance in transcribing speech, even in languages for which little or no annotated data exist. Their strength lies in their ability to extract latent, generalisable acoustic representations from massive multilingual datasets — representations that encode phonetic, phonological, and, to some extent, prosodic information in a shared vector space.

Yet, this very scale raises a central question for linguistics: how deeply do these models understand the fine-grained structure of human linguistic diversity? While they can model speech at a global level, they are not designed to capture the subtle phonetic, morphological, and dialectal distinctions that linguists rely on to analyse variation within a linguistic continuum. In other words, the more these models generalise across languages, the less transparent their internal representations become. What do they actually learn about cross-linguistic similarity? How do they encode distinctions between closely related varieties, such as dialects or contact-induced variants?

For fine-grained linguistic analysis, and particularly for language documentation, we need models that are not only powerful but also linguistically competent — models whose behaviour we can interpret and align with theoretical and empirical knowledge about language structure. Understanding how these models represent similarity and difference between languages is crucial for designing effective adaptation strategies. This involves going beyond performance metrics and exploring how representations learned from massive multilingual training data relate to known typological, phonetic, and acoustic relationships among the world's languages.

With this in mind, the concept of linguistic similarity becomes key. When adapting pre-trained models to low-resource or endangered languages, researchers increasingly rely on transfer learning — the idea that knowledge learned from one or more high-resource "donor" languages can benefit a "target" language with little data. However, not all donors are equally effective: some share more relevant acoustic and structural features with the target than others. Quantifying and predicting these relationships remains an open question. Our study aims to explore how acoustic models capture linguistic similarity and how this understanding can be leveraged to improve ASR performance for minority and dialectally diverse languages. Specifically, we investigate Yiddish — a language situated

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3590–3598
11–16 May 2026.

at the intersection of Germanic and Slavic influences — to test how donor languages contribute to the adaptation of speech models for dialect-aware transcription, and to assess whether the acoustic similarities observed by the models correspond to known linguistic relationships.

The objective of our work is twofold. First, we aim to develop a dialect-sensitive automatic transcription tool for Yiddish, which could facilitate the semi-automated expansion of valuable linguistic resources such as the *Corpus of Spoken Yiddish in Europe (CSYE)* (Bleaman and Nove, 2025) or the project *AHEYM: The Archive of Historical and Ethnographic Yiddish Memories (2002-2009)* (Veidlinger, 2013; Ćavar et al., 2016), both of which include recordings from different Yiddish varieties and often contain code-switched segments that are challenging for automatic processing. At the same time, we address a broader, more fundamental question: the correlation between linguistic proximity, captured in acoustic vector representations, and the performance of fine-tuned acoustic models trained on combinations of genetically or geographically related languages.

From a practical point of view, this study follows a logic of re-use and optimization of existing resources. With the growing number of pre-trained and fine-tuned ASR models available for various languages, a central question emerges: *can these models be effectively repurposed to compensate for the scarcity of data in minority or under-documented languages?* By addressing this question, we aim to assess not only the technical potential of transfer learning across closely related linguistic varieties, but also its implications for the development of more inclusive, resource-efficient ASR systems that can support linguistic documentation and revitalization efforts.

2. State of the Art

Of the approximately 7,000 languages in the world, nearly half are now in danger of disappearing (Bromham et al., 2021). For languages without a written tradition, the creation of oral corpora and their manual transcription is a particularly time-consuming task, often described as the 'transcription bottleneck' (Seifart et al., 2018). In addition, the transcription of low-resource languages presents a number of specific challenges:

1. the limited quantity of annotated data;
2. the absence of standardised orthography;
3. a high degree of dialectal variation;
4. the presence of code-switching and language interference;
5. phonetic and morphosyntactic phenomena that are typologically rare.

The processing of dialectal diversity has only recently become an active area of research. Most existing projects on major languages with substantial dialectal variation, such as Arabic (Djanibekov et al., 2025) or Chinese (Jie et al., 2024) dialects, focus primarily on transcription into a standardised orthography (Joshi et al., 2025). In contrast, within the field of linguistic documentation, field linguists require a much higher level of precision from automatic speech recognition (ASR) systems. Rather than producing standardised orthographic transcriptions, these applications must deliver detailed annotations at the phonemic — and even allophonic — level.

In low-resource settings, transfer learning—the adaptation of models pre-trained on high-resource languages to low-resource linguistic environments—has demonstrated considerable potential in improving performance across a range of speech and language processing tasks (Pakray et al., 2025). The effectiveness of such adaptation, however, depends heavily on the linguistic relationship between the source and target languages, as well as on task-specific and data-related factors such as phonetic coverage, orthographic conventions, and the quantity and quality of available recordings.

The strategy of enhancing the efficiency of acoustic model training by combining data from a low-resource language with that of typologically similar, high-resource varieties was proposed by San et al. (2024), where it was presented and applied in conjunction with Continuous Pre-Training (CPT). CPT allows a pre-existing self-supervised model to be refined by retraining it on a sample of the target language. It has proven effective in improving the performance of models such as `wav2vec` when a relatively small amount of data is available (Nowakowski et al., 2023). However, in the case of most low-resource languages, this having several tens of hours of data is already a rare occurrence. To compensate this lack of data, some studies have resorted to using typologically similar donor languages, so that the acoustic representations acquired from these donor languages can benefit the model when applied to the target language.

The notion of similarity in linguistics encompasses a wide range of interpretations. To operationalize it, San et al. (2024) proposed a metric for quantifying acoustic similarity between languages — the Acoustic Token Distribution Similarity (ATDS) — which serves to evaluate the suitability of a donor language for model adaptation. In their study on Punjabi, training on 10 hours of Punjabi + 60 hours of Hindi achieved performance comparable to training on 70 hours of Punjabi alone. This approach paves the way for an empirical evaluation of linguis-

Phoneme	NEY (Lithuanian)	CEY (Polish)
U2/3	kʊgl	kɪgl
O2	ʃabɛɪsɪm	ʃabɔɪsɪm
A3	pɔnɛm	pʊnɛm
E4	gɛm	gam
U4	hɔɪz	hoːz

Table 1: Phonemes across Yiddish dialects (NEY = Northeastern Yiddish, CEY = Central Yiddish).

tic proximity via ASR performance.

We plan to further develop this line of research in unsupervised learning and adapt it to the field of linguistic documentation, which is characterized by data scarcity, dialectal diversity, and varying levels of transcriptional granularity. In the present study, however, we adapted this framework — the concept of donor languages and the ATDS similarity measure — to the context of supervised training and the practical task of dialect-aware transcription.

3. Methodology

Our experiment was conducted in three stages, which are presented here following the same logic. First, we fine-tuned several types of acoustic models for a dialect-aware Automatic Speech Recognition (ASR) task on Yiddish dialects. Second, we measured the ATDS distance between the Yiddish dialects and their genetically or geographically related languages. Finally, we analysed the correlations between model performance on the validation, in-domain, and out-of-domain test sets and the ATDS values, using both qualitative and quantitative approaches.

3.1. Data

We have conducted experiments on Yiddish, a Germanic Indo-European language that emerged from a historical language shift from German dialects. While Yiddish is not a quintessential low-resource language — it has a standardized spelling system and an abundant literature — its everyday use has declined sharply and the available oral resources (sound archives, testimonies, dialectal narratives) are often noisy, with various dialects and few transcriptions. These characteristics make Yiddish a good starting point for testing the possible contribution of state of the art ASR methods in the context of minority languages.

Due to its extensive contact history, Yiddish exhibits a high degree of linguistic diversity. Its dialectal system is highly differentiated, with most isoglosses being phonetic and primarily affecting vowels and diphthongs (Jacobs, 2005; Falkovitsh, 2024). *Table 1* presents phonetic isoglosses between two Yiddish dialects, studied in this article.

Yiddish dialect	Sex	Nb min.	Nb segments
NEY (Lithuanian)	F	163	1566
CEY (Polish)	M	156	1259

Table 2: Train and validation datasets.

Over the course of migration, new dialects have developed within Hasidic communities in the United States and Israel. These relatively closed groups ensure strong linguistic vitality; however, coexistence with English or Hebrew also leads to interference and innovation.

Hasidic Yiddish is phonetically closer to the Central Eastern Yiddish (CEY) dialect, yet its vowel system shows noticeable shifts due to contact with English. The short high vowels, such as /i/ and /u/, tend to become lower and more centralised within the vowel space. The high (close) vowel [i] often lowers and front-opens, approaching a mid-close quality [e]. Similarly, the back mid vowel [ɔ] undergoes a reduction in rounding, becoming less rounded and more open (Nove, 2021). These developments reflect the influence of bilingualism and language contact within Hasidic communities. For instance, the word kɔnfɪskɪrt 'confiscated' may be realised as kənfɛskɪrt, illustrating both vowel lowering and unrounding.

Native Yiddish speakers typically use their regional dialect rather than the standard form. Consequently, most available data is dialectal, which complicates the creation of robust speech recognition systems.

The data used in our experiments comes from the Reading Electronic Yiddish Documents (REYD; Webber et al.) project. We deliberately limited the total duration to 5 hours and 18 minutes to simulate the conditions of linguistic documentation (*Table 2*).

The transcriptions, originally in standard spelling (Hebrew characters), were converted to the International Phonetic Alphabet. For the Polish dialect, a dialectal conversion was performed according to rules, despite some spontaneous corrections by the reader to the standard norm.

Two test sets were prepared:

1. in-domain (REYD; Webber et al.): NEY dialect, male voice, 157 min., book reading;
2. out-of-domain (Ardila et al., 2020): Hasidic dialect, multiple male voices, 121 min., book reading.

The type of data considered here (read speech) can be considered representative of some types of field data in the case of languages with a writing system and a textual tradition. We also plan to extend these experiments to other types of recordings (spontaneous speech, narration, conversation) in order to analyze the impact of genre and genre

variation on transfer efficiency on closely related languages.

For the second experiment, which focused on measuring linguistic distance, non-transcribed audio corpora from Mozilla Common Voice (Ardila et al., 2020) were used. The selection of languages for these corpora was guided by the linguistic characteristics of Yiddish dialects and their historical context. We evaluated models based on German and English due to the genetic relationship between Yiddish and other Germanic languages, while acknowledging the distinct diachronic distances of German and English to the various Yiddish dialects. Furthermore, the contemporary Hasidic dialect exhibits significant synchronic influence from English. Models fine-tuned on Slavic languages were also included, reflecting the prolonged period of linguistic influence from Polish, Ukrainian, and Russian across different geographical areas and stages of Yiddish dialect development. Hebrew was included in this experiment because approximately 10% of Yiddish vocabulary is borrowed from Hebrew and Aramaic, and the language developed over a long period under diglossic conditions with Hebrew.

We ensured that all corpora were of equal size, approximately 150 minutes each, as the chosen distance measure (ATDS), being statistical in nature is sensitive to data size, a sensitivity that we also observed in our experiments.

3.2. First experiment: Adapting Acoustic Models for ASR Task

For this stage of the experiment, we selected the original Wav2Vec2 model (Baevski et al., 2020) — pre-trained on 53 languages — along with its fine-tuned versions. We also included two versions of a subsequent large-scale Wav2Vec2 model, MMS: the original model, pre-trained on more than 1400 languages, and another version fine-tuned for 1162 languages:

1. Wav2Vec2-large multilingual (XLS-R, pre-trained in 53 languages);
2. Wav2Vec2-large specified for German, English, Polish, Ukrainian, Russian and Hebrew;
3. Wav2Vec2-large multilingual (MMS-all, fine-tuned for 1162 languages);
4. Wav2Vec2-large multilingual (MMS-1b, pre-trained in more than 1400 languages)

To evaluate the task of developing a phonemic transcription system, all models were trained on a corpus of speech utterances paired with their corresponding dialect-specific phonemic transcriptions. The Connectionist Temporal Classification (CTC) loss function was employed, which enables the direct use of these annotations without requiring additional alignment. The model operates by making predictions at the character level, with the target sequences consisting directly of the phonemic transcriptions.

A phoneme vocabulary was constructed, which included a space character but excluded punctuation, and the tokenizer was configured accordingly. For the pre-fine-tuned models, it was necessary to adjust the size of the tokenizer (from 33 to 39, or from 36 to 39 tokens). This expanded tokenizer required the system to retain the existing embeddings and incorporate new, randomly initialised vectors.

3.3. Second experiment: Measuring the ATDS Layer-wise

The Acoustic Token Distribution Similarity (ATDS) measure is designed to quantify the degree of acoustic similarity between two languages or dialects using self-supervised speech representations San et al. (2024). The method relies on embeddings extracted from the encoder of a self-supervised model, such as `wav2vec 2.0 XLSR-53`. Hidden-state representations from a mid-level transformer layer (typically layer 12) are subjected to k-means clustering (with k=300) to generate cluster identifiers, or pseudo-tokens. For each language — both the target and the donor — frequency distributions of these pseudo-tokens are computed from untranscribed speech corpora. The distributions are then compared using cosine similarity, yielding the ATDS score. A higher score indicates that the donor language's acoustic token distribution more closely resembles that of the target language.

The choice of layer from which embeddings are extracted is crucial. Acoustic models do not encode the same type of information across their depth: lower layers tend to capture raw acoustic features, while higher layers encode more abstract and language-specific patterns. Empirical evidence suggests that intermediate layers (approximately layers 9–13) achieve the best trade-off between phonetic precision and language generality, making them optimal for cross-lingual comparison. These layers contain representations that are sufficiently abstract to generalise across languages but not yet over-specialised to the pre-training data.

Probing studies such as Pasad et al. (2021) demonstrate that ABX phone discrimination performance — a proxy for phonetic discriminability — peaks in these middle layers of wav2vec 2.0 and XLSR-53, confirming their suitability for phonologically oriented analyses. Consequently, San et al. (2024) selected layer 12 for their ATDS computations, as it empirically balances phonetic distinctiveness and cross-lingual generality.

In the present study, we extend this approach by systematically observing ATDS values across multiple layers and by comparing their distributions across models with varying linguistic capacities: the “less competent” XLSR-53, the multilingual MMS, and their respective fine-tuned variants. This comparative analysis enables us to evaluate how representational depth and model diversity influence the interpretability and discriminative power of the ATDS metric.

4. Results

4.1. First Experiment

The results of the first experiment, presented in the *Table 3*, yield a pertinent conclusion for the broader task of automating linguistic documentation. The MMS model demonstrates a clear and substantial superiority across all test conditions, particularly on the in-domain data. As expected, the performance of all the models decreases on the out-of-domain voices and dialects; however, the MMS model remains the most robust in this challenging scenario. The superior performance of the original MMS model can be partly attributed to its pre-training data, which almost certainly included Yiddish. The performance gap between the original MMS model and its fine-tuned version for ASR is not large. However, if we consider applying our methodology to even lower-resource languages, the experiments with XLSR-53—which had no prior exposure to Yiddish—prove more informative. In this context, a clear performance difference emerges between the original and fine-tuned models.

The results from the XLSR models (both original and fine-tuned) confirm that leveraging a model pre-adapted to a related task is, in general, more effective than using a generic base model. Furthermore, the choice of fine-tuning language is critical. The model fine-tuned on Hebrew performs considerably worse than the other fine-tuned variants and demonstrates a reduced capacity for generalisation to both in-domain and out-of-domain data. The models fine-tuned on Slavic languages underperformed compared to those based on Germanic languages, underscoring the significant influence of genetic linguistic proximity. The German model proved most accurate on the in-domain data, reflecting its close phylogenetic relationship with Yiddish. Conversely, the English model adapted better to the Hasidic data, likely due to the documented, sustained contact with the English language.

It should be noted that the performances presented in the results could still be improved for each model individually; however, this aspect was intentionally excluded from the present experiment. A hyperparameter search was conducted using the XLRS-53 model. To ensure comparable experimental conditions, the same set of hyperparameters was subsequently applied to all other models. It is acknowledged that this decision may be subject to criticism, as optimal fine-tuning performance typically requires individualised hyperparameter optimisation for each model.

4.2. Second Experiment

Figure 1 and *Figure 2* illustrate the evolution of the Acoustic Token Distribution Similarity (ATDS) values across the transformer layers of the XLSR-53 and MMS acoustic models. Both models display a systematic organisation of linguistic proximity, but their behaviour across layers differs substantially. For greater clarity of visualisation, four language pairs were excluded from the graphs. The complete set of results will be presented in the general discussion in Section 5, where the correlations are discussed in greater detail.

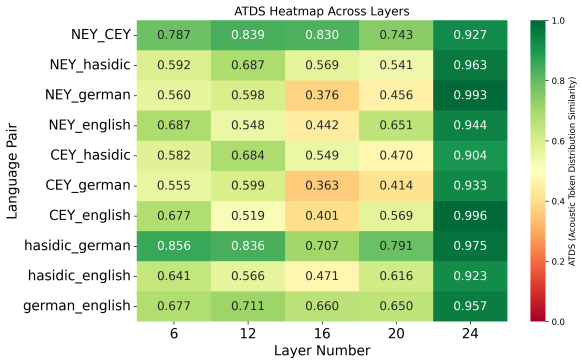

Figure 1: ATDS values across the transformer layers of the XLSR-53.

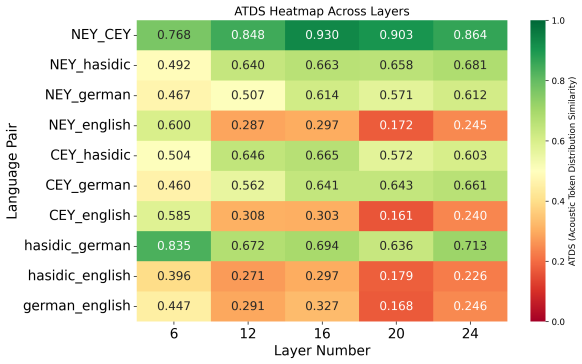

Figure 2: ATDS values across the transformer layers of the MMS.

In the XLSR-53 model, the ATDS values tend to stabilise early and vary less across layers, indicating a relatively homogeneous representation of acoustic space. The middle layers (around layer

Model	Validation		Test (in-domain)		Test (out-of-domain)	
	WER	CER	WER	CER	WER	CER
XLSR-53	0.452	0.114	0.621	0.141	0.662	0.226
German	0.295	0.071	0.508	**0.108**	0.549	0.174
English	0.287	0.068	**0.500**	0.116	**0.520**	**0.160**
Polish	0.320	0.074	0.556	0.118	0.667	0.229
Ukrainian	0.320	0.075	0.570	0.132	0.672	0.219
Russian	0.333	0.076	0.559	0.129	0.701	0.230
Hebrew	0,478	0,107	0,712	0,188	0,784	0,281
MMS	**0.164**	**0.038**	0.448	0.122	**0.451**	0.146
MMS-all	0.179	0.042	**0.388**	**0.080**	0.461	**0.133**

Table 3: Performance of the ASR models on validation, in-domain and out-of-domain test sets (Word Error Rate (WER) / Caracter Error Rate (CER)).

12) achieve moderately high similarities for closely related dialect pairs such as NEY–CEY, but inter-dialect and cross-language distinctions remain less pronounced. This limited variation suggests that XLSR-53 captures some degree of cross-lingual structure, yet its internal representations are less sensitive to fine-grained phonetic variation and dialectal divergence. In other words, its representational space may conflate subtle distinctions between dialects or typologically close languages.

By contrast, the MMS model, trained on a much broader multilingual inventory (over a thousand languages), exhibits greater discriminative capacity. The heatmap shows stronger contrasts both between layers and between language pairs. For typologically close pairs such as NEY–CEY, ATDS values remain consistently high, often exceeding 0.9, confirming that MMS preserves robust phonetic similarity relations. However, for more distant pairs (CEY–English or NEY–English), the similarity drops sharply in mid-layers before rising again in higher layers, indicating layer-dependent reorganisation of linguistic representations.

This increased variation across layers in MMS suggests that multilingual pretraining enhances the model's ability to separate and hierarchise languages according to their acoustic and phonological relatedness. The higher discriminative power reflects the model's richer internal typological mapping—it better distinguishes close from distant varieties while maintaining coherent clusters for dialect continua.

Quantitatively, the MMS model demonstrates a noticeably higher inter-layer variance in ATDS values compared to XLSR-53, reflecting greater sensitivity to phonetic and typological contrasts across languages. The amplitude of variation between minimum and maximum ATDS values within the same language pair often exceeds 0.3 in MMS, whereas it remains below 0.15 for XLSR-53. This enhanced spread suggests that the MMS layers encode progressive abstraction of acoustic information, transitioning from surface phonetic features to language-specific patterns.

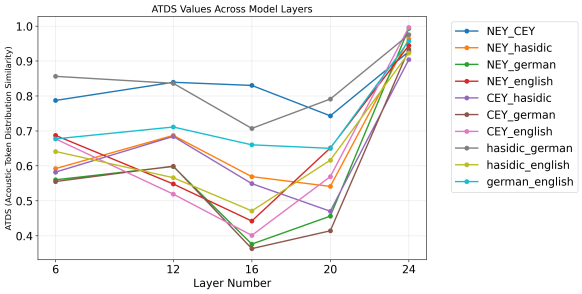

Figure 3: Linear visualisation of ATDS evolution across layers in XLSR-53.

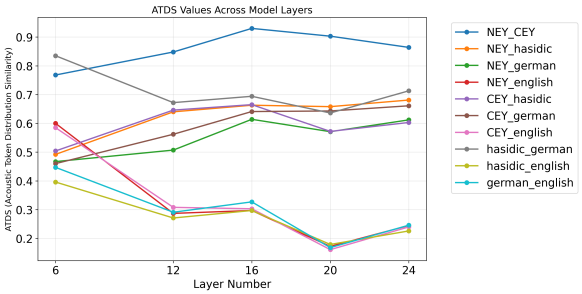

Figure 4: Linear visualisation of ATDS evolution across layers in MMS.

The linear visualisation of ATDS evolution across layers (*Figure 3* and *Figure 4*) provides an additional perspective on the internal organisation of the models. In the MMS representation, the trajectories of ATDS values form distinct clusters corresponding to degrees of linguistic relatedness: closely related dialects (e.g. NEY–CEY) show parallel, consistently high curves; moderately related pairs (e.g. CEY–Hasidic or NEY–German) exhibit intermediate and more fluctuating profiles; while distant pairs, such as Yiddish dialects versus English, follow markedly lower and divergent trajectories. This clear stratification suggests that the MMS model captures the continuum of linguistic similarity in a structured and interpretable way. In contrast, the XLSR model produces flatter and less differenti-

ated patterns, indicating weaker sensitivity to fine-grained dialectal or cross-linguistic variation. Thus, the MMS representation seems to reflect not only higher discriminative power but also a more coherent internal geometry of the multilingual acoustic space.

4.3. Correlation of Model Performances with the Linguistic Similarity ATDS

Preliminary correlation analysis of the results of two experiments indicates that higher ATDS values tend to align with lower WER and CER scores observed in fine-tuned ASR models, confirming that the metric captures meaningful dimensions of linguistic proximity. Therefore, MMS not only provides a more discriminative representational geometry but also a more interpretable mapping between acoustic similarity and empirical ASR performance. We chose the MMS model for further statistical analysis.

For examining the correlation between model performance and linguistic similarity (ATDS), we used Spearman's rank correlation as the primary measure and Pearson's correlation as a robustness check. Spearman's correlation was preferred because ATDS values and CER scores are not necessarily linearly related but are expected to exhibit a monotonic relationship. Moreover, the number of observations is relatively small, and both ATDS and CER are bounded between 0 and 1, which can distort linearity assumptions underlying the Pearson correlation.

We first calculated the correlation between the MMS model's performance and the ATDS values, focusing solely on the one Yiddish dialect (NEY) used for fine-tuning. These results are presented in *Table 4*.

Comparison	**Spearman ρ**	**p-value**
ATDS vs CER validation	-0.77	≈ 0.07
ATDS vs CER in-domain	-0.83	≈ 0.04
ATDS vs CER out-of-domain	-0.54	≈ 0.26

Table 4: Correlation of the MMS model performances with the linguistic similarity ATDS.

We can observe a strong negative correlation between ATDS values across the MMS model layers and the CER on validation set, which is close to significance given only 6 samples; significant negative correlation for in-domain CER test; and a moderate but not significant negative trend for out-of-domain CER.

Layer-wise analysis shows that mid-to-upper layers (16–24) yield stronger correlations—particularly at layer 16 (r = –0.83, p = 0.04 for out-of-domain data)—suggesting that representations in these layers encode more abstract phonological regularities relevant to transfer learning. Lower layers (e.g., 6) show weaker and less consistent associations, likely because they predominantly capture speaker- or channel-specific information rather than linguistic structure. Taken together, these results support the hypothesis that mid-level representations in self-supervised speech models provide the most linguistically meaningful features for predicting transfer success.

These results suggest that acoustic similarity (ATDS) is a good predictor of fine-tuning efficiency, particularly for validation and in-domain conditions. The correlation weakens for out-of-domain evaluation, where other factors (domain shift, speaker variation, lexical divergence) likely dominate. Overall, this supports the hypothesis that donor languages acoustically closer to the target dialect yield more efficient fine-tuning — but the relationship is not strictly linear and depends on evaluation context.

To determine whether the ATDS measure maintains its predictive reliability across dialectal diversity, we calculated the same correlations for the Hassidic Yiddish dialect, which was not seen by the model during fine-tuning. Although the correlations are weaker than for the NEY and CEY dialects used in fine-tuning, the same tendency persists. Higher ATDS values (i.e., stronger acoustic similarity) are associated with lower character error rates, particularly in validation and in-domain settings.

Interestingly, Hasidic Yiddish, being more phonetically divergent due to influences from English and Hebrew, might exhibit greater internal variation. This could blur the correspondence between ATDS and CER across donor languages.

In sum, this dataset still supports the overall hypothesis that ATDS captures meaningful information about transfer potential, even if the predictive relationship weakens as dialectal divergence increases.

5. Discussion and Conclusion

An analysis of the full set of extracted ATDS values (Figure 5) against the backdrop of the Yiddish dialect continuum and its genetically related languages reveals a divergence from established linguistic knowledge and intuition. The values from layers 12 and 16, for instance, suggest that Hassidic Yiddish is acoustically closer to German or Polish than to other Eastern European Yiddish dialects. In contrast, the distances between distinct languages appear more coherent with their genetic relationships: the Germanic languages (German and English) show higher similarity with the likewise Germanic Yiddish than do the Slavic languages or Hebrew, despite the latter's significant lexical contributions to Yiddish.

It is important to note that the absolute values

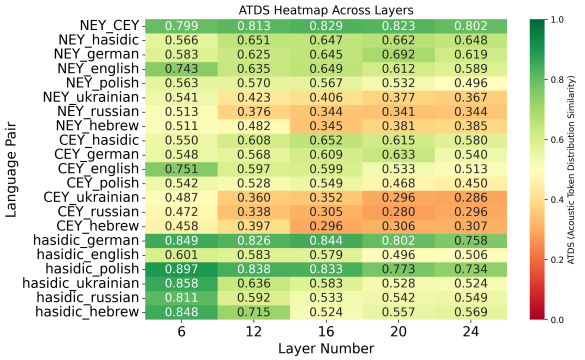

Figure 5: ATDS values across the transformer layers of the MMS for all the linguistic varieties of the experiment.

here differ from those in Figure 2, as they originate from a separate extraction run; however, they remain within the expected confidence interval (CI).

This observation leads to four broader conclusions. First, the ATDS measure is inherently model-dependent, reflecting how a specific model (e.g., XLSR vs. MMS) perceives and encodes acoustic regularities. A model trained or fine-tuned on particular phonetic inventories will inherently bias its latent space representation.

Second, an acoustic model represents the speech signal in a highly complex manner. While experimental evidence suggests middle layers capture language-specific phonetic features, we cannot exclude the influence of other acoustic information. Factors such as recording quality, microphone type, background noise, and speaker-specific characteristics (e.g., voice type, pitch, and speaking rate) can still permeate these layers. The ATDS, therefore, may be capturing a confluence of linguistic and non-linguistic acoustic properties, which can obscure purely phonetic relationships.

Third, phonetic "closeness" in the model's latent space is a distinct concept from genealogical or typological linguistic closeness. For instance, a model like MMS, which is explicitly designed for cross-lingual normalization, might project typologically distant languages into similar regions of its embedding space if they share certain acoustic properties, leading to artificially inflated ATDS scores.

Finally, dialects can appear "further apart" in the model's space than distinct languages. This may occur because the model, lacking explicit training on fine-grained dialectal variation, fails to normalize subtle phonetic differences (e.g., slight vowel shifts or prosodic patterns), thereby over-amplifying them.

Despite these caveats regarding absolute ATDS values, the correlations between ATDS and model performance (CER/WER) remain statistically and practically meaningful. They demonstrate that greater similarity within the model's own latent space is a reliable predictor of better transfer learning performance, regardless of whether that alignment matches external linguistic typologies. In other words, while the absolute distances may not align with linguistic theory, the directional trend—higher ATDS predicts lower CER—confirms the utility of the model's internal "acoustic closeness" for forecasting transfer success. Conceptually, ATDS describes model-internal similarity, not external phonetic typology. Statistically, the presented correlations quantify how well this internal geometry predicts empirical performance. Consequently, the question of adapting this method for multidialectal and fine-grained analysis remains open. Future work must address how to disentangle the confounding acoustic factors mentioned above to isolate a purer phonetic signal. This might involve techniques such as data normalization, adversarial training to remove nuisance variables, or the development of more sophisticated metrics that can separate speaker and channel effects from dialectal identity, thereby unlocking ATDS's full potential for detailed dialectological study.

6. Acknowledgments

The work is supported by the French National Research Agency and Ministry of Higher Education, Research and Innovation (MESR).

7. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *ArXiv*, abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. ArXiv:2006.11477 [cs, eess].

Isaac L Bleaman and Chaya R Nove. 2025. The corpus of spoken yiddish in europe: Goals, methods, and applications. *Language Documentation Conservation*, 19.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Malgorzata Ćavar, Damir Ćavar, Dov-Ber Kerler, and Anya Quilitzsch. 2016. Generating a yiddish speech corpus, forced aligner and basic asr system for the aheym project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4688–4693.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in Arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.

Elye Falkovitsh. 2024. *Yidish. Fonetik, grafik, leksik un gramatik*. De Gruyter, Brill.

Neil G. Jacobs. 2005. *Yiddish: A linguistic introduction*. Cambridge University Press.

Zhou Jie, Gao Shengxiang, Yu Zhengtao, Dong Ling, and Wang Wenjun. 2024. Dialectmoe: An end-to-end multi-dialect speech recognition model with mixture-of-experts. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1148–1159.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Chaya R. Nove. 2021. *Outcomes of language contact in New York Hasidic Yiddish*, pages 43–71. Berlin: Language Science Press.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Inf. Process. Manage.*, 60(2).

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1).

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 100–112, St. Julian's, Malta. Association for Computational Linguistics.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Jeffrey Veidlinger. 2013. *In the Shadow of the Shtetl: Small-Town Jewish Life in Soviet Ukraine*. Indiana University Press.

Jacob Webber, Samuel K. Lo, and Isaac L. Bleaman. 2022. Reyd – the first yiddish text-to-speech dataset and system. In *Interspeech 2022*, pages 2363–2367.