A Quantitative Approach to Understand Self-Supervised Models as Cross-lingual Feature Extractors

Shuyue Stella Li[1⋆], **Beining Xu**[2⋆], **Xiangyu Zhang**[1⋆], **Hexin Liu**[3], **Wenhan Chao**[2], **Leibny Paola Garcia**[1]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]School of Computer Science and Engineering, Beihang University
[3]School of Electrical and Electronic Engineering, Nanyang Technological University
sli136, xzhan233, lgarci27@jhu.edu

Abstract

In this work, we study the features extracted by English self-supervised learning (SSL) models in cross-lingual contexts and propose a new metric to predict the quality of feature representations. Using automatic speech recognition (ASR) as a downstream task, we analyze the effect of model size, training objectives, and model architecture on the models' performance as a feature extractor for a set of topologically diverse corpora. We develop a novel metric, the Phonetic-Syntax Ratio (PSR), to measure the phonetic and synthetic information in the extracted representations using deep generalized canonical correlation analysis. Results show the contrastive loss in the wav2vec2.0 objective facilitates more effective cross-lingual feature extraction. There is a positive correlation between PSR scores and ASR performance, suggesting that phonetic information extracted by monolingual SSL models can be used for downstream tasks in cross-lingual settings. The proposed metric is an effective indicator of the quality of the representations and can be useful for model selection.[1]

1 Introduction

Self-Supervised Learning (SSL) has become a paradigm for learning feature representations from unlabeled data (Liu et al., 2023). In speech processing, self-supervised approaches for learning speech representation are often used to extract features for downstream tasks. These representations can replace the handcrafted feature such as Mel Spectrum or MFCC in many tasks as they are able to extract high-level properties in the speech data (Mohamed et al., 2022; Chung et al., 2019).

English SSL Models take advantage of the high availability of English data and outperform traditional feature extraction methods on a range of downstream tasks in English (Chen et al., 2022; Hsu et al., 2021; Liu et al., 2020). Since the acoustic and phonetic information of human speakers across languages share a level of similarity, it is crucial to study the cross-lingual transfer performance of English SSL models as a feature extractor for non-English audio data (Li et al., 2020; Cho et al., 2018). This will enhance our understanding of the composition of knowledge learned during pre-training, allowing more efficient use of data during model selection. Furthermore, if we are able to use English monolingual models effectively in multilingual downstream tasks, the high cost of training massive multilingual speech models such as XLSR (Babu et al., 2021; Conneau et al., 2021) and mSLAM (Bapna et al., 2022) can be reduced by explicitly incorporating architectural designs promoting cross-lingual transfer. Therefore, the first purpose of this paper is to investigate the factors that improve the ability of monolingual SSL models to extract useful speech representations for ASR tasks in typologically diverse languages.

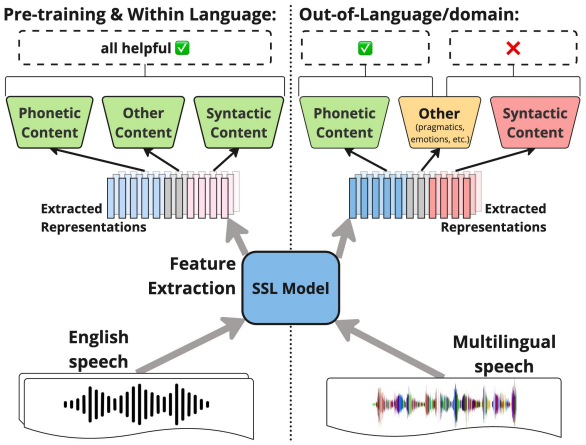

Figure 1: Speech data of English (in-domain) and other languages (out-of-domain) are passed through the SSL models to extract speech representations. All information is expected to aid downstream tasks in English while phonetic content is expected to be useful for out-of-domain downstream tasks; "other" content may include speaker information, etc.

[1]We make our work open-source for further explorations: https://github.com/stellali7/SSL_PSR

The second objective of our study is to analyze the amount of phonetic information versus syntactic information learned by the model during training, and how the phonetic-syntax composition in the model impacts the extracted features. Phonetic content directly impacts the learned phonological structure in the representations. Explicit integration of phonological knowledge has proven to be extremely successful in speech processing (Zhan et al., 2021). On the other hand, semantic and syntactic knowledge learning in the target language during fine-tuning is needed for ASR tasks so that the SSL models do not retain source language semantics and syntax, implying syntactic information might be harmful for cross-lingual feature extraction (Li et al., 2020).

As shown in Figure 1, we expect the pre-trained SSL models to efficiently extract phonetic, syntactic, and other contents to help downstream tasks in English (Chung et al., 2021). At the same time, the extracted phonetic information in out-of-domain and multilingual situations should also aid downstream performance. Therefore, we propose a novel metric to quantify the amount of helpful phonetic information. To the best of our knowledge, this study is the first to quantitatively understand the capabilities and limits of SSL models from a linguistic perspective. Our contributions include:

- We examine five SSL models with different sizes, data preparation methods, and training objectives by analyzing their cross-lingual generalizability as feature extractors on the ASR task.
- We propose a new metric, Phonology-Syntax Ratio (PSR), to measure the phonetic and syntactic content extracted by an SSL model on any given out-of-domain/language dataset. A higher PSR score correlates to a better ASR performance.
- We localize the phonetic content in the SSL model to specific layers using the trained layer-wise weights for the feature representations.

2 Related Work

2.1 Self-Supervised Models

Self-supervised learning (SSL) (Liu et al., 2023; Bengio et al., 2013; Raina et al., 2007) takes advantage of easily accessible unlabeled data to learn a model and then produces universal representations by solving upstream tasks (Liu et al., 2022b). Then, the pre-trained SSL model can be used to process unseen data based on its previous knowledge and handle multiple downstream tasks. SSL models have achieved superior performance in natural language processing (Devlin et al., 2019; Peters et al., 2018), computer vision (Chen et al., 2020; Misra and van der Maaten, 2020), speech processing (Chen et al., 2019; Chi et al., 2021), and especially ASR (Baevski and Mohamed, 2020; Ravanelli et al., 2020; Jiang et al., 2021). In our work, we study a number of SSL models and their feature extraction ability when presented with input from other languages.

2.2 Audio Feature Extraction

Before any downstream speech processing tasks, the audio data is converted to high-dimensional feature vectors through an audio feature extraction system (Moffat et al., 2015). Classic methods, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP) extract cepstral coefficients that contain low-level acoustic features (Dave, 2013; Shanthi and Lingam, 2013). Researchers have also delved into neural-based models, leveraging pre-trained models on large-scale datasets to boost performance (Chi et al., 2021). While progress has been remarkable, challenges such as robustness to noise variations and interpretability of learned features continue to stimulate further research in this domain (Mohamed et al., 2022). In our work, we explore the robustness of the monolingual SSL models when generalized to multilingual settings, from which we interpret the features extracted by these models.

2.3 Automatic Speech Recognition (ASR)

ASR transcribes given audio to text in the script of the spoken language (Malik et al., 2021; Yu and Deng, 2016). Deep neural network (DNN) based techniques (Hinton et al., 2012) have boosted the accuracy of ASR by replacing the traditional Gaussian Mixture Model in cascaded systems involving separate acoustic, language, and lexicon components (Li et al., 2022). End-to-end models (Graves and Jaitly, 2014; Chorowski et al., 2014; Bahdanau et al., 2016; Collobert et al., 2016) have recently become a breakthrough in the speech community, directly translating an input speech sequence into an output text sequence with a single model. Some publicly available and commonly used toolkits include Kaldi (Povey et al., 2011), CMU Sphinx (Lee et al., 1990), SpeechBrain (Ravanelli et al., 2021) and ESPNet (Watanabe et al., 2018).

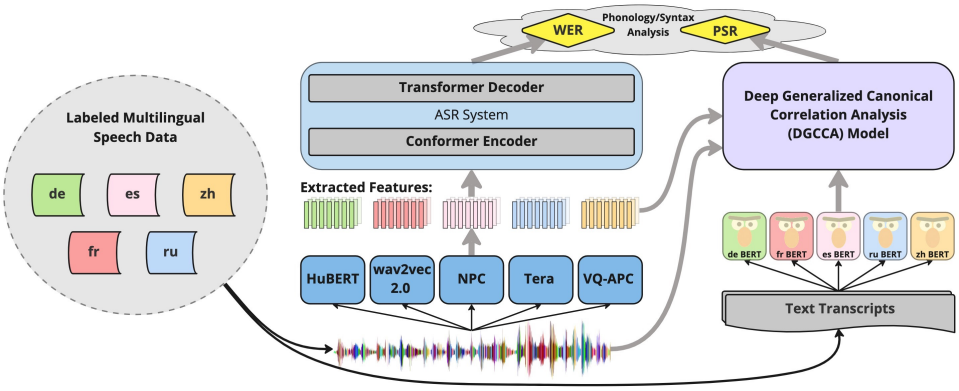

Figure 2: The pipeline to measure the performance of SSL model on different languages. We first use each SSL model as a feature extractor for data in each language and compute a WER score for the ASR task. Then, we calculate the PSR of the representations to analyze the correlation between the ASR performance and the PSR score.

2.4 Analysis Methods of SSL Models

There has been extensive research on analyzing supervised speech models (Belinkov and Glass, 2019; Palaskar et al., 2019; Prasad and Jyothi, 2020). However, research on SSL models, especially in the speech domain, is still relevantly under-explored. Some recent work in this field includes a similarity analysis of self-supervised speech representations, in which they only looked into simpler models such as APA, CPC, and MPC (Chung et al., 2021). Liu et al. (2022a) attempted to distinguish useful representations in SSL models for spoken language identification and reduce spurious information in the representations, but was limited to a specific task. Pasad et al. (2021) and Pasad et al. (2023) analyzed the layer-wise acoustic-linguistic content of pre-trained models by performing layer-independent Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) on English data. However, since the features extracted by DNN models often have high dimensionality (Georgiou et al., 2020), CCA is limited in its ability to freely model complex nonlinear relationships.

2.5 Cross-lingual Knowledge Transfer

Cross-lingual transfer learning has gained attention in the field as it effectively mitigates resource constraints and language-specific challenges, but most importantly to our work, it requires the model to be able to adapt to unseen situations such as a new language (Khurana et al., 2023; Conneau et al., 2020). Effective cross-lingual transfer for speech processing requires the model to have a high-level understanding of both text linguistics and phonetics. Previous work has shown that multilingual models generalize well to target languages (Conneau et al., 2021; Singh et al., 2019; Radford et al., 2023). Lauscher et al. (2020) shows that the quality of the cross-lingual transfer is correlated with the linguistic similarity between the source and target languages. Inspired by this, we use English monolingual models in our work to better compare the linguistic distance between the pre-train data and the target data. Studying the generalization ability of monolingual models to unseen languages allows us to better analyze the learned representations and localize the factors that facilitate cross-lingual transfer for more efficient model design.

3 Analysis Methods

As shown in Figure 2, we first use the SSL models trained on English to extract speech representations on audio data from German (de), French (fr), Spanish (es), Russian (ru), and Chinese (zh). Then, we use the ASR task to evaluate the quality of the extracted features against a Mel Spectrum baseline in Section 3.1. We correlate the WER scores to traditional measures of linguistic distance in Section 3.2. Finally, we quantitatively evaluate the phonetic and syntactic content in the extracted features for each language, as described in Section 3.3.

3.1 Measuring Multilingual Generalizability

We use the standard ASR task on 5 genealogically and typographically diverse languages to evaluate the generalizability of the English SSL models as a cross-lingual feature extractor. To fairly compare the models, we freeze the parameters of the models and use the same downstream architecture (Conformer + Transformer) for all SSL models and the Mel Spectrum baseline feature extractor. We also use the same language model setup and beam size during decoding.

Our pipeline is shown in Figure 2. We select SSL models based on their training methods. These upstream SSL models can be categorized into **masked reconstruction model**: Tera (Liu et al., 2021b) and NPC (Liu et al., 2021a); **masked prediction model**: HuBERT (Hsu et al., 2021); **auto-regressive reconstruction model**: VQ-APC (Chung et al., 2020); and **contrastive model**: wav2vec2.0 (Baevski et al., 2020). Inspired by the setup in SUPERB (wen Yang et al., 2021) and ELMO (Peters et al., 2018), we take the weighted sum from all layers as the extracted representation, and the weight vector is updated during training.

For the downstream model, we use the Conformer (Gulati et al., 2020) as the encoder and the Transformer (Vaswani et al., 2017), which has achieved state-of-the-art (SOTA) results in many speech recognition tasks (Ma et al., 2021). During data analysis, we isolate the effect of the SSL model as a feature extractor by taking the difference (Δ) between the SSL feature extractor and the Mel Spectrum baseline performance. This eliminates any potential noise introduced by data size differences, speech formality levels, and other linguistic differences between languages, allowing a fair comparison between different SSL models. When decoding, we use a simple RNN as a language model and keep the parameters consistent across all tasks.

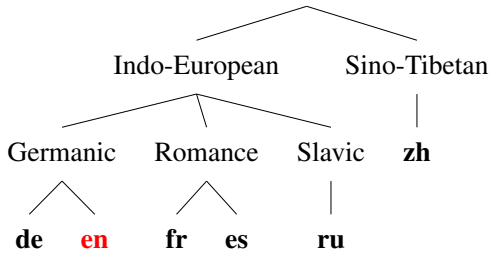

Figure 3: Phylogenetic Tree of Target Langauges

3.2 Measuring Linguistic Distance

We examine the performance of self-supervised models on languages across a diverse range of families and groups in order to investigate the relationship between model performance and linguistic distance. In our analysis, we employ the phylogenetic tree in Figure 3 derived from the theory of language evolution with genetic distance equaling the Levenshtein distance (Serva and Petroni, 2008) as a measure of linguistic distance. Since languages evolve with both their written and spoken forms, the phylogenetic tree will contain the most comprehensive information about the language.

3.3 Measuring Phonetic & Syntactic Content

In this section, we describe approaches to quantify phonetic and syntactic content in the extracted speech representations of SSL models.

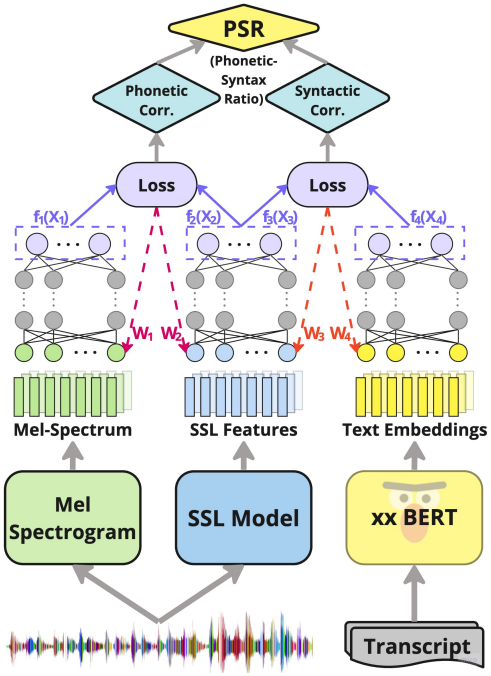

Figure 4: DGCCA pipeline. The model aims to compare the representation extracted by the SSL model to the pure acoustic representation (from Mel Spectrum) and pure syntactic/semantic representation (from BERT).

3.3.1 DGCCA

In order to better analyze the phonetic and syntactic content of the features, we use a tool called Deep Generalized Canonical Correlation Analysis (DGCCA), which is a deep learning technique that measures the nonlinear relationship between arbitrarily many views of the data and learns a view-independent representation (Benton et al., 2019). DGCCA effectively quantifies the phonetic and syntactic content of SSL models when treating the features extracted with different models as different views of the same data.

As shown in Figure 4, DGCCA takes N pairs of data vectors across J views as input and returns a correlation score as a measure of the similarity between the vectors. Using standard backpropagation to optimize the weight matrices $W_j = \{W_1^j, \ldots, W_{K_j}^j\}$, we try to find the linear transformation $U_j \in \mathbb{R}^{d_j \times N}$ of $f_i(X_j) \in \mathbb{R}^{o_j}$ constrained by $GG^T = I_r$ such that:

$$\underset{U_j \in \mathbb{R}^{d_j \times N}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} \sum_{j=1}^{J} \|G - U_j^T f_j(X_j)\|_F^2, \quad (1)$$

where $X_j \in \mathbb{R}^{d_j \times N}$ is the input feature vectors of the j^{th} view; f_j is the function learned using a multilayer perceptron of K_j layers; d_j is the dimension of the j^{th} view and r is the dimension of the learned representation G.

In our case, N is the number of utterances in the test data, where we have the SSL features and Mel Spectrum features of each utterance, as well as the BERT representations of its transcript. The monolingual BERT model in each target language is used when extracting the textual representations.

Features extracted by the SSL models, pure phonetic features (Mel Spectrum), and pure textual features (BERT representations) can be considered as different views (f_i) of the data. The correlation scores between different views are the loss of the converged DGCCA network. We compute the correlation scores between each of the latter two views and the SSL features. The correlation scores between the SSL features and the Mel Spectrum measure the **Phonetic Content** in the extracted features; the correlation scores between the SSL features and the BERT representations measure the **Syntactic Content** in the extracted features.

3.3.2 Phonetic-Syntax Ratio (PSR)

We introduce a new metric: the Phonetic-Syntax Ratio (PSR) in order to quantitatively investigate the phonetic and syntactic content on SSL representation. As described in Section 3.3.1, the similarity to phonetic features and the similarity to syntactic features of the SSL representations are both optimized and quantified as correlation scores when training the DGCCA network. We define the PSR as the ratio between the phonetic correlation score and syntactic correlation score, weighted equally among all data points:

$$PSR = (\frac{1}{n} \sum_{i=1}^{n} \frac{phonetic\ score_i}{syntax\ score_i} - 1) \cdot 100\%, \quad (2)$$

where phonetic scores and syntax scores are the output of DGCCA when the SSL representations are fed in with the Mel Spectrum and BERT contextualized embedding, respectively. The PSR score is model-agnostic and language-agnostic, and can be used for a range of contrastive analysis for inferring cross-lingual transferability.

4 Experimental Setup

4.1 Datasets

We investigate the cross-lingual adaptation capability of English SSL models in five languages. For training the ASR models, we use the Mozilla Common Voice 5.1 dataset (Ardila et al., 2020) for German, French, Spanish, and Russian, and we use the OpenSLR ST-CMDS-20170001_1 Free ST Chinese Mandarin Corpus[2] for Chinese. The Common Voice English test set is used for DGCCA analysis. More details about the datasets are in Table 1.

Lang	hr	voices	train	dev	test
de	751	11,731	196,464	15,341	15,341
fr	605	11,960	254,863	15,621	15,621
es	522	18,906	138,878	14,860	14,860
ru	117	927	13,189	7,242	7,307
zh	-	-	92,280	4,299	4,483
en	1933	61528	435,947	16,029	16,029

Table 1: Dataset description; the number of hours, voices, and utterances for each split. Hour and voice statistics for the Chinese corpus are not available as it is distributed after preprocessing. The number of speakers for the Chinese dataset is 855. Train and dev splits of English were not used.

4.2 Multilingual Generalizability Setup

We use the ASR performance on a range of typologically diverse languages as a metric to infer the models' multilingual generalizability. In order to fairly compare the performance of each SSL model in different language datasets, we use the same downstream model for all languages and features and focus on the within-language difference between the SSL model and the baseline model.

Self-supervised feature extractors We examine a number of English SSL speech models including HuBERT (Hsu et al., 2021), wav2vec 2.0 (Collobert et al., 2016), NPC (Liu et al., 2021a), TERA (Liu et al., 2021b), and VQ-APC (Chung et al., 2020) with model details shown in Table 2. Unlike the baseline model, we use a smaller learning rate considering that self-supervised training usually uses a small learning rate. We use a learning rate of 0.0025 with 40000 warmup steps.

Model architectures After multilingual features are extracted, we use a standard Conformer encoder and a Transformer decoder in our downstream ASR model and a stacked RNN as the language model

[2]http://www.openslr.org/38

Model	architecture	train objective	model size	pre-train	input	stride
HuBERT-BASE HuBERT-LARGE	CNN + Transformer	Predictive	95m 317m	LS-960 LL-60k	wav	20ms
wav2vec2-BASE wav2vec2-LARGE	CNN + Transformer	Contrastive + Diversity	95m 317m	LS-960 LV-53.2k	wav	20ms
NPC	Masked Conv Block	L1 Reconstruction	19.4m	LS-C-360	Mel	10ms
TERA-BASE	Unidirectional LSTM + Prediction Network	L1 Reconstruction	21.3m	LS-C-100	Mel	10ms
VQ-APC	Unidirectional LSTM	L1 Reconstruction	4.63m	LS-C-360	Mel	10ms

Table 2: SSL Model Summary. For the pre-training data description, LS = Librispeech, LS-C = Librispeech-clean, LL = Libri-light, and LV = Libri-vox.

during decoding. More details on the ASR model architecture and training are in Appendix A.

4.3 PSR Computation

We use the DGCCA pipeline shown in Figure 4 to compute the PSR scores for each langauge. The DGCCA model used consists of an MLP network with a Linear layer, a Sigmoid function, and a Batch Norm layer. Each group of tensors has one MLP network, and its output is passed into the DGCCA loss. We used SGD to optimize the network with a learning rate of 1e-6. We use features extracted by the HuBERT model from five different languages (German, French, Spanish, Russian, and English) and also extract its corresponding Mel Spectrum and BERT features. Chinese PSR is not reported because CER was used to evaluate the ASR performance, hence the comparison across languages would not be fair (more details in Section 5.2). When calculating the correlation scores, we use the test set in each target language as input to the DGCCA model with a batch size of 32. Details on the implementation and hardware of the SSL models and the DGCCA model can be found in Appendix B.

5 Results and Analysis

5.1 Multilingual Generalizability

Results from the multilingual ASR tasks are shown in Table 3, with both WER scores and the difference from the Mel Spectrum baseline (Δ).

In the zero-shot setting, it is generally expected that the SSL feature extractor trained on English, without any domain adaptation, performs poorly on the cross-lingual ASR tasks compared to the Mel spectrum baseline. Although it can extract higher-dimensional features, additional English syntactic information in the SSL model can be projected onto the new language (Georgiou et al., 2020). Therefore, the purpose of this experiment is not to improve the SOTA results but rather to probe the SSL models for further phonetic-syntactic analysis.

There are five SSL models being evaluated in this experiment in five languages. The column Avg on the right marginal of Table 3 shows the overall performance of each SSL model in all languages. In general, wav2vec2.0-LARGE significantly outperforms other feature extractors and has a consistent result across languages. There are two instances in which wav2vec2.0-LARGE outperforms the pure acoustic Mel Spectrum baseline. This can be attributed to the cross-lingual phonetic information transfer that the model learned from English pre-training.

5.1.1 Effect of Training Objectives

The HuBERT and wav2vec2.0 models consistently perform better than NPC, TERA, and VQ-APC. HuBERT and wav2vec2.0 both effectively combine CNN encoders with Transformers in their architecture. The attention mechanism allows the models to effectively encode speech features into the latent embedding space and learn contextualized representations. Both HuBERT and wav2vec2.0 use similar architectures and identical pre-training data and setups. However, HuBERT as a cross-lingual feature extractor does not perform as well due to its predictive loss compared to the **contrastive loss** of wav2vec2.0. The masked prediction task during HuBERT pre-training forces the model to learn the language model as well as the acoustic model from continuous English speech inputs (Hsu et al., 2021), so the model might be overfitted to English syntax.

Now we discuss the performance of NPC, TERA, and VQ-APC, which are significantly smaller than wav2vec2.0 and HuBERT both in model and data size. TERA and NPC have comparable model sizes, training objectives, input format, and stride during pre-training, but TERA outperforms NPC with less than one-third of the training data. This is

Model/Lang	de	Δ	fr	Δ	es	Δ	ru	Δ	zh	Δ	Avg.	Δ
Mel (Baseline)	10.0	-	**15.8**	-	**11.5**	-	7.9	-	9.4	-	**10.92**	-
HuBERT-BASE	11.3	1.3	16.5	**0.7**	13.1	1.6	7.8	-0.1	9.8	0.4	11.70	0.78
HuBERT-LARGE	12.4	2.4	16.6	0.8	12.0	**0.5**	8.3	0.4	**9.1**	**-0.3**	11.68	0.76
wav2vec2-BASE	11.8	1.8	16.7	0.9	13.4	1.9	8.5	0.6	9.8	0.4	12.04	1.12
wav2vec2-LARGE	**9.2**	**-0.8**	16.6	0.8	12.3	0.8	**7.6**	**-0.3**	9.4	0	11.04	**0.10**
NPC	16.2	6.2	18.1	2.3	16.1	4.6	11.0	3.1	10.7	1.3	14.42	3.5
TERA-BASE	15.6	5.6	17.1	1.3	14.8	3.3	10.3	2.4	10.0	0.6	13.56	2.64
VQ-APC	13.5	3.5	17.2	1.4	17.3	5.8	12.1	4.2	10.8	1.4	14.18	3.26
Avg.	12.86	2.86	16.97	1.17	14.14	2.64	9.37	1.47	9.94	0.54	-	-

Table 3: Word Error Rate (WER) of German (de), French (fr), Spanish (es), and Russian (ru). For Chinese (zh), we apply Character Error Rate (CER) as the evaluation metric. Δ is the difference from Baseline, the lower the better. wav2vec2.0-LARGE achieves the best performance and the Transformer-based models generally perform better.

due to the alterations in the time, frequency, and magnitude axes of the data during pre-training, which increases **data diversity** and enforces accurate phoneme prediction (Liu et al., 2021b). On the other hand, VQ-APC achieves comparable results as NPC with a much smaller model size. With all the other setups identical, this suggests that the **sequential structure** learned by the Unidirectional LSTM (APC) and the **quantization layers** are more effective at capturing speech representations than convolutional blocks in NPC, implying that speech should be treated as sequential data.

5.1.2 Effect of Model Size

Comparing the HuBERT-BASE / HuBERT-LARGE and wav2vec2.0-BASE / wav2vec2.0-LARGE pairs gives insight into the effect of model size on downstream ASR tasks. The LARGE models generally perform better than the BASE models. This is consistent with a previous study by Pu et al. (2021), in which they empirically showed that scaling SSL models results in improvements in both L1 loss and accuracy on downstream tasks consistent with the power law. Larger models are also more data-efficient when labeled data is scarce. The advantage of the LARGE model over the BASE model is especially apparent on the wav2vec2.0 pair, as wav2vec2.0-LARGE consistently performs better across all languages. As discussed in Section 5.1.1, the more efficient use of data in HuBERT-LARGE may have caused it to learn even more syntactic and semantic representation, which does not benefit cross-lingual speech feature extraction.

5.2 Linguistic Analysis

Now we discuss the performances of all five languages based on their average scores. Smaller Δ indicates better generalizability. According to the phylogenetic tree shown in Figure 3, both German and English belong to the Germanic branch; French, Spanish, and Russian are in different language groups as English; Chinese belongs to another language family. As shown in Table 3, English SSL models have better generalizability in French than in German. This is because French has a profound phonological influence on the development of English (Roth, 2010), and the latter not only borrows some French pronunciation rules, but also shares contextual phonetic similarities of pitch contours (So and Best, 2014). For German, although it appears to have poor SSL performance with high Δ values, the absolute WER is the lowest among German, French, and Spanish, which have similar training sizes. From this, it can be observed that SSL representations has diminishing returns in high-resource situations.

Features extracted by the SSL models also perform well in Russian and Chinese ASR tasks. This might seem surprising, but it is because both Russian and Chinese are low-resource with less than 100k utterances. This demonstrates the robustness of SSL models in low-resource settings and establishes promising directions to generalize to other low-resource languages. Moreover, although Chinese is in the Sino-Tibetan language family, it actually has some phonotactic similarities with English (Ann Burchfield and Bradlow, 2014; Yang, 2021). It is important to note that the CER was used as the metric for Chinese ASR to avoid additional noise introduced by a word segmentation model, so the Chinese results should only be compared across models rather than across languages.

Analysis by linguistic distance can provide some plausible explanations for the results, but there still exist some inconsistencies. These inconsistencies motivate our next section, PSR Analysis, in which we use our novel metric to explain the model performance by categorizing and quantifying linguistic information in the extracted representations.

5.3 PSR Analysis

PSR scores of HuBERT-BASE on English and the target languages are shown in Table 4. As described in Equation 2, the larger the PSR, the more phonetic content in the feature set. First, to validate the PSR scale, we test the SSL features extracted from an English corpus by the SSL model. The PSR value from the English corpus is close to zero, which conforms with the intuition that the English-trained HuBERT model is able to extract useful information in both the phonetic and syntactic fields.

Lang	en	de	fr	es	ru
PSR	.01	.15	.16	.13	.23
WER Δ	-	1.3	0.7	1.6	-0.1

Table 4: PSR Results for Target Languages. A positive PSR means that the phonetic content in the extracted representations is stronger than the syntactic content.

Combined with the information in Table 3, we show that there is a positive correlation between the PSR scores of the feature group and the ASR performance of the model in that language. For example, the Δ value of HuBERT-BASE on German is higher (worse) than that of French and lower (better) than that of Spanish as shown in Table 3, and we see the corresponding relationship of their PSR values in Table 4: German PSR is lower (worse, less phonetic info) than French and higher (better, more phonetic info) than Spanish. This phenomenon indicates that the more phonetic information contained in a set of features, the better the performance of that set of features on cross-lingual or out-of-domain downstream tasks. Therefore, when the SSL model trained with English models is applied to the non-English corpus, *phonetic features are the main contributors to effective information compared with syntactic features*.

5.4 Layer Weights Analysis

All PSR scores shown in Table 4 are positive, suggesting that the features extracted by speech SSL models tend to have more phonetic information than syntactic information. This is partially due to the fact that the weighted sum of layers is used as input features to the ASR model and that the weights are optimized during training to put more emphasis on the phonetic information. Figure 5 shows the magnitude of the weights across all layers of HuBERT-BASE.

First, the layer-wise trend is consistent across all languages, suggesting each layer contains similar information even when trained on different datasets,

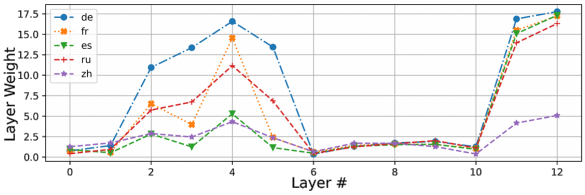

Figure 5: Layer-wise Weight Analysis.

i.e., the weights get updated similarly given the same task. The optimized weights gravitate toward layers that are crucial for the ASR task. The positive correlation between the ASR and PSR scores implies that the layers with large weights contribute to the high PSR scores, i.e. have denser phonetic than syntactic information. From Figure 5, Layers 4, 11, and 12 contribute significantly to the extracted features. Since lower layers contain lower-level information and vise versa, Layer 4 (and its adjacent layers) contain low or intermediate-level information on acoustic and phonetics important for the ASR task. The last two layers are the most salient because they contain high-level information related to human phonetics. Additionally, the weight for Layer 4 is larger in German and French, which are closer to English. This shows that when the pre-training and target languages are highly similar, the low-level phonetic features become more helpful. Our work to localize the phonetic content encoded in specific layers of HuBERT draws similar conclusions with Pasad et al. (2021) and Pasad et al. (2023), which localized various acoustic and linguistic properties in SSL models using CCA.

6 Conclusion

In this work, we studied English self-supervised speech models and probed for the phonetic and syntactic content in the extracted speech representations. We accomplished this using the SSL models as a feature extractor for downstream ASR task in multiple languages. Higher multilingual adaptability of a model is found to be positively correlated to the amount of phonetic information in the extracted representations. Most importantly, we propose a novel metric - the Phonetic-Syntax Ratio (PSR) - to quantify the phonetic and syntactic composition in the representations. PSR can serve as an effective indicator during model selection. We were also able to localize the phonetic information to certain layers in the SSL model. This is a call to other researchers to design smarter objectives when pre-training large models (such as focusing more on phonetic information learning) rather than simply increasing the model size.

Limitations

There are several limitations to our work. First, the value of our PSR was only tested on HuBERT due to limited computing resources. Although the scores reflect the ratio of acoustic and linguistic information in the features extracted by the SSL model, the performance of the corresponding downstream ASR task is not yet empirically shown in every SSL model. Second, the parameters in the SSL models are frozen during ASR training. Multilingual adaptability might be evaluated differently by unfreezing some or all layers of the SSL feature extractor. Finally, we did not calculate the PSR value for Chinese, as we did not find it to be a valuable data point given the Chinese ASR results are reported in CER only. Our choice to evaluate English SSL models is motivated by the abundance or English data, but other monolingual or multilingual models could be used given the abundance of data in the chosen langauge(s). For future directions, we believe that exploring spurious correlations among language pairs (e.g. phonotactical similarities between Chinese and English) is a fruitful direction that might shed light on language selection during cross-lingual transfer in speech models.

References

L. Ann Burchfield and Ann R. Bradlow. 2014. Syllabic reduction in Mandarin and English speech. *The Journal of the Acoustical Society of America*, 135(6):EL270–EL276.

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 Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of self-supervised pre-training for asr. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7694–7698.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philémon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2019. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 1–6, Florence, Italy. Association for Computational Linguistics.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Yi-Chen Chen, Sung-Feng Huang, Hung-yi Lee, Yu-Hsuan Wang, and Chia-Hao Shen. 2019. Audio word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1481–1493.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiát, Shinji Watanabe, and Takaaki Hori. 2018. Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*.

Yu-An Chung, Yonatan Belinkov, and James Glass. 2021. Similarity analysis of self-supervised speech representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3040–3044.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Proc. Interspeech 2019*, pages 146–150.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-Quantized Autoregressive Predictive Coding. In *Proc. Interspeech 2020*, pages 3760–3764.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

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, Online. Association for Computational Linguistics.

Namrata Dave. 2013. Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Theodoros Georgiou, Yu Liu, Wei Chen, and Michael Lew. 2020. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval*, 9(3):135–170.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764–1772, Bejing, China. PMLR.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. 29:3451–3460.

Dongwei Jiang, Wubo Li, Ruixiong Zhang, Miao Cao, Ne Luo, Yang Han, Wei Zou, Kun Han, and Xiangang Li. 2021. A further study of unsupervised pre-training for transformer based speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6538–6542.

Sameer Khurana, Nauman Dawalatabad, Antoine Laurent, Luis Vicente, Pablo Gimeno, Victoria Mingote, and James Glass. 2023. Improved cross-lingual transfer learning for automatic speech translation. *arXiv preprint arXiv:2306.00789*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

K.-F. Lee, H.-W. Hon, and R. Reddy. 1990. An overview of the sphinx speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Song Li, Lin Li, Qingyang Hong, and Lingling Liu. 2020. Improving Transformer-Based Speech Recognition with Unsupervised Pre-Training and Multi-Task Semantic Knowledge Learning. In *Proc. Interspeech 2020*, pages 5006–5010.

Alexander H. Liu, Yu-An Chung, and James Glass. 2021a. Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies. In *Proc. Interspeech 2021*, pages 3730–3734.

Andy T. Liu, Shang-Wen Li, and Hung-yi Lee. 2021b. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423.

Hexin Liu, Leibny Paola Garcia Perera, Andy W. H. Khong, Eng Siong Chng, Suzy J. Styles, and Sanjeev Khudanpur. 2022a. Efficient self-supervised learning representations for spoken language identification. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1296–1307.

Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller. 2022b. Audio self-supervised learning: A survey. *Patterns*, 3(12):100616.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2023. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716.

David Moffat, David Ronan, and Joshua D Reiss. 2015. An evaluation of audio feature extraction toolboxes. *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, pages 277–283.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Shruti Palaskar, Vikas Raunak, and Florian Metze. 2019. Learned in speech recognition: Contextual acoustic word embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6530–6534.

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.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, CONF. IEEE Signal Processing Society.

Archiki Prasad and Preethi Jyothi. 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online. Association for Computational Linguistics.

Jie Pu, Yuguang Yang, Ruirui Li, Oguz Elibol, and Jasha Droppo. 2021. Scaling Effect of Self-Supervised Speech Models. In *Proc. Interspeech 2021*, pages 1084–1088.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 759–766, New York, NY, USA. Association for Computing Machinery.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.

Isabel Roth. 2010. Explore the influence of french on english. *Leading Undergraduate Work in English Studies*, 3:255–261.

M. Serva and F. Petroni. 2008. Indo-european languages tree by levenshtein distance. *Europhysics Letters*, 81(6):68005.

Therese S Shanthi and Chelpa Lingam. 2013. Review of feature extraction techniques in automatic speech recognition. *International Journal of Scientific Engineering and Technology*, 2(6):479–484.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Connie K So and Catherine T Best. 2014. Phonetic influences on english and french listeners' assimilation of mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, 36(2):195–221.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*, pages 2207–2211.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Jing Yang. 2021. Comparison of vots in mandarin–english bilingual children and corresponding monolingual children and adults. *Second Language Research*, 37(1):3–26.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Qingran Zhan, Xiang Xie, Chenguang Hu, and Haobo Cheng. 2021. A self-supervised model for language identification integrating phonological knowledge. *Electronics*, 10(18):2259.

A ASR Model Architecture and Training

The downstream ASR model is composed of a Conformer encoder and a Transformer decoder. The encoder consists of 12 blocks and 4 attention heads with an output size of 256, and the decoder consists of 6 blocks. We use an Adam optimizer with 25000 warmup steps. The model is initialized with Xavier Uniform distribution and trained for 50 epochs with early stopping. We take the average of the best 10 models as the prediction model in the ASR task. To focus on the performance of the SSL feature extractor, we used a simple stacked RNN as the language model during decoding. The RNN language model has 2 layers and each layer has 650 units optimized by the SGD algorithm. We train this language model for 20 epochs and only keep the best one as our language model. During decoding, we use 0.3 as the weight of the language model and decode data with a beam size of 10.

B Implementation and Hardware

We obtain the upstream SSL models and DGCCA model from the S3PRL Speech Toolkit (wen Yang et al., 2021). The ASR training and DGCCA computation were both done on NVIDIA Tesla V100 for all model-language pairs. The average time of each experiment depends on the dataset size but cost about one week to complete on two GPUs for ASR and one day for DGCCA.