A Linguistically Motivated Study to Quantitatively Analyze Self-Supervised Speech Representations

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 explore the cross-lingual adaptability of the English self-supervised learning (SSL) models using Automatic Speech Recognition (ASR) tasks in the target language and propose a new metric to predict the quality of the speech representations. We analyze the effect of size, training objectives, and architecture of the models on their performance as a feature extractor for a set of typologically diverse corpora and study the relationship between the amount of phonetic content in the learned representations and its generalizability across languages. We develop a novel metric, the Phonetic-Syntax Ratio (PSR), to measure the phonetic and synthetic information in the extracted representations on any given out-of-domain/language dataset using deep generalized canonical correlation analysis. Results show that wav2vec2.0 is the most effective cross-lingual feature extractor. Our findings indicate a positive correlation between the PSR score and the ASR performance and in turn, the model's cross-lingual generalizability as a feature extractor. This suggests that phonetic information learned by monolingual SSL models is crucial for downstream tasks in cross-lingual domains. The proposed metric is an effective indicator of downstream task performance and can be useful for model selection.

1 Introduction

Self-Supervised Learning (SSL) has become a paradigm for learning feature representations from unlabeled data (Liu et al., 2021c). 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 (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. Additionally, if we are able to use English monolingual models effectively on multilingual downstream tasks, the high energy cost of training massive multilingual speech models such as XLSR (Babu et al., 2021; Conneau et al., 2020) and mSLAM (Bapna et al., 2022) can be reduced by explicitly incorporating architectural designs that promote 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 Automatic Speech Recognition (ASR) tasks in typologically diverse languages.

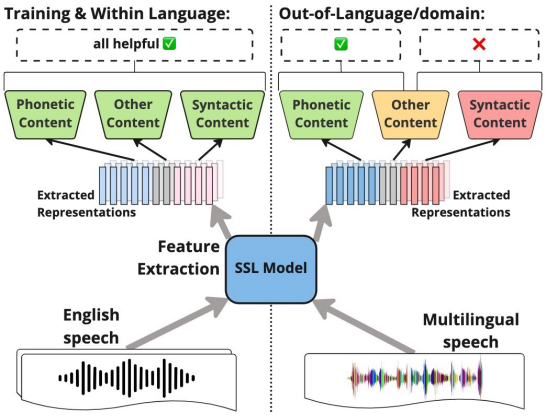

Figure 1: Speech data of English (in-domain) and other languages (out-of-domain) are passed through the SSL models to extract speech representations. All information is expected to aid downstream tasks in English while phonetic content is expected to be useful for out-of-domain downstream tasks; "other"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

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 that 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, while the extracted phonetic and acoustic information in out-of-domain and multilingual situations will aid the downstream performance. Therefore, we propose a novel metric in our study 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 these 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 multilingual generalizability as feature extractors on the ASR task.
- We propose a new metric, Phonology-Syntax Ratio (PSR), to measure 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 the last two layers using the trained layer-wise weights for the feature representation.

2 Related Work

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) transcribes a 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 models 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).

2.2 Self-Supervised Models

Self-Supervised Learning (SSL) (Liu et al., 2021c; 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 performances in natural language processing (Devlin et al., 2018; Peters et al., 2018), computer vision (Chen et al., 2020; Misra and van der Maaten, 2020), speech processing (Chung et al., 2016; Chi et al., 2021), and especially ASR (Baevski and Mohamed, 2020; Ravanelli et al., 2020; Jiang et al., 2021).

2.3 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 underexplored. 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) analyzed the layer-wise acoustic-linguistic content of one pre-trained model by performing layer-independent Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) against the GloVe embedding. However, since the features extracted by deep learning models often have a high dimensionality (Georgiou et al., 2020), it is difficult to express the relationship between two features by a simple linear relationship.

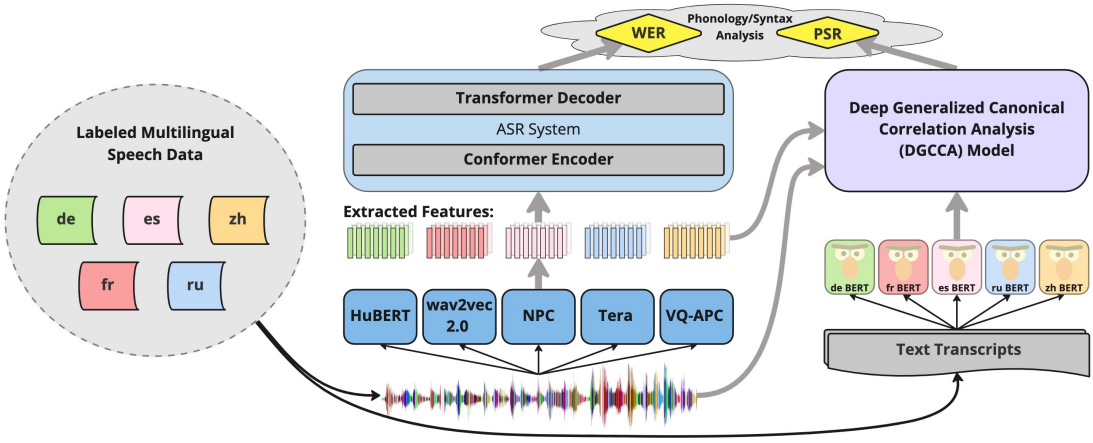

Figure 2: The pipeline to measure the performance of SSL model on different languages. We first use each SSL model as a feature extractor for data in each language and compute a WER score for the ASR task. Then, we calculate the PSR of the representations to analyze the correlation between the ASR performance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

2.4 Linguistic Distance

In linguistics, languages are classified using genetically (or genealogically) or typologically (Gell Mann et al., 2009). The former, which is to group languages into families based on their degree of diachronic relatedness, has classified languages around the world into 424 families. The latter groups languages into types according to their structural characteristics (Hammarström, 2016).

Computationally, when English is used as the intermediate language for measuring similarity, Second Language Acquisition Difficulty with English being either the first or second language can be used to quantitatively calculate the linguistic distance (Chiswick and Miller, 2005; Rabinovich et al., 2018). Bidirectional measures such as Mutual Intelligibility and perplexity remove the English-centric biases but are limited to closely related languages in the same family (Bortoletto et al., 2018; Gooskens, 2007; Gooskens et al., 2018). Another branch of popular algorithms for calculating linguistic distance is based on the Levenshtein distance (Levenshtein et al., 1966), which is defined as the minimum number of edits necessary to transform a word in one language a corresponding word in another language (Bakker et al., 2009). There have been improved metrics such as the Normalized Levenshtein Distance (LDN) (Petroni and Serva, 2010; Yujian and Bo, 2007) and modified normalized Levenshtein Distance (LDND) which a corpus-wide weighted LDN (Holman et al., 2008; Bakker et al., 2009).

3 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 quantitatively analyze the ASR performance against 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 as well as 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)). Following the setup in wen Yang et al. (2021), we take the weighted sum from all layers as the extracted speech representation as follows:

$$F = \sum_{i=0}^{K-1} w_i F_i, \tag{1}$$

where K is the total number of layers, F_i is the representation extracted from the i-th layer, w_i is the weight for the i-th layer. The weight vector $\vec{w}$ is updated during training.

For the downstream model, we use Conformer (Gulati et al., 2020) as the encoder and Transformer (Vaswani et al., 2017) as the decoder. We also use this structure in our baseline model with Mel Spectrum, the structure that has achieved state-of-the-art results in many speech recognition tasks. During data analysis, we isolate the effect of the SSL model as the 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 different data sizes, speech formality levels, and other linguistic differences across languages, allowing for a fair comparison between different SSL models. When decoding, we use a simple RNN as the language model and keep the parameters consistent across all tasks.

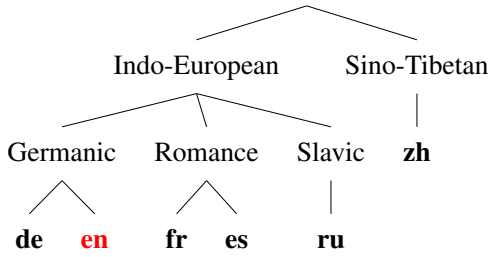

Figure 3: Phylogenetic Tree of Target Langauges

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.

3.3.1 DGCCA

In order to better analyze the phonetic and syntactic content of features extracted from the SSL models,

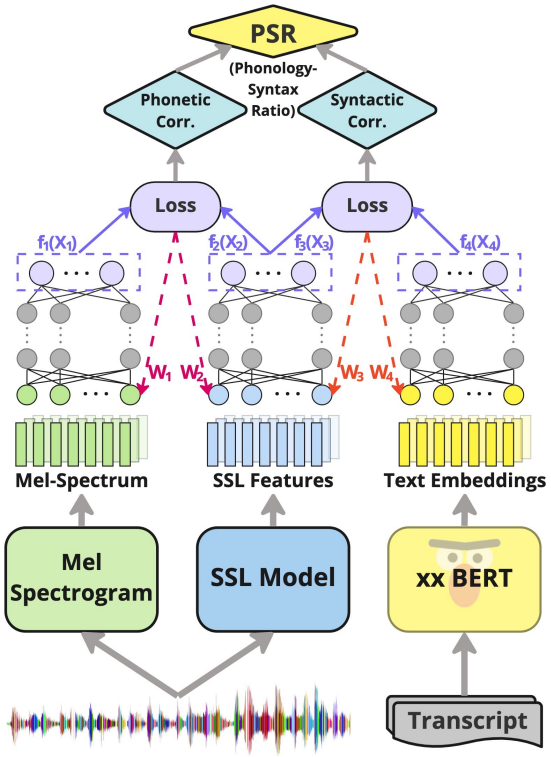

Figure 4: DGCCA pipeline. The model aims to compare the representation extracted by the SSL model to the pure acoustic representation (from Mel Spectrum) and pure syntactic/semantic representation (from BERT).

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 data and learns a view-independent representation (Benton et al., 2017). 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 back-propagation 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:

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

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.

We use DGCCA to learn the correlation between

the SSL representations and the low-level acoustic representations (Mel Spectrum) and between SSL representations and text-based contextualized representations (BERT). Hence, in each computation, we consider $J = 2$ views of the same data with N being the number of utterances in the test set for each language, where each pair of data consists of the representations of the same input utterance (or its transcript). The correlation score of the two views is the converged DGCCA network loss after training.

3.3.2 Phonetic Similarity

As shown in Figure 4, we use DGCCA to measure the Phonetic Similarity by calculating the correlation score between the Mel Spectrum features and the extracted SSL features according to Equation 1 on speech data from different languages. Each utterance has a Mel Spectrum feature vector and its corresponding SSL representation vector, and DGCCA is used to calculate the correlation.

3.3.3 Syntactic Similarity

To calculate the Syntactic Similarity, we use the monolingual BERT model in the target language (xx-BERT) to extract syntactic representations on the transcripts. Analogous to the previous setup, each utterance file corresponds to one syntactic representation from xx-BERT and extracted SSL features. We use DGCCA to calculate the similarity score between these vector pairs..

3.3.4 Phonetic-Syntax Ratio (PSR)

In order to quantitatively investigate the phonetic and syntactic content on SSL representation, we introduce a new metric: the Phonetic-Syntax Ratio (PSR). It is the ratio between the Phonetic Similarity score and Syntactic Similarity score weighted equally among all data points:

$$PSR = (\frac{1}{n}\sum_{i=1}^{n}\frac{phonetic\ score_i}{syntax\ score_i} - 1)\cdot 100\%, \quad (3)$$

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.

4 Experimental Setup

4.1 Datasets

We investigate the cross-lingual adaptation capability of English SSL models in five languages. We train ASR models with Mozilla Common Voice 5.1 dataset (Ardila et al., 2019) for German, French, Spanish, and Russian. We use the OpenSLR ST-CMDS-20170001_1 Free ST Chinese Mandarin Corpus[1] for Chinese ASR training. We use the Common Voice English test set for DGCCA analysis. Train, dev, test splits, and dataset sizes are listed 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 focus on the within-language difference between the performance of the SSL model and the baseline model.

The Baseline 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.

Self-Supervised Models that we examined include include 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.

[1] 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.

The Downstream Model architecture described above is used with both features extracted using the SSL models and the baseline Mel Spectrum features in order to make a fair comparison with the baseline. In addition, we use the same language model in the baseline model during decoding. 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. Also consistent with the baseline model, we used the Adam optimizer.

4.3 DGCCA

When calculating the DGCCA scores, 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. We extract features with a batch size of 32 and pass them into the DGCCA model. We use the respective test sets as input to the DGCCA model. For the DGCCA network, we used the MLP network which consists of 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.

4.4 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.

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, 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 might be projected onto the new language (Georgiou et al., 2020). Therefore, the purpose of this experiment is not to improve state-of-the-art results, but rather to probe the SSL models for further phonetic-syntactic analysis. There are five SSL models being evaluated in this experiment across five languages. The Avg column on the right marginal of Table 3 shows the overall performance of each SSL model in all languages. Overall, wav2vec2.0-LARGE significantly outperforms other feature extractors and has a consistent result across languages. There are two instances where wav2vec2.0-LARGE outperforms the pure acoustic Mel spectrum baseline. This attributes to the cross-lingual phonetic information transfer that the model learned from English pre-training.

5.1.1 Training Objectives

The four HuBERT and wav2vec2.0 models consistently perform better compared to 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. In our particular multilingual ASR task, however, HuBERT does not

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.

perform as well because of its predictive loss compared to the contrastive loss of wav2vec2.0. The masked prediction task during the pre-training of HuBERT forces the model to learn the language model from continuous speech inputs as well as the acoustic model (Hsu et al., 2021).

Now we discuss in detail the performance of NPC, TERA, and VQ-APC, which are significantly smaller than wav2vec2.0 and HuBERT and pre-trained on a smaller dataset. With comparable model sizes, training objectives, input format, and stride during pre-training, TERA outperforms NPC with less than one-third of the training data. This is due to the alterations on 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 compared with 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 is more effective at capturing speech representations than the convolutional blocks in NPC, implying that speech is a sequential data structure.

5.1.2 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 done by Pu et al. (2021), in which they empirically showed that scaling SSL models results in improvements in both the L1 loss and accuracy on downstream tasks consistent with the power law. Large models can also be more data efficient when labeled data is scarce. Moreover, the diversity of pre-training data improves the robustness of the model (Chen et al., 2022). 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 might have caused it to learn even more syntactic and semantic representation which does not benefit cross-lingual tasks.

5.2 Linguistic Analysis

We now discuss the performances of all five languages based on their average score, and 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 have a further distance from 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 the pitch contours (So and Best, 2014). For German, although it seems to have a poor SSL performance with high Δ values, the absolute WER is the lowest among the German, French, and Spanish which have similar training sizes. From this, it can be observed that SSL representations give diminishing returns in high-resource situations, motivating the adaptation of SSL methods to low-resource settings.

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 shows the robustness of SSL models in low-resource settings and establishes

promising directions in the generalization 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 to avoid additional noise introduced by a Chinese word segmentation model, so the Chinese results should only be analyzed 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 investigating linguistic information in the extracted representations.

5.3 PSR Analysis

We now analyze the PSR scores of HuBERT-BASE on each of the target languages and English as shown in Table 4. As described in Equation 3, the larger the PSR value, the more acoustic and phonetic content there is in the feature set. First of all, to validate the PSR scale, we test the SSL features extracted from the English corpus by the HuBERT 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

It is worth noting that 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 partially attributes to the fact that the weighted sum of layers is used as the input features to the ASR model as in Equation 1 and that the weights are optimized during training. The optimized weights gravitate toward layers with more phonetic information after ASR training since higher PSR (stronger phonetic content) is more helpful for the task as discussed in Section 5.3. Therefore, visualizing the weights gives insights into which layers contribute more to the phonetic information. Figure 5 shows the magnitude of the weights across all layers of HuBERT-BASE.

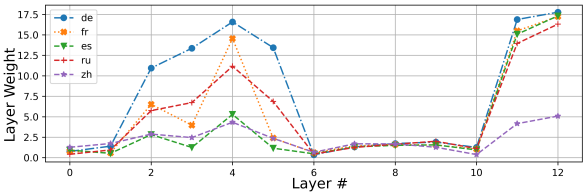

Figure 5: Layer-wise Weight Analysis.

First, the sign of each layer is consistent across all languages, suggesting that each layer contains similar information even when trained on different datasets, i.e., the weights get updated similarly given the same task. As shown in Figure 5, the last two layers are the most salient. This means that the last two layers provide richer phonetic information in the HuBERT model. Layer 4 also shows strong weights for some languages, especially German and French which are closer to English. Our work to localize the phonetic content encoded in specific layers of HuBERT draws similar conclusions with Pasad et al. (2021), which localized various acoustic and linguistic properties in wave2vec2.0.

6 Conclusion

In this work, we examined self-supervised speech models trained on monolingual English and probed for the phonetic and syntactic content in the speech representations. We accomplished this using the SSL models as a feature extractor for downstream ASR tasks in multiple languages. Higher multilingual adaptability of a model is found to be posi-

tively 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. This metric can be easily adapted to any out-of-domain situations as a preliminary evaluation during model selection. We were also able to localize the phonetic information to the last two layers of HuBERT. Our study is the first quantitative review of the existing SSL models and provides a model/data-agnostic evaluation of the domain adaptability without the need for compute-intensive training.

Limitations

There are a number of 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.

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 Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

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. IEEE.

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

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon 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. IEEE.

Dik Bakker, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*.

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 transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2017. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*.

Giovanni Bortoletto, Leena Manninen, Emma McKenzie, and Oona Raatikainen. 2018. Measuring language distance using perplexity. *Natural Language Engineering*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 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 *International conference on machine learning*, pages 1597–1607. PMLR.

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. IEEE.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, 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. IEEE.

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. IEEE.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-quantized autoregressive predictive coding. In *Proc. Interspeech 2020*.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *arXiv preprint arXiv:1603.00982*.

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. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Murray Gell Mann, Ilia Peiros, and George Starostin. 2009. Distant language relationship: The current perspective. *J. Lang. Relatsh.*, 1:1–41.

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.

Charlotte Gooskens. 2007. The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and multicultural development*, 28(6):445–467.

Charlotte Gooskens, Vincent J van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Harald Hammarström. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1):19–29.

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, et al. 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.

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica, v.42, 331-354 (2008)*.

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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 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 pretraining for transformer based speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6538–6542. IEEE.

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.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

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 *Interspeech*, 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. IEEE.

Hexin Liu, Leibny Paola Garcia Perera, Andy WH 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 Bjoern 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. 2021c. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

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. *computer vision and pattern recognition*.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. 2022. Self-supervised speech representation learning: A review. *arXiv preprint arXiv:2205.10643*.

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. IEEE.

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. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *north american chapter of the association for computational linguistics*.

Filippo Petroni and Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, 389(11):2280–2283.

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 *IEEE 2011 workshop on automatic speech recognition and understanding*, 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.

Jie Pu, Yuguang Yang, Ruirui Li, Oguz Elibol, and Jasha Droppo. 2021. Scaling effect of self-supervised speech models. In *Interspeech*, pages 1084–1088.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

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*, pages 759–766.

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. IEEE.

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

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

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

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.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

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