Towards Building a Language-Independent Speech Scoring Assessment

Shreyansh Gupta*, Abhishek Unnam*, Kuldeep Yadav, Varun Aggarwal

SHL Labs, Gurugram, India

Shreyansh.Gupta@shl.com, Abhishek.Unnam@shl.com, kdyadav.ce@gmail.com, Varun.Aggarwal@gmail.com

Abstract

Automatic speech scoring is crucial in language learning, providing targeted feedback to language learners by assessing pronunciation, fluency, and other speech qualities. However, the scarcity of human-labeled data for languages beyond English poses a significant challenge in developing such systems. In this work, we propose a $Language\text{-}Independent$ scoring approach to evaluate speech without relying on labeled data in the target language. We introduce a multilingual speech scoring system that leverages representations from the wav2vec 2.0 XLSR model and a force-alignment technique based on CTC-Segmentation to construct speech features. These features are used to train a machine learning model to predict pronunciation and fluency scores. We demonstrate the potential of our method by predicting expert ratings on a speech dataset spanning five languages - English, French, Spanish, German and Portuguese, and comparing its performance against Language-Specific models trained individually on each language, as well as a jointly-trained model on all languages. Results indicate that our approach shows promise as an initial step towards a universal language independent speech scoring.

Introduction

Effective communication and spoken language skills are important in multiple business functions. Previous research (Singh and Harun 2020; Alyammahi et al. 2021) has shown that clear and concise communication can lead to an increase in productivity, efficiency, and understanding in the workplace. Increasingly, organizations are using automated speech scoring techniques to screen job candidates as well as to provide computer-assisted language learning (CALL) (Ai 2015). Typically, speech scoring systems comprise measurements of different competencies including Pronunciation, Fluency, Automatic Listening, etc.

Previous work in automated speech scoring (Witt and Young 1997; Ai 2015; Evanini and Wang 2013) have examined phone level scores derived from GMM-HMM (Gaussian Mixture Model − Hidden Markov Model) based speech recognizer outputs. With the proliferation of deep learning techniques, more recent studies (Ying 2019; Hu et al. 2015; Sudhakara et al. 2019) have used acoustic models trained using a Deep Neural Network to improve mispronunciation detection & diagnosis (MDD). Further, deep models evolved into audio transfer-based architecture to develop an end-to-end MDD system (Peng et al. 2021; Li and Lai 2022). One of the major limitations of these existing approaches is that they require training separate Language-Specific acoustic models (and labeled datasets) for each language and do not leverage the existing labeled data from adjacent languages. It has become a major barrier to making speech scoring systems more accessible, ubiquitous, and scalable to low-resource languages.

Recent work in NLP and speech recognition (Gu et al. 2018; Xu, Baevski, and Auli 2022) used transfer-learning strategies to transfer lexical, sentence-level representations from numerous source languages to a single target language and learn cross-lingual audio representations for transcription of unknown languages respectively. These models provide a viable route towards more pervasive speech processing technologies by enhancing the task performance of low-resource languages by utilizing data from high-resource languages. In addition, they demonstrate that it is sufficient to maintain a single multilingual model as opposed to a multitude of monolingual models.

Building upon the existing work, we present a multilingual speech scoring system that generates pronunciation and fluency scores. We propose the first framework of its kind for $Language\text{-}Independent$ scoring that can exhibit cross-lingual generalization without requiring labeled data in the target language. Our method employs self-supervised representations that are pre-trained on diverse multilingual speech data and fine-tuned for phoneme recognition across languages. Next, a Connectionist Temporal Classification (CTC) based segmentation technique is used to force-align the text to be spoken with speech and obtain the various word and phoneme boundaries along with their respective probability scores. We introduce a feature set based on force alignment and automatic speech recognition to predict pronunciation and fluency scores. The entire framework is evaluated on unseen target languages. While not achieving full language independence yet, this work represents an important first step which allows instantly bootstrapping scoring capabilities, which can be iteratively improved as additional data gets collected in the target language. In particular, the

*These authors contributed equally.

paper makes the following contributions.

- We utilize the self-supervised representations from XLSR and a CTC-Segmentation algorithm to construct a feature set that predicts pronunciation and fluency across different languages.
- The performance evaluation on the dataset of five different languages demonstrates that it is possible to build a scalable $Language\text{-}Independent$ speech scoring system for comparatively low-resource languages.
- The inclusion of typographically similar languages during training notably enhances the effectiveness of the language-independent speech scoring system, highlighting the importance of considering linguistic similarities when building robust and scalable language-independent speech scoring systems.

Proposed Approach

In this section, we describe the XLSR model and CTC-Segmentation based force-alignment technique to extract the feature set. A detailed workflow diagram of the proposed approach is in Figure 1.

XLSR

We made use of XLSR (Conneau et al. 2021), a multilingual wav2vec 2.0 model comprising of 3 components - a CNN-based encoder network, transformer-based context network, and vector quantization module. The encoder network converts the raw audio signal X into latent speech representation $Z_1, Z_2, ...Z_t$. These representations are then fed into a transformer model G to build context representations $C_1, C_2, ...C_n$. During training, a certain proportion of time steps are masked in the latent representation Z and a contrastive task is designed to distinguish the correct representations from a set of distractors sampled from other masked time-steps. Before the training step, the quantized representations $q_1, \dots q_T$ are built using the quantization module $g : Z \mapsto Q$ which represents the targets used in the contrastive task.

The model was pre-trained using publicly available data from 53 distinct languages. One of the XLSR models was fine-tuned to recognize phonetic labels, while the other was fine-tuned for Automatic Speech Recognition (ASR) using the Common Voice dataset (Ardila et al. 2020), which contains speech samples from 56 languages. The model fine-tuned on phoneme recognition task was used for implementing the force-alignment algorithm and the ASR model was used in ASR-based feature set (described in the section - Feature Set).

CTC-Segmentation Based Force-Alignment

We used the CTC-Segmentation algorithm, as proposed in (Kürzinger et al. 2020), which provides an effective method for precisely aligning audio with text transcripts. Given an audio's word transcript $W = \{w_1, w_2, w_3...w_n\}$ and its corresponding phoneme sequence $\{ph_1, ph_2, ph_3...ph_M\}$, we aim to align each phoneme ph_j to it's timestamp in the audio frames $t \in \{1, 2, ...T\}$. We leverage a phonemizer (Bernard and Titeux 2021) to automatically generate phonetic transcriptions of the text into International Phonetic Alphabet (IPA) phonemes without needing Language-Specific dictionaries. Next, a trellis diagram is constructed where the x-axis corresponds to audio frames t, the y-axis indexes phonemes j, and each point (t, j) represents a possible alignment between frame t and phoneme ph_j. The joint probabilities k(t, j) stored at each point is computed recursively as in equation 1.

$$k_{t,j} = \begin{cases} max(k_{t-1} \cdot P(bl|t), k_{t-1,j-1} \cdot P(ph_j|t)) & \text{if } t > 0 \wedge j > 0 \\ 0 & \text{if } t = 0 \wedge j > 0 \\ 1 & \text{if } j = 0 \end{cases} \quad (1)$$

where bl represents the blank token from CTC formulation and $P(ph_j|t)$ is the posterior probability of phoneme ph_j at frame t extracted from the XLSR model (trained on phoneme detection, explained in the section - XLSR). The maximum probability path through this trellis, found efficiently via backtracking (as in equation 2), provides the optimal forced alignment $A = \{a_1, a_2, a_3...a_T\}$ between the phonemes and audio frames.

$$a_t = \begin{cases} M-1 & \text{if } t > argmax_{t'}(k_{t',M-1}) \\ a_{t+1} & \text{if } k_{t,a_{t+1}} \cdot P(bl|t+1) > k_{t,a_{t+1}-1} \cdot P(ph_j|t+1) \\ a_{t+1}-1 & else \end{cases} \quad (2)$$

The alignment a_t of the phoneme from the transcript for the audio frame t is determined using the transitions with the highest probability. The forced alignment provides start and end timestamps for each aligned phoneme segment s_{seg}. The posterior probability for segment s_{seg} is computed as in equation 3.

$$P(s_{seg}) = \frac{1}{L_{seg}} \sum_{t \in s_{seg}} P(ph_{a_t}|t) \quad (3)$$

where L_{seg} is the number of frames in the segment s_{seg} and $P(ph_{a_t}|t)$ is the posterior probability of the phoneme aligned to the frame t.

Feature Set

We used two categories of features described below. All of these features were extracted from the speech data collected (described in section - Dataset and Ratings). These features were subsequently used in training regression models to predict pronunciation and fluency scores in a supervised learning setting.

Force-Alignment Based Using the method described in the section - CTC-Segmentation Based Force-Alignment, the speech sample is force-aligned against the sequence of phonemes (retrieved from the sentence to be uttered). For each (audio, sentence) pair, an alignment file containing information on the position, duration, and probability of each phoneme was generated. These alignment files were used to derive several speech quality features predictive of pronunciation and fluency, such as rate of speech, position, and

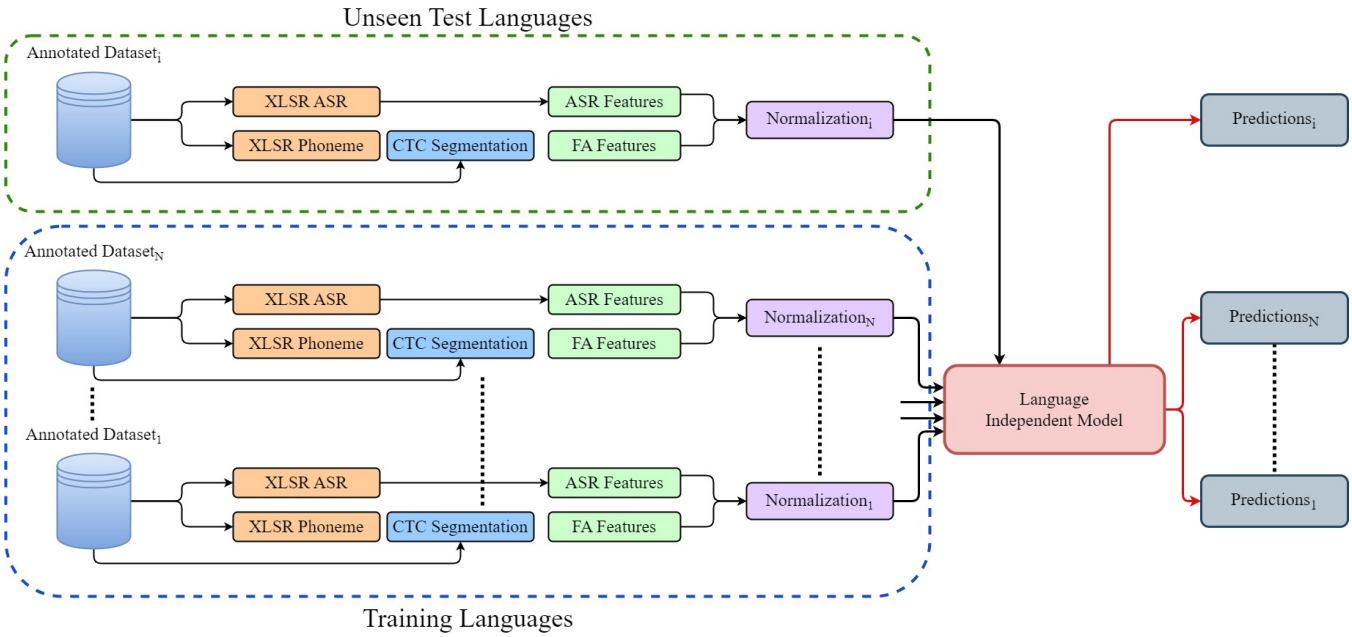

Figure 1: Workflow diagram of the language-independent scoring approach. The languages below the dotted line represent those used for training the machine learning models. The unseen target languages above the dotted line are held out for model evaluation. "FA Features" refers to the Force-Alignment Based features, while "ASR Features" refers to the ASR-Based features.

length of pauses, likelihood of recognition, posterior probabilities, phonation to time ratio, Log-Likelihood Ratio etc. (Franco et al. 1997, 1999; Zechner et al. 2009; Neumeyer et al. 1996; Cucchiarini, Strik, and Boves 2000). Several additional distance-based features were extracted by comparing the probability of phonemes in the real sequence to the probability of phonemes in a random sequence of the same length.

ASR-Based Word-error (WER), character-error (CER), and phoneme-error rates (PER) were extracted for the speech sample using the XLSR model fine-tuned on automatic speech recognition (ASR) and phoneme recognition tasks (explained in section - XLSR) respectively. These error rates were calculated with respect to the actual sentence to be spoken.

Normalization Although the stated features (a few examples have been described in Table 1) ensure that the broad association with the pronunciation and fluency ratings across languages is preserved, the magnitude (or contribution) of the features may differ across languages and can have an impact on pronunciation or fluency scoring. For example, the duration of pauses in English varies from those in Spanish, as well as the accuracy of XLSR models for speech transcription. This can cause an issue specifically when we enhance these features with different scales for a joint learning task. To mitigate this, we use a simple yet effective strategy of Language-Specific normalization (also described in Figure 1) of each feature value where the unlabeled dataset of that language is used to learn distributional properties.

Feature Tagging We observed that the expert ratings for pronunciation and fluency of a candidate are highly correlated; It is imperative that we take the necessary steps to ensure that the trained model specifically predicts the competency it is intended to and avoids the use of spurious non-causal features which might be predictive of the related competency. We followed a Hybrid method of feature selection (Chen et al. 2018) where language training experts, using information about feature correlation with expert ratings, feature collinearity, and feature normality identified each feature as pronunciation or fluency-based. All the selected features for e.g., rate-of-speech and duration of pauses were classified as fluency-based, while PER, log-likelihoods, likelihood-ratio and posterior probabilities etc. were used to train the machine learning model for the respective competency.

Experiments

We conducted detailed experiments to evaluate the efficacy of a Language-Independent approach for speech scoring (i.e. Pronunciation, Fluency) using real-world datasets.

Dataset and Ratings

In our dataset, there were 5 distinct languages: English, French, Spanish, German and Portuguese. We collected data from 2,500 job-seekers from multiple countries including the US, UK, India, and several other European countries. Each candidate had to complete 2 distinct sections.

- **Read & Speak**: They were required to read the sentences on the screen

Name	Category	Description
PER	Pronunciation	Edit distance between recognized and reference phoneme sequences.
probspeakCorrectNosWords	Pronunciation	Total likelihood of the correct phoneme sequences, normalized by the number of words in the utterance.
diffNumPh	Pronunciation	Difference between the likelihood of the correct phoneme sequence and an incorrect sequence, normalized by the total number of phonemes in the utterance.
rateOfSpeech	Fluency	Rate at which words are spoken, dividing the number of words by the duration of speech.
meanLongPause	Fluency	Average duration of long pauses in the speech sample by taking pauses above a certain threshold (250-500ms) and averaging their lengths.
llspNumph	Fluency	Calculates the total log-likelihood of pauses or silences between words, normalized by the total number of phonemes in the utterance.

Table 1: This table highlights and describes some example features derived from the force-aligned and automatic speech recognition model outputs.

- **Listen & Repeat**: They were required to listen to a sentence and repeat it.

Score	Description
5	Completely fluent speech at normal speed. Any hesitation is appropriate and there is no sign of searching for words or structures.
4	Fluent speech at normal speed, with only occasional repetition or self-correction. Hesitation may occasionally indicate searching for words or structures but is generally appropriate.
3	Uneven flow, with some repetition, especially in longer utterances. Some evidence of searching for words, which does not cause serious strain.
2	Very uneven. Frequent pauses and repetitions indicate searching for words or structures. Excessive use of fillers and difficulty sustaining longer utterances cause serious strain on the listener.
1	Extremely uneven. Long pauses, numerous repetition and self-corrections make speech difficult to follow.

Table 2: Rubric used by expert raters for rating fluency on a scale of 1 to 5.

Each candidate recorded a total of 21 audio responses over the two sections with each response having a time limit of 10 seconds. The sample was created to include candidates of varying ages, gender, race, and educational background. We collected nearly 50,000 audio files from the 2,500 job-seekers. The responses were collected using our proprietary spoken language assessment platform i.e. SVAR (SHL 2012).

Our expert raters were recruited online. All raters have prior experience as soft skills or language trainers. We selected the sample of raters to be representative of various nationalities and genders. Each audio response is rated on $Pronunciation$ and $Fluency$ by 3 raters. Both Pronunciation and Fluency were rated based on a 5-point rubric, an example of the rubric used has been presented in Table 2. The inter-rater agreement was measured using the Pearson coefficient of correlation (r). Our inter-rater agreement was high with r more than 0.8 for both pronunciation and fluency. We used the mean ratings of raters, as the final rating per audio for each of the competencies. Finally, we derive a candidate-level rating by averaging the audio-wise rating from the 21 responses.

Modeling

We trained an audio-specific machine-learning model with the help of our feature extraction pipeline described in the section - Feature Set. Candidate-level scores were derived by averaging the predicted audio-level scores for each candidate. For each model, the corresponding dataset was divided into train and test sets. We used a stratified 70-30 split for train-test sets. We used linear regression, linear regression (Linear) with L1 regularization (Lasso), linear regression with L2 regularization (Ridge regression), decision trees, random forests (R.F), and support vector machines (SVM). The models with the best cross-validation (4-fold) correlation were selected (see Table 3 for exact details). We trained the following models to perform a comprehensive evaluation across different approaches:

- $Language\text{-}Specific$: Separate models are trained for all five languages. The set of features[1] discussed in the section - Feature Set was used as inputs for the corresponding pronunciation and fluency models. This is the best-case scenario where the training data need to exist for the language in question.
- $Language\text{-}Combined$: We jointly trained a combined model for all five languages (each of the languages was present in the train and test set) after normalizing the features using an unlabeled distribution set.
- $Language\text{-}Independent$: The model was trained using normalized features (refer to section - Normalization) from a subset of languages explained in Figure 1), while holding out certain unseen target languages in the validation and test set. We reported results for the scenario of training on all languages except one held-out unseen target language. Additionally, we trained variants using different combinations of languages in training to ana-

[1]No feature normalization was performed as we were training the models separately for each language

Language	Language-Specific						Language-Combined					
	Pronunciation			Fluency			Pronunciation			Fluency		
	Model	r	MAE	Model	r	MAE	Model	r	MAE	Model	r	MAE
English	R.F	0.70	0.36	R.F	0.75	0.66	SVM	0.69	0.35	R.F	0.72	0.29
French	R.F	0.74	0.44	Linear	0.66	0.38	SVM	0.70	0.61	R.F	0.66	0.61
Spanish	R.F	0.79	0.57	R.F	0.74	0.58	SVM	0.77	0.60	R.F	0.69	0.63
Portuguese	Lasso	0.84	0.10	R.F	0.86	0.19	SVM	0.84	0.20	R.F	0.66	0.30
German	Ridge	0.91	0.12	SVM	0.88	0.31	SVM	0.93	0.31	R.F	0.91	0.20

Language	Language-Independent					
	Pronunciation			Fluency		
	Model	r	MAE	Model	r	MAE
English	SVM	0.65	0.58	R.F	0.66	0.88
French	SVM	0.72	0.81	R.F	0.63	0.61
Spanish	SVM	0.77	0.66	R.F	0.64	0.74
Portuguese	Lasso	0.77	0.23	R.F	0.65	0.26
German	Ridge	0.92	0.33	Lasso	0.83	0.51

Table 3: Candidate-level evaluation metrics on the test set for different modeling strategies. This table compares the performance of Language-Specific, Language-Combined, and Language-Independent models on predicting pronunciation and fluency scores at the individual candidate level.

Language		Language-Specific		Language-Combined		Language-Independent	
		P-Actual	F-Actual	P-Actual	F-Actual	P-Actual	F-Actual
English	P-Pred	0.70	0.68	0.69	0.69	0.65	0.64
	F-Pred	0.57	0.75	0.55	0.72	0.52	0.66
French	P-Pred	0.74	0.73	0.70	0.68	0.72	0.71
	F-Pred	0.55	0.66	0.51	0.62	0.52	0.63
Spanish	P-Pred	0.79	0.79	0.77	0.75	0.77	0.75
	F-Pred	0.70	0.74	0.66	0.69	0.40	0.64
Portuguese	P-Pred	0.84	0.77	0.84	0.76	0.77	0.71
	F-Pred	0.75	0.86	0.57	0.66	0.55	0.65
German	P-Pred	0.91	0.90	0.93	0.91	0.92	0.90
	F-Pred	0.85	0.88	0.89	0.91	0.81	0.83

Table 4: Candidate-level cross-correlation matrices on the test set for different modeling strategies. The diagonal values show the correlation between actual and predicted scores for the same competency, e.g. actual (P-Actual) and predicted (P-Pred) pronunciation. The off-diagonal values show cross-correlation between actual scores for one competency and predicted scores for the other, e.g. actual pronunciation scores and predicted fluency scores.

lyze the impact on generalizability, some of the results are explained in the section - Results.

The Pearson correlation coefficient (r) is one of the widely used metrics for speech scoring systems and therefore, we use it to benchmark accuracy between predicted scores and the experts' ratings. Different scales of feature values across languages may lead to distribution shift as described in the section - Normalization, therefore we consider mean absolute error (MAE) $\sum \frac{|y_{pred}-y|}{n}$) as the other evaluation metric. We also evaluated the cross-correlation between the predicted pronunciation model scores and the fluency ratings to ensure that the trained model captures the unique proposition for each competency and has minimal overlap.

Results

Table 3 presents the performance comparison of the Language-Specific, Language-Combined, and Language-Independent models on candidate-level scores. We find that the r for each of the languages was close to 0.70 for pronunciation, while it was more than 0.66 for fluency when trained in a Language-Specific and Language-Combined approach. Meanwhile, the results for the Language-Independent model showed that the mean r was more than 0.65 and 0.63 for pronunciation and fluency, respectively, in each of the languages. There is a performance drop of 8.33% in the r of pronunciation scores for Portuguese (highest) when compared to Spanish, French, English (all fell by almost 2%) when we switch from Language-Specific modeling

Language	Training Set	Pronunciation			Fluency		
		Model	r	MAE	Model	r	MAE
Portuguese	English, French, Spanish	Lasso	0.77	0.24	SVM	0.64	0.27
	English, French	Lasso	0.78	0.25	Linear	0.49	0.35
German	English, French, Spanish	Ridge	0.92	0.53	Lasso	0.83	0.61
	English, French	Linear	0.93	0.59	R.F	0.79	1.40

Table 5: Candidate-level evaluation metrics on the test set when modifying training language set. This table provides pronunciation and fluency prediction results for the target languages - Portuguese and German under different compositions of training languages.

to Language-Independent, while for German the same increased by 1%. When we look at the results for fluency, there is a drop of 24% in r for Portuguese while for English, Spanish, French and German it was 12%, 13%, 4% and 5% respectively. Expectantly, Language-Specific and Language-Combined outperformed the Language-Independent modeling approaches in terms of r. We also observed a higher MAE when using Language-Independent modeling in certain cases, particularly fluency. This demonstrates that scaling challenges do arise when designing Language-Independent models, and there is room for improvement. In Table 4, we looked at the cross correlations - how the trained model for pronunciation correlated with the ratings for fluency, and vice-versa. We see that the predicted scores correlate with the actual ratings more than that of the related competency. This justifies the manual tagging of features with respect to each competency. In practice, we found that r of above 0.65 for pronunciation and 0.60 for fluency works fine for several use cases i.e. job candidate screening where the bottom 30-40% candidates need to be automatically screened using an assessment. The proposed Language-Independent approach can be an effective tool to build such speech-scoring applications.

Training Language Selection Analysis

In order to explore the effect of training languages on unseen target languages, we analyzed model performance when systematically modifying the composition of languages used during training. We report one such analysis in Table 5, where the selection of training languages had a significant impact on the model performance for fluency prediction in Portuguese as the target language. Specifically, when we included the closely related language Spanish (Posner and Sala 2023) in the training data, there was an increase of 30% in the fluency score correlation compared to the setting without Spanish. In contrast, for the more distant language - German, modifying the training languages had a smaller impact, with only a 5% drop in fluency prediction. Nonetheless, these results indicate the importance of incorporating typologically similar languages to the target language in the training data for fluency prediction. Interestingly, a similar impact of related languages was not observed for pronunciation score prediction, likely attributable to the fact that pronunciation features were relying on phoneme likelihoods and distances rather than speech patterns. For Portuguese, the presence of closely related Spanish in the training data was clearly beneficial for the model to learn shared speech patterns that generalized to improved fluency prediction. In summary, carefully selecting appropriate training languages, especially linguistically proximal neighbors can maximize the generalization capabilities of language-agnostic speech scoring models for fluency.

Conclusion

Traditional development of speech scoring systems requires building customized models for each target language. This process involves extensive data collection and annotation, model training and optimization per language, often taking many months to deploy a new language. We present a multilingual speech scoring system that uses self-supervised representations from XLSR and a force-alignment technique based on CTC-Segmentation to construct speech features. We proposed a first of it's kind $Language\text{-}Independent$ speech evaluation system, that can evaluate speech without any labeled dataset. We evaluated the performance of the Language-Independent model along with other approaches, and results showed that Language-Independent model achieves strong performance, though not on par with Language-Specific models yet. However, they provide a solid starting point for low-resource languages without any labeled datasets. Additionally, empirical results demonstrate that including typologically related languages within the training set improves performance on unseen target languages for predicting fluency, we achieved close to 30% increase in correlation for Portuguese when adding Spanish to the training data. However, no such boost was observed for pronunciation scoring. The proposed methodology enables expeditious bootstrapping of fundamental scoring capabilities within weeks, rather than months required for conventional Language-Specific systems. This accelerated time-to-production can catalyze proliferation of speech assessments for applications in computer-assisted language learning, and screening contact center workers. In the future, we plan to scale this to other Asian, and tonal languages and investigate further how cross-language knowledge and dependencies can further enhance the accuracy of speech scoring systems and also, study the bias implications across different demographics.

References

Ai, R. 2015. Automatic Pronunciation Error Detection and Feedback Generation for CALL Applications. In Zaphiris, P.; and Ioannou, A., eds., *Learning and Collaboration Technologies*, 175–186. Cham: Springer International Publishing. ISBN 978-3-319-20609-7.

Alyammahi, A.; Alshurideh, M.; Kurdi, B. A.; and Salloum, S. A. 2021. The Impacts of Communication Ethics on Workplace Decision Making and Productivity. In Hassanien, A. E.; Slowik, A.; Snášel, V.; El-Deeb, H.; and Tolba, F. M., eds., *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020*, 488–500. Cham: Springer International Publishing. ISBN 978-3-030-58669-0.

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Bernard, M.; and Titeux, H. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68): 3958.

Chen, L.; Zechner, K.; Yoon, S.-Y.; Evanini, K.; Wang, X.; Loukina, A.; Tao, J.; Davis, L.; Lee, C. M.; Ma, M.; Mundkowsky, R.; Lu, C.; Leong, C. W.; and Gyawali, B. 2018. Automated Scoring of Nonnative Speech Using the SpeechRaterSM v. 5.0 Engine. *ETS Research Report Series*, 2018(1): 1–31.

Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; and Auli, M. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, 2426–2430.

Cucchiarini, C.; Strik, H.; and Boves, L. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107 2: 989–99.

Evanini, K.; and Wang, X. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *INTERSPEECH*, 2435–2439.

Franco, H.; Neumeyer, L.; Kim, Y.; and Ronen, O. 1997. Automatic pronunciation scoring for language instruction. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 1471–1474 vol.2.

Franco, H.; Neumeyer, L.; Ramos, M.; and Bratt, H. 1999. Automatic detection of phone-level mispronunciation for language learning. In *Sixth European Conference on Speech Communication and Technology*.

Gu, J.; Hassan, H.; Devlin, J.; and Li, V. O. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 344–354. New Orleans, Louisiana: Association for Computational Linguistics.

Hu, W.; Qian, Y.; Soong, F. K.; and Wang, Y. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67: 154–166.

Kürzinger, L.; Winkelbauer, D.; Li, L.; Watzel, T.; and Rigoll, G. 2020. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. In Karpov, A.; and Potapova, R., eds., *Speech and Computer*, 267–278. Cham: Springer International Publishing. ISBN 978-3-030-60276-5.

Li, R.; and Lai, X. 2022. W2V-ATT: research on text-dependent MDD method based on wav2vec2.0. In Tiwari, R., ed., *International Conference on Neural Networks, Information, and Communication Engineering (NNICE 2022)*, volume 12258, 1225804. International Society for Optics and Photonics, SPIE.

Neumeyer, L.; Franco, H.; Weintraub, M.; and Price, P. J. 1996. Automatic text-independent pronunciation scoring of foreign language student speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 3: 1457–1460 vol.3.

Peng, L.; Fu, K.; Lin, B.; Ke, D.; and Zhang, J. 2021. A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Interspeech*.

Posner, R.; and Sala, M. 2023. *Romance languages*.

SHL. 2012. Language proficiency test: Language assessment.

Singh, A. K. J.; and Harun, R. N. S. R. 2020. Industrial trainees learning experiences of English related tasks at the workplace. *Studies in English Language and Education*, 7(1): 22–42.

Sudhakara, S.; Ramanathi, M. K.; Yarra, C.; and Ghosh, P. K. 2019. An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities. In *Proc. Interspeech 2019*, 954–958.

Witt, S.; and Young, S. 1997. Computer-assisted pronunciation teaching based on automatic speech recognition. *Language Teaching and Language Technology Groningen, The Netherlands*.

Xu, Q.; Baevski, A.; and Auli, M. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Proc. Interspeech 2022*, 2113–2117.

Ying, W. 2019. English Pronunciation Recognition and Detection Based on HMM-DNN. In *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 648–652.

Zechner, K.; Higgins, D.; Xi, X.; and Williamson, D. M. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Commun.*, 51: 883–895.