Recognition of Different Accents and Speech Sentiments Using Zero-Shot Learning

Fahmida Ahmed
Dept. of Computer Science
BRAC University
Dhaka, Bangladesh
fahmida.ahmed@g.bracu.ac.bd

Syed Ziaul Bin Bashar
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
syed.ziaul.bin.bashar@g.bracu.ac.bd

Md. Muhtadee Faiaz Khan Soumik
Dept. of Computer Science
BRAC University
Dhaka, Bangladesh
md.muhtadee.faiaz.khan.soumik@g.bracu.ac.bd

Md Farhadul Islam
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

Md Sabbir Hossain
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

Annajiat Alim Rasel
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—In the modern world, Machine Learning (ML) has become very popular at the forefront of many advancements in technology. It is used in many industries for solving multitudes of various problems. In recent years, a popular ML trend has become the industry standard - Zero-Shot Learning (ZSL). Nowadays different versions of zero-shot learning are used in many industries. We will discuss how the ZSL can recognize different accents to pronounce words of the same language in this modern world. We will also discuss how ZSL can detect speech sentiments in a different conversations. We will consider the pros and cons of using ZSL in this industry and some of the obstacles. We will also shed light on our new idea of using zero-shot learning on accent detection and speech sentimentation. We talk about the system model, the steps in particular, and aggregation pseudocode. We also discuss some issues with our model idea and provide solutions. Lastly, we reach a conclusion and discuss some future work. We believe ZSL will bring forth a data revolution in the modern world.**

***Index Terms*—Zero-Shot Learning (ZSL), Speech Sentimentation, Accent Recognition, Machine Learning (ML), Natural Language Processing (NLP), Solution**

I. Introduction

Many people have significant difficulty communicating because of their own or other people's accents. Others find it distracting when characters in an American film are speaking with a British accent. Therefore, the development of a system that can convert an accent while maintaining the speaker's original voice identity would have far-reaching implications for many fields. [1] One of Natural Language Processing's most fascinating uses is in the field of sentiment analysis. Analysis of user-generated content on social media sites and review sites for consumer goods are both included in this field. [2]After identifying emotional states, the zero-shot model classifies them as positive, negative, or neutral. From the high-dimensional input of the sentence transformer, dimensionality is gradually reduced as the input is mapped into probability values of various emotions, and then the probability values are mapped into the sentiment labels. [3] The entire sentiment analysis process is sped up because the second-stage input does not require laborious feature extraction or complex machine learning methods capable of catching sentiments directly from the text. Three benchmark datasets are used to test the effectiveness of the proposed method with different classifiers such as machine learning, neural networks, and ensemble learning. When compared to conventional Sentiment analysis detection, the proposed emotion-sentiment detection model requires less data for training. Pre-trained transformer models have a wide variety of applications in sentiment analysis tasks, including as text classifiers, for evaluation, and as zero-shot models that can assess the significance of a given word in the context of a given text. Emotion and sentiment words can be searched for explicitly and implicitly by zero-shot models, with the resulting probabilities indicating how closely the words are related to the text. Despite not needing training data, zero-shot models are not optimally tuned for sentiment analysis tasks and may require supplementary mechanisms to overcome their shortcomings. [3] In this paper, we present a graphical representation of speaker and accent embedding distributions for accent-converted and natural speech; the findings demonstrate the ability of accentron syntheses to capture the identity of the speaker's voice while also conveying the target

accent.

II. Literature Review

Researchers have turned to both indirect approaches, such as compatibility learning frameworks, and direct ones, such as learning intermediate attribute classifiers, to overcome the difficulty posed by zero-shot learning's split-second test and training sets. A new voice with the vocal identity of a specific L2 speaker and the accent of an L1 speaker is the goal of foreign accent conversion, as described in a publication published by science direct. The foreign-accented computer (FAC) can operate as a "golden speaker" to help a second-language speaker perfect their pronunciation by listening to their own voice spoken in a native accent. FAC is also used in voice recognition enhancement, customised TTS synthesis, and movie dubbing. Voice morphing, frame pairing, articulatory synthesis, and sequence-to-sequence (seq2seq) modeling are only some of the methods proposed for FAC. However, there are two significant shortcomings with traditional FAC methods. First, they function on a one-to-one basis, meaning that for each pair of L1 and L2 speakers, a new model must be trained. Second, a huge amount of speech data (about a thousand utterances per L2 speaker) is required. This means that L2 learners utilizing traditional FAC approaches for practical applications like pronunciation training have to record a large number of utterances and then wait for a dedicated model to be trained, which may be a time-consuming and discouraging process. They gave Accentron a comprehensive test on the L2-ARCTIC data set. Accent-converted speech and natural speech's speaker and accent embedding distributions were first shown; the results confirm that Accentron syntheses are capable of accurately reproducing both the L2 voice identity and the L1 accent. Second, they administered a battery of listening tests in two distinct environments: (1) a traditional FAC setup, where the test L2 speakers were present throughout training, and (2) a zero-shot FAC setup, which implies the test L2 speakers were unavailable throughout training. When compared to two other state-of-the-art FAC systems in normal FAC settings, Accentron improves accentedness by 27% while preserving acoustic quality and voice uniqueness. Additionally, Accentron's performance suffered no noticeable drop when tested with the zero-shot FAC configuration. [4]

Another paper from the MDPI study claims that people have spent decades modifying their speech patterns so that computers can "understand" them, but that speaking in natural language is now the norm. The vast majority of texts found on the internet are in an unorganized and unannotated format, making them largely useless. Only via processing can such noisy data be transformed into actionable intelligence. But processing by hand is laborious and slow. In contrast, automatic methods can reduce the need for human intervention, hasten the delivery of the intended machine output, and eliminate or reduce the need for large amounts of irrelevant data. In order to solve issues with language technology, natural language

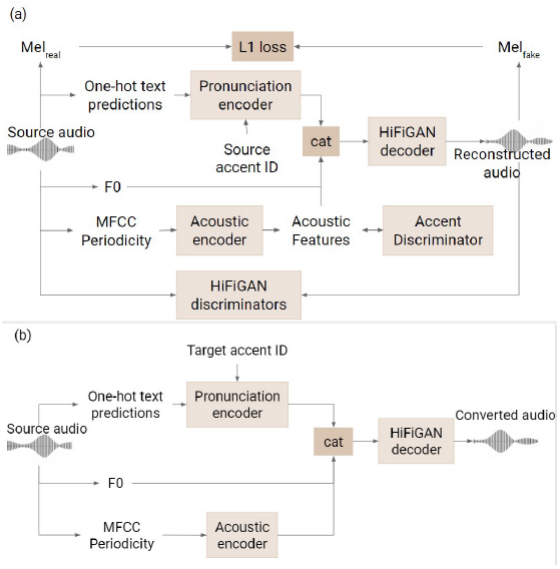

Fig. 1. (a) Training (b) During inference [4]

processing (NLP) implements AI techniques for smart human-machine communication. With the use of data mining, pattern recognition, and natural language processing, computers can perform cognitive tasks traditionally performed by humans. Machine translation systems, web search engines, natural language assistants, and opinion analysis are just some of the NLP applications helping to solve societal issues. The tone (emotions, thoughts) of a document is just as significant as its actual words in modern times. Understanding the emotional state of a group depends heavily on the ability to identify and categorize people's thoughts and feelings. [3] The purpose of sentiment analysis is to , generalize, and anticipate whether the text contains subjectivity and reflects the , apart from which feeling is the dominating. Most sentiment analysis research focus on assigning positive, negative, and sometimes neutral sentiment evaluations to the given text. A less investigated direction in sentiment analysis is to transfer the emphasis from studying sentiment toward a single item to the internal mood of the text itself. Zeroshot cross-lingual experiments show the evaluation of monolingual models applied to another language. Most models perform well with regularly used languages such as English; nevertheless, applying these algorithms straight to low-quality corpora typically delivers unsatisfactory results. Cross-lingual sentiment analysis intends to exploit highquality and rich resources in English to improve the classification performance of resourcescarce languages. These methods handle the challenge of training separate models for each language, but despite that, they lack mechanisms allowing to adjust traditional training methods (classifiers) for the Sentiment analysis task. [3]

III. Federated Learning Types and Uses

Zero-shot learning is a machine learning approach in which a model is able to classify items into categories that it has not

seen during training. This is achieved by using a pre-trained model that has been trained on a large dataset of categories and has learned a generalizable representation of the features that distinguish those categories. When presented with a new item that belongs to a category that it has not seen before, the model can use this learned representation to make a prediction about which category the new item belongs to.

Accent detection is the task of identifying the accent or dialect of a speaker based on their spoken language. Zero-shot learning could potentially be used for accent detection by training a model on a large dataset of speech samples from a variety of different accents, and then using that model to make predictions about the accent of a new speaker based on their speech patterns. However, this approach would likely require a very large and diverse dataset to be effective, and the accuracy of the model's predictions would depend on the degree to which the features that the model has learned to associate with different accents are generalizable to new accents.

To do this, you would first need to decide on the machine learning model that you want to use for the task. There are many different models and algorithms that can be used for zero-shot learning, including neural networks, support vector machines, and decision trees. You would then need to decide on a suitable dataset for training the model, and a method for representing the data in a way that the model can learn from it. A hidden Markov model (HMM) is a type of statistical model that can be used to represent the probability distribution over a sequence of observations. HMMs are commonly used for tasks such as speech recognition, language modeling, and bioinformatics, where the goal is to model the underlying sequence of events or states that generate a set of observations. Once the model has been trained, it can be used to make predictions about the underlying series of events or states based on a set of observations. This is done by using the trained model to calculate the probability of each possible sequence of hidden states given the observations and choosing the most likely sequence as the prediction.

BERT (Bidirectional Encoder Representations from Transformers) is a type of transformer-based machine learning model that has been widely used for natural languages processing tasks such as language translation, text classification, and question answering. BERT is able to process language by considering the context of words in a sentence, rather than just considering the individual words in isolation, which allows it to perform well on tasks that require an understanding of the context in which language is used.

BERT processes data in the form of text and is able to understand the context in which words are used by considering the relationships between words in a sentence. In contrast, HMMs process data in the form of a sequence of observations, and model the underlying sequence of events or states that generate the observations. This means that BERT may be more suitable for tasks that involve understanding the context in which language is used, while HMMs may be more suitable for tasks that involve modeling the underlying sequence of events or states that generate a set of observations. In terms of zero-shot learning, both BERT and HMM can potentially be used for the task by training the model on a large dataset of speech samples from a variety of different accents and using the trained model to make predictions about the accent of a new speaker based on their speech patterns. The accuracy of the model's predictions will depend on the quality and diversity of the training data, as well as the complexity and effectiveness of the model itself.

Once the model has been trained, you can use it to make predictions about the accent of a new speaker by inputting their speech patterns into the model and seeing which accent the model predicts. The accuracy of the model's predictions will depend on the quality and diversity of the training data, as well as the complexity and effectiveness of the model itself.

IV. Results and Analysis

There have been experiments with three systems (that were previously trained only on English classification tasks) to perform zero-shot classification on French text. [5] These tasks are Amazon Reviews (English & French), SNLI, and SST. The systems are compared to a "bridged" system, which involves translating the French text to English and then running it through the English classifier. It is found that the performance of the zero-shot systems was close to that of the bridged systems, and in some cases outperformed other methods that used bilingual or multilingual embeddings.

Model	Amazon (Fr)		SST (Fr)		SNLI (Fr)	
	Bridged	Zero-Shot	Bridged*	Zero-Shot	Bridged	Zero-Shot
Proposed model: *Encoder-Classifier*	73.30	51.53	79.63	59.47	74.41	37.62
+ Pre-trained Encoder	79.23	75.78	84.18	81.05	80.65	72.35
+ Freeze Encoder	83.10	81.32	84.51	83.14	81.26	73.88

Fig. 2. Zero-Shot performance on all French test sets. (Note that we use English accuracy in the bridged column for SST) [5]

The performance of the zero-shot systems improved significantly when using the pre-trained NMT encoder, and further improvement was observed when the encoder was frozen. On the Amazon Review task, the zero-shot system performed within 2% of the best-bridged system, and on the SST task, it had an accuracy of 83.14%, which was within 1.5% of the bridged equivalent.

Model	SNLI (Fr)
Our best zero-shot *Encoder-Classifier*	**73.88**
INVERT (Søgaard et al. 2015)	62.60
BiCVM (Hermann and Blunsom 2014)	59.03
RANDOM (Vulić and Moens 2016)	63.21
RATIO (Vulić and Moens 2016)	58.64

Fig. 3. Comparison of our best zero-shot result on the French SNLI test set to other baselines. [5]

On the SNLI task, the best zero-shot system was compared to other methods that used bilingual or multilingual embeddings, which were evaluated on the same French test set. [6] The zero-shot system performed the best, with an accuracy

of 73.88%. Other methods, such as INVERT [7], BiCVM [8], RANDOM [9], and RATIO, used techniques like inverted indexing, learning bilingual compositional representations, and training bilingual embeddings on parallel sentences with randomly shuffled tokens, but the system still outperformed them by a significant margin of 10.66% to 15.24

V. Conclusion and Future Work

The experiments presented in this paper demonstrate that zero-shot classification of French text is possible using systems that have only been trained on English classification tasks. The performance of these zero-shot systems was found to be comparable to that of bridged systems, which involved translating the French text to English before running it through an English classifier. In some cases, the zero-shot systems outperformed other methods that used bilingual or multilingual embeddings. The zero-shot systems showed particularly strong performance on the SNLI task, outperforming other methods by a significant margin. Further improvement in the performance of the zero-shot systems was observed when using a pre-trained NMT encoder, and freezing the encoder resulted in even better performance. These results suggest that it may be possible to perform zero-shot classification on other languages using systems trained only on English tasks and that pre-trained NMT encoders may be useful for this purpose.

References

[1] M. Jin, P. Serai, J. Wu, A. Tjandra, V. Manohar, and Q. He, "Voice-preserving zero-shot multiple accent conversion," *arXiv preprint arXiv:2211.13282*, 2022.

[2] G. Thakkar, N. M. Preradovic, and M. Tadic, "Multi-task learning for cross-lingual sentiment analysis," *arXiv preprint arXiv:2212.07160*, 2022.

[3] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, and R. Damaševičius, "Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning," *Applied Sciences*, vol. 12, no. 17, p. 8662, 2022.

[4] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, p. 101302, 2022.

[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Zero-shot cross-lingual classification using multilingual neural ... (n.d.)," Jan 2023. [Online]. Available: https://www.researchgate.net/publication/327643772_Zero-Shot_Cross-lingual_Classification_Using_Multilingual_Neural_Machine_Translation

[6] Ž. Agić and N. Schluter, "Baselines and test data for cross-lingual inference," *arXiv preprint arXiv:1704.05347*, 2017.

[7] A. Søgaard, Ž. Agić, H. M. Alonso, B. Plank, B. Bohnet, and A. Johannsen, "Inverted indexing for cross-lingual nlp," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1713–1722.

[8] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," *arXiv preprint arXiv:1404.4641*, 2014.

[9] I. Vulić and M.-F. Moens, "Bilingual distributed word representations from document-aligned comparable data," *Journal of Artificial Intelligence Research*, vol. 55, pp. 953–994, 2016.