Interspeech 2022
18-22 September 2022, Incheon, Korea

Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms

Mariia Lesnichaia[1,†], *Veranika Mikhailava*[2,†], *Natalia Bogach*[1], *Iurii Lezhenin*[1,3], *John Blake*[2], *Evgeny Pyshkin*[2]

[1]Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia
[2]The University of Aizu, Aizu-Wakamatsu, Japan
[3]Speech Technology Center, St. Petersburg, Russia

pyshe@u-aizu.ac.jp, bogach@kspt.icc.spbstu.ru

Abstract

Automatic speech recognition is hindered by the linguistic differences occurring in accented speech. This paper advances a classification method for accented speech using a CNN-based model trained and tested on English with Germanic, Romance and Slavic accents. The input feature set was examined to find the optimal combination of time-frequency and energy characteristics of speech fed into the machine learning model. We also tuned model hyperparameters and the dimensionality of input features. We argue that mel-scale amplitude spectrograms on a liner scale appear more powerful in accent classification tasks compared to conventional feature sets based on MFCCs and raw spectrograms. Our models used only sparse data from the Speech Accent Archive, yet produced state-of-the-art classification results for English with Germanic, Romance and Slavic accents. The accuracy of our models trained on linear scale amplitude mel-spectrograms ranged from 0.964 to 0.987, outperforming existing models classifying accents using the same dataset.

Index Terms: Automatic accent identification, Convolutional neural networks (CNN), Mel-frequency cepstral coefficients (MFCC), Amplitude mel-spectrogram.

1. Introduction

Features associated with speakers, such as geographic region, gender, age, social class, and mother tongue, combine to create distinctive accents [1]. The compounded effect of phonemic and prosodic contact between L1 and L2 phonological systems may be perceived as foreign accents [2, 3]. Conventional acoustic language models adapted to fit the standard language corpora fail to fulfill the recognition requirements when applied to accented speech [3, 4]. Adding more pronunciation samples to the training dataset for speaker-independent speech recognition scheme is inappropriate as it increases the processing time for audio recordings and creates additional noise, degrading the performance [5]. In contrast, accent detection improves the robustness of automatic speech recognition systems (ASR), since it helps to overcome this unwanted variability [6, 7, 1, 8, 9, 10, 11, 12]. Being a sub-task of speech and language recognition, in terms of classification models, accent detection is based on the same machine learning architectures [13, 14, 5, 15, 10, 1, 16], e.g. CNN [17, 1, 18, 14, 13, 19], FFNN [10], HMM [20], KNN [21], Logistic Regression [15, 20], GMM [22, 23], LSTM and bLSTM [3, 8], Random Forest, SVM [24, 25, 20, 23, 21]. Accent classification accuracy depends upon the input feature set. The best results to date have been achieved using mel-frequency cepstral coefficients (MFCC), while classification was also successful using other types of input features, such as: spectrogram (SG), chromagram (CG), spectral centroid (SC), and spectral rolloff (SR), mel-weighted single filtered frequency (SFF) spectrogram [1, 25]. Previous works addressing speech accent identification that inspired and shaped our current research either in terms of classification model, audio descriptors or training dataset are summarized in Table 1.

This study investigated whether using time-frequency and energy features could improve the accuracy when used jointly with MFCC as input features in the task of automatic accent detection. We demonstrate that the greatest contribution to recognition has been made by the presence of stable time-frequency patterns of energy distribution, represented by amplitude mel-spectrograms on a linear scale, which alone could be fed into the classification model as mel-spectrogram captures all of the relevant pronunciation-specific details [28]. The accuracy of our model ranged from 0.964 to 0.987 when working with 9 classes of accented speech in English. A similar result using mel-spectrogram with CNN model accent was achieved when discriminating between 5 accent classes of spoken Kashmiri showed an accuracy of 0.9866 [26].

2. Feature Extraction

A common approach to speech signal processing is to use short-term analysis, under the assumption that signal characteristics within a short-term frame remain unchanged. Speech utterances are compared to feature vectors, presumably differing in their distribution with different L1s. For speech signal analysis, the frame length is near 10–30 ms, with an overlap between frames approximately equal to half their length [10].

2.1. Audio descriptors

The six additional features used to extend the MFCCs are described here. **Spectral centroid** (SC) indicates the frequency at which the energy of the spectrum is concentrated, or where the center of mass of the sound is located. **Spectral roll-off** (SR) is a measure of the asymmetry of the spectral shape of the signal. SR represents the frequency below which a given percentage (85%) of the total energy of the spectrum lies. This value is used to determine vocalized sounds in speech since unvoiced sounds have a large proportion of the energy contained in the high frequency range of the spectrum. **Chromagram**, which is usually a 12-dimensional feature vector, represents the amount of energy for each of the signal's height classes (C, C#, D, D#, E, etc.). **Zero Crossing** (ZCR) represents the number of signal

[†] M.L. and V.M. are both the first authors of this work.

10.21437/Interspeech.2022-462

Table 1: *Related works*

Paper, year	Feature set	Model	Classes	Accents	Dataset
[26], 2022	Mel spectrogram	CNN	5	5 Kashmiri accents	Custom
[18], 2021	SG	CNN (LeNet)	5	DU, FR, JA, NS, PO	IViE, Cambridge English Corpus
[19], 2021	SG	CNN	5	AR, FR, GE, IN, NS	SAA
[1], 2020	MFCC, SG, CG, SC, SR		5	AR, FR, NS, SP, ZH	
			3	AR, NS, ZH	
[13], 2020	MFCC	CNN with attention	2	IN, NS	Custom
			4	IN	
			9	IN, NS	
[15], 2020	MFCC	Logistic Regression	3	HA, IG, YO	
[20], 2019	MFCC	LSTM, RF	4	NS, SP	
[10], 2017	MFCC, LPCC	FFNN	6	GA, IN, IT, JA, KO, NS	Wildcat
[14], 2017	SG	CNN (AlexNet)	3	NS, SP	SAA
[27], 2017	MFCC	GMM	3	ML	Custom
[16], 2012	Mel-spectrogram statistics	FF-MLP	3	IN, MS, ZH	
[5], 2005	2nd and 3rd formants	GMM	2	IN, NS	Custom (SAA subset)

sign changes within a segment. ZCR can be helpful in describing the noisiness of the signal. For unvoiced speech, the ZCR characteristic takes on higher values due to unvoiced speech being associated with turbulence. **Root mean square** (RMS) represents the average signal strength. **Fundamental frequency** (F0) is the lowest frequency at which a person's vocal cords vibrate when making voiced sounds. F0 makes a significant contribution to the perception of foreign accents [6], which is especially noticeable for Germanic and Romance languages [29]. Estimation of the fundamental frequency of the signal is carried out using the autocorrelation-based YIN algorithm [30]. The first input feature set includes 30 audio descriptors, namely: 13 MFCCs, 12 chroma coefficients, SC, SR, ZCR, RMS and F0.

2.2. Amplitude mel-spectrograms

An alternative input feature set was formed by **amplitude mel-spectrograms on a linear scale**. The audio signal frequencies f were converted to mel-spectrograms $M(f)$ as follows:

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

Linear scale amplitude mel-spectrograms were chosen because they performed better at classifying accents than logarithmic amplitude mel-spectrograms, power mel-spectrograms, and SFF mel-spectrograms. By experimenting with mel-spectrograms with 32, 64 and 128 bands as input features, we found that the optimal balance between learning rate and recognition accuracy can be achieved using mel-spectrograms with 64 bands (see Section 3 for details).

3. Experiments

3.1. Experiment set-up

We used a subset of 9 groups from the Speech Accent Archive [31]. These groups are labelled according to L1 as Germanic languages (English (EN), German (GE), Dutch (DU), Swedish (SW)), Romance languages (Spanish (SP), Italian (IT), French (FR)) and Slavic languages (Polish (PO), Russian (RU)). At the time of experiments, the number of recordings by group were as follows: $\{$EN: 650, GE: 44, DU: 53, SW: 23, SP: 233, IT: 39, FR: 85, PO: 39, RU: 81$\}$. To compensate for the unequal distribution, we used the first 80 samples from the larger groups.

The classification model for accent detection is built on CNN used in [1]. The output value is the probability distribution vector which attributes the speech sample to a specific accent class. The model consists of two convolutional layers with ReLU activation function and two-dimensional filters. The first and second convolutional layers contain 32 and 64 blocks, respectively. After each convolutional layer batch normalization and pooling are applied. The flatten layer is followed by two dense layers of direct propagation. We use 128 neurons and ReLU activation function in the first dense layer. We set the number of neurons equal to the number of accents and use the softmax activation function in the second layer. The input of the model is a feature matrix extracted from audio signals. For the basic implementation of the model, we chose convolutional filters with size (3, 3) and pooling layers (2, 2) with a stride of 2 following [1]. To prevent overfitting, we use the dropout method with a variable probability of any neuron turning to zero – depending on the type of input data, a value from 10 to 50% is used. We used categorical cross-entropy as a loss function during training. Learning loss function is minimized using the adaptive moment estimation (Adam) algorithm, where the constant learning rate coefficient is 0.001, and the parameters β_1 and β_2 are 0.9 and 0.999, respectively.

3.2. Experiments on model adjustment

3.2.1. Audio processing

Audio recordings with a sampling rate of 22050 Hz were split into multiple consecutive frames of 25 ms, each with an overlap of 10 ms. To determine whether to keep or remove the silence fragments (pauses) from the input to improve recognition quality, we performed the experiments for both situations while applying the set of characteristics including 13 MFCCs and fundamental frequency $F0$.

Keeping the fragments of silence resulted in higher accuracy and so audio files processed in all the subsequent accent classification experiments are used in their original form, i.e., with all the pauses preserved.

3.2.2. Data augmentation, regularization and filter size

The optimal maximum percentage of horizontal shift found experimentally during data augmentation is 20%. Horizontal shifts from 5 to 30% were tested for the Italic group.

Following the recommendations in [1], we experimented with the 2D filter configurations for 30 characteristics of the MFCC-based feature set (Section 2.1). Thus, we used two configurations (3,3) (3,3) and (5,5) (3,3) for kernel size and pool size in the convolutional and pooling layers respectively. The (3,3) (3,3) filters performed better than (3,3) (2,2).

For amplitude mel-spectrograms ona linear scale (Section 2.2) four 2D filter configurations were tried for classification among the 3 Romance accents {FR, IT, SP}. The length of the input feature matrices used to represent the input data was 100. The learning process was completed when the change in the recognition accuracy did not exceed 1% within 10 epochs. The highest recognition accuracy of 99.04% with a relatively short model training time were achieved when using filters of size (3, 3) in both the convolutional and pooling layers. Thus, filters of size (3, 3) in hidden layers are the most universal (Table 2).

Table 2: *Filter sizes for amplitude mel-spectrograms*

French, Italian, Spanish (Italic group)				
Kernel size	Pool size	Learning Time (mm:ss)	Accuracy	Error
(3, 3)	(2, 2)	41:06	0.9889	0.0614
(3, 3)	(3, 3)	20:01	0.9904	0.0261
(5, 5)	(3, 3)	17:57	0.9852	0.0564
(7, 7)	(3, 3)	26:14	0.9867	0.0468

3.2.3. Input matrix dimension

When working with speech signals, it is necessary to consider the patterns of change in the characteristics describing these signals over time. Thus, it is essential to consider the sequences of vectors of features or input matrices, not vectors at discrete points in time. Dividing the input features into larger chunks allows for longer speech patterns that are more likely to be accent-dependent. This could be performed, however, at the expense of training set decrease and longer computation time. Shorter fragments, on the contrary, allow for larger training sets; but should input matrices be too small, it may be impossible to capture information about the accent. To find the optimal size of the input feature, matrices feature vectors of MFCC were grouped into blocks containing 30 to 500 vectors per block. The training stops when the change in accuracy is less than 0.5% for an interval of 20 epochs or when 300 epochs is reached among five accents and 170 epochs in other cases. The probability of a neuron going to zero when using the thinning method is 50%.

As an effect of modifying the dimension of input features and the maximum percentage of horizontal image shift during data augmentation, classification accuracy among five classes increases by about 7% compared to [1] (60.95% and 53.92%) for recognition among five accents.

We used a dropout of 0.25, the size of the filters in the convolution layers is (5, 5), and (3, 3) in the pooling layers. The training stops when the recognition accuracy ceased to change by at least 1% for ten epochs. Mel-spectrograms, consisting of 64 frequency bands, proved to be the most effective and were chosen as input characteristics for recognition. Although the use of 128-band mel-spectrograms can slightly increase the recognition accuracy, training time increases severalfold. Contrariwise, using mel-spectrograms consisting of 32 mel-frequency bands (being naturally less computationally expensive) leads to a significant increase of error while testing the classifier. The optimal length of the input feature matrices in the case of using amplitude mel-spectrograms on a linear scale is 75. Thus, this value is used when classifying using amplitude mel-spectrograms.

Table 3: *Classification results using different types of input features for Slavic and Italic accents*

Features	Test Accuracy	Test Loss
Russian, Polish (Slavic group)		
Threshold Accuracy – 0.72		
MFCC	0.84	0.37
MFCC + F0	0.83	0.4
MFCC + spectral centroid	0.85	0.39
MFCC + spectral decay	0.84	0.4
MFCC + chromagram	0.79	0.44
MFCC + ZCR	0.84	0.38
MFCC + RMS	0.83	0.41
All	0.81	0.41
French, Italian, Spanish (Italic group)		
Threshold Accuracy – 0.43		
MFCC	0.75	0.6
MFCC + F0	0.69	0.71
MFCC + spectral centroid	0.67	0.73
MFCC + spectral decay	0.68	0.72
MFCC + chromagram	0.63	0.84
MFCC + ZCR	0.71	0.68
MFCC + RMS	0.7	0.7
All	0.66	0.8

3.3. Experiments on accent detection feature sets

3.3.1. MFCC and other audio descriptors

We investigated which characteristics for MFCC extension would positively impact classification accuracy while maintaining the basic filter sizes in the hidden layers of the classifier. The training stops when either training accuracy of 90% or 120 epochs is reached for all accent sets except for {EN, RU, SP, SW}. The training process continues until 350 epochs is reached for the cases of {EN, RU, SP, SW}.

Recognition accuracy of our model outperformed pure MFCC in half of the cases with filter sizes (3, 3) in convolutional layers and (2, 2) in pooling layers (Table 3). In the case of the accent group {EN, GE, IT, PO}, adding the fundamental frequency to the MFCC helped to increase the recognition accuracy by about 3%. For the set {EN, RU, SP, SW}, the most effective selection was to use all types of additional characteristics: the increase in classification accuracy also turned out to be about 3% compared to the usage of pure MFCC.

3.3.2. Mel-spectrograms

Linear scale mel-amplitude spectrograms extracted from audio signals were also tried as input to the classifier model under the settings for filter size, input matrices dimensions and learning proved to be optimal in previous sections. The length of the input feature matrices was 75 elements. The number of epochs was limited to 60, while the preliminary termination of the learning process was set when the change in recognition accuracy stopped by at least 1% within ten epochs. Regularization was applied to the training set or to test set, but not for both at a time. Its values ranged from 10 to 25%.

4. Discussion

By the end of the training, the model is able to achieve similar accuracy and loss values for the training and test data. For a smaller number of epochs compared to previous experiments, it

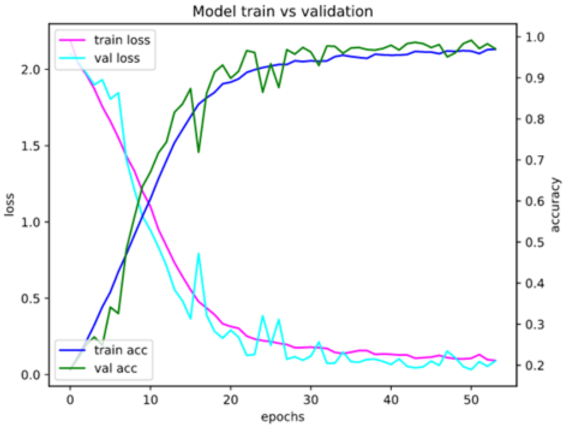

Figure 1: *Training curves of the proposed model for 9 classes of accents (DU EN FR GE IT PO RU SP SW).*

Table 4: *Accuracy and loss for trained classification models*

Accents	Accuracy	Loss
PO RU	0.987	0.039
FR IT SP	0.986	0.052
DU EN GE SW	0.982	0.075
EN RU SP SW	0.988	0.042
EN GE IT PO	0.985	0.053
DU EN FR RU	0.984	0.039
EN FR GE RU SP	0.978	0.071
DU EN FR GE RU SP	0.964	0.097
DU EN FR GE IT PO RU SP SW	0.986	0.044
Average	0.982	0.056

was possible to achieve a much smaller error and greater accuracy, which means that using amplitude mel-spectrograms on a linear scale allows the model to place broad boundaries between classes (Table 4). Linear scale amplitude mel-spectrograms gave much better accuracy and loss results than using MFCC alone or combined with additional features. The average number of training epochs was 46, with an average duration of 37.18 sec. It took 52 epochs with an average duration of 64.09 sec for the model to learn to classify the 9 accents (Fig. 1). Table 5 presents the results obtained during testing of the model for the same number of classes against the previously reviewed publications.

Intonation makes a significant contribution to the recognition of foreign accents. Based on the fact that the F0 contour in most experiments did not improve the classification results, we can conclude that intonation features are subsumed within MFCC. When extracting MFCC, information about F0 is partially preserved due to the close distance between the low-frequency channels of the mel-filters [32].

Amplitude mel-spectrograms on a linear scale showed high efficiency in recognizing foreign accents in English speech. However, the results turned out to be slightly lower compared to [13]. This may be explained by the variation in recording equipment using within datasets. All entries in [13] were created with the same recording equipment while this was not the case in the Speech Accent Archive dataset.

Compared to other solutions based on the Speech Accent Archive dataset – [14, 1, 19] and with [5], the implemented model achieved better recognition accuracy with no additional computational overhead by tuning hyperparameters and dimensionality of input features, as well as selecting amplitude mel-spectrograms on a linear scale as input features. The better recognition quality compared to [5] can be explained, among other things, by the fact that the authors of [5] removed silence fragments from audio recordings before extracting characteristics. During this research, we found that pauses in speech have a positive effect on the ability to detect accent.

Table 5: *Accuracy of existing solutions and the results obtained*

Source	Classifier	Number of classes	Accuracy of existing solution	Accuracy of proposed model
[26]	CNN	5	0.987	0.978
[18]	CNN (LeNet)	5	0.923	0.978
[19]	CNN	5	0.902	0.978
[1]		3	0.703	0.986
		5	0.539	0.978
[14]	CNN (AlexNet)	3	0.61	0.986
[13]	CNN with attention	2	1.0	0.987
		4	0.99	0.984
		9	0.995	0.986
[15]	Logistic Regression	3	0.82	0.986
[10]	FFNN	6	0.914	0.964
[16]	FF-MLP	3	0.99	0.986
[27]	GMM	3	0.89	0.986
[5]		2	0.862	0.987
[20]	LSTM, RF	5	0.947	0.978

Amplitude mel-spectrograms on a linear scale, carrying information about the energy of the audio signal, showed good results when classifying up to 9 accents. In more than half of the cases, when comparing the obtained results with those described in the literature we reviewed, it was possible to achieve higher recognition accuracy results, namely 98.6%. When using amplitude mel-spectrograms as input data, our proposed model demonstrates equally high accuracy and completeness of classification both on average and separately for each class, despite the sparsity of the dataset used.

Thus, the amplitude mel-spectrograms on a linear scale showed effectiveness in determining the speech accent in a foreign language using a CNN-based classifier. Further studies of this approach may expand the number of recognition classes, using an intermediate classifier to determine the L1 language group of the speaker before classifying a particular accent and using a dataset with a variety of spoken content.

Though the techniques and features used are known in the speech processing domain, exhaustive experiments involving their combination and application to a specific problem of accent recognition have not been reported so far. Due to space limitations, we will prepare a separate publication for in-depth description of the classification experiments conducted for different sets of languages using a variety of applied features along with an approach to selecting the optimal parameters for CNN filters.

5. Acknowledgment

This research was in part funded by the Japan Society for the Promotion of Science (JSPS), grant number 20K00838.

6. References

[1] Y. Singh, A. Pillay, and E. Jembere, "Features of speech audio for accent recognition," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 2020, pp. 1–6.

[2] P. Boula de Mareüil and B. Vieru, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, pp. 247–267, 02 2006.

[3] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss, "Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features." in *Interspeech*, 2016, pp. 2388–2392.

[4] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *Interspeech*, 2018, pp. 2454–2458.

[5] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, 2005, pp. 139–143.

[6] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 836–839 vol.1.

[7] T. C. C. Huang and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, pp. 141–153, 04 2004.

[8] F. Weninger, Y. Sun, J. Park, D. Willett, and P. Zhan, "Deep learning based Mandarin accent identification for accent robust asr." in *INTERSPEECH*, 2019, pp. 510–514.

[9] Z. S. F. Zheng, G. Zhang, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, pp. 582–589, 11 2001.

[10] E. Tverdokhleb, H. Dobrovolskyi, N. Keberle, and N. Myronova, "Implementation of accent recognition methods subsystem for elearning systems," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, 2017, pp. 1037–1041.

[11] H. Huang, X. Xiang, Y. Yang, R. Ma, and Y. Qian, "Aispeech-sjtu accent identification system for the accented english speech recognition challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6254–6258.

[12] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *Interspeech*, 2019, pp. 2140–2144.

[13] A. Ahamad, A. Anand, and P. Bhargava, "Accentdb: A database of non-native English accents to assist neural speech recognition," 2020.

[14] A. Ensslin, T. Goorimoorthee, S. Carleton, V. Bulitko, and S. Poo Hernandez, "Deep learning for speech accent detection in video games," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 13, no. 1, Sep. 2017.

[15] A. E. M. Francisca Oladipo, Rahmon A Habeeb, "Accent identification of ethnically diverse Nigerian English speakers," *SSRN Electronic Journal*, 9 2020.

[16] Y. Ma, M. Paulraj, S. Yaacob, A. Shahriman, and S. K. Nataraj, "Speaker accent recognition through statistical descriptors of mel-bands spectral energy and neural network model," in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, 2012, pp. 262–267.

[17] Q. T. Duong *et al.*, "Development of accent recognition systems for Vietnamese speech," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2021, pp. 174–179.

[18] C. Graham, "L1 identification from l2 speech using neural spectrogram analysis," *Interspeech*, 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-1545

[19] P. Berjon, A. Nag, and S. Dev, "Analysis of French phonetic idiosyncrasies for accent recognition," *Soft Computing Letters*, vol. 3, p. 100018, 12 2021.

[20] J. Bird, E. Wanner, A. Ekárt, and D. Faria, "Accent classification in human speech biometrics for native and non-native english speakers," in *Pervasive Technologies Related to Assistive Environments (PETRA)*, 06 2019, pp. 554–560.

[21] G. R. Krishna, R. Krishnan, and V. K. Mittal, "A system for automatic regional accent classification," in *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 2020, pp. 1–5.

[22] J. Cheng, N. Bojja, and X. Chen, "Automatic accent quantification of Indian speakers of English." in *Interspeech*, 2013, pp. 2574–2578.

[23] A. Lazaridis, E. el Khoury, J.-P. Goldman, M. Avanzi, S. Marcel, and P. N. Garner, "Swiss French regional accent identification." in *Odyssey*, 2014.

[24] G. Işik and H. Artuner, "Turkish dialect recognition using acoustic and phonotactic features in deep learning architectures," *Journal of Information Technologies*, vol. 13, no. 3, pp. 207–216, 2020.

[25] R. Kethireddy, S. R. Kadiri, P. Alku, and S. V. Gangashetty, "Mel-weighted single frequency filtering spectrogram for dialect identification," *IEEE Access*, vol. 8, pp. 174 871–174 879, 2020.

[26] S. S. Malla, "Acoustic features based accent classification of Kashmiri language using deep learning," *Global Journal of Computer Science and Technology*, 2022.

[27] M. Aswathi Sanal, "Accent recognition for malayalam speech signals," *International Journal of Innovative Research in Computer and Communication Engineering*, 2017.

[28] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu *et al.*, "End-to-end accent conversion without using native utterances," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6289–6293.

[29] L. Rasier and P. Hiligsmann, "Prosodic transfer from l1 to l2. theoretical and methodological issues," *Nouveaux Cahiers de Linguistique Française*, vol. 28, 01 2007.

[30] H. K. Alain de Cheveigné, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[31] George Mason University. (2021) Speech accent archive. [Online]. Available: https://accent.gmu.edu/

[32] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 24–33, 2007.