Teager Energy Cepstral Coefficients for Spoken Language Identification

Arth J. Shah*, Savita H. Yadav† and Hemant A. Patil*
* Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat, India
* E-mail: {202101154, hemant_patil}@daiict.ac.in
† Uka Tarsadia University, Bardoli, Gujarat, India
† E-mail:savita22125@gmail.com

***Abstract*—Spoken Language Identification (SLID) is a key component in audio processing that facilitates the recognition and understanding of audio clips of spoken languages. Various applications, such as automatic speech recognition, multilingual voice assistants, and real-time translation services need SLID capabilities. Effective SLID enhances the transcription and processing of spoken language data and improves user experiences by providing personalized and relevant responses in the correct language. This study propose Teager Energy Cepstral Coefficients (TECC) features to capture the characteristics of spoken language. To evaluate how well TECC performs for SLID, we conducted a comparative analysis of its performance across various spectral features, namely, Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC). Two classifiers, deep residual networks (ResNet-50) and Time Delay Neural Networks (TDNN), were employed in this comparison. To maximize the performance of SLID system, we applied feature-level fusion and score-level Fusion techniques to advance the state-of-the-art. Additionally, latency analysis assesses time optimization to ensure the system operates efficiently. We obtained 98.25 % accuracy in this study for two languages, and 84.25 % on 10 different languages using TECC features.**

***Index Terms*—Spoken Language Identification, Teager Energy Cepstral Coefficients, ResNet-50, Feature-Level Fusion, Score-Level Fusion.**

I. Introduction

The task of automatically identifying language being used by a user from a sample of speech, regardless of the speaker's accent, gender, or identity is called as Spoken Language Identification (SLID) task. Solving the issue of SLID, has enormous potential in real-life cases, such as multilingual spoken translation, human-machine communication systems, spoken language retrieval, multilingual speech transcription system, and many other language-based tasks. Beside solving all these issues, also helps users in many international affairs, such as smooth communications between people of other countries. The proposed system has the potential to work seamlessly with automatic speech recognition, multilingual voice assistants, and real-time translation services, enhancing communication across different languages. The capability of proposed method enhances the transcription and processing of spoken language data, ensuring accurate interpretation. In our study, we employed a system that can identify 10 international languages from raw speech waveforms. We employ signal processing-based front-end feature extraction, and deep learning classifiers as back-end for SLID task. Numerous contemporary studies employ characteristics that provide insights into the structure of the vocal tract system, utilizing the acoustics of speech, which can be derived from Mel Frequency Cepstral Coefficients (MFCC) features [1], [2]. Such acoustic attributes encompass distinct phonemes for each language, thereby assisting in differentiating between multiple languages. Nevertheless, MFCC features are notoriously susceptible to noisy surroundings and variations in speaking styles.
In this paper, we propose use of Teager Energy Operator (TEO)-based features, namely, Teager Energy Cepstral Coefficients (TECC) features for SLID task. While the MFCC features align with human auditory perception and the LFCC features provide an equal frequency representation, TECC captures the vocal tract energy and the non-linear characteristics of speech signals. Deep learning classifiers, such as ResNet-50, and TDNN have been used in this study, in order to leverage the proposed methodology with recent technological advancement. Experimental results demonstrate that the combination of TECC features with the ResNet-50 classifier achieves better accuracy for LID as compared to the TDNN classifier (to be discussed in Section IV-A). Similar observations can be observed on MFCC, and LFCC features, with both the classifiers. Traditional classification methods have been used previously for many studies, such as in [3], [4], authors employ acoustic features with statistical methods, such as the Support Vector Machine (SVM), Hidden Markov Model (HMM), Universal Background Model (UBM), and Gaussian Mixture Model (GMM) for classification of acoustic features. Motivated by employment of residual-based features with TDNN classifier for SLID task [5], we employ TECC-based features with TDNN and ResNet-50 classifier.

In order to obtain optimal results by exploring variety of features, we also reported the results using two types of data fusion strategies (namely, feature-level, and score-level), which resulted in a further increase in accuracy. Although the proposed TECC feature vector does not perform at its optimal capacity, we successfully provide logical reasoning and analysis for the results obtained.

A. Related Works

Many existing studies in recent days have boost the research potential on SLID tasks. Some factor, that remains common in most of the studies on SLID tasks, till date are dataset

(VoxLingua107), and use of recent deep learning models. In [5], authors employ linear prediction residual-based features for SLID task for classification of Indian languages. For countries with multilingual and large population, such as India, SLID has key importance. For diversified countries, such as India, where there are more than 720 languages, and Papua New Guinea (840 languages), SLID plays a critical role. In [6], authors employ CNN-based method to identify 5 languages, with CNN classifier and spectrograms as features. However, study in [6] fails to demonstrate hard-link relationship between spectrograms and speech acoustics. In [7], authors use LSTM as a classifier and obtained 86 % accuracy. ECAPA-TDNN and MFCC30 have also been employed for SLID task, which reported an accuracy of 76.8 % for 8 language classes. Many other studies, were also reported for SLID task, with advancement in self-supervised learning based pre-trained models, such as whisper [8], wav2vec2.0 [9], HuBERT [10], and XLS-R [11]. Proposed methodology provides the following novelty :

- This study discussed how the energy of vocal tract system, and non-linearity in speech signal contribute to language-specific information for SLID task.
- Technical reasoning on energy difference of signals of different languages.
- As the SLI task should be robust against smaller speech segments, we show our model's latency period for the proposed TECC features and also for fair comparison with existing works [5].
- For improvement of results, we also did various types of fusion, namely, feature-level, and score-level.

II. Proposed Approach

In this Section, we explain the motivation for TEO, development of TEO, and computational detail's of TECC feature extraction.

A. Energy of Real Physical System

In [12], authors proposed that within a single pitch period, speech can be represented as a combination of signals, where the amplitude and frequency vary over time, known as AM-FM signals. This implies that Simple Harmonic Motion (SHM) can help to explain the concept of energy in speech wave. By applying Newton's second law of motion to the mass-spring system with spring constant C and mass N, it gives the dynamics as the 2nd order linear differential equation [13] :

$$\frac{d^2y}{dt^2} + \frac{C}{N}y = 0, \tag{1}$$

whose solution is recognized as Simple Harmonic Motion (SHM) :

$$y(t) = A\cos(\Omega_0 t + \phi), \tag{2}$$

where A is amplitude, and Ω_0 denotes the angular frequency (in radians) of oscillations. The previous solution can explained in following way. Any periodic function can be decomposed into Fourier series, which consists of an infinite sum of sine waves of varying frequencies. The general solution of such a 2nd order linear differential equation, where C and N are positive constant, form $e^{\pm j\sqrt{(C/N)}}$ which ultimately results the form $y(t) = A\cos(\Omega_0 t + \phi)$ [14].

The total energy consists the combined potential energy in the spring and the kinetic energy of the mass, i.e,

$$E = P.E. + K.E. = \frac{1}{2}Cx^2 + \frac{1}{2}N\left(\frac{dy}{dt}\right)^2. \tag{3}$$

Substituting $y(t) = A\cos(\Omega_0 t + \phi)$ and using trigonometry,

$$E = \frac{1}{2}NA^2\Omega^2, \tag{4}$$

$$\text{Or,} \quad E \propto A^2\Omega_0^2. \tag{5}$$

Eq. (2) is the true energy required to generate $y(t)$ in SHM.

B. Development of TEO

Kaiser proposed following algorithm for continuously estimating the energy level present in a signal, or the energy needed to generate the signal [3]. In discrete-time domain Eq. (2) can be expressed as:

$$y(n) = A\cos(\omega_0 n + \phi). \tag{6}$$

From (6), we can write:

$$y(n+1) = A\cos(\omega_0(n+1) + \phi), \tag{7}$$

$$y(n-1) = A\cos(\omega_0(n-1) + \phi). \tag{8}$$

Multiplying (7) and (8) and using trigonometry,

$$\begin{aligned} &y(n+1)y(n-1) \\ &= A^2\cos(\omega_0(n+1) + \phi)\cos(\omega_0(n-1) + \phi), \\ &= [A\cos(\omega_0 n + \phi)]^2 - A^2\sin^2\omega_0. \end{aligned} \tag{9}$$

Using Eq. (6) in Eq. (9),

$$A^2\sin^2\omega_0 = y^2(n) - y(n+1)y(n-1). \tag{10}$$

For low values of $\omega_0, \sin\omega_0 \approx \omega_0$, hence Eq. (10) can be written as

$$A^2\omega_0^2 \approx y^2(n) - y(n+1)y(n-1) = \psi\{y(n)\}, \tag{11}$$

where $\psi\{.\}$ is the TEO, which gives a running estimate of the energy of discrete-time signal $x(n)$. Fig. 1 represents functional block diagram of proposed SLID system using TECC features. TECC is known to capture energy of the vocal tract system and the non-linearities in the speech signal, which plays an crucial roles in extracting language-specific information. This energy is reflected in the formant frequencies, which are the resonant frequencies of the vocal tract system. Different languages utilize distinct sets of vowel and consonant sounds, which are characterized by their formant frequencies. The energy distribution across different frequency bands (spectral envelope) provides information about the articulation of phonemes. Different languages have unique spectral patterns due to their specific phoneme inventories, which may be key acoustic cue for SLID problem. The energy patterns also reflect

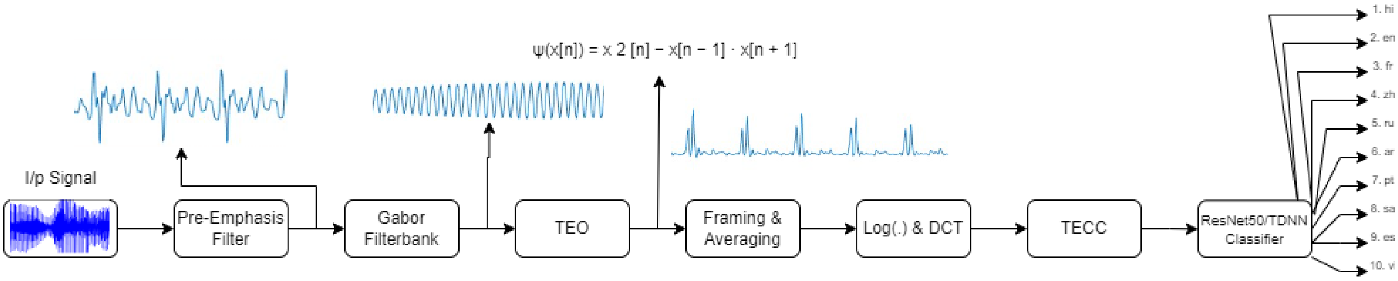

Fig. 1. Functional block diagram of proposed SLID system using TECC features, with ResNEt-50 / CNN classifier.

prosodic features, such as stress, rhythm, and intonation, which vary between languages and convey important linguistic information. The non-linear nature of speech production mechanism creates harmonics and overtones, which are crucial for the perception of timbre and the distinction of different phonetic elements. These patterns are also language-specific and can be used to identify phonemes and prosodic features. Hence, TECC works better with MFCC, and LFCC combined after data fusion for SLID task (to be discussed soon).

III. Experimental Setup

A. Dataset Used

We employed statistically meaningful VoxLingua107 dataset, which is a comprehensive multilingual speech corpus, designed for training language identification models. It comprises over 6,800 hours of speech data spanning 107 different languages. The recordings in this dataset are sourced from YouTube, ensuring a diverse range of accents, dialects, and speaking styles. The dataset is annotated at the segment-level, allowing for precise language identification. It includes a variety of speech contexts, such as news, conversations, and lectures. We selected 10 languages, namely (labels), Hindi (hi), English (en), French (fr), Mandarin Chinese (zh), Russian (ru), Arabic (ar), Portuguese (pt), Sanskrit (sa), Spanish (es), and Vietnamese (vi). 4000 random samples from each language were taken to conduct experiments, with average time spam of 9.4 seconds for each sample (1K samples ranging from each range {0-5, 5-10, 10-15, and 15-20 seconds}). We accounted around 11.3 hours of data for each of 9 languages (for Sanskrit, 9 hours data was taken), resulting into a total of 111.16 hours of vast data.

B. Classifiers Used

1) ResNet-50: ResNet contains four blocks within each block. The first block has three convolutional layers, followed by four, six, and three convolutional layers, respectively. Batch normalization and ReLU activation functions are applied after each convolutional layer. After the main blocks, there is a global average pooling layer that reduces the spatial dimensions of the feature maps. This is followed by a fully-connected layer with a softmax activation function, which produces the final output probabilities for different classes. This architecture is also known as *ResNet-50* [15].

2) Time Delay Neural Networks (TDNN): Motivated by previous works on SLID, authors in this study also decided to employ TDNN as classifier [5]. The TDNN architecture's ability to process sequential data and capture long temporal contexts proves advantageous for SLI tasks [16]. The incorporation of an attention mechanism facilitates the emphasis on more prosodically and linguistically significant frames within an utterance, enhancing the model's performance. The input shape to both the classifiers was constant, i.e., 20 * 750 (number coefficients*maximum number of frames). In SLID task, some frame-level features within an utterance exhibit more pronounced and unique prosodic characteristics. To tackle this, study reported in [17] has integrated attention mechanism to assign weights to these frame-level features. Specifically, [18] employs this attention mechanism on TDNN for SLID. With similar motivation, this study also employs an attention mechanism.

C. Cepstral Features Used

We employed MFCC and LFCC as baseline features for comparison with TECC. MFCC, and LFCC well known to capture human perspective-based characteristics, on Mel scale (human auditory scale), as well as linear scale, gives an perfect case, which can be compared with TECC, that captures energy of vocal tract system, and the non-linearity as in speech signal. MFCC has high resolution on low-frequency region, and has low resolution at high-frequency regions. LFCC on other hand, has linear frequency resolution across entire frequency axis. Such characteristics help in covering all necessary aspects for classification of speech signals based on a particular language. We employed static features for both MFCC and LFCC, of variable dimension (soon to be discussed in sub-Section IV-A), and for each feature sets, the frame length was kept to be 30 *ms*, and shift was 15 *ms*. Number of subband filters was taken as 40 (fixed for all the feature sets for fair comparison), and NFFT was taken to be 512.

IV. Experimental Results

A. Effect of Dimension of Feature Vector

The dimension of feature vector is the number of coefficients per audio frame (i.e., speech segment) obtained from speech signal after framing and windowing. As number of coefficients per frame increase, the extracted amount of features per unit file increases, resulting in more storage, space, and time complexity (for both feature extraction and classification task). However, with more amount of data per frames, we can't get more clarity on data, that which data belongs to which class, thereby leading to cutting edge classification of classes. On the other hand, while number of coefficients increase than maximum number of required coefficients, then it starts *over-fitting* the model, leading to declassification of the model. For fair comparison with cepstral features, we also fine-tuned both comparative cepstral features. For MFCC, we obtained optimal feature dimension as 20, whereas for LFCC, and TECC, we obtained optimal feature dimension as 16. We obtained accuracy of 96.37 % for LFCC, and 76.37 % for TECC on ResNet-50 classifier as shown in Fig. 2. All the experiments of this sub-Section were performed using TDNN and ResNet-50 classifiers. Results of TDNN classifier were not as much good as for ResNet classifier, thereby the results of TDNN feature dimention can be found on[1]. We got results of TECC feature vector better on ResNet-50 classifier and hence, we selected ResNet as optimal classifier. Although we are getting much less accuracy on proposed TECC vector, we decided to explore data fusion of features to obtain possible complimentary observations.

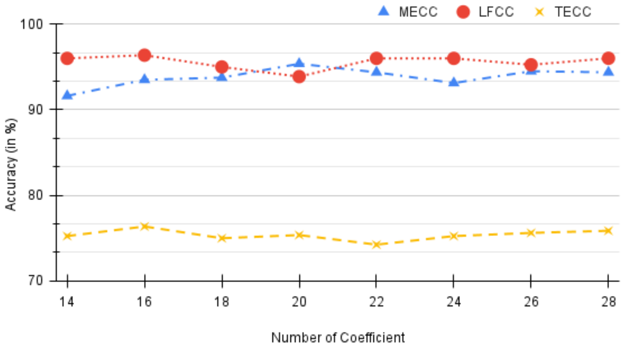

Fig. 2. Effect of dimension of feature vector using ResNet-50 classifier.

B. *Effect of Number of Language Classes*

With increase in number of languages, the complexity of data points increases, i.e., the probability of classifier classifying the data point correctly decreases. With a logistical observation, as the number of class increase, the number of clusters of data points formed increases, resulting in a more degraded classification performance. As the data points that have closely resembling phonemes, vowels may result into unequal classification of languages. Similar observations have been made around 17 years ago with two similar-sounding languages (namely, Hindi and Urdu) with TEO operator itself [19]. Fig. 3 denotes the decrease in accuracy with increase in number of languages. Similar observations can be made for both the comparative cepstral features, as well.

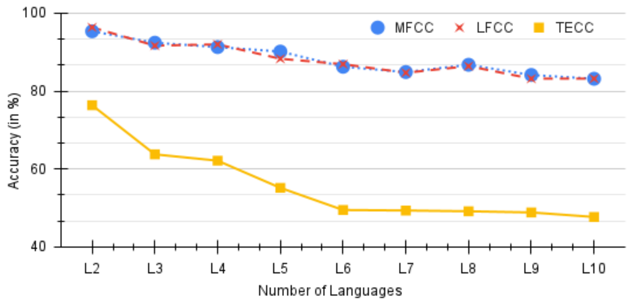

Fig. 3. Variation in accuracy with increase in number of language classes.

C. *Score-Level Fusion*

We examine score-level fusion in order to increase the accuracy of TECC features in this sub-Section. For comparison, we fused all three features, i.e., LFCC+MFCC, LFCC+TECC, and MFCC+TECC on both classifiers. As we obtained better results on ResNet-50 classifier and hence, we did all the experiments of this sub-Section on ResNet-50 classifier. Score-level fusion includes parameter, namely, α. Score-level fusion can be calculated by following formula :

$$Score_{TECC}(\alpha) + Score_{MFCC}(1-\alpha) = Score_{fusion}. \quad (12)$$

We obtained highest of 86.62 % of accuracy for 10 language classes, which is 3.42 % higher than individual optimal accuracy obtained for 10 languages. Fig. 4 denotes score-level fusion on MFCC and LFCC features. Fusion of all other features can be found at[1].

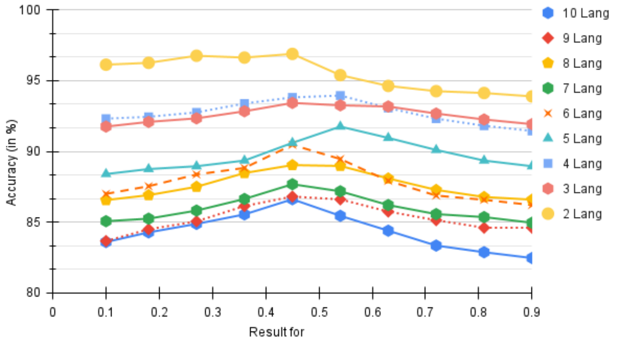

Fig. 4. Score-level fusion of MFCC and LFCC on ResNet-50 classifier.

D. *Feature-Level Fusion*

This sub-Section exploits feature-level fusion after exploring score-level fusion. Motivated by increase in accuracy at higher number of languages on score-level fusion, we explored feature-level fusion in order to examine the change in accuracy. Feature-level fusion can be also as stack fusion, as it is classification of two or more features, by stacking it one over other. We did double fusion as well as triple fusion in this study.

[1]”https://github.com/ARTHARKING55/TDNN_dimention/tree/main”

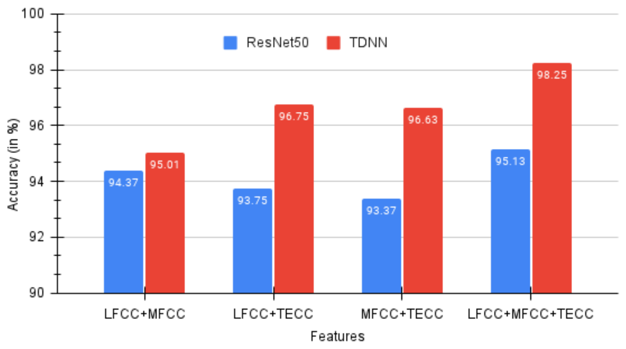

Fig. 5. Feature-level fusion of features on both TDNN and ResNet-50 classifier.

In Fig. 5, an increase in accuracy of $\approx 4\%$ can be observed for two languages as compared to individual accuracies. We can further say from these observations that, combined information of signal, i.e., energy of vocal tract system, nonlinearity in speech signal (TECC), human perception of audio information (MFCC), and linear frequency resolution-based human perception (LFCC), combine perform better than MFCC, LFCC, and TECC alone. Due to limited resources and time, we could not do feature-level fusion of more than two languages (Hindi and English), thereby leaving it to future tasks.

E. Analysis of latency Period

Latency Analysis (LA) of SLID leverages capability of system to its real-time development and usability. LA helps in finding out that in how much less time and storage, we can obtain promising results with acceptable performance. LA thereby helps in saving storage and time both. Moreover, from practical system deployment perspective, latency analysis is one of the key factor for evaluation of any systems' capability. LA have also been performed on SLID task previously in [5]. For fair analysis, we performed latency analysis on both cepstral features also and have reported them in Fig. 6. Observation of Fig. 6, denotes that even after selecting 350 frames instead of 750 frames, we see an accuracy decrease of only 0.62 %, whereas for MFCC and LFCC, accuracy decreased by 1.99 % and 2.25 %. This decrease in frames and minor decrease in results denote the superiority of proposed TECC-based model over the other models in realistic scenarios.

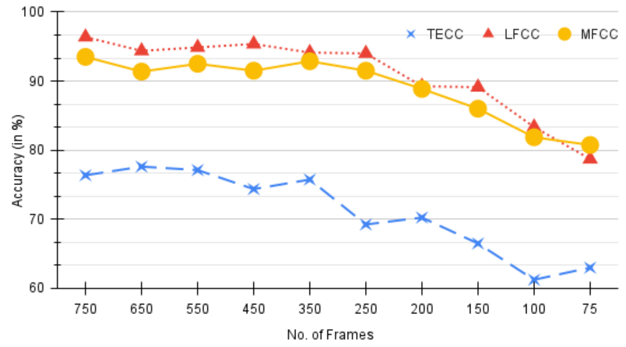

Fig. 6. Analysis of latency period using ResNet-50 classifier.

F. Comparison with Existing Studies

This sub-Section compares the proposed work with existing works on SLID task. Most of studies do not employ full dataset of VoxLingua107, due to limitations of resources and storage. We compared our results with various studies that have variable languages. Table I denotes the comparison of proposed methods with existing methods. For fair comparison, we compared our results with same number of language classes as in existing works.

TABLE I
COMPARISON WITH EXISTING WORKS AND PROPOSED WORK ON DIFFERENT NUMBER OF LANGUAGES.

Source	Number of Languages	Existing Results (in %)		Proposed Results (in %)	
		Accuracy	EER	Accuracy	EER
[20]	2	-	2.26	**98.25**	**1.75**
[20]	3	-	53.27	**92.67**	**5.5**
[21]	4	80.21	-	**92.88**	**4.75**
[22]	6	83	-	**87.54**	**7.47**
[23]	8	70.9	-	**86.72**	**7.58**
[24]	10	-	50	**84.25**	**8.75**
[25]	3	83.50%	-	**92.67**	**5.5**
[26]	5	88.41	7.24	**91.75**	**5.15**
[27]	10	-	11.35	**84.25**	**8.75**
[28]	6	-	8.4	**87.54**	**7.47**

V. SUMMARY AND CONCLUSIONS

In this study, we employed TECC feature vector, which captures vocal tract energy from the audio speech of spoken languages. TECC offers competitive performance compared to traditional MFCC and LFCC features for SLID tasks. ResNet-50 demonstrates superior effectiveness in utilizing TECC features, achieving higher accuracy than TDNN, For better real implementation of model proposed, we also performed various fusion techniques, in which it will enhance the classification accuracy of SLID. In order to compare our study to other existing approaches, in Section IV, we compared our results with existing optimal approaches for SLID, and found to be better.

REFERENCES

[1] D. Deshwal, P. Sangwan, and D. Kumar, "Feature extraction methods in language identification: A survey," *Wireless Personal Communications*, vol. 107, pp. 2071–2103, 2019.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[3] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative study on spoken language identification based on deep learning," in 26^{th} *European Signal Processing Conference (EUSIPCO)*, 2018, Rome, Italy, pp. 2265–2269.

[4] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "I-vector-based prosodic system for language identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, Kyoto, Japan, pp. 4861–4864.

[5] B. S. Hora, K. Parmar, S. Machhar, H. A. Patil, K. Praveen, and B. Radhakrishnan, "Exploring residual cepstral features for spoken language identification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, Taipei, Taiwan, pp. 131–138.

[6] S. S. Sapkota, A. Shakya, and B. Joshi, "Spoken Language Identification Using Convolutional Neural Network In Nepalese Context," in 26^{th} *Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, 2023, Delhi, India, pp. 1–6.

[7] G. Satyanarayana, D. Varun, D. Sandeepika, and B. Uma, "Spoken language detection using deep learning,"

[8] M. Wang, Y. Li, J. Guo, *et al.*, "Whislu: End-to-end spoken language understanding with whisper," in *Proc. INTERSPEECH*, vol. 2023, Dublin, Ireland, 2023, pp. 770–774.

[9] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020, {Last Accessed Date : 2^{nd} July, 2024}.

[10] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification, and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021, {Last Accessed Date : 2^{nd} July, 2024}.

[11] S. Gupta, K. S. S. Motepalli, R. Kumar, V. Narasinga, S. G. Mirishkar, and A. K. Vuppala, "Enhancing language identification in indian context through exploiting learned features with wav2vec2. 0," in *International Conference on Speech and Computer, LNCS*, Springer, vol. 14339, 2023, pp. 503–512.

[12] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[13] V. Tomar and H. A. Patil, "On the development of variable length Teager energy operator (VTEO)," in *INTERSPEECH*, 2008, Brisbane, Australia, pp. 56–59.

[14] H. A. Patil, "Speaker recognition in indian languages: A feature based approach," *Indian Institute of Technology Kharagpur (IIT-K), Ph. D Thesis*, 2005.

[15] B. Koonce and B. Koonce, "ResNet-50," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021.

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015, Dresden, Germany, pp. 3214–3218.

[17] F. R. rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, Alberta, Canada, pp. 5359–5363.

[18] T. Mandava and A. K. Vuppala, "Attention based residual-time delay neural network for indian language identification," in *12^{th} International Conference on Contemporary Computing*, 2019, Noida, India, pp. 1–5.

[19] H. A. Patil and T. Basu, "Cepstral domain Teager energy for identifying perceptually similar languages," in *Pattern Recognition and Machine Intelligence: Second International Conference, PReMI, Kolkata, India, December 18-22, 2007. LNCS, Springer*, pp. 455–462.

[20] H. Kim and J.-S. Park, "Automatic language identification using speech rhythm features for multi-lingual speech recognition," *Applied Sciences*, vol. 10, no. 7, p. 2225, 2020.

[21] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 5 123 671, 2021.

[22] A. Draghici, J. Abeßer, and H. Lukashevich, "A study on spoken language identification using deep neural networks," in *Proceedings of the 15^{th} International Audio Mostly Conference*, 2020, pp. 253–256.

[23] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (lstm) recurrent neural networks," *PloS one*, vol. 11, no. 1, e0146917, 2016.

[24] Z. Ma and H. Yu, "Language identification with deep bottleneck features," *arXiv preprint arXiv:1809.08909, Sep 18*, 2018, {Last Accessed Date : 2^{nd} July, 2024}.

[25] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009, Vancouver, Canada, pp. 1–4.

[26] K. Parmar, B. S. Hora, S. Machhar, H. A. Patil, K. Praveen, and B. Radhakrishnan, "Spoken language identification using linear frequency residual cepstral coefficients," in *International Conference on Pattern Recognition and Machine Intelligence, LNCS*, Springer, 2023, pp. 724–733.

[27] H. S. Das and P. Roy, "Optimal prosodic feature extraction and classification in parametric excitation source information for Indian language identification using neural network based Q-learning algorithm," *International Journal of Speech Technology*, vol. 22, pp. 67–77,

[28] G. Gelly, J.-L. Gauvain, V. B. Le, and A. Messaoudi, "A divide-and-conquer approach for language identification based on recurrent neural networks.," in *INTERSPEECH*, San Francisco, USA, 2016, pp. 3231–3235.