IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 October 2023, accepted 17 October 2023, date of publication 14 November 2023, date of current version 1 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333039

RESEARCH ARTICLE

Enhancing Biometric Speaker Recognition Through MFCC Feature Extraction and Polar Codes for Remote Application

NILASHREE WANKHEDE AND SUSHAMA WAGH, (Senior Member, IEEE)
Electrical Engineering Department (EED), Veermata Jijabai Technological Institute (VJTI), Mumbai 400019, India

Corresponding authors: Nilashree Wankhede (nswankhede_p20@el.vjti.ac.in) and Sushama Wagh (srwagh@ee.vjti.ac.in)

This work was supported by the All India Council for Technical Education (AICTE) under Quality Improvement Program (QIP) for the Ph.D. degree with the Veermata Jijabai Technological Institute (VJTI), Mumbai University.

ABSTRACT While extensive research has been conducted in the field of biometrics, particularly in face and fingerprint recognition, remote speaker recognition has yet to gain global acceptance due to challenges related to accuracy and data integrity. Previous studies in speaker recognition have explored techniques such as Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN), yielding accuracy rates of 90.4% and 92.8%, respectively over a fixed and small database with a standalone system. To address the data integrity and accuracy issues for enhancement in remote speaker recognition, a novel approach is proposed in this paper. Initially, remote speaker recognition is implemented using a client-server setup, but the presence of channel noise hindered any noticeable improvement in accuracy compared to existing methods. The new approach involves extracting MFCC parameters from voice samples and subsequently applying polar error-correcting coding techniques for storage as well as transmission to achieve fidelity. Using a code rate of 1/2 and a block length of 1024 bits, the transmission of polar-coded MFCC features over a noisy channel yielded a lower bit error rate when coupled with successive list decoding. Simulation results demonstrate a reduction in bit error rate, resulting in an accuracy of 95.2% in the implemented remote speaker recognition system. This represents a significant 5% improvement over the existing standalone system that uses uncoded MFCC features. These findings highlight that the Polar codes can be effectively utilized in speaker recognition systems to enhance their robustness and reliability, especially in scenarios with noisy channels or challenging conditions.

INDEX TERMS Biometric, bit error rate, mel-frequency cepstral coefficient, polar codes, recognition rate, speaker recognition.

I. INTRODUCTION

With the biometric authentication as a remote login procedure, many important tasks like telephone banking, authoritative access and other major enterprises are keen to introduce the technology for their employees or customers who can log into workplace systems through internet networks anytime with ease and can access restricted areas or files or mark their remote presence using a client server model [1]. Also, in the field of Internet of Things (IoT), where remote device monitoring and control using Internet and cloud usage is involved, applications requiring biometric verification can be used with ease if sufficient research is done on keeping the biometric features intact over Ethernet or Wi-Fi connections. When user authentication is required at a remote place and it is difficult for the user to access the biometric device like a fingerprint machine or a camera, voice recognition [2] and authentication can help in a great way in providing remote access. For voice biometrics, the user need not be present physically and can authenticate themselves using just a mobile device with an inbuilt mike. According to a poll by a security enterprise named

The associate editor coordinating the review of this manuscript and approving it for publication was Wen-Sheng Zhao.

Veridium [3], around 70% of consumer base felt the need of biometric authentication into the workplace, the primary reason being not having to remember passwords. About 40 % of the organizations have been using fingerprint reader technology for attendance monitoring and verification procedures. According to Unisys report [4], a security solution enterprise, their survey in 2018 revealed that the biometric technologies ranked from high to low by consumer preference are: voice recognition, fingerprint scan, facial scan, hand geometry and iris scan respectively. Many companies related to security and solutions are known to have been working on voice biometric solutions and related recognition and verification systems in order to enhance the overall speed and security in authentication systems at their respective sites.

In recent speaker recognition related implementations [5], [6], [7], voice biometrics exhibit a lower recognition accuracy, which underscores the need for supplementary authentication methods rather than full reliance on voice biometrics. Though the voice features can be considered as a unique characteristic of an individual, they need to be used along with a multilevel verification system. For example, the existing biometric authentication systems such as face or fingerprint recognition used in some applications could be supported by speaker recognition feature. Voice recognition, when it co-exists with face or fingerprint recognition system will help to provide multiple level security [7] to any authentication systems and thus the accuracy and integrity of voice features is one of the research areas to work upon.

Various organizations have adopted biometric authentication to streamline customer account access, thereby replacing traditional methods such as passwords or requiring in-person presence for account or locker operations. For example, the iris recognition was implemented towards mobile banking by the Bank of America, and the British bank and Wells Fargo were trying to implement voice recognition in the year 2017 [9] but had to face issues of low accuracy and spoof attacks. In 2013, biometric identification with the iPhone fingerprint sensor was used by Apple iPhone. Voice biometric authentication and verification would be helpful to customers in a way that they need not remember their passwords or pin numbers. A customer's real-time voice creates a distinct and protected identity that remains impervious to theft or misuse by unauthorized individuals [10]. It can turn out to be the cheapest among all other biometric authentication means, as it does not need any readers or special devices when compared to fingerprint or iris or any other biometric tool. Efforts are been made to provide a multilevel-security system for authentication process in addition to fingerprint or iris or face detection and verification process in military, security and confidential areas. Multi-Factor Authentication (MFA) is a unique multi-layer approach customizable to each organization's security requirements [1], [11].

Remote login authentication for all in telephone banking and in bank transactions for server access is another challenge. Remote on-field attendance monitoring systems and forensic investigation can also be studied and implemented. Though fingerprint, iris and face recognition applications are already in place, a lot of research work is required in the area of speaker recognition to make voice biometric as one of the parameters in MFA process with increased reliability over noisy channel [12].

Some banking systems have tried to incorporate authentication systems where the user needs to report to the bank and provide his or her voice signatures to operate the account. However, the adoption of voice authentication for online transactions and remote login procedures has been delayed due to several security-related considerations and as observed in [8]. Also, integrity of the voice signature is very important to differentiate one voice sample from other within or beyond the databases maintained by the authentication systems. For the 4G long term evolution cellular systems, the Turbo code was selected to provide channel coding for mobile broad band data. However, the 3GPP standardization group, after a careful analysis for 5G new radio [13], replaced the Turbo code by Low density parity check code (LDPC) and the Polar code as in [14].

This research work presents the Mel-Frequency Cepstral Coefficients (MFCC) extraction [15] done towards voice feature matching and transmitting the features over a noisy channel. Initially Remote speaker recognition was implemented considering a client-server scenario, and no improvement in accuracy was observed over existing methods. This reduction in accuracy is observed due to channel noise in remote recognition scenario. It has been proposed to use the Polar coding technique which is one of the recently developed error correcting codes as it will aid in keeping the MFCC features intact over an AWGN channel.

The key contributions of the research work are as follows:

1) The Polar encoding technique applied on the MFCC features as a part of remote speaker authentication system.
2) Comparative analysis in terms of bit error rate (BER) for Polar encoded voice features with that of uncoded ones.
3) Comparative analysis of frame error rates (FER) obtained using successive cancellation and successive cancellation list decoding of MFCC coefficients, with block length of $N = 1024$ and code rate 1/2.
4) Comparing the accuracy obtained in speaker recognition with uncoded MFCC coefficients with that of Polar coded coefficients used for feature matching at the receiver end.

The implementation suggests that as wireless multimedia communication continues to advance, leveraging Polar codes in remote multimodal authentication applications can thrive. This approach supports seamless and noise-free transmission and reception of MFCC vectors, ensuring precise customer identification through their voice biometric feature.

The rest of the paper is prepared as follows: Section II provides a discussion of speaker recognition and describes

the MFCC parameter extraction procedure, followed by section III which briefly introduces the Polar codes. Section IV provides proposed methodology and Section V presents the comparative results of BER performance analysis obtained from Polar coding technique carried out on the speaker related extracted MFCC parameters, Section V provides the recognition accuracy calculated after the implementation with coded MFCC features followed by conclusions derived from this research work in section VII.

Abbreviations and Acronyms: The abbreviations used throughout the article are listed in Table 1.

TABLE 1. Lists of abbreviations.

Acronym	Definition
ASR	Automatic Speaker Recognition System
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
CNN	Convolutional Neural Networks
CRC	Cyclic Redundancy Check
Eb	Signal Energy Per Bit
FEC	Forward Error Correction
FER	Frame Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
FER	Frame Error Rate
LPC	Linear Predictive Coding
MFA	Multi-Factor Authentication
MFCC	Mel- Frequency Cepstral Coefficient
No	Noise Power Spectral Density
RR	Recognition Rate
SC	Successive Cancellation Decoding
SCL	Successive Cancellation List
VQ	Vector Quantization
3GPP	3rd Generation Partnership Project
5G	Fifth Generation

II. LITERATURE REVIEW ON REMOTE SPEAKER RECOGNITION AND MFCC FEATURE EXTRACTION FROM VOICE

Every individual has a unique anatomical structure of vocal tract and hence to differentiate between various speakers and identify them, extraction of certain voice parameters is required [8], [10]. Remote authentication of a person using his or her voice can be an application in todays wired and wireless communication systems where the speaker need not be physically present in front of an authentication system. Speaker recognition has two aspects: speaker verification and speaker identification. The speaker identification helps in finding an individual person from the given group of speakers. If N is the number of speakers, then the input speaker coefficients are compared with all N speaker's coefficients stored in the database whereas the speaker verification is a different technique. Based on the narrowed down matching, the verification is a fast procedure which authenticates a person by either accepting or rejecting him.

The implementation procedure for an Automatic Speaker Recognition System (ASR) as seen from the Figure 1 has

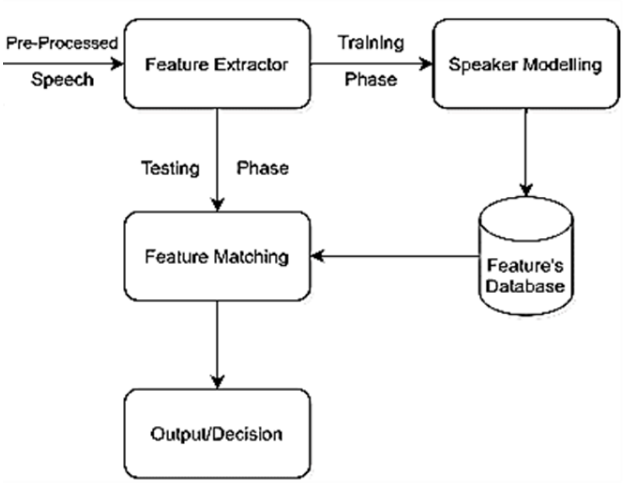

FIGURE 1. Structure of an automatic speaker recognition system.

been discussed with extraction of physiological-based features from an individual's speech.

Speaker identification and speaker verification both can either be implemented as being text-dependent to keep simple algorithm or the check points could be made more complex by running a text-independent authentication system [20]. A particular text is to be spoken and corresponding feature vectors are stored in a text-dependent approach. During matching phase, the speaker has to speak the same text. But, if text-independent implementation is used, the speaker can randomly speak anything in real time and feature vectors are extracted from his or her random spoken content. The recognition rate is better in text dependent than text independent techniques as mentioned in [25] and [26] for voice authentication process, the reason being there is not much randomness in feature extraction process if a particular sentence or word has specific pattern, and there is less complexity in matching process. Accuracy and complexity of a speaker recognition system varies with a closed or open system as stated in [15]. Voice information can be stored in the database as training set when limited number of speakers form the closed loop system. In contrast, an open system can accommodate a multitude of speakers, regardless of whether they are included in the authentication system's registration list. There are various techniques used for feature extraction [2] which include use of Pitch extraction, Formant extraction, Linear Predictive Coding, Mel Frequency Cepstral Coefficient and Perceptual Linear Predictive Coefficients. MFCC feature extraction [17] is the process of extracting the voice features from the voice sample while doing frequency domain analysis. In training phase, each registered speaker has to provide samples of their voice for generating a reference model for that speaker. Mel-frequency cepstral coefficient extraction includes the pre-emphasis step, framing, windowing, fast Fourier transform, Mel-frequency filter bank and Direct cosine transform [17]. Voice sample is pre-emphasized and passed through high pass filter due to the requirement

of the voiced section of the voice signal which falls off at high frequencies. The framing helps in to take constant signal for short time span. To minimize the signal discontinuities, hamming window has been used in the implementation as it provides better frequency resolution. Equation (1) shows the expression used for windowing where, $u(m)$ is the voice signal input, $W(m)$ is the hamming window, $v(m)$ is the signal output and the vale m has to be less than (M-1) with M as number of samples as given as in (2).

$$v(m) = u(m) * W(m) \tag{1}$$

$$w(m) = 0.54 - 0.46\cos(2\pi m/(M-1)), 0 \leq n \leq N \tag{2}$$

The Fast Fourier transform procedure is used to convert the time domain frame of M samples to frequency domain and hence for each frame the information about magnitudes in the frequency response is obtained. The Mel-frequency-cepstrum is the representation of short-term power spectrum of sound, based on linear cosine transform of a log power spectrum on a nonlinear Mel scale frequency. The Mel-filter bank is used to filter an input power spectrum and length of output array is the number of filters created. The output of this fast Fourier transform is multiplied by a set of triangular band-pass filters to get log energy of each triangular band-pass filter. Discrete cosine Transform (DCT) is the final step in which the Mel spectrum coefficients are converted back to time domain. Generally, there are 13 MFCC features used for extraction as mentioned in [17], but this implementation has considered 26 more related to the first order derivative and second order derivatives of the MFCC extracted features. First and second order derivatives are calculated by taking the difference of the MFCC coefficients between the samples of the audio signal and it will help in understanding how the transition is occurring. The extraction of the cepstrum via the Inverse DFT from the previous section results in 12 cepstral coefficients for each frame. A thirteenth feature: the energy from the frame correlates with phone identity and it represents the sum over time of the power of the samples in the frame. No two consecutive frames are similar in a speech signal and hence a small change, such as, the nature of the change from a stop closure or the slope of a formant at its transitions can provide a useful cue for phone identity.

Thus, MFCC parameters are obtained for every speaker from their voice signal. The MFCC parameters vary from human to human for the same spoken content and can be considered as unique characteristics. These MFCC parameters can be transmitted or processed ahead to match with the existing databases. For a speaker identification mechanism, the samples stored in an already existing database are matched with the features obtained from the real-time input audio sample. The advantage of using MFCC parameters as feature vectors, is that, the reduction in the number of bits produced corresponding to the spoken voice signal and providing reduced message for transmission over channel. Feature matching includes various techniques such as mean square error, hidden Markov model, Vector quantization (VQ), dynamic time wrapping, Gaussian mixture model and artificial neural networks. For the feature matching purpose VQ model is used, where this technique is mostly used for text dependent systems [8], [19]. VQ method uses centroiding for classifying a set of feature vectors per speaker and its popular method used in many applications such as voice recognition, lossy data compression which includes voice and image compression. The feature vectors extracted from each speaker based on procedure of MFCC can be considered as code-words and each code-word is used to construct a code-book for each speaker, who is the part of enrollment procedure. In the speaker recognition the real time voice parameters are compared with the codebook of each speaker and the differences are calculated for which the Linde-Buzo-Gray (LBG) algorithm [19] as a part of vector quantization has been implemented over text independent system. In addition to MFCC, the inclusion of vocal tract parameters aids in the computation of LPC coefficients as voice features. Each of these features contributes to capturing distinct aspects of vocal characteristics. Post the speaker recognition implementation on a set of speakers overall performance evaluation is done to get accuracy. The recognition ratio (RR), the False Acceptance Ratio (FAR), the False Rejection Ratio(FRR) and Equal Error Rate(EER), determine the overall performance of any speaker recognition system. The recognition ratio is a measure of the system's ability to correctly identify or authenticate genuine users. It represents the percentage of genuine users who are correctly recognized or accepted by the system. A higher recognition ratio indicates better performance. FAR is a measure of the system's vulnerability to false acceptance. It represents the percentage of unauthorized or impostor attempts that are incorrectly accepted as genuine by the system. FRR is a measure of the system's likelihood to falsely reject genuine users. It represents the percentage of legitimate access attempts that are incorrectly rejected by the system. The EER is a specific point where the FAR and FRR are equal. It is a crucial metric as it provides a balanced assessment of the system's performance. At the EER threshold, the system is making an equal number of false acceptances and false rejections. Lower EER values indicate better overall system performance. In essence, these metrics help evaluate the trade-off between security and convenience in biometric authentication systems. A good system aims for a high recognition ratio, a low FAR, a low FRR, and a low EER, striking a balance between security and user convenience.

III. LITERATURE REVIEW ON POLAR CODING

This section presents the discussion on Polar coding technique towards keeping the voice features extracted from speakers to be intact over an AWGN channel. Polar codes were originally defined by the researcher Arikan as in [23] and it was suggested that for a given binary discrete memoryless channel, the Polar codes do achieve channel capacity. The channels can be categorized into good and bad by channel polarization and based on the reliability sequences, the information bits are transmitted over good channels, whereas

the frozen bits are transmitted over bad channels. 3GPP standards define these reliability sequences for different code lengths [21]. It has been mentioned in literature [22], that for long block length latency issue exists but still for a finite block length the polar codes are practically feasible.

Polar codes (n, k) are one of the linear block codes with code rate k/n, k being number of message bits and 'n' the codeword bits. Mathematically, as in (3), the input vector information and G_N the matrix generator [21] can be multiplied as

$$y = xG_N \tag{3}$$

where $x = (x_1, x_2 \ldots x_{N-1})$ denotes the input vector.

Let F denote the Kronecker product, then the matrix generator can be represented by (4) where, $F^{\otimes n}$ denotes the n^{th} tensor power of F and Π denotes the permutation matrix known as bit reversal.

$$G_N = \Pi F^{\otimes n} \tag{4}$$

Equation (5) is the Kronecker polarizing matrix which is the base for the Polar encoding in this implementation. The polarization effect introduced by the polar codes allows dividing the 'n-bit' input vector 'x' as either reliable or unreliable bit-channels. The 'k' information bits are assigned to the most reliable bit channels of 'x', while the remaining 'n-k' bits, called frozen bits, are set to a predefined value which is taken as '0'.

$$F = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{5}$$

Frozen bits are assigned to the most unreliable bit-channel. Codeword 'y' is transmitted through the channel, and the decoder receives the output sequence $r = (r0, r1, \ldots, r\ n\text{-}1)$ which is the noisy version of $y = (y0, y1, \ldots, y\ n\text{-}1)$. For decoding of these Polar codes, two different methods, i.e., Successive cancellation decoding(SC) and Successive cancellation list (SCL) decoding, whose comparative analysis and algorithms are explained in the implementation section V of this paper.

IV. PROPOSED METHOD OF USING POLAR CODING ON VOICE BIOMETRIC MFCC FEATURES

The proposed implementation as seen from the given blocks in Figure 2 has been discussed in this paper. Input voice samples with either text dependent or independent scenario can be processed to obtain the 13 MFCC coefficients as described earlier in section II.

In this implementation, text independent scenario is used. For example, a particular voice sample from an utterance of 3 seconds, sampled at 22 KHz, consisting of around 66150 bits after digitization, was analyzed. After calculating the MFCC features, it was found that around 4368 bits are obtained for transmission from 13 MFCC features for that voice signal. Hence a 93.93 % of reduction of the number of bits was obtained. The MFCC coefficients acquired from each voice sample are subsequently organized into feature

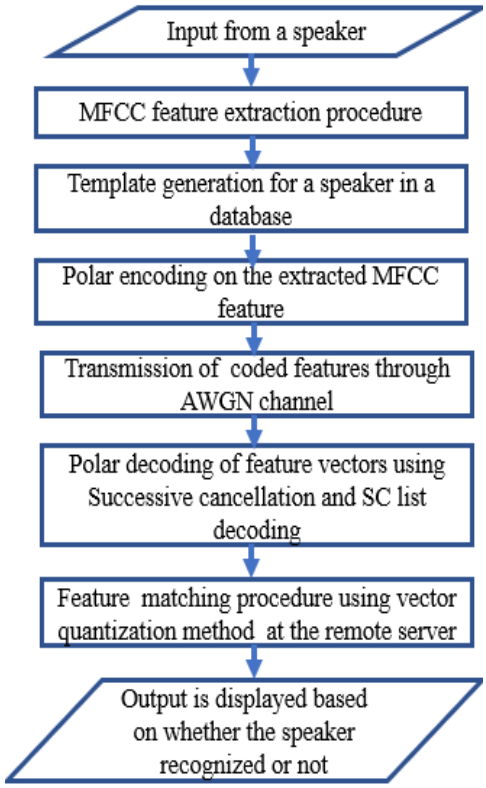

FIGURE 2. Implementation procedure for obtaining the integrity of MFCC coefficients towards speaker Recognition.

vectors specific to individual speakers. These feature vectors are given distinct labels, effectively serving as a ''template'' for each speaker. The feature matching procedure requires enrollment procedure, creation of database and vector matching techniques. The research work presented here highlights the procedure used for checking the integrity of the extracted MFCC parameters over AWGN channel using Forward Error Correcting Codes (FEC) currently deployed in 5G applications such as Polar codes. These implementation results can be further used towards the remote speaker recognition or voice as biometric identity for authentication to compare with other FEC's as well.

V. IMPLEMENTATION RESULTS OF POLAR CODING ON MFCC PARAMETERS OBTAINED FROM SPEAKERS

This section gives insight of the implementation results using polar coding for the integrity of MFCC coefficients. Figure 3 shows the snapshot for the MFCCs extracted for three out of the twenty one speakers. These 13 coefficients serve as feature vector for every individual speaker and used for further processing and encoding before transmission. The polar encoding technique used is as given in section III of this paper. For decoding two different methods have been used and comparative analysis is performed. This research work is about application of polar codes to speaker authentication method, employing both successive cancellation decoding

and successive-cancellation list decoding of MFCC coefficients. Our goal is to maintain the integrity of voice features for every speaker enrolled in the remote authentication database and get accuracy of the speaker recognition system. Below is the Successive cancellation algorithm used for polar decoding in this implementation and the Figure 3 gives the basic path flow of decoding. The output of the algorithm is the decoded vector for a given length N of the codeword. The likelihood ratios are calculated and based "0" bit, "1" bit or "frozen" bit based on known reliability sequences used in Polar encoding according to 3GPP as discussed in [24].

Name	mfccs.shape	MFCC coefficients	Delta_MFCC	Delta2_MFCC
Anupama	(13, 346)	[-425.88885 - 396.01346 - 394.73492 - 406.59872 - 398.12372 - 398.93362 -399.39517 - 397.82343 - 399.95346 - 396.09269 - 393.28882 - 394.77408 -403.39786]	[1.610938 1.610938 1.610938 1.610938 1.610938 0.00511068 0.57922107 1.115448 0.08875885 - 0.21439157 - 0.6068644 - 1.0621948 -1.0970317]	[-1.4521168 - 1.4521168 - 1.4521168 - 1.4521168 - 1.4521168 0.4900091 0.5062469 - 0.15740372 - 0.3494439 - 0.6653965 - 0.75978136 - 0.6191632 -0.1750759]
Joshanu	(13, 155)	[-435.74057 - 359.16678 - 320.66565 - 323.53644 - 310.2396 - 291.41617 -297.76868 - 297.497 - 292.02103 - 295.58633 - 291.55478 - 288.68082 -296.0257]	[13.963362 13.963362 13.963362 13.963362 13.963362 6.746757 3.8441682 3.2147374 1.3233353 0.16851705 0.20028839 0.18626505 1.013058]	[-7.288984 - 7.288984 - 7.288984 - 7.288984 - 7.288984 - 3.2232952 -1.4957768 - 1.4767501 - 0.862948 - 0.10052794 - 0.6905175 - 0.22096853 1.1260024]
Karmesh	(13, 273)	[-590.5787 - 576.0367 - 574.4388 - 575.09625 - 568.8631 - 569.3822 -573.48334 - 572.474 - 571.35913 - 570.5947 - 568.8857 - 563.1838 -560.54785]	[1.5865194 1.5865194 1.5865194 1.5865194 1.5865194 0.52718407 0.46055195 0.78802186 1.0488474 1.2662923 1.2322682 0.81141865 0.48025513]	[-1.205759 - 1.205759 - 1.205759 - 1.205759 - 1.205759 - 0.2990679 -0.06476583 0.42323613 0.9671746 0.36908486 - 0.44994813 - 0.756598 -0.7280067]

FIGURE 3. Snapshot of the MFCC coefficients obtained from 3 out of 21 speakers in the database.

Implemented Successive Cancellation Decoding Algorithm

```
// Y: output vector
//cal: likelihood ratio
// O: decoded vector
//n: length of Y
For i =1...n
  1 If Y[i] is frozen bit
      O[i] ← Y[i]
  2 else
    If (cal (i) > = 1)
      O[i] ← 0;
  3 Else
      O[i] ← 1
```

The polar encoding technique can be employed with different code rates by proper selection of N (block length) and K (message bits). So, keeping the code rate same and varying the input nits and corresponding codeword bits, the frame error rate was calculated for combinations of N and K as shown in the below simulation results. The main reason of not increasing the code rate to some other value was that the feature vectors specific to speakers are considered to be finite and of small dimensions. The increase in code rate will increase the payload at the subsequent data link layer and thus may introduce latency in the authentication procedure. For higher values of N, that is,for indirectly higher block lengths chosen, the decoding performance using Successive cancellation decoding has been plotted.

From Figure 4, it is observed that as N increases to 512 or 1024 for a given fixed code rate chosen here as 1/2, the feature vectors were showing more integrity at the decoding stage, than while using the N = 32 or 64 as the BER performance was observed to be better with increasing value of N. Table 2 summarizes the findings for the BER values obtained at various Eb/No, N and K values as specified.

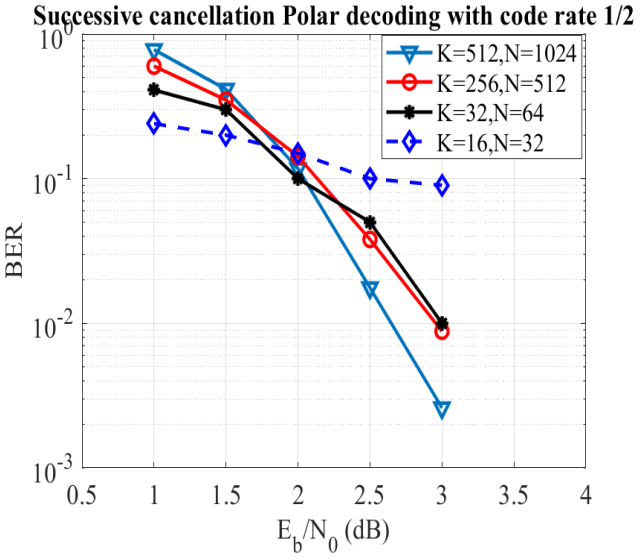

FIGURE 4. Comparative BER performance of Polar codes used for MFCC coefficients using SC coding with variable (N, K), code rate 1/2.

Table 2 provides the frame error rate values obtained at Eb/No of 2.5 dB. The FER for a scenario of uncoded MFCC coefficients is obtained at the receiver side and its quite high

TABLE 2. BER analysis of Polar Coding, code rate =1/2, N=1024.

Eb/No in dB	BER values for N=1024, K=512	BER values for N=512, K=256	BER values for N=64, K=32	BER values for N=32, K=16
1	7.80 e^{-1}	6.0 e^{-1}	4.1 e-1	2.4 e-1
1.5	4.15 e^{-1}	3.5 e^{-1}	3.0 e-1	2.0 e-1
2	1.17 e^{-1}	1.5 e^{-1}	0.1 e-1	1.5 e-1
2.5	1.76 e^{-3}	3.8 e^{-2}	5.0 e-2	1.0 e-2
3	2.60 e^{-3}	8.8 e^{-3}	1.0 e-2	9.0 e-2

up to 7×10^{-1}. With Polar encoding done and then transmission and decoding of these MFCC coefficients the FER values are quite lower up to 2.16×10^{-4}. With the code rate of 1/2 used in every case, for higher code lengths the Successive cancellation decoding algorithm works optimally, but when looked into the FER values obtained, its error-correction performance not effective for short code lengths. Also, it suffers from the drawback of latency in decoding and if there is a decision error in one of the estimated vectors, the error propagates for all estimated vectors.

Therefore, in the context of employing the polar coding technique alongside the SC decoding method, while maintaining a consistent code rate (K/N) across the entire process, it has been noted that, at specific Eb/No levels, opting for larger values of N and K leads to lower bit error rates. Additionally, with an increase in Eb/No, the bit error rate decreases for all selected values of N and K. Consequently, the values N=1024 and K=512 have been set as constants for all subsequent implementations. It is important to note that the complexity of the decoder escalates with higher N values, thus necessitating a careful consideration of computational time.

Implemented Successive Cancellation List Decoding Algorithm:
// l < L max and Lmax Is fixed
1 SCL with l
2 IF CRC is verified, the codeword is the most probable
Else
3 IF (2∗Lo< Lmax)
4 then calculate L=2∗Lo
5 Go to 1
Else
6 consider codeword is the most probable

In order to investigate the performance and reliability of polar codes, Successive Cancellation List decoding algorithm was introduced by Tal et al in [20] and it was used with optimum Lo value. While implementing this method, two bits, bit 0 and bit 1 are generated in each iteration of decoding and finally Lo most probable best sequences are used. This decoder can also use Cyclic Redundancy Check (CRC) algorithm. For the verification purpose, highest probability codeword is chosen. If CRC is not verified, the value of Lo is doubled and the CRC aided SCL decoding are repeated. Figure 5 displays the comparative results for polar coding technique used with decoding done by two different methods. As shown the after 20 iterations the FER performance is better for SC list decoding method with N=1024 and K=512 bits per frame, than the SC decoding method whereas Figure 6 gives similar results as that of FER for BER performance with coding gain higher in SC list decoding.

It has been observed that the FER performance with N = 1024 bits is better in terms of lower FER with increase in Eb/No values for Successive cancellation list decoding method rather than for only SC decoding procedure. Though

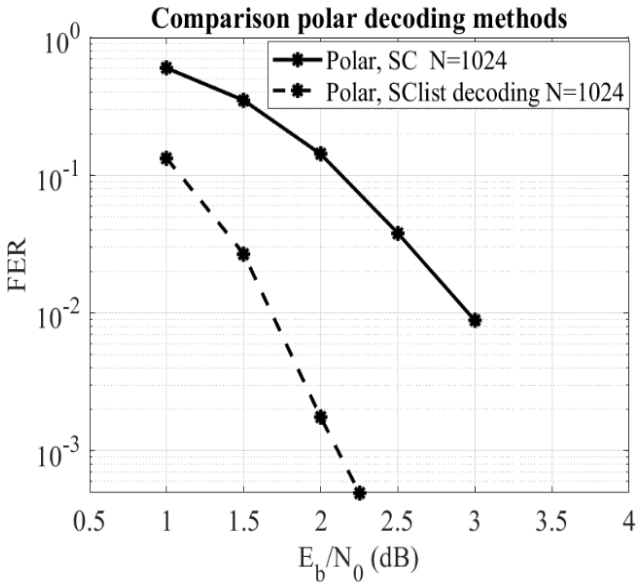

FIGURE 5. Comparative FER performance of Polar codes used for MFCC coefficients using SC and SCL decoding methods at code rate of 1/2.

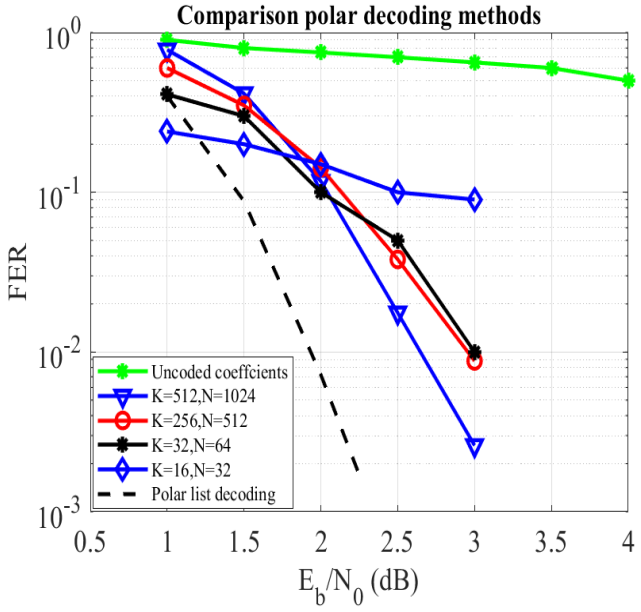

FIGURE 6. Comparative FER performance for all the Polar decoding methods used for coded MFCC versus uncoded MFCC coefficients.

the complexity of SCL algorithm relies on CRC calculations and double decoding and CRC checks, the FER performance curves at code rate of 1/2 are indicating that SC list decoding can be the best decoding method as for now in applications involving forward error correcting polar codes.

Table 3 illustrates the compatibility of the SC list decoding method with a speaker recognition system, demonstrating its capability over SC method to accurately retrieve the original MFCC coefficients even when transmitted through a noisy channel.

TABLE 3. FER analysis of Polar Coding, code rate = 1/2 N=1024 at Eb/No = 2.5 b.

Channel Coding Method	Decoding method	FER values
Using the uncoded MFCC coefficients	No Encoding used	$7\ X10^{-1}$
Proposed method of Polar Coding the MFCC's	Successive cancellation decoding	$1.76\ X10^{-2}$
Proposed method of Polar Coding the MFCC's	SC list decoding	$2.16X10^{-4}$

VI. ACCURACY FINDINGS FOR THE PROPOSED SPEAKER RECOGNITION METHOD

The method used for calculating accuracy or overall performance can be applied to many applications such as voice recognition, lossy data compression which includes voice and image compression. As discussed in Section II, FAR is the measure of the system's vulnerability to false acceptance and FRR the measure of the system's likelihood to falsely reject genuine users is evaluated in this implementation. A lower FAR is desirable because it means the system is less likely to mistakenly grant access to unauthorized users. A lower FRR is preferred because it means the system is less likely to deny access to authorized users. The threshold value for classification can be determined by estimating the score distribution where EER minimizes to a value point where both FAR and FRR are equal. Accuracy is defined as the proportion of correctly verified speakers among the total number of enrolled speakers within a speaker authentication system. In the context of a locally created database in this research work, comprising 21 speakers, Table 4 presents the obtained accuracy percentages, reflecting the success rates of speaker authentication for these 21 individuals.

The approach adopted was text-independent, where each speaker produced a random short sentence for a duration of 3 seconds. The feature vectors were derived from the 13 MFCC coefficients extracted from the 3-second voice samples. Results obtained are compared with uncoded MFCC scenarios with those results provided by researchers in [25] and [26]. As shown in Table 2, the comparison of accuracy percentages in terms of number of speakers correctly recognized out of the database of 21 speakers, around 95.2 % accuracy was obtained when polar coded MFCC parameters used for voice authentication. Speaker verification without the channel coding being used gives accuracies percentages of about 80 to 90 % [25]. When CNN is used along with uncoded MFCC, recognition accuracy increases above 92 % as shown in the comparative analysis in [26]. Thus, the accuracy percentage obtained as shown in Table 2 is indicative that using Polar codes, the MFCC coefficients obtained across a noisy channel are still useful and able to provide successful voice authentication for a text independent scenario which has been implemented.

TABLE 4. Accuracy percentages for successful speaker authentication using uncoded and coded MFCC coefficients.

Speaker recognition Method	System type	Coded /Uncoded feature vectors	Accuracy percentage
Feature extraction using MFCC [25]	Standalone	Uncoded MFCCs	90.4 %
Feature extraction MFCC [26] and trained using CNN	Standalone	Uncoded MFCCs	92.8 %
Feature extraction using MFCC and use of MFCCs for remote speaker authentication	Remote	Uncoded MFCCs	60 to 80%
Feature extraction using MFCC and proposed use of Polar coded MFCCs for remote authentication	Remote	Polar coded MFCCs	95.2 %

The database consists of 21 distinct speakers that includes both male and female speakers. Sound files are stored beforehand corresponding to 21 speakers. Testing is done in real time for the speaker recognition in a normal environment. Recognition rate of the trained VQ codebook model is defined by (6), where, RR is the recognition rate, N *correct* is the number of correct recognitions of testing speech samples per digit, and N *total* is the total number of testing speech samples.

$$RR = (\mathrm{N}total/\mathrm{N}correct)\mathrm{X}100 \tag{6}$$

Results obtained are thus compared with those recognition systems which use 13 MFCC vectors as provided by researchers in speaker recognition domain as given in [25] and [26]. The FAR and FRR values were found to be 0.09 and 0.19 respectively. In cases where a speaker recognition attempt initially fails, it typically takes a maximum of three subsequent attempts to achieve a successful recognition. High value of FAR is related with security whereas FRR is related to convenience of the end user. Ideally these values should be lower to achieve at a minimum equilibrium point and recognition rate or accuracy of 95.2% is obtained for the implemented MFCC based remote speaker recognition system.

In our study, we found that the inclusion of Polar coding in the MFCC transmission process resulted in an average total authentication time of 4 seconds. This represents a 2-second increase compared to the use of uncoded MFCCs, attributable to the Polar decoding procedure. Given our emphasis on accuracy percentage, we are keen to address the need for reducing computational time and complexity, which will be a key aspect to be further investigated.

VII. CONCLUSION

As a part of conclusions derive from this research work, the primary objective was met to ensure the fidelity of MFCC parameters extracted from voice signals originating from diverse speakers. To achieve this goal, the performance of polar coded MFCC coefficients in contrast to their unencoded counterparts was meticulously examined and compared in terms of Bit error rate. Polar encoded MFCC coefficients sent to a distant system, with a code rate of 1/2, employing a block length of 1024, were extracted back using the successive list decoding method. To evaluate the effectiveness of the approach used, accuracy has been calculated for the speaker recognition system using a modest database consisting of 21 speakers and vector quantization for feature matching procedure. This allowed to benchmark the accuracy rate obtained through the implemented system against already existing research in speaker recognition that did not employ Polar codes. The results of this research, which utilized text-independent speech and MFCC coefficients derived from speech signals, demonstrated a significant improvement in remote speaker recognition accuracy, reaching 95.2%. Moreover, the incorporation of polar coded MFCC coefficients into the authentication process led to impressive performance metrics, with a False Acceptance Rate (FAR) of 0.09 and a False Rejection Rate (FRR) of 0.19. The reduction in bit error rates achieved through the use of Polar coded MFCC coefficients translated directly into improved recognition rates for remote speaker authentication. Hence, in the context of remote voice biometric authentication, exploring additional contemporary forward error-correcting codes within noisy channel environments emerges as a promising avenue for enhancing the accuracy and dependability of authentication systems.

ACKNOWLEDGMENT

Nilashree Wankhede would like to thank VJTI, her QIP Institute, for fostering an excellent research environment, facilitating the successful execution of the research work, also would like to thank the Fr. C. Rodrigues Institute of Technology, her parent Institute, for providing a valuable opportunity to engage in research activities and contribute to professional growth, and also would like to thank the doctoral colleague, Neha Septa, for her valuable inputs in preparing this manuscript.

REFERENCES

[1] R. Ryu, S. Yeom, S.-H. Kim, and D. Herbert, "Continuous multimodal biometric authentication schemes: A systematic review," *IEEE Access*, vol. 9, pp. 34541–34557, 2021, doi: 10.1109/ACCESS.2021.3061589.

[2] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.

[3] Veridium Enterprise. (2019). *How Your Biometric Data is Different From Your Password—Veridium*. [Online]. Available: https://veridiumid.com/case-studies/?_ga=2.182866598.1698325735.1656058419-1850179022.1656058419

[4] C. Burt. (2019). *More Than 4 in 5 Americans Support Airport Biometrics, Unisys Survey Shows—Biometric Update*. [Online]. Available: https://www.biometricupdate.com/201906/more-than-4-in-5-americans-support-airport-biometrics-unisys-survey-shows

[5] X. Mu and C.-H. Min, "MFCC as features for speaker classification using machine learning," in *Proc. IEEE World AI IoT Congr. (AIIoT)*, Seattle, WA, USA, Jun. 2023, pp. 566–570, doi: 10.1109/AIIoT58121.2023.10174566.

[6] A. Sedik, L. Tawalbeh, M. Hammad, A. A. A. El-Latif, G. M. El-Banby, A. A. M. Khalaf, F. E. A. El-Samie, and A. M. Iliyasu, "Deep learning modalities for biometric alteration detection in 5G networks-based secure smart cities," *IEEE Access*, vol. 9, pp. 94780–94788, 2021, doi: 10.1109/ACCESS.2021.3088341.

[7] H. Mandalapu, P. N. A. Reddy, R. Ramachandra, K. S. Rao, P. Mitra, S. R. M. Prasanna, and C. Busch, "Audio-visual biometric recognition and presentation attack detection: A comprehensive survey," *IEEE Access*, vol. 9, pp. 37431–37455, 2021, doi: 10.1109/ACCESS.2021.3063031.

[8] R. Anand, J. Singh, V. Jains, and S. Rathore, "Biometrics security technology with speaker recognition," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 232–236, 2012.

[9] J. Lee. (2017). *Bank of America to Pilot Samsung Iris Recognition for Mobile Banking—Biometric Update*. [Online]. Available: https://www.biometricupdate.com/201708/bank-of-america-to-pilot-samsung-iris-recognition-for-mobile-banking

[10] L. Wu, J. Yang, M. Zhou, Y. Chen, and Q. Wang, "LVID: A multimodal biometrics authentication system on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1572–1585, 2020, doi: 10.1109/TIFS.2019.2944058.

[11] P.-H. Lee, L.-J. Chu, Y.-P. Hung, S.-W. Shih, C.-S. Chen, and H.-M. Wang, "Cascading multimodal verification using face, voice and iris information," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 847–850, doi: 10.1109/ICME.2007.4284783.

[12] T. M. Alsultan, A. A. Salam, K. A. Alissa, and N. A. Saqib, "A comparative study of biometric authentication in cloud computing," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Jun. 2019, pp. 1–6, doi: 10.1109/ISNCC.2019.8909117.

[13] K. Arora, J. Singh, and Y. S. Randhawa, "A survey on channel coding techniques for 5G wireless networks," *Telecommun. Syst.*, vol. 73, no. 4, pp. 637–663, Apr. 2020, doi: 10.1007/s11235-019-00630-3.

[14] M.-C. Chiu, "Analysis and design of polar-coded modulation," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1508–1521, Mar. 2022, doi: 10.1109/TCOMM.2022.3142280.

[15] J. Wang, A. Ji, and M. T. Johnson, "Features for phoneme independent speaker identification," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Jul. 2012, pp. 1141–1145, doi: 10.1109/ICALIP.2012.6376788.

[16] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[17] D. R. Gowda, M. H. Pereira, and V. S. Venkatesh, "Speaker recognition system using MFCC and vector quantization," in *Proc. 2nd Int. Conf. Signal Process., Image Process. VLSI*, 2015, pp. 116–121, doi: 10.3850/978-981-09-6200-5_d-55.

[18] J. H. Bae, A. Abotabl, H.-P. Lin, K.-B. Song, and J. Lee, "An overview of channel coding for 5G NR cellular communications," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, 2019, p. e17, doi: 10.1017/atsip.2019.10.

[19] H. Yao, A. Fazeli, and A. Vardy, "List decoding of Arıkan's PAC codes," *Entropy*, vol. 23, no. 7, p. 841, Jun. 2021, doi: 10.3390/e23070841.

[20] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015, doi: 10.1109/TIT.2015.2410251.

[21] J. Gałka, M. Mąsior, and M. Salasa, "Voice authentication embedded solution for secured access control," *IEEE Trans. Consum. Electron.*, vol. 60, no. 4, pp. 653–661, Nov. 2014, doi: 10.1109/TCE.2014.7027339.

[22] E. Kiktova and J. Juhar, "Speaker recognition for surveillance application," *J. Elect. Electron. Eng.*, vol. 8, no. 2, pp. 19–22, 2015.

[23] M. A. Nematollahi, M. A. Akhaee, S. A. R. Al-Haddad, and H. Gamboa-Rosales, "Semi-fragile digital speech watermarking for online speaker recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, pp. 1–15, Dec. 2015, doi: 10.1186/s13636-015-0074-5.

[24] H. Hentilä, Y. Y. Shkel, and V. Koivunen, "Secret key generation using short blocklength polar coding over wireless channels," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 144–157, Jan. 2022.
[25] D. Harjani, M. Jethwani, and M. Roja, "Speaker recognition system using MFCC and vector quantization," *Int. J. Sci. Res. Develop.*, vol. 1, no. 9, pp. 1935–1937, 2013.
[26] P. Budiga, B. Bhavana, G. Gunisetty, N. D. Moka, and G. V. S. Reddy, "CNN trained speaker recognition system in electric vehicles," in *Proc. Int. Virtual Conf. Power Eng. Comput. Control*, May 2022, pp. 978–983, doi: 10.1109/PECCON55017.2022.9851029.
[27] Y. Sutcu, S. Rane, J. S. Yedidia, S. C. Draper, and A. Vetro, "Feature transformation of biometric templates for secure biometric systems based on error correcting codes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–6, doi: 10.1109/cvprw.2008.4563111.
[28] W. Xia and J. H. L. Hansen, "Attention and DCT based global context modeling for text-independent speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2668–2679, 2023, doi: 10.1109/TASLP.2023.3284521.

NILASHREE WANKHEDE received the B.E. degree in electronics and telecommunication engineering and the Master of Engineering degree in speech processing from the University of Mumbai, India, in 2002 and 2013, respectively, where she is currently pursuing the Ph.D. degree with the Electrical Engineering Department, Veermata Jijabai Technological Institute (VJTI).

From 2003 to 2005, she was with the Larsen and Toubro Institute of Technology as a Teaching Faculty Member. Since 2005, she has been an Assistant Professor with the Department Electronics and Telecommunication Engineering, Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, Maharashtra, India. She has to her credit more than 20 research paper publications in her 17 years of experience in the field of digital communication, wireless communication, optical communication, statistical signal analysis, and speech processing. Her current research interests include channel encoding techniques and performance analysis toward integrating her knowledge in speech signal analysis and remote authentication systems. Her research work is supported by All India Council for Technical Education (AICTE) under Quality Improvement Program (QIP) for the Ph.D. degree with VJTI, Mumbai University.

Prof. Wankhede had been with the IETE Navi Mumbai Centre in the capacity of an Executive Member and her professional society activities as a Life Member of IETE include organization of various seminars, conferences, and workshops. She is also a Life Member of ISTE Society and a Student Member of IEEE Society.

SUSHAMA WAGH (Senior Member, IEEE) received the B.E. degree from WCE, Shivaji University, the M.E. degree in electrical engineering with a specialization in power systems from the Veermata Jijabai Technological Institute (VJTI), Mumbai University, Mumbai, India, and the Ph.D. degree in electrical engineering from the University of Western Australia, Perth, WA, Australia, in 2012. Since 1998, she has been a Faculty Member with VJTI. She is currently a Visiting Scientist with the SLAC National Accelerator Laboratory, Grid Integration and Mobility Group, Menlo Park, CA, USA. Before joining SLAC, she was a Visiting Researcher with Tufts University, Medford, MA, USA, from 2015 to 2016, where she was involved in designing dynamic phasor-based controllers for solid-state transformers. Her current research interests include hybrid grid component modeling, analysis, stability, and control.