

Article

Exploring the Effectiveness of the Phase Features on Double Compressed AMR Speech Detection

Aykut B ker ^{*,†}  and Cemal Hanil i [†]

Department of Electrical and Electronics Engineering, Bursa Technical University, 16310 Bursa, Turkey; cemal.hanilci@btu.edu.tr

* Correspondence: aykut.buker@btu.edu.tr

† These authors contributed equally to this work.

Abstract: Determining whether an audio signal is single compressed (SC) or double compressed (DC) is a crucial task in audio forensics, as it is closely linked to the integrity of the recording. In this paper, we propose the utilization of phase spectrum-based features for detecting DC narrowband and wideband adaptive multi-rate (AMR-NB and AMR-WB) speech. To the best of our knowledge, phase spectrum features have not been previously explored for DC audio detection. In addition to introducing phase spectrum features, we propose a novel parallel LSTM system that simultaneously learns the most representative features from both the magnitude and phase spectrum of the speech signal and integrates both sets of information to further enhance its performance. Analyses demonstrate significant differences between the phase spectra of SC and DC speech signals, suggesting their potential as representative features for DC AMR speech detection. The proposed phase spectrum features are found to perform as well as magnitude spectrum features for the AMR-NB codec, while outperforming the magnitude spectrum in detecting AMR-WB speech. The proposed phase spectrum features yield 8% performance improvement in terms of true positive rate over the magnitude spectrogram features. The proposed parallel LSTM system further improves DC AMR-WB speech detection.

Keywords: DC AMR speech detection; AMR-NB; AMR-WB; audio forensics



Citation: B ker, A.; Hanil i, C. Exploring the Effectiveness of the Phase Features on Double Compressed AMR Speech Detection. *Appl. Sci.* **2024**, *14*, 4573. <https://doi.org/10.3390/app14114573>

Academic Editor: Yan Wang

Received: 26 March 2024

Revised: 9 May 2024

Accepted: 24 May 2024

Published: 26 May 2024



Copyright:   2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Verifying the integrity of an audio recording or speech signal is a critical task in audio forensics, aiming to ascertain whether the signal remains authentic or has been tampered with or edited since its initial generation [1]. The majority of multimedia devices available on the market, such as smartphones and voice recorders, record and store audio files in compressed formats. Any attempt to edit or manipulate a compressed audio recording requires first converting it into a pulse code modulation (PCM) waveform, followed by the manipulation of the PCM signal. Subsequently, the tampered PCM signal is re-compressed back into its original format to conceal any traces of manipulation, resulting in an audio recording compressed twice, as illustrated in Figure 1a. Another scenario leading to double compression involves creating a deceptive high-quality audio recording to misrepresent the true quality of the signal, as depicted in Figure 1b. Here, an original signal compressed at a low bit rate is initially decoded into PCM format and then re-compressed at a higher bit rate, again resulting in a double-compressed audio recording. Additionally, techniques like audio watermarking or steganography often result in a signal being compressed twice, as shown in Figure 1c. Therefore, distinguishing between single-compressed (SC) and double-compressed (DC) audio recordings is a crucial task in audio forensics, closely linked to the integrity of the recording. Particularly in legal contexts, such as courtrooms, double compression can serve as a trace to verify the integrity of the recording.

Various audio codecs are employed to encode audio signals, including MP3 [2], AAC [3], AC3, WMA, Vorbis, among others. However, the adaptive multi-rate (AMR)

codec [4,5] stands out in several respects. Firstly, while the majority of encoders are perceptual codecs designed to compress audio signals (often music signals), based on the characteristics of the human ear, the AMR codec is primarily tailored for speech signals. Secondly, AMR speech coding has emerged as a mandatory codec for GSM services, making it the most prevalent speech codec in use. Thirdly, unlike other perceptual audio codecs, the AMR codec employs the code excited linear prediction (CELP) model instead of considering the masking effect on the human ear. Lastly, most smartphones and audio recorders save recordings in AMR format, and numerous freely available audio editing software tools can convert AMR-compressed signals into PCM waveform or other formats. Consequently, individuals lacking prior experience in digital signal processing, such as forgers or fraudsters, can easily modify AMR signals. As a result, this study focuses on detecting double-compressed (DC) AMR speech signals.

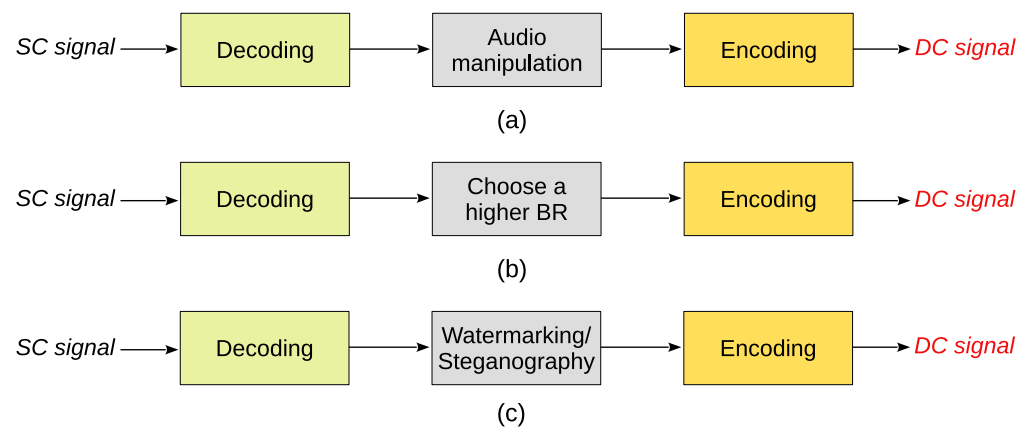


Figure 1. Cases of DC signal generation: (a) Creating a manipulated/tampered signal. (b) Generating a fake-quality signal. (c) Audio watermarking/steganography.

The magnitude of the discrete Fourier transform (DFT) of short speech segments, commonly known as the spectrogram representation of speech signals, has found widespread application in various tasks such as speaker recognition [6], speech recognition [7], emotion recognition [8], and audio event detection [9]. Its effectiveness stems from its ability to capture the spectral content variation of the signal over time, making it suitable for use with deep neural networks (DNNs), such as deep convolutional neural networks (CNNs), and long-short-term memory (LSTM) networks. In our previous work, we introduced the use of spectrograms for the first time in detecting narrowband (sampling frequency of 8 kHz) DC AMR (AMR-NB) speech signals [10], achieving state-of-the-art performance. However, the spectrogram representation only utilizes the magnitude spectrum, discarding the phase information of the Fourier transform. Numerous studies in the literature have highlighted the utility of the phase spectrum in tasks such as synthetic/converted speech detection [11,12], speech recognition [13], and speaker recognition [14]. Additionally, research has emphasized the perceptual importance of phase information in speech coding [15,16] and human speech perception [17]. Previous studies have indicated that preserving the original phase spectrum results in more natural sounding reconstructed speech, underscoring the potential value of phase spectrum features in distinguishing SC and DC speech signals. Motivated by the significance of the phase spectrum in other speech-related tasks and its impact on human speech perception, we propose to explore and evaluate the performance of the phase spectrum in DC AMR speech signal detection. To the best of our knowledge, this study represents the first investigation into the effect of phase information on DC audio detection. In addition to proposing the use of phase features for DC AMR speech detection, we examine the influence of phase information on both narrowband AMR (AMR-NB) and wideband AMR (AMR-WB)-coded signals, comparing their performance with the traditional magnitude spectrum (spectrogram). Given that prior

research has primarily addressed the detection of DC AMR-NB speech signals, comparing the results of DC AMR-NB and AMR-WB speech detection will be crucial for future studies on DC speech detection. In summary, our main contributions are as follows:

- We propose utilizing the phase spectrum as a feature for detecting DC speech signals. As far as we know, the performance of phase features in DC audio detection remains unexplored. Therefore, this study represents the first investigation into the effect of phase information on this task. We examine the impact of three different representations of the phase spectrum on DC speech signal detection.
- We assess the performance of the proposed phase features on both AMR-NB and AMR-WB-coded speech signals. Given that existing studies have focused solely on AMR-NB-coded signals, understanding the performance of the AMR-WB codec in DC audio detection and comparing it with the AMR-NB codec will be valuable for the audio forensics community.
- We introduce the use of LSTM networks for DC AMR speech detection, marking the first instance in the literature. Both spectrogram and phase spectra can be viewed as time-series data, with the former capturing spectral energy variations over time and the latter representing phase variation over time. Leveraging LSTM networks, known for their effectiveness with time-series data, we propose their application for DC AMR speech detection for the first time in the literature.

2. Related Work

Considering existing studies, research on DC audio signal detection can be divided into three categories based on the audio codec used: (i) DC MP3 audio detection, (ii) DC AAC audio detection, and (iii) DC AMR speech detection. The majority of early studies focused on DC audio detection employed the MP3 codec, with the aim of distinguishing whether an MP3 audio file belongs to the SC or DC class. Similarly, interest in detecting DC AAC-encoded audio, another widely used perceptual audio codec found in platforms like Apple iTunes, YouTube, and DVD standards, grew after numerous studies were conducted on the AAC codec. DC AMR speech detection, on the other hand, has gained considerable attention in recent years due to its extensive usage in smartphones and digital voice recorders. In this section, we provide a brief overview of the existing literature for each category of studies.

In [18], it was demonstrated that high-frequency components (above 16 kHz) were attenuated when a CD-quality music signal was compressed using an MP3 codec with a low bit rate (128 kbps). However, it was observed that these high-frequency components became perceptible when a higher bit rate was used. Consequently, the authors proposed employing a support vector machine (SVM) classifier with spectral features extracted from the 16 kHz to 20 kHz band of the audio spectrogram to detect the original compression bit rate of the MP3 audio signals. Detection of “fake-quality” MP3 recordings, where an MP3 recording compressed at a low bit rate is decoded and re-encoded at a higher bit rate, using the number of small valued modified discrete cosine transform (MDCT) coefficients, was addressed in [19]. Similarly, in [20], statistical features extracted from zero-valued and nonzero-valued MDCT coefficients were used with SVM classifiers to detect DC MP3 audio signals. The use of statistical features derived from quantized MDCT (QMDCT) coefficients for DC MP3 audio detection was proposed in [21,22]. In [23], the distribution of the nine most significant digits of the MDCT coefficients was employed to detect MP3 double compression using an SVM classifier. In [24,25], authors computed the Chi-square distance between the histograms of observed and simulated MDCT coefficients and then applied a threshold to the distance value to determine whether the MP3 audio is SC or DC.

In [26], the authors applied the Markov model to the Huffman codebook indices to extract features and utilized an SVM classifier for detecting DC AAC signals. It was reported that the proposed features gave higher performance when the first compression bit rate (BR1) was lower than that of the second compression bit rate (BR2). In [27], QMDCT coefficients were used with the SVM classifier to detect DC AAC signals. It was

found that QMDCT coefficients provided reasonable performance when the BR2 value was higher than BR1. In [28], it was demonstrated that the audio scale factor decreases as the number of compression iterations increases, making it suitable for detecting DC AAC signals. A detection rate above 99% was achieved using the proposed features with the SVM classifier. In [29], authors proposed using QMDCT coefficient features with a CNN classifier to detect the compression bit rate of SC AAC audio signals.

From the brief literature review on MP3 and AAC compression detection presented above, it is evident that features computed from MDCT coefficients were widely utilized, with reported detection rates exceeding 95% in most studies. The remarkable performance of MDCT-based features in detecting MP3 or AAC compression is unsurprising, considering that both MP3 and AAC codecs utilize MDCT coefficients for encoding audio signals. Therefore, extracting features with prior knowledge about the encoder is expected to yield high performance without a doubt.

In the initial study addressing DC AMR-NB speech signal detection, statistical features based on spectral energy distribution and correlation of spectral components were proposed and used with the SVM classifier [30]. In [31], stacked autoencoder (SAE) and dropout networks were utilized for DC AMR-NB speech detection. Here, the long speech signal was first divided into short, non-overlapping segments, which were then fed into the classifier network. A majority voting strategy was employed for determining whether input speech is SC or DC AMR-NB signal. Similarly, Ref. [32] employed a SAE network for feature extraction, followed by modeling extracted features using a Gaussian mixture model (GMM) classifier for DC AMR-NB speech detection. In [33,34], statistical features extracted from linear predictive coding (LPC) coefficients and line spectral pairs (LSP) of the speech signals were utilized with the SVM classifier for DC AMR-NB speech detection. Given that AMR codec employs LPC analysis for speech signal encoding, the efficacy of LPC coefficient-based features for DC AMR detection, as reported in [30,31,33,34], is unsurprising. Moreover, Ref. [35] demonstrated that reasonable DC AMR-NB speech detection performance can be achieved using traditional spectral features without prior knowledge of encoding algorithms, where simple but effective long-term average spectra (LTAS) features were employed with traditional fully connected DNNs. In [36], the spectrogram of the speech signal was divided into temporal segments, each consisting of 10 speech frames, and LTAS of each segment was computed. These LTAS vectors were then inputted into a DNN classifier for DC AMR-NB speech detection. The introduction of deep convolutional neural networks (CNNs) in [10], utilizing speech spectrogram representation as input features, marked a significant advancement in DC AMR-NB detection. It was demonstrated that both CNNs and spectrogram representations achieved state-of-the-art performance. Subsequently, Ref. [37] proposed the use of angular margin softmax loss functions over vanilla softmax in CNN-based detection systems. It was reported that angular margin losses exhibit superior performance compared to traditional softmax loss.

Inspired by the superior performance of spectrogram features on DC AMR-NB speech signal detection, as reported in [10,37], which are extracted without prior knowledge about the AMR encoder, this paper aims to investigate the potential of phase information in DC AMR speech detection. While previous studies [10,35,36] have demonstrated the considerable impact of compressing a speech signal twice on its high-frequency components in the magnitude spectrum, the effects of compression on the phase spectrum remain unexplored. The phase spectrum of a speech signal is known to convey crucial cues not present in the magnitude spectrum and has been extensively utilized in various speech-related recognition tasks, including speaker recognition, speech recognition, language recognition, and synthetic speech detection. However, to the best of our knowledge, the phase spectrum has not been previously utilized for DC speech/audio detection. Therefore, this paper proposes the use of phase features for DC AMR-NB and DC AMR-WB speech signal detection.

3. AMR Speech Codec

The narrowband AMR speech codec (AMR-NB) encodes speech signals sampled at 8 kHz using eight different bit rates, where $BR \in \{4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20\}$ kbits/s (kbps). It employs the code-excited linear prediction (CELP) model for encoding. The CELP model parameters are extracted through 10th order LP analysis of 20 ms speech frames, with each frame containing 160 samples. These extracted CELP model parameters are then encoded and transmitted during the encoding process. On the decoder side, the transmitted parameters are first decoded to reconstruct the excitation signal. This reconstructed excitation signal is subsequently applied to the LP synthesis filter to obtain the speech signal. For more technical details regarding the AMR-NB speech codec, readers are referred to [4].

The Wideband AMR speech codec (AMR-WB) operates on speech signals with a sampling frequency of 16 kHz and consists of nine bit rates: 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, and 23.85 kbits/s. During encoding, the WB AMR codec first applies a pre-emphasis filter to the speech signal, followed by dividing the signal into short frames of 20 ms each. Each frame is then analyzed using a 16th-order LP analysis to obtain the code-excited linear prediction (CELP) model parameters. Unlike the AMR-NB codec, which applies LP analysis once per frame, the AMR-WB encoder further divides each frame into four subframes. For each subframe, adaptive and fixed codebook parameters are transmitted. During decoding, the transmitted parameters are decoded first, followed by synthesizing the speech signal by applying the reconstructed excitation signal to the LP synthesis filter. More details on the AMR-WB encoding and decoding steps can be found in [5].

4. Magnitude and Phase Spectra of AMR-Encoded Speech Signals

While speech is inherently a non-stationary signal, it is assumed to be quasi-stationary when analyzed over a short duration (typically 20–40 ms) [38]. Therefore, speech signals are processed through short-term analysis. In this approach, the signal is initially segmented into short, overlapping frames and then the desired speech processing techniques are applied to each of these individual frames. The spectrum of speech signals is obtained using the short-term Fourier transform (STFT), which involves applying the discrete Fourier transform (DFT) to these short speech frames. Concretely, given a set of short speech frames denoted as $x(n, t)$ where $n = 0, 1, \dots, N - 1$ and $t = 1, 2, \dots, T$ representing the sample and frame indices, respectively, the STFT is computed by:

$$X(k, t) = \sum_{n=0}^{N-1} x(n, t)w(n)e^{-j2\pi kn/N} \quad (1)$$

where $w(n)$ is the data-tapering window (e.g., Hamming or Hanning window) and k represents the DFT bin. Since $X(k, t)$ is complex-valued, Equation (1) can be written in the polar form as

$$X(k, t) = |X(k, t)|e^{j\theta(k, t)} \quad (2)$$

where $|X(k, t)|$ is the short-term magnitude spectrum (also known as the spectrogram of the signal and visualized by intensity plot on log-scale) and $\theta(k, t)$ is the short-term phase spectrum of the speech signal.

Spectrogram features ($|X(k, t)|$) are widely used in the majority of the current state-of-the-art DNN-based speech-related recognition tasks, such as speaker recognition and speech recognition. Hence, usually phase information ($\theta(k, t)$) is discarded. Similarly, to the best of our knowledge, the short-term phase spectrum has not been used for DC audio detection tasks previously. However, the importance of phase information in human perception has been reported in several studies [15–17]. For example, in [16], it was shown that the reconstructed speech signal sounds less natural when phase information is discarded during speech coding. Similarly, in [15], it was reported that human perception of phase varies with frequency especially for low-pitched speakers. In [17], authors showed that the STFT phase spectrum is as important as the STFT magnitude spectrum in speech intelligibil-

ity when analysis-synthesis parameters are properly selected. Since phase information was found to have a considerable impact on human speech perception and speech intelligibility, intuitively, it can be considered a reasonable feature representation for DC audio detection. Hence, in this paper, we aim at investigating the effect of the STFT phase spectrum on DC AMR audio detection.

In order to analyze the effect of AMR compression on the phase spectrum of a speech signal, the phase spectrum of the selected original (uncompressed) speech signal and its compressed (SC and DC) counterparts are first computed. Suppose that $\theta_O(k, t)$, $\theta_{SC}(k, t)$ and $\theta_{DC}(k, t)$ denote the phase spectrum of the original SC and DC speech signals, respectively. Then the average difference phase spectrum between the original speech and SC AMR speech is computed for each DFT bin k by

$$\Delta\theta_{O,SC}(k) = \frac{1}{T} \left[\sum_{t=1}^T \theta_O(k, t) - \sum_{t=1}^T \theta_{SC}(k, t) \right] \quad (3)$$

similarly, the average difference phase spectrum between the SC and DC speech signals is computed by

$$\Delta\theta_{SC,DC}(k) = \frac{1}{T} \left[\sum_{t=1}^T \theta_{SC}(k, t) - \sum_{t=1}^T \theta_{DC}(k, t) \right] \quad (4)$$

where T is the total number of speech frames. Equation (3) reveals the impact of the first compression on the phase spectrum compared to the original (uncompressed) signal for each frequency (k), while Equation (4) delves into the effect of subsequent compression on the phase spectrum relative to the SC speech signal. Figure 2 displays the average difference phase spectra $\Delta\theta_{O,SC}(k)$ (left panel) and $\Delta\theta_{SC,DC}(k)$ (right panel) calculated using the AMR-NB codec across various bit rates. When computing $\Delta\theta_{SC,DC}(k)$, which represents the average difference phase spectra between the SC and DC signals, the first compression bit rate (BR1) of the DC signals is set to match the value used for obtaining the SC signal (BR = BR1 = 5.9 kbps). This adjustment allows us to assess the effect of the second compression bit rate (BR2) on the phase spectrum. From the figure, it is observed that compressing the narrowband speech signal considerably affects the phase spectrum above 500 Hz, and the difference between the phase spectrum of the original and SC signals becomes larger as the compression bit rate increases. For the DC signals in turn, compressing the signal for the second time notably affects the phase spectrum at higher frequencies.

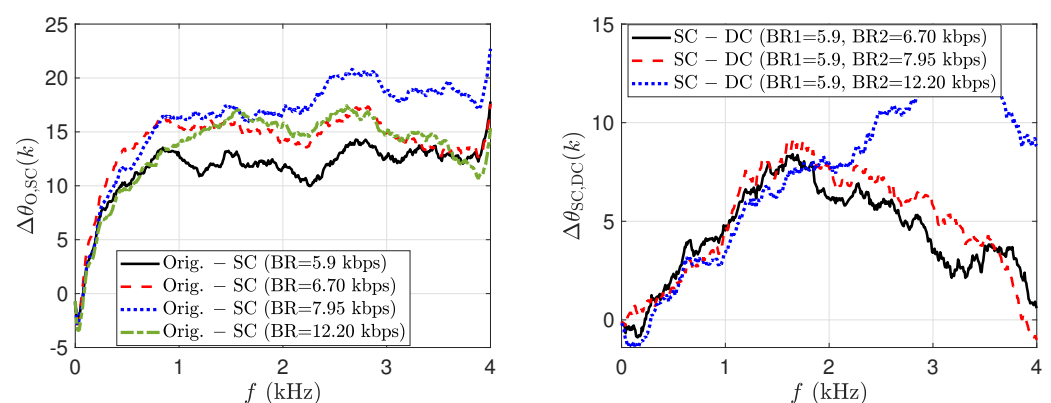


Figure 2. Average difference phase spectra between the original (uncompressed) and SC speech signals (**left**) and between the SC and DC speech signals (**right**). The AMR-NB codec is used to generate the SC and DC signals. For the DC signals, the first compression bit rate (BR1) matches with the compression bit rate of the SC speech signal (BR = 5.9 kbps) to assess the effect of second compression bit rate (BR2) on DC signals.

Figure 3 shows the average phase spectra $\Delta\theta_{O,SC}(k)$ and $\Delta\theta_{SC,DC}(k)$ computed using the AMR-WB codec. Similar to the observations with AMR-NB-coded signals (as shown

in Figure 2), compressing the speech signal for the first time using the AMR-WB codec has a considerable impact on the phase spectrum. As the compression bit rate increases, the phase spectrum exhibits notable variations compared to the uncompressed signal spectrum. These variations become more pronounced when the signal is compressed for the second time, particularly at higher frequencies.

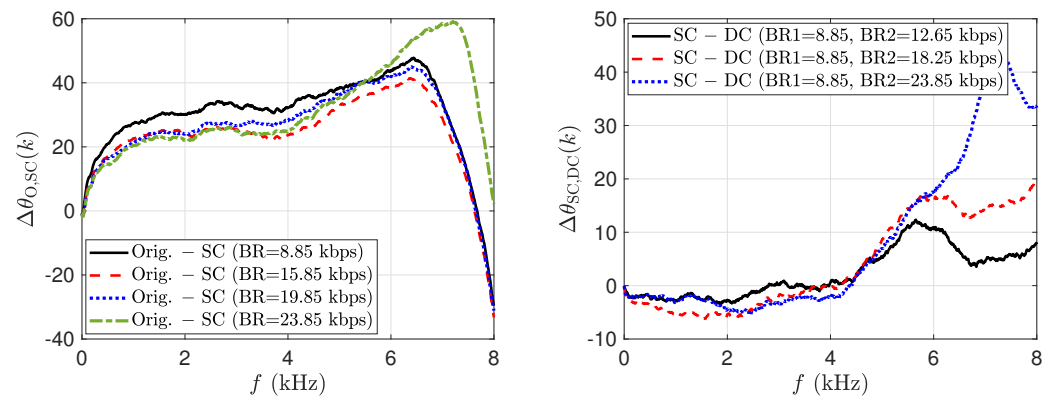


Figure 3. Average difference phase spectra between the original (uncompressed) and SC speech signals (**left**) and between the SC and DC speech signals (**right**). The AMR-WB codec is used to generate the SC and DC signals. For the DC signals, the first compression bit rate (BR1) matches with the compression bit rate of the SC speech signal (BR = 8.85 kbps) to assess the effect of second compression bit rate (BR2) on DC signals.

Based on the average difference phase spectra visualized in Figures 2 and 3, it is evident that the phase spectra of the SC and DC signals exhibit different behaviors. Therefore, the phase spectrum can serve as a discriminative feature between SC and DC signals. In this study, we explore the effect of the STFT phase spectrum on DC AMR speech detection by proposing the use of three different phase spectrum-based feature representations:

- **STFT Phase Spectrum:** In this approach, the STFT phase spectrum $\theta(k, t)$ is unwrapped to obtain a continuous function of k , in order to remove the discontinuities by adding an integer multiple of 2π to the phase spectrum at each DFT bin k . The unwrapped phase spectrum is used as the feature for DC AMR speech detection.
- **Cosine Normalized Phase Spectrum (CosPhase):** After unwrapping the phase spectrum, the dynamic range of the phase spectrum may exhibit large variations, making it challenging to model the phase information accurately. To address this, we apply the cosine function to the unwrapped phase spectrum, resulting in features within the range $[-1, 1]$.
- **CosPhase with discrete cosine transform (DCT + CosPhase):** To obtain a more compact representation of the phase information, DCT is applied to the CosPhase features to obtain uncorrelated features.

In addition to the three phase spectrum-based features described above, we also utilize the magnitude spectrum (spectrogram) representation of the speech signal as a baseline feature in our experiments, enabling us to compare the performance of the phase spectrum features.

5. DC AMR Detection System

Distinguishing between SC and DC speech signals constitutes a binary classification task due to the presence of only two possible classes. As a result, previous research predominantly relied on traditional pattern recognition techniques such as support vector machines (SVM) to tackle this challenge. However, the effectiveness of such methods hinges on pre-processing the speech signals to extract pertinent information for the given task. It is imperative that the extracted features are highly representative to ensure reasonable performance. In earlier studies, feature extraction often involved the use of handcrafted fea-

tures, leveraging prior knowledge about the compression algorithm. For instance, features derived from MDCT coefficients were utilized for MP3 and AAC codecs, while feature vectors obtained through LP analysis were employed for the AMR codec. Subsequently, these extracted features were fed into classification algorithms to either estimate the distributions of training features (generative models) or draw a decision boundary between the two classes in the feature space (discriminative models). In the case of generative models, decisions are made based on the posterior probabilities of potential classes given the test feature vector. Conversely, discriminative models employ distance measures to make decisions. Nevertheless, traditional pattern classification techniques suffer from certain limitations, notably the challenge of determining which features are most pertinent for the task and selecting an appropriate classification method. These decisions often necessitate profound domain expertise and entail a protracted trial-and-error process.

With the rapid advancement of deep learning methodologies, the concept of end-to-end learning has emerged. In end-to-end learning, DNN possesses the capability to automatically learn the most pertinent features and make decisions based on input data. Consequently, both feature extraction and classification tasks are seamlessly integrated into the DNN system. In this study, we leverage LSTM-based DNNs for the task of DC AMR speech detection. LSTM networks, belonging to the family of recurrent neural networks (RNNs), are renowned for their ability to capture long-term dependencies. These networks comprise a series of memory blocks that are interconnected recurrently. They acquire knowledge of long-term dependencies by retaining previous information within their memory blocks and utilizing it to compute the output of the current block based on both previous and current input. Hence, LSTM networks are well-suited for modeling temporal sequences, making them particularly appropriate for tasks involving time-series data such as spectrograms and phase features extracted using STFT. These features inherently represent a time-series sequence of frequency domain features, rendering LSTM networks suitable for learning the long-term dependencies essential for DC AMR speech detection.

In this study, we propose two different LSTM-based systems for DC speech detection. The baseline LSTM system, illustrated in Figure 4, comprises two LSTM layers, followed by one fully connected layer and a classification layer. Each LSTM layer encompasses 256 units with tanh nonlinearity, while the fully connected (FC) layer consists of 512 units with a sigmoid activation function. The classification layer employs the softmax activation function, featuring two units corresponding to the SC and DC speech classes, respectively. Given the input features X (spectrogram or phase spectrum features), the outputs of the classification layer yield the posterior probabilities $p(C_1|X)$ and $p(C_2|X)$, where C_1 and C_2 denote the SC and DC speech classes, respectively. Employing the posterior probabilities of SC and DC classes, the Bayes decision rule, ensuring the minimum classification error rate, is utilized to make decisions. This rule is defined as follows:

$$Decision = \begin{cases} \text{Decide } C_1, & p(C_1|X) > p(C_2|X) \\ \text{Decide } C_2, & p(C_2|X) \geq p(C_1|X) \end{cases} \quad (5)$$



Figure 4. Proposed baseline LSTM system DC AMR speech detection.

The second system proposed in this study is the parallel LSTM system, illustrated in Figure 5. The parallel LSTM system comprises two LSTM networks operating in parallel, where the deep features learned by each branch are fused through concatenation and subsequently applied to the fully connected (FC) and classification layers. The motivation behind proposing the parallel LSTM system is to explore whether the magnitude spectrum (spectrogram) and phase features convey complementary information when simultaneously utilized in each branch of the parallel system. Each branch of the parallel system is designed to learn different levels of information due to the diverse inputs, enabling the learned deep

embeddings to be fused to enhance the performance of DC AMR detection. Each branch of the parallel LSTM system consists of two LSTM blocks, each containing 256 units with a tanh activation function. The outputs of the second LSTM block of each branch are then flattened, resulting in 25,600 -dimensional deep features from each branch. These features are concatenated to form a single 51,200 -dimensional super vector, which is subsequently fed into two FC layers comprising 512 and 256 units with rectified linear unit (ReLU) activations, respectively. The classification layer consists of 2 units representing the SC and DC classes, utilizing softmax activation

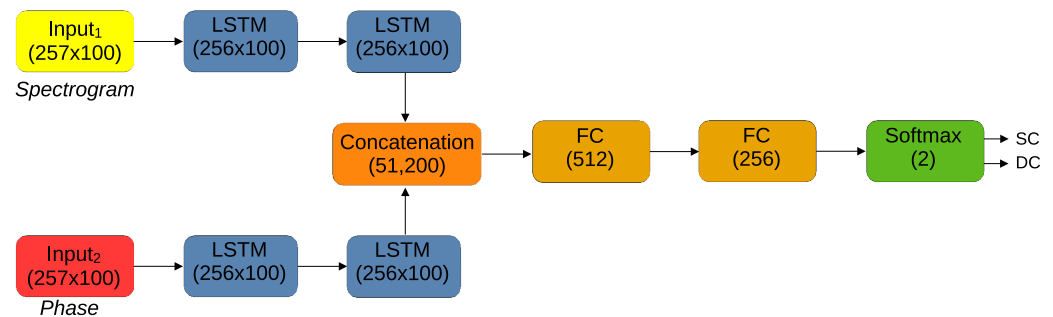


Figure 5. Proposed parallel LSTM system for DC AMR speech detection.

All parameters in both systems, such as the number of LSTM and FC blocks, the number of units in each layer, and the activation functions, are optimized based on preliminary experiments. The best-performing parameters are then employed in the proposed methods.

6. Experimental Setup

6.1. Database

Double-compressed AMR audio detection experiments are conducted on the TIMIT dataset [39]. TIMIT is a well-known audio dataset commonly used for speech and speaker recognition. It comprises 6300 speech recordings sampled at 16 kHz from 630 different speakers, with each speaker having 10 different utterances. Among these utterances, two speech recordings (SA1 and SA2) are shared by all speakers.

For the AMR-NB codec, all speech signals in the database are initially downsampled to 8 kHz, as the AMR-NB codec operates on narrowband signals. Subsequently, all speech signals in the database are compressed using the AMR-NB codec at eight different bit rates, ranging from 4.75 kbps to 12.2 kbps. This process yields a total of 50,400 SC AMR-NB speech recordings ($6300 \times 8 = 50,400$). These SC AMR-NB files are then decoded and recompressed at eight different bit rates to generate DC AMR-NB signals. Consequently, we obtain a total of 403,200 double-compressed AMR-NB audio files ($6300 \times 8 \times 8 = 403,200$).

A similar procedure to the one used for the AMR-NB database generation is applied to generate the SC and DC AMR-WB speech signals. The speech signals in the TIMIT database, sampled at 16 kHz, are utilized to generate the AMR-WB-coded signals. Since the AMR-WB codec employs nine different bit rates ranging from 6.6 kbps to 23.85 kbps, the AMR-WB database comprises a total of 56,700 SC speech signals ($6300 \times 9 = 56,700$) and 510,300 DC speech signals ($6300 \times 9 \times 9 = 510,300$).

In the experiments, one-second-long speech signals are utilized. If a signal's duration exceeds 1 s, its first one-second portion is cropped. Conversely, if the signal is shorter than 1 s, its samples are duplicated and appended to the signal to achieve a one-second duration. For each dataset, 25% of the speech signals are allocated for training the DC signal detection system, while the remaining signals are used for validation and testing. Since SA1 and SA2 utterances are shared by all speakers in the dataset, the validation set comprises only these signals. To ensure the effectiveness of the model in distinguishing DC signals from SC signals, the training and test sets are disjoint in terms of speaker allocation.

This prevents the system from potentially learning to differentiate between speakers rather than detecting DC signals.

6.2. Feature Extraction and Classifier Setup

The proposed LSTM-based systems in the experiments utilize magnitude spectrum (spectrogram) and phase spectrum representations of the speech signals as input. These feature representations are computed using STFT analysis of the signal. To achieve this, speech signals are segmented into short frames lasting 25 ms each (containing 200 and 400 speech samples for narrowband and wideband signals, respectively) with a frame shift of 10 ms, resulting in a frame rate of 100 frames per second. Each frame is then multiplied by a Hamming window, and a 512-point DFT of each windowed frame is computed to obtain complex-valued Fourier transforms. Subsequently, the spectrogram is computed by taking the logarithm of the magnitude of the STFT spectrum of the signal. The phase spectrum representation is obtained by computing the angle of the STFT spectrum and then applying phase unwrapping to avoid discontinuities. The resulting spectrogram and phase spectrum representations are of dimensions 257×100 , where each dimension corresponds to the DFT bin and frame index, respectively.

The features (spectrogram or phase spectrum features) of the SC and DC speech signals from the training set are then fed into the LSTM-based DC detection systems as described in Section 5. All models (LSTM and parallel LSTM networks) are trained using the Adam optimizer [40]. The networks undergo training for 100 epochs, with an initial learning rate set to 0.001. If the validation loss increases compared to the previous epoch, the learning rate is reduced by a factor of 0.95. To prevent overfitting, an early stopping approach is employed during model training.

6.3. Performance Criteria

Given that DC AMR speech signal detection involves a binary classification task aimed at distinguishing between DC signals and those of SC, the evaluation of system performance revolves around true positive and true negative rates (TPR and TNR, respectively). In this context, the DC speech class is designated as the positive class, while the SC speech class is regarded as the negative class. TPR, also referred to as sensitivity, represents the probability of a DC signal being correctly classified as such by the system. It is computed by

$$\text{TPR}[\%] = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100. \quad (6)$$

where TP and FN are true positive and false negative values, respectively. TNR, also known as specificity, in turn denotes the probability of classifying a signal as belonging to the SC class given that the signal actually is SC. It is computed by

$$\text{TNR}[\%] = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (7)$$

where TN and FP correspond to true negative and false positive values, respectively.

In addition to the TPR and TNR evaluation metrics, the accuracy criterion is employed to demonstrate the overall performance of the system. It is computed by

$$\text{Acc}[\%] = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (8)$$

7. Experimental Results

In the experiments, we initially assess the performance of individual feature representations using the baseline LSTM system for AMR-NB and AMR-WB speech detection. The average SC and DC detection rates (represented by TNR and TPR values, respectively) for each feature representation are summarized in Table 1. For each feature, the average SC detection rates are computed by averaging eight TNR values for the AMR-NB codec (8 BR

values) and nine TNR values for the AMR-WB codec (9 BR values), respectively. Similarly, the average DC detection rates (TPR) are determined by aggregating 64 TPR values for the AMR-NB codec ($8 \text{ BR1} \times 8 \text{ BR2} = 64$ combinations) and 81 TPR values for the AMR-WB codec ($9 \text{ BR1} \times 9 \text{ BR2} = 81$ combinations), respectively. The results presented in the table reveal considerably higher detection rates achieved using the AMR-NB codec compared to the AMR-WB encoder, regardless of the feature set used. For instance, while the magnitude spectrum (spectrogram) features yield a 99.94% SC detection rate with the AMR-NB codec, it reduces to 76.73% for the AMR-WB codec. Similarly, employing spectrogram features results in a TPR value of 99.98% with the AMR-NB codec, whereas a degradation of approximately 20% is observed in the DC detection rate (TPR of 80.06%) when utilizing the AMR-WB codec. This trend persists for the phase spectrum features as well, indicating that detecting AMR-WB-coded speech signals is inherently more challenging than detecting narrowband signals, regardless of the feature representation employed. An important observation from the table is that, for narrowband AMR-coded speech signals, the proposed phase-spectrum-based features (except the raw phase spectrum) yield comparable results to spectrogram features. For instance, while spectrogram features achieve a 99.94% average SC detection rate, CosPhase and DCT+CosPhase features yield average TNR values of 99.93% and 98.72%, respectively. Similarly, the TPR values obtained using CosPhase features and DCT + CosPhase features closely match the DC AMR-NB detection rate of spectrogram features. For wideband-coded speech signals, CosPhase features exhibit higher detection rates compared to spectrogram features for both SC and DC speech signals. For instance, while spectrogram features achieve 76.73% and 80.06% average TNR and TPR values, respectively, cosPhase features further enhance the SC and DC detection rates to 84.03% (approximately 10% relative improvement) and 86.51% (approximately 8% relative improvement), respectively. The elevated detection rates achieved using phase spectrum features on both AMR-NB and AMR-WB compressed speech signal detection indicate that phase spectrum conveys useful information for DC speech detection. Among the three proposed phase-spectrum-based features, CosPhase features yield the highest detection rates for both SC and DC speech signals, regardless of the codec. Therefore, CosPhase features will be utilized in the remaining experiments.

Table 1. Average TNR and TPR values for AMR-NB and AMR-WB codecs obtained using different feature representations and the baseline LSTM system

Input	AMR-NB		AMR-WB	
	TNR (%)	TPR (%)	TNR (%)	TPR (%)
Spectrogram	99.94	99.98	76.73	80.06
Phase Spectrum	83.40	81.82	68.65	62.46
CosPhase	99.93	99.84	84.03	86.51
DCT + CosPhase	98.72	97.95	57.47	42.81

To further analyze the DC detection rates obtained using spectrogram and CosPhase features, we present the TPR values for each combination of first and second compression bit rates (BR1-BR2). Table 2 summarizes the TPR values obtained using spectrogram features and the baseline LSTM system on the AMR-NB codec. Although there are 64 possible combinations of BR1-BR2 pairs for DC AMR-NB-coded speech signals, only 6 cases yield TPR values lower than 100%. Hence, in Table 2, bit rate combinations that provide 100% detection rates are omitted for clarity. Interestingly, among these six BR1-BR2 combinations where detection rates are lower than 100%, the BR1 value is consistently 4.75 kbps, and as the BR2 value increases, the detection rate slightly decreases in general. However, the reduction in detection rates is negligible.

Table 2. Detection rates (TPR in %) for DC AMR-NB speech using spectrogram features and the baseline LSTM network. BR1: first compression bit rate and BR2: second compression bit rate.

Feature	System	BR1 (kbps)	BR2 (kbps)	TPR
Spectrogram	LSTM	4.75	4.75	99.47
Spectrogram	LSTM	4.75	5.9	99.94
Spectrogram	LSTM	4.75	6.7	99.83
Spectrogram	LSTM	4.75	7.95	99.80
Spectrogram	LSTM	4.75	10.2	99.97
Spectrogram	LSTM	4.75	12.2	99.72

The DC AMR-NB detection rates obtained using the proposed CosPhase features and the baseline LSTM system for each possible BR1-BR2 combination are reported in Table 3. The results summarized in the table demonstrate the effectiveness of phase spectrum features in detecting DC AMR-NB speech signals. Nearly all of the detection rates reported in the table (except for three entries) exceed 99.5%, indicating the efficacy of the proposed features. The lowest detection rate (98.82%) is observed when BR1 = 12.2 kbps and BR2 = 4.75 kbps, corresponding to a radical down-transcoding procedure.

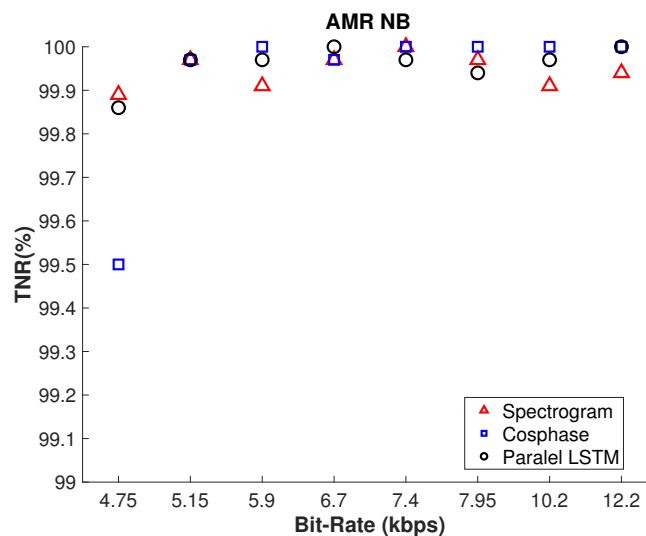
Table 3. Detection rates (TPR in %) for DC AMR-NB speech using CosPhase features and the baseline LSTM network.

		BR2 (kbps)							
		4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
BR1 (kbps)	4.75	99.89	99.83	99.39	99.20	99.97	99.97	99.83	99.94
	5.15	99.47	99.58	99.86	99.89	99.91	99.97	100	99.94
	5.75	99.53	99.86	99.97	99.86	99.94	99.97	99.97	99.97
	6.7	99.61	99.97	100	99.89	99.97	100	100	99.97
	7.4	99.80	99.97	100	99.72	99.86	99.97	100	100
	7.95	99.89	99.91	99.89	99.78	99.97	99.97	99.97	100
	10.2	99.78	99.97	99.75	99.91	99.94	100	99.94	100
	12.2	98.82	99.75	99.50	99.61	99.53	99.80	99.89	99.83

We next investigate the impact of combining information learned from spectrogram and CosPhase features using the proposed parallel LSTM network structure outlined in Section 5. Figure 6 illustrates the SC AMR-NB detection rates (TNR values) obtained using the parallel LSTM system, along with the results obtained using single features with the baseline LSTM system for comparison. From the figure, it is evident that the parallel LSTM system outperforms stand-alone spectrogram or phase spectrum features in most cases. However, it is noteworthy that the TNR values depicted in the figure are very close to each other, making it challenging to draw a general conclusion solely based on the TNR values. Hence, the DC AMR-NB speech detection rates obtained using the parallel LSTM system are reported in Table 4. Notably, perfect detection rates (100%) are observed for the majority of BR1-BR2 cases. Comparing the detection rates in the table with the results obtained using stand-alone CosPhase features given in Table 3, it is evident that phase spectrum and spectrogram features complement each other. Consequently, combining the learned deep features from the spectrogram and CosPhase features through parallel LSTM further enhances the detection rates for the AMR-NB codec.

Table 4. TPR detection rates (in %) for DC AMR-NB speech utilizing spectrogram and CosPhase features with the proposed parallel LSTM network.

		BR2 (kbps)							
		4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
BR1 (kbps)	4.75	99.42	99.83	99.75	99.26	99.80	99.58	99.83	99.80
	5.15	100	100	100	100	100	100	100	100
	5.95	100	100	100	100	100	100	100	100
	6.7	100	100	99.97	100	100	100	100	100
	7.4	100	100	99.94	100	100	100	100	100
	7.95	100	100	99.97	100	100	99.97	100	100
	10.2	100	99.89	100	100	100	100	100	100
	12.2	100	99.97	100	99.97	99.94	99.94	99.97	99.97

**Figure 6.** Detection rates (TNR in %) for SC AMR-NB compressed speech obtained using the baseline LSTM system with spectrogram and CosPhase features and the proposed parallel LSTM system.

In the experiments, we next explore the impact of phase features on DC AMR-WB speech detection. The DC AMR-WB speech detection rates obtained using individual spectrogram and CosPhase features with the baseline LSTM network are summarized in Tables 5 and 6, respectively. The detection rates where CosPhase features are superior to the spectrogram features are underlined in Table 6 for ease of comparison. Firstly, one can observe that unlike the AMR-NB codec (Tables 2 and 3), where the detection rates are consistently higher than 99.5%, DC AMR-WB detection rates are notably lower regardless of the features. This confirms that DC AMR-WB speech detection is indeed a more challenging task compared to detecting DC AMR-NB speech. For spectrogram features, when the first compression bitrate (BR1) is fixed, DC speech detection rates generally decrease as the BR2 value increases. Similarly, a similar observation is noted when BR1 increases for a fixed value of BR2. Regarding the CosPhase features, higher detection rates are observed as BR1 (or BR2) increases for a fixed value of BR2 (or BR1). Upon comparing the results obtained using the magnitude spectrum (Table 5) and phase spectrum (Table 6), it becomes evident that in most cases, CosPhase features exhibit superior performance to spectrogram features. For instance, while a TPR value of 71.46% is obtained using spectrogram features when BR1 = 23.85 kbps and BR2 = 18.25 kbps, a considerable relative improvement of approximately 35% is observed using CosPhase features, yielding a TPR value of 96.82%. Similarly, when BR1 = 23.85 kbps and BR2 = 23.05 kbps, CosPhase features yield approximately 38% higher detection rates than spectrogram features. The results summarized in Tables 5 and 6 underscore the considerable impact of phase spectrum features on AMR-WB-coded speech detection, significantly enhancing DC speech detection performance compared to magnitude spectrum features.

Table 5. Detection rates (TPR in %) for DC AMR-WB speech using spectrogram features and the baseline LSTM network.

		BR2 (kbps)								
		6.6	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
BR1 (kbps)	6.6	92.76	95.14	94.72	95.25	94.87	94.56	94.99	95.41	95.14
	8.85	93.09	92.53	93.93	94.03	93.17	94.18	93.87	93.97	96.06
	12.65	89.68	89.12	86.31	86.46	84.64	85.81	86.19	87.48	90.56
	14.25	89.43	88.91	86.50	85.46	84.03	84.62	85.83	85.64	89.49
	15.85	89.30	87.63	85.31	85.39	83.26	84.33	85.41	85.83	89.18
	18.25	88.30	87.09	84.16	83.91	81.35	82.90	83.22	83.32	88.13
	19.85	87.59	87.02	84.03	83.41	80.94	82.25	82.38	83.34	87.63
	23.05	86.50	85.87	82.25	81.63	79.89	80.87	81.19	81.71	86.92
	23.85	79.81	78.13	73.55	73.87	70.02	71.46	72.17	71.92	73.20

Table 6. Detection rates (TPR in %) for DC AMR-WB speech using CosPhase features and the baseline LSTM network. The underlined entries in the table indicate the BR1-BR2 combinations with higher detection rates in comparison to results obtained using spectrogram features in Table 5.

		BR2 (kbps)								
		6.6	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
BR1 (kbps)	6.6	62.90	56.35	85.35	88.28	89.95	93.61	94.81	<u>97.07</u>	94.03
	8.85	65.12	57.23	86.08	89.56	92.19	<u>95.75</u>	<u>96.48</u>	<u>98.49</u>	95.75
	12.65	65.96	55.35	84.56	<u>89.37</u>	<u>90.71</u>	<u>95.14</u>	<u>96.19</u>	<u>98.03</u>	<u>95.60</u>
	14.25	66.40	56.77	86.35	<u>88.26</u>	<u>89.95</u>	<u>95.41</u>	<u>96.61</u>	<u>98.22</u>	<u>95.46</u>
	15.85	66.54	57.30	<u>85.69</u>	<u>89.58</u>	<u>90.96</u>	<u>94.70</u>	<u>95.92</u>	<u>98.41</u>	<u>95.27</u>
	18.25	66.40	57.17	<u>88.45</u>	<u>90.56</u>	<u>91.42</u>	<u>95.20</u>	<u>96.35</u>	<u>98.32</u>	<u>95.64</u>
	19.85	65.96	58.24	<u>86.98</u>	<u>90.08</u>	<u>91.50</u>	<u>95.35</u>	<u>96.31</u>	<u>98.32</u>	<u>95.64</u>
	23.05	68.11	59.22	<u>89.24</u>	<u>91.96</u>	<u>92.63</u>	<u>95.79</u>	<u>96.82</u>	<u>98.32</u>	<u>95.96</u>
	23.85	71.90	70.10	<u>88.47</u>	<u>92.00</u>	<u>93.22</u>	<u>96.82</u>	<u>97.99</u>	<u>99.26</u>	<u>94.60</u>

Finally, we investigate the performance of combining deep features learned through the proposed parallel LSTM system on AMR-WB-coded speech signals. The SC AMR-WB speech detection rates obtained using the parallel LSTM system are shown in Figure 7. Additionally, the TNR values obtained using spectrogram and CosPhase spectrum features with the baseline LSTM system are depicted in the figure to facilitate comparison between the proposed parallel system and the baseline system. It is evident from the figure that the parallel LSTM system outperforms the stand-alone systems regardless of the compression bit rate value. Similarly, CosPhase features yield higher TNR values than spectrogram features, irrespective of the bit rate. These two important observations suggest that (i) phase spectrum features are more representative than magnitude spectrum features for detecting SC AMR-WB speech, and (ii) phase spectrum features and spectrogram features convey complementary information. Combining the deep embeddings learned from these two features through the proposed parallel LSTM network further enhances SC AMR-WB speech detection.

The DC AMR-WB speech detection rates obtained using the proposed parallel LSTM system for each BR1-BR2 combination are summarized in Table 7. In the table, underlined entries indicate BR1-BR2 combinations where the parallel LSTM outperforms the baseline LSTM system with CosPhase features (Table 6), while bold numbers highlight cases where the parallel LSTM yields higher detection rates than the baseline LSTM system with spectrogram features (Table 5). The results reveal that combining deep features learned by the parallel LSTM network from two different sources (magnitude and phase spectrum) considerably improves DC speech detection rates compared to using single CosPhase features (Table 6) for almost all BR1-BR2 combinations. For instance, while a 56.35% detection rate is achieved when BR1 = 6.6 kbps and BR2 = 8.85 kbps using stand-alone CosPhase features, employing the parallel LSTM network, which combines two deep features, increases the

detection rate to 74.64%, representing an approximately 32% performance improvement. The only exception where the parallel LSTM results are slightly lower than the phase spectrum features is when BR2 = 23.85 kbps. However, the performance difference between the two systems (parallel LSTM and LSTM with phase spectrum) is not considerably large. Similarly, the parallel LSTM outperforms the spectrogram features for almost every second compression bit rate, except for 6.6 and 8.85 kbps. Interestingly, when the BR2 value is either 6.6 kbps or 8.85 kbps, the spectrogram features yield better DC speech detection rates than the phase spectrum features or the parallel LSTM system. This observation suggests that magnitude spectrum features are more effective when the BR2 value is very low. With the exception of these two very low BR2 values, the parallel LSTM results are advantageous over stand-alone spectrogram features.

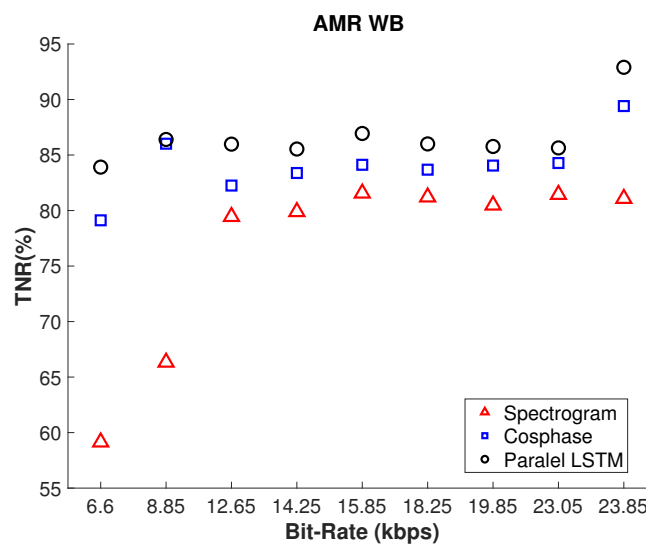


Figure 7. Detection rates (TNR in %) for SC AMR-WB speech obtained using the baseline LSTM system with spectrogram and CosPhase features and the proposed parallel LSTM system.

Table 7. TPR detection rates (in %) for DC AMR-WB speech utilizing spectrogram and CosPhase features with the proposed parallel LSTM network. Underlined entries indicate conditions where the parallel LSTM outperforms the baseline LSTM system with CosPhase features (Table 6), whereas bold numbers indicate cases where the parallel LSTM is superior to the baseline LSTM system with spectrogram features (Table 5).

		BR2 (kbps)								
		6.6	8.85	12.65	14.25	15.85	18.25	19.85	23.05	23.85
BR1 (kbps)	6.6	<u>73.78</u>	<u>74.64</u>	<u>91.31</u>	<u>93.70</u>	<u>93.76</u>	96.33	97.30	98.26	93.61
	8.85	<u>77.09</u>	<u>75.56</u>	<u>93.64</u>	94.89	95.85	97.55	98.47	98.99	96.21
	12.65	<u>75.04</u>	<u>73.01</u>	90.96	<u>93.30</u>	<u>93.43</u>	96.08	97.11	98.45	94.64
	14.25	<u>75.89</u>	<u>73.32</u>	91.29	<u>93.05</u>	<u>93.20</u>	96.56	97.19	98.80	94.49
	15.85	<u>76.31</u>	<u>74.47</u>	91.61	94.03	93.36	96.50	97.25	98.84	94.84
	18.25	<u>76.19</u>	<u>73.53</u>	92.57	93.99	94.58	96.56	97.23	98.74	94.70
	19.85	<u>76.42</u>	<u>75.64</u>	91.48	93.24	93.70	96.86	97.28	98.68	94.81
	23.05	<u>77.40</u>	<u>75.12</u>	93.26	94.72	94.74	96.88	97.63	98.93	95.31
	23.85	<u>79.45</u>	79.79	91.86	94.16	93.95	96.90	98.03	99.22	95.27

8. Discussion and Conclusions

In this study, we investigated phase-spectrum-based features for DC AMR narrowband (AMR-NB) and wideband (AMR-WB) speech detection. To the best of our knowledge, this is the first study to utilize phase spectrum features for DC audio or speech detection tasks. We demonstrated that compressing a speech signal for the first time using the AMR-NB codec significantly affects phase spectrum components above 500 Hz compared to the

uncompressed (original) signal. Additionally, SC speech signals compressed at different bit rates exhibit distinct phase spectra (left panel of Figure 2). Similarly, we observed differences in phase spectra of DC narrowband speech signals at higher frequencies (right panel of Figure 2). Likewise, phase spectra of SC AMR-WB speech signals differ from the original speech signal, and compressing the speech signals for the second time has a considerable impact on the phase spectrum compared to the SC speech spectrum (Figure 3). Based on these observations, it is evident that the phase spectrum can serve as a discriminative feature to distinguish SC and DC-AMR-coded speech signals for both AMR-NB and AMR-WB codecs. We proposed three different phase spectrum-based feature representations for DC AMR speech detection: (i) conventional STFT phase spectrum, where the unwrapped phase spectrum of the STFT of the speech signal was used as the features; (ii) CosPhase features, where cosine normalization was applied to the unwrapped phase spectrum to reduce the dynamic range of the features; and (iii) DCT + CosPhase features, where discrete cosine transform was applied to the cosine normalized features to obtain uncorrelated feature representations. The magnitude spectrum (spectrogram) was also used as the baseline feature to compare the performance of the proposed phase spectrum features.

In addition to proposing phase-spectrum-based features for DC AMR speech detection, we introduced a novel DNN-based classification system based on LSTM networks. LSTM networks are renowned for their exceptional performance on time-series data, and since both the spectrograms and phase spectra serve as time-series representations of magnitude and phase information, respectively, we utilized LSTM networks for DC AMR detection. To the best of our knowledge, this is the first study to propose the use of LSTM networks for DC audio detection. Furthermore, alongside traditional LSTM networks, we proposed a parallel LSTM system for DC AMR speech detection. This parallel LSTM system simultaneously receives both spectrogram and phase spectrum as inputs and processes these features separately to learn the most representative deep features. Subsequently, the learned deep feature representations are concatenated to form a single feature vector, encompassing both magnitude and phase information. This concatenated feature vector is then applied to the fully connected (FC) and classification layers. Our goal with this approach was to investigate whether magnitude and phase spectra convey complementary information for the DC AMR speech detection task.

Experiments conducted on the TIMIT database showed that CosPhase and DCT + CosPhase features yield as good performance as magnitude spectrum features on the AMR-NB codec (Table 1). For example, while the spectrogram feature gave a 99.98% true positive rate (TPR) for AMR-NB-coded DC speech signals, a TPR of 99.84% was obtained using CosPhase features. Conversely, for the AMR-WB codec, CosPhase features were found to be superior to the spectrogram features (86.51% vs. 80.06% TPR values) using the baseline LSTM-based classifier (Table 1).

The superior performance of the CosPhase features was expected because the dynamic range of the unwrapped phase spectrum features exhibits large variations, which possibly makes it difficult to accurately model the phase information. However, applying the cosine function to the phase spectrum reduces the dynamic range of the features and normalizes them within the range $[-1, 1]$. Hence, this helps the classifier learn the discriminative features more accurately. Utilizing the proposed parallel LSTM network using spectrogram and CosPhase features further improved the DC AMR-NB speech detection performance compared to the stand-alone CosPhase features with the baseline LSTM network. For example, stand-alone CosPhase features yielded a 100% TPR for only 9 out of the possible 64 BR1-BR2 combinations (Table 3), whereas the parallel LSTM system with spectrogram and CosPhase inputs achieved a perfect detection rate (100% TPR) in 45 BR1-BR2 cases (Table 4). This observation suggests that the CosPhase features convey complementary information to the spectrogram features; hence, fusing the learned deep features further improves the detection rates for the AMR-NB codec. Additionally, the computational speeds of the proposed baseline LSTM and parallel LSTM systems for processing a single trial (one input signal) were compared. It was found that the baseline LSTM system completes

a prediction in 0.05 s, while the parallel LSTM system requires 0.07 s. The difference in computation time is expected since the parallel LSTM simultaneously propagates two input features (spectrogram and phase spectrum) through two distinct LSTM branches, thereby extending the prediction time. Nonetheless, the prediction times for both systems are relatively short, demonstrating that the proposed system is capable of operating in real time.

Experimental results carried out using the AMR-WB codec showed that detecting DC AMR-WB speech signals is a more challenging task than detecting AMR-NB speech signals. The detection rates obtained using the AMR-WB codec (TNR or TPR values) were found to be considerably lower than those obtained using the AMR-NB codec. However, CosPhase features were found to be significantly superior to spectrogram features in detecting AMR-WB-coded speech signals. For instance, while a TPR value of 71.92% was obtained using spectrogram features when BR1 = 23.85 kbps and BR2 = 23.05 kbps (Table 5), CosPhase features brought approximately a 38% relative improvement, yielding a TPR value of 99.26% (Table 6). The use of both spectrogram and CosPhase features with parallel LSTM further improved SC (Figure 7) and DC AMR-WB speech detection (Table 7) performances independently of the compression bit rates. This again highlights that CosPhase features complement spectrogram features, and the combination of both phase and magnitude spectrum information significantly enhances the detection performance for the AMR-WB codec.

In conclusion, this research makes a significant contribution to DC AMR speech detection by introducing phase spectrum features with deep LSTM networks, focusing on the AMR-NB and AMR-WB speech codecs. Previous studies primarily focused on the narrowband AMR codec (AMR-NB), and while high detection rates were independently reported by other researchers, the performance of the AMR-WB speech codec in DC speech detection remained unknown. Our study reveals that detecting DC AMR-WB speech presents a greater challenge compared to detecting AMR-NB speech. However, the utilization of phase spectrum-based features notably enhanced the detection rates. Furthermore, integrating magnitude and phase spectrum features through the proposed parallel LSTM architecture further boosted the detection rates. These findings underscore the potential of phase spectrum features in DC audio or speech detection tasks, which hold significant relevance from an audio forensics standpoint. Future studies may explore the performance of phase spectrum features on perceptual audio codecs such as MP3 and AAC, expanding beyond speech codecs. Additionally, investigating the efficacy of phase spectrum features on DC AMR speech signals using various DNN architectures and angular margin softmax losses presents promising avenues for further research.

Author Contributions: Conceptualization, A.B. and C.H.; methodology, A.B. and C.H.; software, A.B.; validation, A.B. and C.H.; formal analysis, A.B.; investigation, A.B.; resources, C.H.; data curation, A.B.; writing—original draft preparation, A.B.; writing—review and editing, C.H.; visualization, A.B.; supervision, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Maher, R.C. Audio forensic examination. *IEEE Signal Process. Mag.* **2009**, *26*, 84–94. [[CrossRef](#)]
2. Brandenburg, K.; Stoll, G. ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio. *J. Audio Eng. Soc.* **1994**, *42*, 780–792.

3. Bosi, M.; Brandenburg, K.; Quackenbush, S.; Fielder, L.; Akagiri, K.; Fuchs, H.; Dietz, M. ISO/IEC MPEG-2 Advanced Audio Coding. *J. Audio Eng. Soc.* **1997**, *45*, 789–814.
4. 3GPP TS 26.090-Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions. 2015. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1392> (accessed on 26 March 2024)
5. 3GPP TS 26.190-Speech Codec Speech Processing Functions; Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec; Transcoding Functions. 2022. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1424> (accessed on 26 March 2024)
6. An, N.N.; Thanh, N.Q.; Liu, Y. Deep CNNs With Self-Attention for Speaker Identification. *IEEE Access* **2019**, *7*, 85327–85337. [\[CrossRef\]](#)
7. Abdel-Hamid, O.; Mohamed, A.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional Neural Networks for Speech Recognition. *Ieee/Acm Trans. Audio Speech Lang. Process.* **2014**, *22*, 1533–1545. [\[CrossRef\]](#)
8. Toyoshima, I.; Okada, Y.; Ishimaru, M.; Uchiyama, R.; Tada, M. Multi-Input Speech Emotion Recognition Model Using Mel Spectrogram and GeMAPS. *Sensors* **2023**, *23*, 1743. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Papadimitriou, I.; Vafeiadis, A.; Lalas, A.; Votis, K.; Tzovaras, D. Audio-Based Event Detection at Different SNR Settings Using Two-Dimensional Spectrogram Magnitude Representations. *Electronics* **2020**, *9*, 1593. [\[CrossRef\]](#)
10. Bükér, A.; Hanilçi, C. Deep convolutional neural networks for double compressed AMR audio detection. *Iet Signal Process.* **2021**, *15*, 265–280. [\[CrossRef\]](#)
11. Saratxaga, I.; Sanchez, J.; Wu, Z.; Hernaez, I.; Navas, E. Synthetic speech detection using phase information. *Speech Commun.* **2016**, *81*, 30–41. [\[CrossRef\]](#)
12. Wu, Z.; Chng, E.S.; Li, H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In Proceedings of the Interspeech 2012, Portland, OR, USA, 9–13 September 2012; pp. 1700–1703.
13. Shi, G.; Shanechi, M.; Aarabi, P. On the importance of phase in human speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1867–1874.
14. Nakagawa, S.; Wang, L.; Ohtsuka, S. Speaker Identification and Verification by Combining MFCC and Phase Information. *IEEE Trans. Audio, Speech, Lang. Process.* **2012**, *20*, 1085–1095. [\[CrossRef\]](#)
15. Kim, D.S. Perceptual phase quantization of speech. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 355–364.
16. Pobloth, H.; Kleijn, W. On phase perception in speech. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99), Phoenix, AZ, USA, 15–19 March 1999; pp. 29–32.
17. Paliwal, K.K.; Alsteris, L.D. On the usefulness of STFT phase spectrum in human listening tests. *Speech Commun.* **2005**, *45*, 153–170. [\[CrossRef\]](#)
18. D'Alessandro, B.; Shi, Y.Q. MP3 bit rate quality detection through frequency spectrum analysis. In Proceedings of the 11th ACM Workshop on Multimedia and Security (MM&Sec), Princeton, NJ, USA, 7–8 September 2009; pp. 57–62.
19. Yang, R.; Shi, Y.Q.; Huang, J. Defeating Fake-Quality MP3. In Proceedings of the 11th ACM Workshop on Multimedia and Security (MM&Sec), Princeton, NJ, USA, 7–8 September 2009; pp. 117–124.
20. Qiao, M.; Sung, A.H.; Liu, Q. Revealing real quality of double compressed MP3 audio. In Proceedings of the Proceedings of the 18th ACM International Conference on Multimedia, New York, NY, USA, 25–29 October 2010; pp. 1011–1014.
21. Liu, Q.; Sung, A.H.; Qiao, M. Detection of Double MP3 Compression. *Cogn. Comput.* **2010**, *2*, 291–296. [\[CrossRef\]](#)
22. Qiao, M.; Sung, A.H.; Liu, Q. Improved detection of MP3 double compression using content-independent features. In Proceedings of the 2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2013), Kunming, China, 5–8 August 2013; pp. 1–4.
23. Yang, R.; Shi, Y.Q.; Huang, J. Detecting double compression of audio signal. In *Media Forensics and Security II*; Memon, N.D., Dittmann, J., Alattar, A.M., III, E.J.D., Eds.; International Society for Optics and Photonics; SPIE: San Jose, CA, USA, 2010; Volume 7541, p. 75410K.
24. Bianchi, T.; De Rosa, A.; Fontani, M.; Rocciolo, G.; Piva, A. Detection and Classification of Double Compressed MP3 Audio Tracks. In Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), Montpellier, France, 17–19 June 2013; pp. 159–164.
25. Bianchi, T.; Rosa, A.D.; Fontani, M.; Rocciolo, G.; Piva, A. Detection and localization of double compression in MP3 audio tracks. *Eurasip J. Inf. Secur.* **2014**, *2014*, 10. [\[CrossRef\]](#)
26. Jin, C.; Wang, R.; Yan, D.; Ma, P.; Zhou, J. An efficient algorithm for double compressed AAC audio detection. *Multimed. Tools Appl.* **2016**, *75*, 4815–4832. [\[CrossRef\]](#)
27. Huang, Q.; Wang, R.; Yan, D.; Zhang, J. AAC Audio Compression Detection Based on QMDCT Coefficient. In *Cloud Computing and Security*; Sun, X., Pan, Z., Bertino, E., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 347–359.
28. Huang, Q.; Wang, R.; Yan, D.; Zhang, J. AAC Double Compression Audio Detection Algorithm Based on the Difference of Scale Factor. *Information* **2018**, *9*, 161. [\[CrossRef\]](#)
29. Seichter, D.; Cuccovillo, L.; Aichroth, P. AAC encoding detection and bitrate estimation using a convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2069–2073.

30. Shen, Y.; Jia, J.; Cai, L. Detecting Double Compressed AMR-format Audio Recordings. In Proceedings of the 10th Phonetics Conference of China (PCC), Shanghai, China, 18–20 May 2012.
31. Luo, D.; Yang, R.; Huang, J. Detecting double compressed AMR audio using deep learning. In Proceedings of the ICASSP, Florence, Italy, 4–9 May 2014; pp. 2669–2673.
32. Luo, D.; Yang, R.; Li, B.; Huang, J. Detection of Double Compressed AMR Audio Using Stacked Autoencoder. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 432–444. [\[CrossRef\]](#)
33. Sampaio, J.F.P.; Nascimento, F.A.O. Double compressed AMR audio detection using linear prediction coefficients and support vector machine. In Proceedings of the 22th Brazilian Conference on Automation, João Pessoa, Brazil, 9–12 September 2018.
34. Sampaio, J.F.; de O. Nascimento, F.A. Detection of AMR double compression using compressed-domain speech features. *Forensic Sci. Int. Digit. Investig.* **2020**, *33*, 200907.
35. Büker, A.; Haniçlı, C. Double Compressed AMR Audio Detection Using Long-Term Features and Deep Neural Networks. In Proceedings of the ELECO, Bursa, Turkey, 28–30 November 2019; pp. 590–594.
36. Büker, A.; Haniçlı, C. Double Compressed AMR Audio Detection Using Spectral Features With Temporal Segmentation. In Proceedings of the ELECO, Bursa, Turkey, 25–27 November 2021; pp. 284–288.
37. Büker, A.; Haniçlı, C. Angular Margin Softmax Loss and Its Variants for Double Compressed AMR Audio Detection. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec’21), Virtual, 22–25 June 2021; pp. 45–50.
38. Rabiner, L.; Schafer, R. *Theory and Applications of Digital Speech Processing*, 1st ed.; Prentice Hall Press: Upper Saddle River, NJ, USA, 2010.
39. TIMIT Acoustic-Phonetic Continuous Speech Corpus. 1993. Available online: <https://catalog.ldc.upenn.edu/LDC93S1> (accessed on 26 March 2024).
40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.