ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447419

TOWARDS FASTER END-TO-END DATA TRANSMISSION OVER VOICE CHANNELS

Weijun Zhang, Hao Han, Mingwei Li, Yulong Tian

College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, China

ABSTRACT

As new technologies spread, phone fraud crimes have become master strategies to steal money and personal identities. Inspired by website authentication, we propose an end-to-end data modem over voice channels that can transmit the caller's digital certificate to the callee for verification. Without assistance from telephony providers, it is difficult to carry useful information over voice channels. For example, voice activity detection may quickly classify the encoded signals as non-speech signals and reject the input waveform. To address this problem, we propose a chirp-based modulation method and corresponding demodulator using deep learning technology. We demonstrate that FastDOV works in the real world and outperforms the existing methods in terms of goodput.

***Index Terms*—** Deep Learning, Voice Activity Detection, Modulation, Chirp, Cellular

1. INTRODUCTION

Phone scams and voice phishing (a.k.a. vishing) have mainly increased in the past several years due to the global spread of the COVID-19 pandemic and evolutionary AI technologies. As shown in Fig. 1, scammers and fraudsters are trying to take advantage of the pandemic to trick people into believing that they come from a trusted institution, hospital, or government agency. Some calls may use a fake caller ID [1] to appear more legitimate, as well as explore AI-empowered chatbot [2] and deepfake [3] voices to mimic known persons.

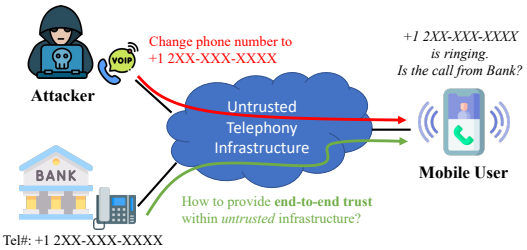


Fig. 1. Use case for FastDOV to prevent phone frauds

To stop phone scams, it is essential to provide end-to-end authentication of conversation parties over traditional (untrusted) phone networks. Like Internet websites, an SSL certificate ensures the authenticity of each website's identity. However, the modern telephony infrastructure provides no means for a callee to reason about the caller's identity except the called ID, which is unfortunately forged. Thus, we need the caller to have the capability to transmit its digital certificate to the callee for authentication. This transmission should be end-to-end without assistance from telephony providers, compatible with existing infrastructure, and not rely on mobile data such as 4G/5G, given that some callees are dumb phones.

One possible solution is to use dial-up modems that have been available for decades to transmit data through telephone lines. However, this approach does not work for mobile phones. That is because the baseband in a smartphone phone that connects to cellular networks is a black box for end users. Without smartphone vendors or network providers, it is nearly impossible for users to implement their "own" data modem on their smartphones. Although a data plan offers an alternative solution to transfer data over cellular networks, it will incur an extra financial cost.

To address these issues, some academic studies have proposed approaches enabling data transmission over the voice channel of cellular networks. Hermes [4] can transmit data over unknown voice channels with the idea of frequency shift keying coding in the voice channel. Work [5] used a single codebook to transmit the voice and designed an efficient low-bit-rate speech coder. The work [6] proposes a DoV technology based on a short harmonic waveform codebook that relies on linear predictive coded speech compression (LPC speech coding) and has a high transmission rate and robustness. AuthLoop [7] provides a strong cryptographic authentication protocol inspired by TLS 1.2 to determine the identity of the entity at the other end of the call (that is, the caller ID). However, those approaches either cannot be implemented in an end-to-end manner or reach the claimed data rate in our experiments.

To this end, we revisit an old topic of enabling data transmission over voice channels and present FastDOV, a fast data transmission mechanism that provides reliable goodput over unknown cellular channels. In FastDOV, we propose a chirp-based modulation/demodulation scheme to resist signal distortion in the complex network infrastructure since chirp signals are proven to be robust to channel noise. We also lever-

Hao Han is the corresponding author (hhan@nuaa.edu.cn). This work was supported in part by NSFC (#61972200).

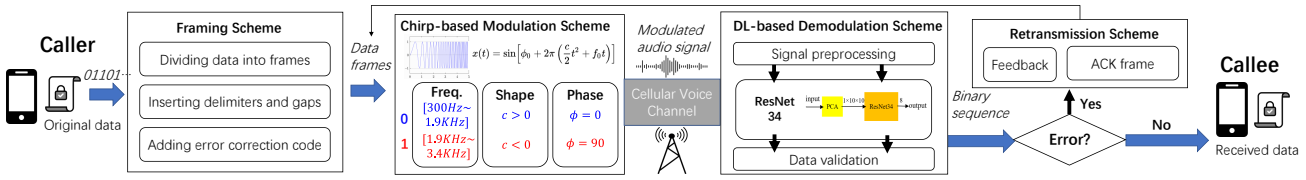


Fig. 2. System framework

age deep learning (DL) technology to demodulate distorted chirp signals to reduce the error rate. To avoid the impact of VAD (Voice activity detection), we propose a dedicated data link protocol with a stop/resume mechanism, time synchronization, and a retransmission scheme between the caller and callee. We implemented a prototype of FastDOV using COTS smartphones and evaluated it through extensive experiments. The results show that FastDOV can reliably achieve the data goodput of 1291.04 bit/s on average over the voice channel in multiple mobile networks, which is a good improvement compared to the state-of-the-art approaches. These advances also testify to the power of carefully introducing Deep Learning (DL) pipelines in advanced communication systems.

The contributions of this work are summarized as follows:

- We propose a novel chirp-based modulation scheme and a DL-based demodulation scheme robust to distortions and interruption in ordinary cellular voice channels.
- We design a dedicated data link protocol for our modulation/demodulation schemes to synchronize the data communication and provide retransmission support.
- We present a working prototype of FastDOV and extensive empirical study under various environmental factors. The evaluation results show that FastDOV achieves higher goodput than state-of-the-art approaches.

2. SYSTEM DESIGN

The objective of FastDOV is to build an end-to-end data modem on top of existing (unmodified) cellular voice channels. The overview of FastDOV is presented in Fig. 2, illustrating the workflow between a transmitter and a receiver. The transmitter first divides the data under transmission into multiple frames according to a pre-defined size. A specially-designed delimiter is inserted at the beginning and the end of each frame. Each frame is appended with the error correction code and converted into an audio signal using a chirp-based modulation. The modulated chirps are then transmitted over the cellular voice channel.

When the phone call is answered, the receiver relies on the delimiter to locate each frame's beginning and end within the received audio stream. These frames are demodulated and decoded using a Deep Learning (DL) algorithm, followed by the error correction check. If any error is found, the proposed feedback mechanism is used to notify the transmitter to retransmit frames in error. Lastly, all successful frames are combined to restore the origin data.

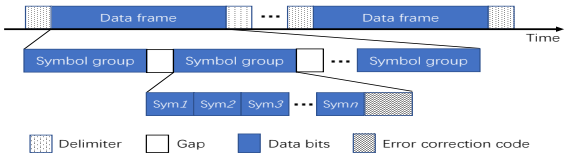


Fig. 3. The format of data frames

2.1. Framing Scheme

In FastDOV, data are divided into *frames* and transmitted over the voice channel. The format of the data frames is presented in Fig. 3. A specially-designed signal is inserted to separate each data frame. A data frame comprises multiple *symbols* and parities for error correction. Each symbol carries a fixed length of data bits. To avoid the impact of VAD which is a technology used for detecting whether a voice signal exists [8] and to remove non-voice clips during audio sessions. we add a gap after a group of symbols within a data frame. Such a gap is an empty signal without carrying any information, removing the memory of the VAD algorithm.

Determining the start/end of transmission. We designed a unique chirp as a delimiter, inserted at the start of every frame and appended to the last frame to indicate when the data frames begin and end. It ensures synchronization between the receiver and transmitter. We also find that the sudden energy change on a voice channel may cause signal fluctuation, so the frontmost and backmost data signals may not be received completely. Thus, the inserted delimiter is also treated as a guard to protect our data signals.

To detect the exact position of a delimiter on the receiver side, we adopt a cross-correlation-based method, where the known delimiter signal is correlated with the received audio stream in a sliding window. Let the received signal stream as $\{u_i\}$ where $i = 1, 2, \cdots, n$ and each u_i is an audio sample. The delimiter emitted by the transmitter is denoted as $\{v_i\}$ where $i = 1, 2, \cdots, m$ and $n >> m$. A sliding window with a length equal to m is extracted from $\{u_i\}$ using matched filtering. The sample correlation coefficient r is computed as follows:

$$r = \frac{\sum_{i=1}^{m}(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \overline{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \overline{v})}}$$

where $\overline{u}$ and $\overline{v}$ are the sample means of the sliding window

and $\{v_i\}$, respectively. For each r, Welford's one-pass algorithm can achieve the computational complexity $O(m)$. As the sliding windows move from the beginning to the end of the audio stream sample by sample, each r is computed with the total complexity $O(nm)$. A large value of r means a high similarity between two sequences. The position of the delimiter can be found at the maximal peak of these coefficients.

We define some known symbols in an agreement with the sender and receiver. Based on that, we can adjust the position of delimiters to avoid mis-synchronization.

Error correction and retransmission. Since some bits could be transmitted incorrectly, in FastDOV, we use Reed-Solomon code (RS) [9] as our error correction code which will be added at the end of each symbol group. When the error correction mechanism cannot handle the exception, FastDOV will simply retransmit the signal.

2.2. Chirp-based Modulation Scheme

In FastDOV, we adopt a chirp signal rather than sinusoidal waves used in [4] and [10] for modulation. Chirp has a wide range of applications in communication, sonar, radar, and other fields due to its strong anti-interference feature. We use linear-frequency chirp since this chirp is enough to tolerate noise and signal distortion on the cellular voice channel. We modulate 3-bit information by varying the frequency, shape, and phase of our chirp signal (1 bit for frequency, 1 bit for shape, and 1 bit for phase).

Frequency. The voice channel of the GSM is designed to transfer speech only, and the energy that reflects the main characteristics of human speech mainly concentrates in the range of $[300Hz, 3400Hz]$. Hence, we use $f_0 = 300Hz$ and $f_0 = 1.9KHz$ to encode bits 0 and 1, respectively. Bit 0 is represented by the frequency range of $[300Hz, 1900Hz]$, and bit 1 is modulated by the range of $(1900Hz, 3400Hz]$.

Shape. In the frequency range of either $[300Hz, 1900Hz]$ or $(1900Hz, 3400Hz]$, we can switch the starting frequency f_0 and finish frequency f_1 to encode extra 1-bit information. In other words, we change the shape of the chirp without changing its frequency band. For example, the chirp with $300Hz \rightarrow 1.9KHz$ represents the bit 0, and $1.9KHz \rightarrow 300Hz$ for the bit 1.

Phase. Without changing the frequency and shape of a chirp signal, we can modulate an extra bit by using different initial phases. Whether the signal carries 0 or 1 depends on whether the initial phase ϕ_0 equals zero. Thanks to the proposed DL-based demodulation scheme, we can reliably detect the phase difference of each symbol.

2.3. DL-based Demodulation Scheme

Feature extraction. We use time-domain, frequency-domain, and phase-angle features of each symbol for demodulation. In order to improve recognition accuracy, we extract frequency-domain features from three sources, including all the sampling points, the first half sampling points, and the latter half sampling points. The frequency-domain features consist of three parts. The first part is extracted from all the sampling points. The second and third parts are extracted from the first half of the sampling points and the latter half of the sampling points, respectively.

Demodulation preprocessing. Since the raw features could be high dimensional (could exceed 1000), we first use PCA (Principal Component Analysis) [11] to reduce the number of the dimensions of the features to 100. PCA is a commonly used method for dimension reduction. It works by performing linear projections and can mitigate overfitting problems. We then use a deep learning model (e.g., ResNet34 [12]) to demodulate the features generated by PCA. The deep learning model takes an input of $1 \times 10 \times 10$ (the output of PCA will be resized into this size) and ouputs 8 classes (8 different symbols described in Section 2.2).

3. EVALUATION

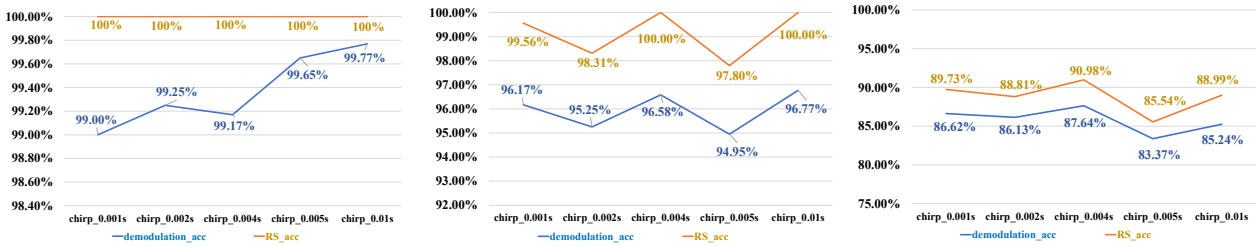
3.1. Experimental Methodology

Experimental settings. Our experiments involve two smartphones, which communicate through cellular networks, and two laptops that help process the data. Specifically, we use a Huawei P60 to call a Xiaomi Redmi Note 7. Due to the restriction of security policies of the Android system, we cannot directly manipulate phone calls without rooting the phone. Hence, on the transmitter side, we use a laptop (Lenovo Y7000) as an external Bluetooth speakerphone of the Huawei phone leveraging the HFP (Hands-free profile) protocol [13]. We directly modulate the voice signals on the laptop and then feed those modulated signals into the external speakerphone. On the receiver side, the Xiaomi phone records phone calls, and another Lenovo Y7000 laptop will demodulate those recorded phone calls.

In our experiments, FastDOV was used to transmit a predefined certificate. The audio sampling rate is 48000Hz by default. Each symbol group contains interleaved symbols and gaps with durations of 0.6s and 0.5s, respectively. RS level is set to 3:1 (one parity bit per 3 data bits).

Model training. The ResNet34 model used for demodulation will firstly be initialized by parameters pre-trained on ISLVRC2012 dataset [14] and then fine-tuned by 28800 modulation symbols collected in our experiments. 70% of those symbols are used for training, and the remainder are used for test. We use cross-entropy loss and SGD (stochastic gradient descent) optimizer with the weight decay set to 0.8 in the model training. We fine-tune the model for 40 epochs using a batch size of 64. The learning rate is 0.0005 for the first 30 epochs and decreases to 0.00005 for the latter 10 epochs.

Performance metrics. We use accuracy, throughput, and



(a) When cellular signal is **strong** (-90dBm< RSRP <-85dBm, 15< SINR <20)

(b) When cellular signal is **medium** (-100dBm< RSRP <-90dBm, 11< SINR <15)

(c) When cellular signal is **weak** (-106dBm< RSRP <-100dBm, 0< SINR <6)

Fig. 4. Accuracy of DL-based demodulation

Table 1. Goodput (bit/s) when using different cellular carriers

Receiver \ Transmitter	China Mobile	China Telecom	China Unicom
China Mobile	1270.69	1265.43	1267.14
China Telecom	1266.72	1271.83	1260.28
China Unicom	1268.47	1262.13	1269.56

goodput to evaluate our implementation. *Accuracy* is the demodulation accuracy of classifying modulation symbols. *Throughput* is the amount of all the received bits, including protocol overhead bits and duplicated bits per unit of time. *Goodput* is the amount of useful information delivered to the receiver per unit of time and is the size of the transmitted file divided by the time it takes to transfer the file.

3.2. Experimental Results

The pattern and length of a delimiter are important parameters of FastDOV. The delimiter pattern can be quadratic, linear, hyperbolic, or logistic; the delimiter length can be any value. By conducting multiple experiments, we confirm that these two factors do not significantly affect the performance of FastDOV. Therefore, in our main experiments, we will simply use the quadratic delimiter with a length of 0.1s. We also test FastDOV on networks provided by China's three major cellular carriers: China Mobile, China Telecom, and China Unicom. Our experiments find that FastDOV has similar goodput in all the cellular networks (Table 1). Hence, we only use China Mobile for our main experiments.

Real-world Performance. We tested the reliability of FastDOV without retransmission using different symbol lengths by varying signal strengths (strong, medium, and weak in Fig. 4). Fig. 4 summarizes the results. It is seen that the demodulation accuracy decreases as the cellular signal strength decreases. FastDOV achieves high accuracy in most cases. When the signal strength is strong or medium (Fig. 4(a) and Fig. 4(b)), the DL model always achieves very high demodulation accuracy ($> 94\%$, the blue line). If we additionally use an RS-based error correction code to process the output of the DL model, the demodulation accuracy will increase to $> 97\%$ (the red line). When the signal strength is weak (Fig. 4(c)), the demodulation accuracy is less than 90%. We note that when the signal is weak, the channel is unstable and even unusable for human voice communication. Thus, a demodulation accuracy above 83% is acceptable.

Table 2. Throughput and goodput under different symbol lengths (bit/s)

Symbol length	0.001s	0.002s	0.004s	0.005s	0.01s
Throughput	1785.12	853.75	427.72	347.83	173.49
Goodput(Strong)	1338.84	640.32	320.79	260.87	130.12
Goodput(Medium)	1332.95	629.50	320.79	255.13	130.12
Goodput(Weak)	1201.34	568.67	291.85	223.15	115.79

Table 3. Performance comparison between various methods

Method	Throughput/bps	Goodput/bps	Modulator	Synchronous	Retransmission
FastDOV	1785.12	1291.04	hybrid	✓	✓
Hermes	1200	NA	FKS	✗	✗
Authloop	NA	500	3-FKS	✓	✓

Table 2 reports the throughput and goodput of the experiments in Fig. 4. The results show that the throughput and goodput decrease as the symbol length increases. Hence, the signal length should be set as a small value in practice. Our design achieves excellent goodput, 1338.84 bit/s when the signal is strong and 1201.34 bit/s when the signal is weak.

Performance comparison. We compared FastDOV with other representative data transmission over voice channel methods, including Hermes [4] and Authloop [7]. Table 3 shows the result. Our proposed FastDOV outperforms Hermes and Authloop by a large margin. The throughput of FastDOV is 1785.12 bit/s, while Hermes only achieves a throughput of 1200bit/s. The averaged goodput of FastDOV is 1291.04 bit/s (the symbol length is 0.001s), while Authloop only have a goodput of 500 bit/s. Additionally, FastDOV and Authloop have synchronization and retransmission designs which can ensure the signal received is complete and correct, but Hermes overlooks those designs.

4. CONCLUSION

We propose a deep-learning-based acoustic modulation scheme named FastDOV to transmit data over mobile voice channels without infrastructure support. FastDOV consists of a novel chirp-based modulation and a tailored ResNet34 model for demodulation. FastDOV can achieve 1291.04 bit/s goodput on average, which is better than existing approaches. In addition, we demonstrate that our system can transfer digital certificates over voice channels to prevent telecom fraud.

5. REFERENCES

[1] H. A. Mustafa, "Secure and reliable wireless communication through end-to-end-based solution," *Dissertations & Theses - Gradworks*, vol. 42, no. 2, pp. 555–63, 2014.

[2] Msi Bhuiyan, A. Razzak, M. S. Ferdous, Mjm Chowdhury, and S. Tarkoma, "Bonik: A blockchain empowered chatbot for financial transactions," in *19th IEEE Intl. Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2020)*, 2020.

[3] Eric Tjon, Melody Moh, and Teng-Sheng Moh, "Effynet: A dual task network for deepfake detection and segmentation," in *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 2021, pp. 1–8.

[4] Aditya Dhananjay, Ashlesh Sharma, Michael Paik, Jay Chen, Trishank Karthik Kuppusamy, Jinyang Li, and Lakshminarayanan Subramanian, "Hermes: Data transmission over unknown voice channels," in *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2010, MobiCom '10, p. 113–124, Association for Computing Machinery.

[5] Fauzia I Abro, Farzana Rauf, BS Chowdhry, and Muttukrishnan Rajarajan, "Towards security of gsm voice communication," *Wireless Personal Communications*, vol. 108, no. 3, pp. 1933–1955, 2019.

[6] Piotr Krasnowski, Jerome Lebrun, and Bruno Martin, "Introducing a novel data over voice technique for secure voice communication," *Wireless Personal Communications*, vol. 124, no. 4, pp. 3077–3103, 2022.

[7] Bradley Reaves, Logan Blue, and Patrick Traynor, "{AuthLoop}:{End-to-End} cryptographic authentication for telephony over voice channels," in *USENIX Security Symposium*, 2016, pp. 963–978.

[8] T. Chmayssani and G. Baudoin, "Data transmission over voice dedicated channels using digital modulations," in *Radioelektronika, International Conference*, 2008.

[9] Tadao Kasami and Shu Lin, "The binary weight distribution of the extended (2m,2m - 4) code of the reed-solomon code over gf(2m) with generator polynomial $(x-\alpha)(x-\alpha 2)(x-\alpha 3)$," *Linear Algebra and its Applications*, vol. 98, no. 1, pp. 291–307, 1988.

[10] T. Zhang, S. Li, and B. Yu, "A universal data transfer technique over voice channels of cellular mobile communication networks," *IET Communications*, 2020.

[11] Takio Kurita, "Principal component analysis (pca)," *Computer Vision: A Reference Guide*, pp. 1–4, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] SOURCEFOURGE, "Hfp for linux," `https://nohands.sourceforge.net/`.

[14] Princeton University Stanford Vision Lab, Stanford University, "Imagenet large scale visual recognition challenge 2012 (ilsvrc2012)," `https://www.image-net.org/challenges/LSVRC/2012/index.php`, 2020.