Transcoding: A New Strategy for Relay Channels

Chih-Chun Wang, David J. Love, and Dennis Ogbe
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907, USA

Abstract— **The relay channel is a traditional information-theoretic problem which has important applications in the Internet of Things (IoT) and other future communication networks. In this work, we focus on the simplest possible relaying model: the so-called *separated* (or *two-hop*) relay channel where there is no direct link between the source and the destination. Previous work has shown that for this channel, the decode-and-forward (DF) relaying strategy is capacity-achieving under the assumption of asymptotic block lengths. In this paper, however, we are interested in the finite-delay regime where simpler sub-optimal techniques like amplify-and-forward (AF) can be used to avoid the need for buffering at the relay. We present a new strategy called *transcoding* which presents a tradeoff between the low-latency advantages of amplify-and-forward and the high-rate, high-latency decode-and-forward scheme. Our results indicate that our simple, intuitive transcoding schemes outperform traditional relaying schemes in the finite-delay regime.**

I. Introduction

Arguments about the Shannon capacity of a communication channel typically involve the assumption of a channel code with infinite blocklength. Practical communication systems, however, are always bound to coding schemes with a finite number of symbols per block. The impact of this restriction on the achievable rate was studied in [1] for single-hop communication systems. It is of no surprise that the tight bound on the block error probability discovered in [1] increases as the blocklength decreases. In addition to the interest from the academic community, the performance characteristics of finite-blocklength systems are of interest to industry, with low latency systems being one of the key technology goals for the upcoming fifth-generation (5G) wireless standards [2].

This paper seeks to apply the lessons learned from the analysis of finite-blocklength systems to relay channels. The relay channel model applies in some form to most of the commercially available communication systems today and has been of considerable interest to the academic community since the 1970s [3], [4]. Throughout the years several different relaying techniques have emerged in the literature, including compress-and-forward [5], hash-and-forward [6], compute-and-forward [7] and noisy network coding [8]. However, the two most well-known techniques are decode-and-forward [4], [9], where the relay node decodes the data it receives and forwards the decoded data, and amplify-and-forward [10], where the relay simply scales and retransmits its received symbols. For simple relaying models it is known that decode-and-forward is optimal from a capacity perspective and amplify-and-forward is optimal in terms of delay.

The analysis of relay channels in the context of finite block lengths became a topic in academia fairly recently, with only a limited number of publications on the topic. Works such as [11] and [12] explore the capacity of these systems under the decode-and-forward scheme and a cooperative two-hop setting, which implies a direct link between the source and the destination. Work in this area has also considered multi-hop scenarios [13], [14]. The commonalities of most of the prior work in this field are twofold:

1) Most of the work was written in the context of wireless communication systems and Rayleigh fading channel models.
2) To the best of our knowledge, all of the work in the field has centered around the analysis of existing relaying strategies (DF, AF, etc.) which are optimal only in asymptotic blocklength regimes.

In contrast, this paper presents a new strategy for the relaying problem, designed with finite blocklengths in mind. This strategy, dubbed *transcoding*, presents a new middle ground between the low-latency amplify-and-forward approach and the high-latency, computationally intensive decode-and-forward technique. Our focus is on the tradeoff between achievable rate and delay across the relay channel and we present results indicating increased performance over DF and AF at equal delay. Our strategy is based on intuitive ideas from the concatenated coding literature applied to the relay channel model.

The rest of the paper is organized as follows. In Section II, we give an outline of our system model and problem statement. In Section III we give a general description of the transcoding idea. Section IV presents an example of transcoding over a relay channel consisting of binary symmetric channels compared to the decode-and-forward scheme. Finally, we conclude and discuss future work in Section V.

II. Problem Set-up

This work focuses on the two-hop relay channel model (sometimes referred to as a *separated* relay), which is depicted in Figure 1. We assume direct links between the source and the relay as well as between the relay and the destination. There is no direct link between the source and the destination. We further assume discrete memoryless channels between all nodes. We suppose that the source wishes to transmit L information bits to the destination, which we write in vector notation as $\mathbf{s} = [s_0, \cdots s_{L-1}]^{\mathsf{T}}$. This information is transmitted by the source in the form of blocks of channel symbols $x_{S,t} \in \mathcal{X}_S$, where $\mathcal{X}_S$ represents the input alphabet of the source-to-relay channel and t denotes a time index. We assume that these coded blocks $\mathbf{x}_S = [x_{S,t}, x_{S,t+1}, \cdots, x_{S,t+\delta_S-1}]^{\mathsf{T}}$ contain δ_S symbols and that each symbol is transmitted over a period of T_1 seconds. The source obtains $\mathbf{x}_S$ from its information symbols by the application of some encoding operation $\mathcal{S}(\cdot)$, i.e., $\mathbf{x}_S = \mathcal{S}(\mathbf{s})$. The transmit blocks enter the source-to-relay channel and emerge at the relay as input vector $\mathbf{y}_R$, where each element, denoted $y_{R,t}$ is drawn from the channel's output alphabet $\mathcal{Y}_R$. The relay operates on blocks of symbols of size $\delta_{R,Q}$, which we will refer to as "sub-block". We denote the k-th received sub-block at the relay node as $\mathbf{y}_{R,k}$. We further assume that a transmit block consists of K sub-blocks, i.e. $\delta_S = K\delta_{R,Q}$ and $\mathbf{y}_R = [\mathbf{y}_{R,0}^{\mathsf{T}}, \cdots \mathbf{y}_{R,K-1}^{\mathsf{T}}]^{\mathsf{T}}$.

For every received sub-block, the relay forwards a deterministic mapping $\mathbf{x}_{R,k} = \mathcal{R}(\mathbf{y}_{R,k})$ through the next channel to the destination, where $\mathbf{x}_{R,k}$ consists of $\delta_{R,C}$ symbols drawn from some input alphabet $\mathcal{X}_R$ with a transmit period T_2. We thus have for a full block transmitted from the relay to the destination,

$$\begin{aligned}\mathbf{x}_R &= \left[x_{R,t}, \cdots, x_{R,t+K\delta_{R,C}-1}\right]^{\mathsf{T}} \\ &= \left[\mathbf{x}_{R,0}^{\mathsf{T}}, \cdots, \mathbf{x}_{R,K-1}^{\mathsf{T}}\right]^{\mathsf{T}} \\ &= \left[\mathcal{R}(\mathbf{y}_{R,0})^{\mathsf{T}}, \cdots, \mathcal{R}(\mathbf{y}_{R,K-1})^{\mathsf{T}}\right]^{\mathsf{T}}. \qquad (1)\end{aligned}$$

The destination then observes its channel output vector $\mathbf{y}_D$, drawn from some alphabet $\mathcal{Y}_D$. It applies a decoding function $\mathcal{D}(\cdot)$ to this received vector to obtain its estimate of the information symbols, denoted as $\widehat{\mathbf{s}} = \mathcal{D}(\mathbf{y}_D)$. We declare a codeword error when $\widehat{\mathbf{s}} \neq \mathbf{s}$.

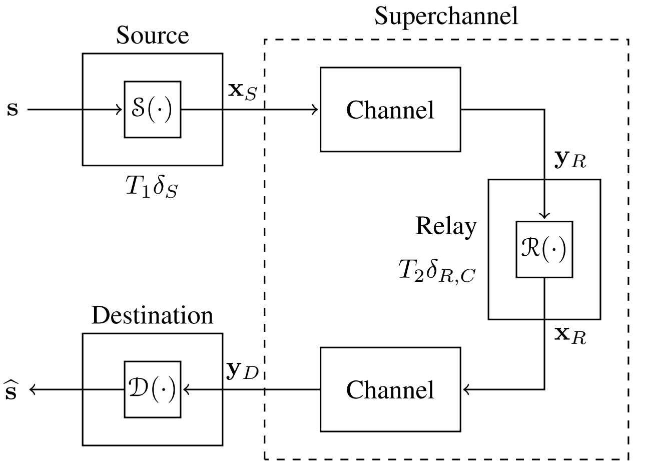

Fig. 1. Two-hop (separated) relay channel

As indicated in Fig. 1, one straightforward abstraction in this set-up is to view the concatenation of the source-to-relay channel, the mapping at the relay $\mathcal{R}(\cdot)$, and the relay-to-source channel as an effective "superchannel", a term borrowed from the literature on concatenated coding [15]. The goal of this work is to present a new way of thinking about the design of channel codes for models like this. Our focus is furthermore on the tradeoff between the achievable rate and the total delay over the superchannel. The first question to ask is then, "What are the sources of delay in this model?" The answer is that, clearly, delay is introduced at two points in the model.

1) At the source, which transmits a block of δ_S channel symbols with symbol period T_1.
2) At the relay, which first accumulates $\delta_{R,Q}$ symbols and then transmits a block of $\delta_{R,C}$ symbols with symbol period T_2.

A straightforward definition of the overall delay in our system model is then given as

$$\Delta = T_1\delta_S + T_2\delta_{R,C}, \qquad (2)$$

where we note that the symbol periods T_1 and T_2 must satisfy $T_1\delta_{R,Q} = T_2\delta_{R,C}$, i.e., the transmission periods for received and transmitted sub-blocks at the relay must be of the same length. Furthermore, the achievable rate in units of bits per channel use is given as

$$R_{\mathsf{total}} = \frac{L}{T_1\delta_S}. \qquad (3)$$

The model from Fig. 1 is held intentionally general and can be applied to a variety of relaying techniques. For example, with the decode-and-forward relaying strategy, we are limited to choices of block lengths that satisfy $\delta_S = \delta_{R,Q}$. The relay forwards decoded and re-encoded sub-blocks to the destination as soon as they are received. This limitation causes increased delay over the superchannel, a concept illustrated in Fig. 2. Here, the total delay between the start of the encoding operation at the source and the time when the destination extracts its message satisfies

$$\begin{aligned}\Delta_{DF} &= T_1\delta_S + T_2\delta_{R,C} \\ &= 2T_1\delta_S. \qquad (4)\end{aligned}$$

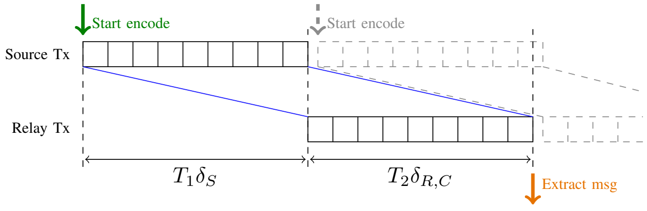

Fig. 2. Sources of delay with the decode-and-forward strategy

In contrast, the amplify-and-forward strategy (shown in Fig. 3) fixes $\delta_{R,Q} = \delta_{R,C} = 1$, which leads to a low overall delay. The strength of the model given in this section is that it allows us to describe more general relaying schemes than amplify-and-forward and decode-and-forward. The scheme from Section III exploits this flexibility and presents a middle ground between the high throughput, high delay extremum (decode-and-forward) and the low throughput, low delay extremum (amplify-and-forward).

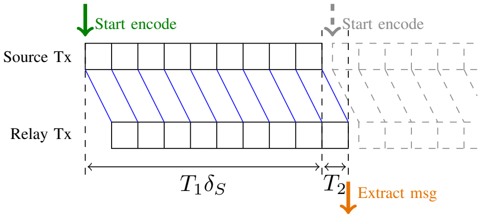

Fig. 3. Sources of delay with the amplify-and-forward strategy

At this point we should note that the relationship between the achievable rate and the total delay is more subtle than (2) and (3) seem to indicate. The achievable rate is, of course, coupled to the desired average probability of block error (denoted as ϵ), which decreases as δ_S increases. Thus, the main question that this work seeks to answer is then, "Given a desired average probability of error ϵ, what is the tradeoff between the achievable rate R_{total} and the maximum allowed delay Δ across the relay channel for different relaying strategies $\mathcal{R}(\cdot)$?"

Some intuitive insight into this question can be gained from Cover and El Gamal's seminal paper on relay channels [4]. Suppose we are interested in the extreme case with $\epsilon \to 0$. Theorem 1 in [4] shows that we can communicate at capacity with a decode-and-forward scheme where $\delta_S \to \infty$. The case for fixed $\epsilon > 0$ and fixed delay is less clear and our discovery of the transcoding principle shows that there is room to increase our rate with smart relaying techniques.

III. The Transcoding Principle

The main idea of the transcoding principle is to allow the relay to perform an *arbitrary multidimensional mapping* from its input symbols to its output symbols. More specifically, the relay function can be any function satisfying

$$\mathcal{R}(\cdot) : (\mathcal{Y}_R)^{\delta_{R,Q}} \mapsto (\mathcal{X}_R)^{\delta_{R,C}}. \qquad (5)$$

This can be seen as a more general approach than decode-and-forward, which is restricted to decoding and possibly re-encoding its received sub-blocks. When designed properly, these arbitrary mappings have the potential to increase the achievable rate over the superchannel. Fig. 4 illustrates the total delay over the superchannel with the transcoding strategy. Here, the relay processes sub-blocks of sizes $1 \leq \ \delta_{R,Q} \ \leq \delta_S$, which results in better delay performance than decode-and-forward without sacrificing the ability to correct some errors at the relay.

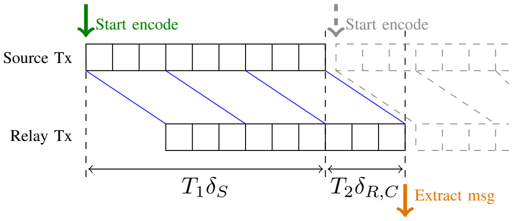

Fig. 4. Sources of delay with the transcoding strategy

To enable these mappings at the relay, the source first encodes its L message symbols with a capacity-achieving block code, the *outer code*, of rate $R_{\mathsf{outer}} = L/L_1$ and then encodes K sub-blocks of its length-L_1 codeword with a rate $R_{\mathsf{inner}} = L_2/\delta_{R,Q}$ *inner code*, where $L_1 = KL_2$ and $\delta_S = K\delta_{R,Q}$. This operation is visualized in Figure 5.

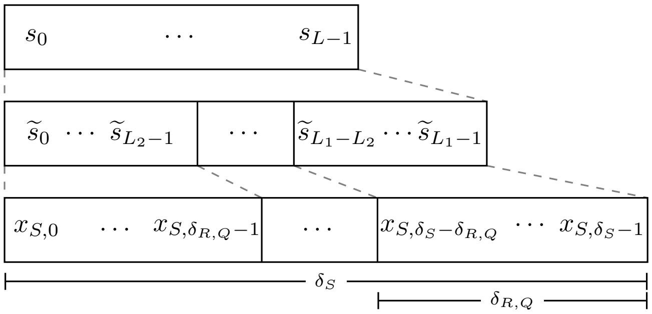
s_0
$\cdots$
s_{L-1}
$\tilde{s}_0 \ \cdots \ \tilde{s}_{L_2-1}$
$\cdots$
$\tilde{s}_{L_1-L_2} \cdots \tilde{s}_{L_1-1}$
$x_{S,0} \quad \cdots \quad x_{S,\delta_{R,Q}-1}$
$\cdots$
$x_{S,\delta_S-\delta_{R,Q}} \quad \cdots \quad x_{S,\delta_S-1}$
δ_S
$\delta_{R,Q}$

Fig. 5. Block coding at the source

The relay then processes codewords of the inner code and, instead of simply decoding or amplifying, forwards an arbitrary deterministic mapping of the received codeword to the destination. The idea here is that a smart joint design of the outer code and the mapping of the inner code will outperform less sophisticated schemes. Furthermore, similar to decode-and-forward schemes, the mapping can be designed to match the two channels at each end of the relay. As a further point of intuition, a well-designed mapping function could be used to correct only a subset of errors introduced by the first channel, leaving the rest to the outer code at the destination.

The transcoding principle is best explained with a specific example of a mapping function. Suppose, for simplicity, that all channel alphabets are binary and both component channels of our superchannel are binary symmetric channels. Furthermore, suppose that the codeword space of the inner code is a subspace of space of all sequences of binary digits of length $\delta_{R,Q}$ with the Hamming distance metric. Fig. 5 shows as an example a space with three codewords (the filled black circles) separated by some minimum distance d_{min}. The crosses represent two different received sub-blocks, corrupted by the source-to-relay channel. In our example mapping, the relay transmits the *codeword* corresponding to the received sub-block, as long as the received sub-block is within some radius d_{TC} (dashed circle around codeword 1) of a valid codeword. In the depicted example, the relay would forward codeword 1 upon reception of the "blue" sub-block, correcting the errors introduced by the source-to-relay channel.

For cases in which a received sub-block falls exactly between two valid codewords (red cross in Fig 6), there are two potential courses of action:

1) *Forced decoding*
With this strategy, the relay always transmits a valid codeword, i.e. for the "red" sub-block, the relay would choose between codewords 1 and 2 with equal proba-

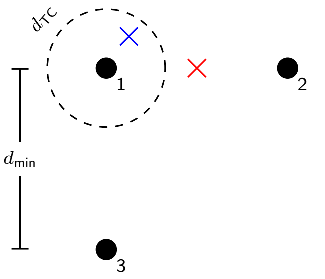

Fig. 6. Sub-block mapping at the relay

bility.

2) *Partial decoding*
With this strategy, the relay transmits sub-blocks not within d_{TC} of any valid codeword as-is, without any correction. This results in a propagation of those errors to the destination. Sub-blocks close enough to valid codewords are corrected as described above.

It is evident that with this example mapping our delay distribution is $\delta_{R,Q} < \delta_S$. As the example in Section IV will show, the partial decoding technique provides better performance than the forced decoding technique. This observation seems to verify our intuition: If the relay cannot decode a sub-block (as is the case for the red sub-block in Fig. IV), letting the longer outer code take care of the errors is less costly than decoding the wrong codeword.

Note that the example given in this paper is of course only one out of many possible options for the mapping function $\mathcal{R}(\cdot)$. In practice, deriving a good mapping is a design problem which will consist of some combination of heuristics and computer-based search methods. Properties of "good" mapping functions will most likely involve a notion of algebraic structure and, depending on the application, a short sub-block length.

IV. Transcoding for the Binary Symmetric Channel

The example in this section seeks to illustrate the advantages of simple transcoding schemes when compared to the traditional DF and AF techniques. We model both channels (source → relay and relay → destination) as binary symmetric channels with transition probabilities p_1 and p_2, respectively. The input and output alphabets on all nodes thus consist of binary digits, i.e. $\mathcal{X}_S = \mathcal{Y}_R = \mathcal{X}_R = \mathcal{Y}_D = \{0, 1\}$. In Figs. 7 and 8, we compare the rate-blocklength tradeoff using both Gallager's random coding exponent [16] and Polyanskiy's normal approximation of the coding theorem [1]. Using either framework, we compute the achievable rate over the superchannel as follows:

1) Fix the desired average probability of error ϵ
2) For a given relaying strategy $\mathcal{R}(\cdot)$ with sub-block sizes $\delta_{R,Q}$ and $\delta_{R,C}$, compute the transition probabilities of the superchannel $p(j \mid k)$
3) Find the block size δ_S and input distribution $\mathbf{Q}$ which give the desired error probability over the superchannel
4) The delay over the superchannel is then given by (2) and the achievable rate by (3)

In Fig. 7 (Gallager bound), the result of step 3 was obtained by assuming a uniform input distribution $\mathbf{Q}$ and finding the δ_S which maximizes [16]

$$\Pr(\widehat{\mathbf{s}} \neq \mathbf{s}) = \epsilon \leq \exp\left[-\delta_S E_r(R_{\text{total}})\right], \tag{6}$$

where the error exponent $E_r(R_{\text{total}})$ is given as

$$E_r(R_{\text{total}}) = \max_{0 \leq \lambda \leq 1} \max_{\mathbf{Q}} E_o(\lambda, \mathbf{Q}) - \lambda R_{\text{total}}. \tag{7}$$

Here, an optimization has to be performed over λ, the channel input distribution $\mathbf{Q}$, and

$$E_o(\lambda, \mathbf{Q}) = -\ln \sum_{j=0}^{J-1} \left[\sum_{k=0}^{K-1} Q(k) p(j \mid k)^{1/(1+\lambda)} \right]^{1+\lambda}. \tag{8}$$

In Fig. 8 (Gaussian approximation), the result of step 3 was obtained by again assuming a uniform input distribution $\mathbf{Q}$ and then finding δ_S which satisfies

$$L = \delta_S C - \sqrt{\delta_S V} Q^{-1}(\epsilon) + \frac{1}{2}\log_2(\delta_S) \tag{9}$$

(see [1], Sec. IV). Here, L is, as defined earlier, the number of information bits at the source, C is the capacity of the superchannel, $Q^{-1}(\cdot)$ is the inverse of the Gaussian tail distribution function $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ dt, and the channel dispersion V is used as defined in [1], eqn. (239). We plot our comparison for two different bounds in order to show that the trends we observe hold for different approximations.

In both figures, we use the (8,4) extended Hamming code [17] as inner code and plot the rate-blocklength tradeoff for the partial decoding technique (with $d_{\text{TC}} = 1$) as well as the forced decoding technique. We fix the error probability to $\epsilon = 10^{-3}$ and the component channel transition probabilities to $p_1 = 0.04$ and $p_2 = 0.13$ for the source-to-relay and the relay-to-destination channels, respectively. To put the performance of our proposed techniques in context, we plot curves for the decode-and-forward scheme and the single-hop upper bound for this superchannel, which considers a transmission over a single BSC with transition probability $p = 0.13$.

The results presented in Figs. 7 and 8 characterize the increased performance of the transcoding principle in the low-delay regime. Even though we observe a crossover point between the transcoding curves and the ones for decode-and-forward in both plots, transcoding delivers significant performance improvements for short blocklengths. In Fig. 7, we can observe an approx. 16% increase in achievable rate for $100 \leq \Delta \leq 200$ when comparing decode-and-forward with our transcoding example using partial decoding. In addition, when using the Gaussian approximation (Fig. 8), we observe an approx. 29% advantage of the partial decoding scheme over decode-and-forward for values of Δ around 100 bits. We believe that examples like these show the potential of smart transcoding techniques, which seek to match the component channels to the input to achieve greater performance in the low-delay regime. Note that we omitted the

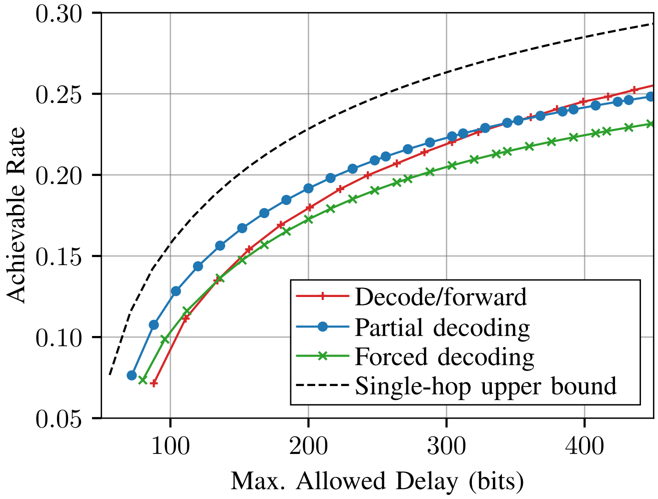

Fig. 7. (Gallager bound) Rate-blocklength tradeoff for $p_1 = 0.04$, $p_2 = 0.13$, and average error probability $\epsilon = 10^{-3}$. The capacity (achievable with infinite blocklength and DF) for this superchannel is $C \approx 0.4426$.

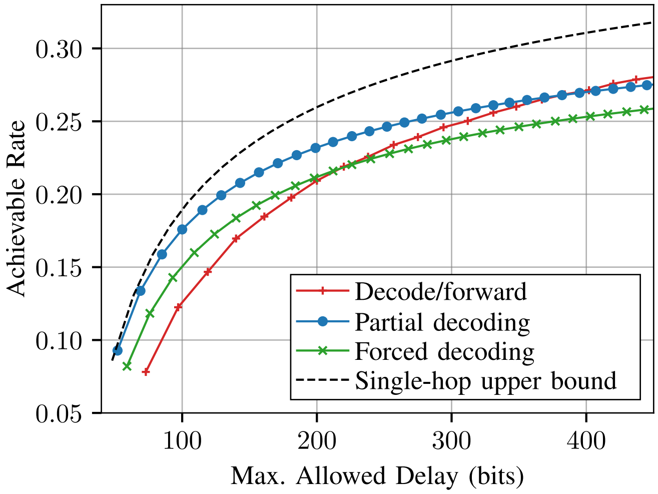

Fig. 8. (Gaussian approximation) Rate-blocklength tradeoff for $p_1 = 0.04$, $p_2 = 0.13$, and average error probability $\epsilon = 10^{-3}$. The capacity (achievable with infinite blocklength and DF) for this superchannel is $C \approx 0.4426$.

curves for amplify-and-forward from both plots for clarity; both curves were consistently below any of the alternatives, which was expected.

The dominance of the partial decoding scheme over the forced decoding scheme satisfies our intuition that *error propagation* is the key to a well-designed transcoding scheme. With partial decoding only the sub-blocks which the relay can decode with high likelihood are decoded at the relay, the rest of the errors are left for the outer code. In contrast, with forced decoding there is increased potential to decode *whole codewords* incorrectly, leading to a decrease in performance of the outer code. The correct propagation of transmission errors sets the partial decoding technique apart.

V. Conclusion

This paper presented a new strategy for relay channels with a finite blocklength constraint. The transcoding principle is based on simple ideas and presents a tradeoff between the decode-and-forward and amplify-and-forward strategies from the literature. We numerically compared our techniques to the state-of-the art and observed increased performance for the example case of relaying over binary symmetric channels. This paper serves as an introduction to a whole new class of previously undiscovered relaying techniques and is thus meant to inspire future work in many different directions. There are plenty of unanswered questions about the transcoding idea at this point. They include the application of this idea to AWGN and Rayleigh fading models or models with a direct connection between source and destination. Other avenues for further research include the construction of optimal mapping functions and the analysis of their structure.

References

[1] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] Nokia Networks, “Looking ahead to 5G,” 2016.

[3] E. C. V. D. Meulen, “Three-terminal communication channels,” *Advances in Applied Probability*, vol. 3, no. 1, pp. 120–154, 1971. [Online]. Available: http://www.jstor.org/stable/1426331

[4] T. Cover and A. E. Gamal, “Capacity theorems for the relay channel,” *IEEE Trans. Inform. Theory*, vol. 25, no. 5, pp. 572–584, September 1979.

[5] S.-H. Lee and S.-Y. Chung, “When is compress-and-forward optimal?” in *Information Theory and Applications Workshop (ITA), 2010*. IEEE, 2010, pp. 1–3.

[6] T. M. Cover and Y. H. Kim, “Capacity of a class of deterministic relay channels,” in *2007 IEEE International Symposium on Information Theory*, June 2007, pp. 591–595.

[7] B. Nazer and M. Gastpar, “Compute-and-forward: Harnessing interference through structured codes,” *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct 2011.

[8] S.-H. Lee and S.-Y. Chung, “Noisy network coding with partial df,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 2870–2874.

[9] G. Kramer, M. Gastpar, and P. Gupta, “Cooperative strategies and capacity theorems for relay networks,” *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3037–3063, Sept 2005.

[10] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec 2004.

[11] Y. Hu, J. Gross, and A. Schmeink, “On the performance advantage of relaying under the finite blocklength regime,” *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 779–782, May 2015.

[12] ——, “On the capacity of relaying with finite blocklength,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1790–1794, March 2016.

[13] L. Dickstein, V. N. Swamy, G. Ranade, and A. Sahai, “Finite block length coding for low-latency high-reliability wireless communication,” in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 908–915.

[14] A. Chaaban and A. Sezgin, “Multi-hop relaying: An end-to-end delay analysis,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2552–2561, April 2016.

[15] G. D. Forney, *Concatenated codes.* Cambridge, MA, USA: MIT Press, 1966.

[16] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.