DEEP DIGITAL JOINT SOURCE-CHANNEL BASED WIRELESS SPEECH TRANSMISSION

Mohammad Bokaei[1], Jesper Jensen[1,2], Simon Doclo[3], Jan Østergaard[1]

[1] Aalborg University, Aalborg, Denmark [2] Oticon A/S, Copenhagen, Denmark
[3] Carl von Ossietzky Universität, Oldenburg, Germany

ABSTRACT

In this paper, we study low-latency speech transmission over a wireless channel based on deep digital Joint Source-Channel Coding (JSCC). Inspired by recent advances in quantization techniques in the realm of JSSC problems, we explore the feasibility of employing constellation-constrained deep JSCC for speech transmission. Our proposed system leverages a single DNN to jointly handle source coding, channel coding, and direct output mapping to specific constellation points. We demonstrate the ability of the joint system to operate effectively under various latency constraints while outperforming separate systems, especially in adverse channel conditions. Simulation results validate the efficacy of our approach, highlighting its potential for real-world applications requiring low-latency speech transmission over wireless channels.

***Index Terms*—** low-latency, joint source-channel coding, speech transmission, edge communication.

1. INTRODUCTION

In conventional communication systems, particularly those concerning speech transmission, the prevailing paradigm operates on the basis of separate source and channel coding [1]. Shannon's foundational work [2] has established the asymptotic optimality of separate systems for extended data streams, yet its efficacy decreases in low-latency communication scenarios, necessitating the use of short-length data blocks. Recently, numerous DNN-based joint source-channel image and audio transmission systems [3, 4] have shown superior performance compared to separate systems, especially in low latency scenarios [5].

In digital communications, the conventional practice involves mapping channel-encoded bits to elements within a two-dimensional finite constellation diagram, with popular schemes including quadrature amplitude modulation, phase shift keying, and amplitude shift keying. In contrast, most DNN-based joint systems [3–5], directly map input signals to complex or real-valued arbitrary symbols, which are subsequently transmitted while adhering to power constraints. The current infrastructure of commercial communication systems and hardware is predominantly designed around established communication standards, which dictate the use of predefined constellations and specific design orders [6]. Consequently, Deep JSCC-based transmission systems necessitate customized hardware for implementation due to their departure from these standardized practices.

Numerous studies have addressed the challenge of embedding quantization within the design of DNNs [7–9]. However, within the context of wireless transmission, these approaches often resemble digital image or speech coders [3,7,9], lacking explicit consideration of channel coding and wireless channel transmission aspects in their models. A notable investigation into the problem of deep JSCC-based channel output-constrained image transmission is presented in [10]. Here, the authors demonstrate that by constraining the output of the encoder to a finite number of codewords, they can achieve performance close to the performance of non-constrained systems. The study encompasses joint source coding, channel coding, and direct mapping of the output to specific constellation points, effectively addressing the challenges inherent in deep JSCC-based image transmission.

In this paper, we investigate deep digital JSCC-based low-latency speech transmission over an Additive White Gaussian (AWGN) wireless channel, drawing inspiration from the quantization technique proposed in [10]. To the best of our knowledge, this represents the first study of constellation-constrained deep JSCC-based speech transmission. Our analysis aims to demonstrate that our proposed system, leveraging a single DNN, can operate effectively under varying latency constraints. Additionally, we aim to compare the performance of our proposed system to the separate systems, particularly under challenging channel conditions. Simulation results will be presented to validate these claims and provide empirical evidence supporting the efficacy of our approach.

2. SYSTEM MODEL

This section presents the proposed deep digital JSCC-based speech transmission system. The system comprises a digital JSCC encoder, a wireless Gaussian channel, and a dig-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.956369.

ital JSCC decoder, as illustrated in Fig. 1. The encoder and decoder utilize DNNs, whose parameters are determined through training to achieve optimal performance. The architecture of the encoder and decoder resembles the JSCC-based analog speech transmission system proposed in [8]. The main difference lies in the purpose and type of quantizer and the size of the proposed DNN network. The purpose of the quantizer in the proposed system is to map the analogue embedding at the output of the encoder directly to the transmitted symbols. In the next subsection, we first introduce the system architecture.

2.1. System architecture

Fig. 1 depicts the encoder architecture, comprising a convolutional layer, three encoding blocks, another convolutional layer, and a quantization layer. The encoding blocks include dilated convolutions with skip connections as in [11]. The Parametric Rectified Linear Unit (PReLU) activation function [12] is utilized between all layers, except preceding the quantization layer, where the Hyperbolic Tangent (Tanh) activation function is utilized. This choice is motivated by the desire to confine the quantization layer's boundaries, thereby aiding training convergence. The total downsampling level is defined as the multiply of each layer's downsampling level $S = \Pi_{i=1}^{d} s_i$, where s_i denotes layer downsampling level of the ith layer and d denotes the number of encoder layers ($d = 5$ in the proposed system). The quantization scheme in the propsed system directly maps compressed and channel-coded analog data to constellation points, in contrast to more conventional quantization schemes that map data to a string of bits before assigning constellation points. The output of the encoder thus consists of constrained modulated data. Further details regarding the quantizer and its rationale are discussed in Subsection 2.2.

The subsequent layer in the system model is the wireless transmission channel, modelled as an AWGN channel in this paper. The wireless channel receives the modulated quantized data from the encoder and introduces Gaussian noise. The final layer is the JSCC decoder, which receives the transmitted data from the wireless channel and reconstructs the input speech signal to encoder. Notably, the decoder performs joint demodulation, source decoding, and channel decoding. The architecture of the decoder mirrors that of the encoder without the quantizer layer. It starts with a convolutional layer, followed by three decoding blocks and another convolutional layer as the final layer, as illustrated in Fig. 1. The decoding and encoding blocks have similar architecture as described in [11].

The proposed system utilizes a fully convolutional DNN, accommodating speech inputs of varying dimensions. Owing to its fully convolutional architecture, the latency of the system is determined by the total downsampling level S, and the dimension of the input signal. The input dimension cannot

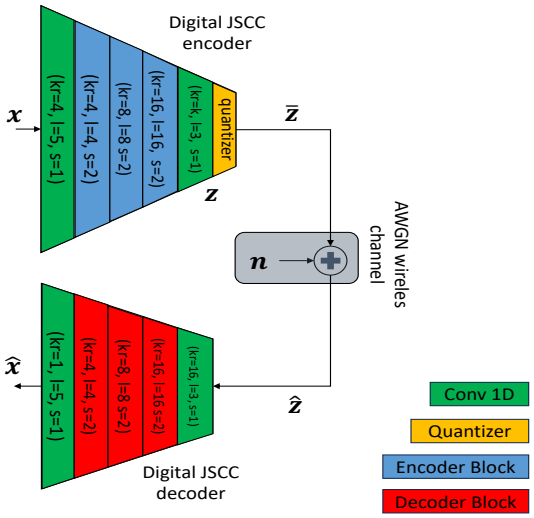

Fig. 1: Overview of the proposed digital deep JSCC-based speech transmission system. In each block, kr, l, and s are mean number of kernels, kernel size, and stride level, respectively. The details of the structure of the encoder block and decoder block are presented in [11].

be shorter than S, representing the minimum latency, while longer input dimensions determine total latency.

Let us denote the input speech signal as $\mathbf{x} \in \mathbb{R}^m$ and the encoder output $\bar{\mathbf{z}} \in \mathbb{R}^k$, where m and k represent the dimension of the input speech and encoder output, respectively. Consequently, the output of the AWGN channel is $\hat{\mathbf{z}} = \bar{\mathbf{z}} + \mathbf{n}$, where $\mathbf{n} \in \mathbb{R}^k$ denotes Gaussian noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2)$, with σ^2 representing the noise variance. The decoder, in turn, receives $\hat{\mathbf{z}}$ and estimates the output speech signal $\hat{\mathbf{x}} \in \mathbb{R}^m$. We define the compression ratio as $R = k/m$, indicating the degree of compression applied to the input speech signal, including any redundant information added by channel coding. The ratio R, in conjunction with the number of quantization levels, determines the total transmitted bitrate. In the subsequent subsection, we provide more details about the quantization layer.

2.2. Quantizer

In the last layer of the encoder, a quantizer is utilized to convert constrained analog source-channel coded data into a limited number of constellation points. This design is inspired by prior work in [10]. Let $\mathbf{z} \in \mathbb{R}^k$ denote the input of the quantization layer and $\bar{\mathbf{z}} = Q_{\mathcal{C}}(\mathbf{z})$ the output of the quantizer, where $Q_{\mathcal{C}} : \mathbb{R} \rightarrow \mathcal{C}^k$ denotes the quantization function mapping analog compressed and channel-coded data to the modulation constellation set $\mathcal{C}$, with constellation points $\mathbf{c}_i \in \mathbb{R}$. The soft-to-hard quantizer method, as introduced in [13] and employed in [10], is used. In the forward path, hard quantization is applied to input data $\mathbf{z}$, mapping each element to its nearest neighbour in the constellation set $\mathcal{C}$, yielding $\bar{\mathbf{z}}$. Since the hard quantization operation is non-differentiable, a differentiable approximation termed soft quantization is used to

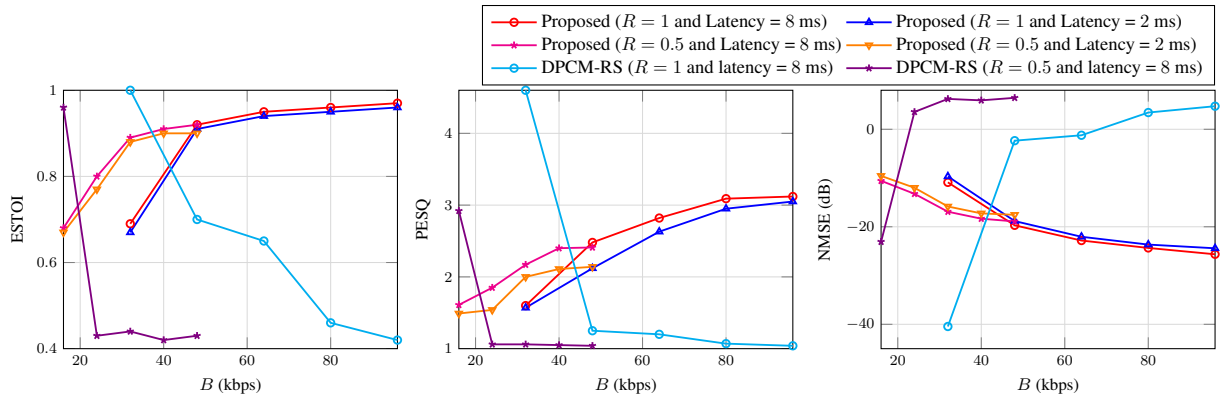

Fig. 2: Performance of the proposed deep joint source channel system and separate source-channel coding systems in terms of speech intelligibility (ESTOI), speech quality (PESQ), and NMSE for SNR = 20 dB.

approximate the gradient during the backward pass. The soft quantization function $\tilde{Q}_{\mathcal{C}}$ is defined as follows [13].

$$\tilde{\mathbf{z}}_i = \tilde{Q}_{\mathcal{C}}(\mathbf{z}_i) = \sum_{j=1}^{M} \frac{e^{-\tau d_{ij}}}{\sum_{l=1}^{M} e^{-\tau d_{il}}} \mathbf{c}_j, \tag{1}$$

where the index i indicates the ith element in a vector, M is the number of constellation points, and d_{ij} represents the squared ℓ_2-distance between $\mathbf{z}_i$ and $\mathbf{c}_j$, i.e., $\|\mathbf{z}_i - \mathbf{c}_j\|_2^2$. The parameter τ determines the accuracy of the approximation: for small values of τ, the approximation deviates from the hard quantization function with a smooth gradient, while for large values of τ, the approximation is more accurate and closer to the hard quantizer but with a sharp gradient. In the backward pass, we approximate $\frac{\partial \bar{\mathbf{z}}}{\partial \mathbf{z}} = \frac{\partial \tilde{\mathbf{z}}}{\partial \mathbf{z}}$. The total bitrate of the system denoted as B, is a function of M and the sampling frequency F_s of the input speech signal and given by

$$B = RF_s\lceil \log_2(M) \rceil, \tag{2}$$

where $\lceil . \rceil$ is the ceiling function.

As proposed in VQ-EMA [7], a simple alternative to address the non-differentiability issue of the hard quantizer is to approximate it with an identity function. In this approach, the hard quantizer is applied in the forward path, while in the backward path, the quantizer gradient $\frac{\partial \bar{\mathbf{z}}_i}{\partial \mathbf{z}_i} = 1$. Our empirical observations have shown that utilizing the approximation in Equation. (1) outperforms the simple identity approximation of the gradient.

Although vector quantization is more commonly used in deep learning-based quantization approaches [7, 8], it should be noted that in the proposed method we opted for one-dimension modulation, which entails mapping a real number in the latent space to quantized real constellation points. Increasing the dimension of the modulations is equivalent to employing a vector quantizer. Empirical observations indicate that increasing the dimension of the modulation, thereby employing a vector quantizer, did not yield performance improvements beyond those achieved by using a simple scalar value quantizer. Another significant advantage of the proposed digital transmission system over the analog counterpart is its ability to operate effectively across various latencies, as demonstrated in the simulation section.

The proposed encoder and decoder are trained end-to-end by minimizing the distortion cost function between the input signal $\mathbf{x}$ and the reconstructed signal $\hat{\mathbf{x}}$. Here, we have chosen the Mean Square Error (MSE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$ as the distortion cost function. Similar to the approach in [10], we do not include the embedding loss in the cost function. This decision is made to avoid the need for meticulous tuning of the loss between embedding and distortion, as highlighted in [10].

3. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed deep digital JSCC-based speech transmission system and compare it with separate speech transmission systems using differentiated pulse coded modulation (DPCM) as a low-latency speech coder. As performance metrics, We consider the perceptual evaluation of speech quality (PESQ) metric [14], the extended short-time objective intelligibility (ESTOI) metric [15], and the normalized mean squared error (NMSE).

3.1. Simulations setup

We used the Librispeech dataset [16] for both training and evaluating the proposed digital speech transmission system with a sampling frequency of 16 kHz. For the training phase, 2400 FLAC files with a total duration of 13100 s were used, of which 90% for training and 10% for the validation. For the test phase, 200 FLAC files with a duration of 1300 s were used. The Adam optimizer [17] was employed to optimize the DNN with a learning rate of 0.001 and $\beta_1 = 0.9$ and $\beta_2 = 0.99$. A learning rate scheduler was used that decreases the learning rate by a factor of 0.8 when the validation loss does not improve for three consecutive epochs. An early stopping

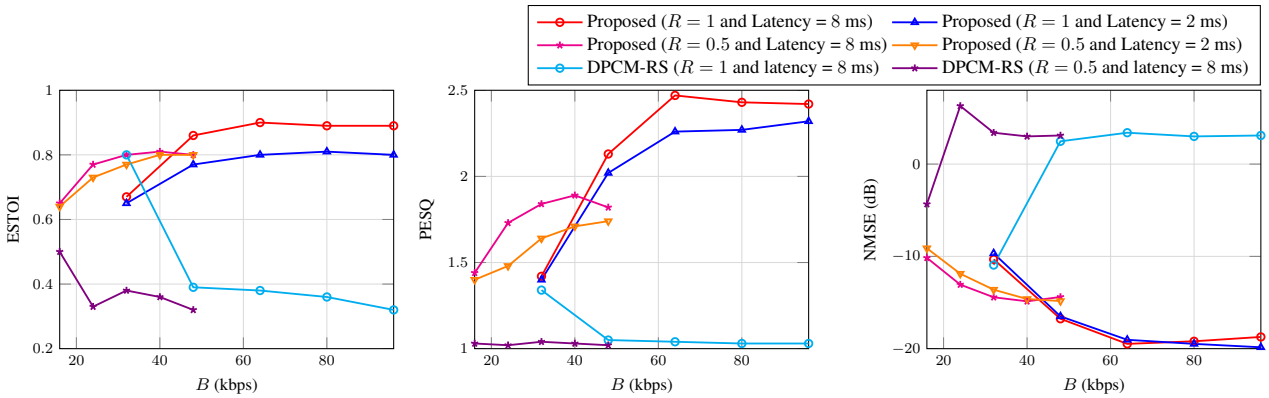

Fig. 3: Performance of the proposed deep joint source-channel coding system and a separate source-channel coding system in terms of speech intelligibility (ESTOI), speech quality (PESQ), and NMSE for SNR = 10 dB.

strategy with an eleven-epoch patience threshold is employed, the maximum number of epochs is 300, and the batch size was set to 1028. The encoder of the proposed DNN has a total downsampling level $S = 8$, which is the minimum possible input signal dimension. This input length is equal to 0.5 ms latency with a sampling frequency of 16 kHz. However, the performance is poor at this extremely low latency. We trained the system for latency of 8 ms and tested its performance for latencies of 8 ms and 2 ms since the system is able to perform under different latencies. The total number of the parameters is dependent on R; the maximum number of parameters is 16841 for $R = 1$. The average constellation power is set to $P = \frac{1}{M}\sum_{i=1}^{M} |\mathbf{c}_i|^2 = 1$. Similar to [10], the controlling parameter τ is first set to 5, and after each 3000 parameter update step, it is increased with 5 until it reaches 100.

3.2. Performance comparison

In this part, we compare the proposed deep joint source-channel coding transmission system with a separate source-channel coding system. We compare the systems at the same SNR of the wireless channel, the same bitrate, and the same modulation size. The separate system utilizes DPCM [18] for the speech coder and Reed-Solomon coding as the channel coder. We denote the separate system DPCM-RS. DPCM is an efficient source coder that can run at extremely low latencies. The channel coding rate determines the bitrate allocation between source coding and channel coding and impacts the performance of the DPCM-RS system. Therefore, in each scenario, the channel coding rate is chosen by a grid search and based on the NMSE performance. Although the latency of the DPCM-RS system is flexible, we only consider a latency of 8 ms since the performance at the latency of 2 ms is poor.

For two different SNRs (10 dB and 20 dB), Figures 3 and 2 depict the performance metrics (ESTOI, PESQ, NMSE) for the proposed system and the DPCM-RS system. As already mentioned, for the proposed system, we consider two latencies (2 ms and 8 ms), whereas for the DPCM-RS system, we only consider a latency of 8 ms. For both systems, we consider two different compression ratios: $R = 0.5$ and $R = 1$ and $M = [4, 8, 16, 32, 64]$ resulting in different bitrates (see 2). As expected from 2, for $R = 0.5$, the bitrates are halved compared to $R = 1$. For bitrates between 32 and 48 kbps, there are two versions of each system which can be compared at the same bitrate. For example, both $R = 1$ and $M = 4$ as well as $R = 0.5$ and $M = 8$ results in a total bitrate $B = 32$ kbps. However, it should be realized that the systems with higher M need more bandwidth or a larger symbol transmission rate.

We can see in Fig 3 at bitrate 32 kbps $R = 0.5$ yields better than $R = 1$ and at bitrate 48 kbps $R = 1$ yields a better performance than $R = 0.5$.

Notably, for both SNRs and compression ratios, the proposed method demonstrates comparable performance for latencies of 2 ms and 8 ms. This observation shows the effectiveness of the proposed system under different latency constraints.

For the DPCM-RS system, performance degradation can be observed with increasing bitrates for both SNRs. This can be attributed to the design principle of DPCM-RS, namely a separate source and channel coding system typically tailored based on channel capacity considerations. Such systems only perform properly when the total bitrate remains below the channel capacity threshold, where the transmitted data can be recovered completely (at least theoretically). However, as the bitrate surpasses the channel capacity, a significant decline in performance is observed. This decline underscores the limitations of traditional separate source-channel coding approaches when faced with bitrate demands exceeding channel capacity constraints.

Comparing the deep JSCC-based system with the DPCM-RS system reveals notable insights. While the DPCM-RS system exhibits superior performance at low bitrates and favourable channel conditions (SNR = 20 dB), its efficacy diminishes as bitrates exceed channel capacity. In contrast, the proposed deep JSCC-based system demonstrates improved

performance at high bitrates, outperforming the DPCM-RS system under adverse channel conditions (SNR = 10 dB) and for lower bitrates, especially with $R = 0.5$ even with lower latency.

4. CONCLUSION

In this paper, we investigated the feasibility of low-latency digital speech transmission over an AWGN wireless channel based on deep source-channel coding. By leveraging recent advances in soft quantization techniques and employing constellation-constrained deep JSCC, we have demonstrated the ability of the proposed system to effectively handle different latency constraints while outperforming separate source-channel coding system, particularly in challenging wireless channel conditions. Our findings underscore the potential of our approach for real-world applications requiring low-latency speech transmission over wireless channels.

5. REFERENCES

[1] J. G. Proakis, *Digital communications*, McGraw-Hill, Higher Education, 2008.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[4] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[5] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Deep joint source-channel analog coding for low-latency speech transmission over gaussian channels," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 426–430.

[6] "IEEE standard for information technology–telecommunications and information exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*, pp. 1–4379, 2021.

[7] A. Razavi, A. Van den Oord, and O. inyals, "Generating diverse high-fidelity images with VQ-VAE-2," *Advances in neural information processing systems*, vol. 32, 2019.

[8] N. Zeghidourand, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.

[9] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[10] T. Tung, D. Kurka, M. Jankowski, and D. Gündüz, "DeepJSCC-Q: Constellation constrained deep joint source-channel coding," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 4, pp. 720–731, 2022.

[11] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.

[13] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," *Advances in neural information processing systems*, vol. 30, 2017.

[14] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001*, 2001, pp. 749–752.

[15] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] A. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.