PYRAMID CELP: EMBEDDED SPEECH CODING FOR PACKET COMMUNICATIONS

Christoph Erdmann, David Bauer, and Peter Vary

Institute of Communication Systems and Data Processing
Aachen University of Technology
Templergraben 55, D-52056 Aachen, Germany
e-mail: erdmann@ind.rwth-aachen.de

ABSTRACT

In this paper we present a novel speech coding algorithm for transmission over packet networks based on CELP (Code-Excited-Linear-Prediction). While various CELP type speech coders like the ITU-T G.729 or the Adaptive Multi-Rate (AMR) codec are applicable to packet voice communication systems such as the Internet, there are some fundamental reasons that may limit their performance in case of packet loss. We investigate the alternative of *embedded coding* in the CELP framework, focussing on an efficient quantization of the excitation signal to generate a hierarchically-structured bit stream by means of pyramid coding [1]. Thus, the receiver can reconstruct speech at a basic quality by decoding only a subset of the entire bit stream. The quality of the decoded speech increases with the amount of received bits. We demonstrate the performance by an experimental hierarchical wideband (0.05-7 kHz) speech coder, applied to an IP (Internet Protocol) channel simulation.

1. INTRODUCTION

The extremly time-variant channels offered by packet networks feature signifciant transmission delay and delay variation (so-called *jitter*) leading to packet loss. To mitigate the influence of packet loss is one of the major challenges for real-time voice communications over packet networks, in the following subsumed under the acronym Voice-over-IP (VoIP).

An important feature of a speech coding algorithm designed for VoIP applications is bit rate scalability. Variable bit rate (VBR) coding is particularly advantageous for speech coding applications, where channel impairments play a major role in the design of the coder. The additional degree of freedom by adapting the bit rate to the changing statistical character of the channel, allows it to maintain real-time speech transmission even under network conditions, where fixed rate coders fail to operate.

In this paper we investigate a special case of VBR coding in which a single encoding algorithm generates a fixed-rate hierarchically-structured data stream from which reduced-rate data streams can be extracted. The embedded bit stream for a given rate contains valid bit streams for each lower rate. Depending on the number of bits received in a finite time intervall, the decoder can select the rate and fills in the missing bits with zeros prior to decoding the modified full-rate data signal with its fixed decoding algorithm.

This paper is organized as follows. Section 2 describes the methodology of pyramid encoding and its application to hierarchical image and speech coding. In Section 3 we propose a novel embedded speech codec called P-CELP (pyramid CELP). In Section 4 we demonstrate the performance of a wideband P-CELP speech coder, applied to an IP channel simulation, and Section 5 presents our conclusions.

2. PYRAMID CODING

Presently, pyramid coding is much more explored in the field of image coding [1] than for speech coding applications, which led to a far better understanding of pyramid coding in the context of image compression. A so-called *image pyramid* means the representation of an image with multiple resolutions. The different resolution-levels are denoted as layers of the pyramid. These layers are ordered and stacked one upon the other, so that the highest resolution level establishes the top-layer while the level with the lowest resolution marks the base-layer. The lower resolution levels consist of less pixels than the higher resolution levels. Thus, the upper layer pictures are smaller than the lower ones, yielding a formation that can be visualized as a pyramid.

2.1. Decomposition

A simple yet important structure of a pyramid is the so-called *Gauss* pyramid $\mathbf{g}$, which is depicted on the left side of Fig. 1. A complete description of the Gauss pyramid is given by a set of reduce operators R_i.

Let $\mathbf{s} \in \mathbb{R}^m$ be the m-dimensional input vector. L denotes the depth of the pyramid-decomposition, i.e. the number of pyramid-layers, and m_i ($i = 1, \ldots, L$) the number of samples in the i-th layer. Further on, the following decimation condition

$$m_1 = m, \quad m_i > m_{i+1} \tag{1}$$

holds for all pyramid-layers. By defining a set of reduce operators

$$R_i : \mathbb{R}^{m_i} \rightarrow \mathbb{R}^{m_{i+1}}, \quad i = 1, \ldots, L-1 \tag{2}$$

and applying R_i sequentially to the input signal $\mathbf{s}$, one obtains the different layers $\mathbf{g}_i$ of the Gauss pyramid $\mathbf{g}$, given recursively by the equation:

$$\mathbf{g}_i = R_{i-1}(\mathbf{g}_{i-1}), \quad i = 1, \ldots, L. \tag{3}$$

In fact, R_0 (depicted in the dashed box) corresponds to the $m \times m$ identity matrix $\mathbf{I}^{m \times m}$, so that for the top-layer $\mathbf{g}_1 = \mathbf{s}$ holds.

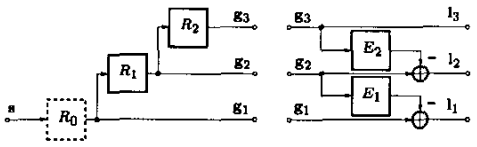

Fig. 1. Three-stage Laplace pyramid $\mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3)$, based on a Gauss pyramid $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3)$.

Using Gauss pyramids is most self-evident, when representing signals that should be applied to scalable compression algorithms.

When combined with interpolation the hierarchical structure yields the so-called *Laplace* pyramid.

Based on the decomposition of the signal into a Gauss pyramid, the representation of the Laplace pyramid is derived in a second step by defining a set of expand operators

$$E_i : \mathbb{R}^{m_{i+1}} \rightarrow \mathbb{R}^{m_i}, \quad i = 1, \ldots, L-1 \quad . \tag{4}$$

Applied to the $(i+1)$-th layer $\mathbf{g}_{i+1}$ of the Gauss pyramid, E_i provides a signal mapped to the size m_i of the i-th layer. As shown on the right side of Fig. 1, the layers $\mathbf{l}_i$ of the Laplace pyramid $\mathbf{l}$ are given by the recursive formula:

$$\begin{aligned} \mathbf{l}_L &= \mathbf{g}_L \,, \\ \mathbf{l}_i &= \mathbf{g}_i - E_i(\mathbf{g}_{i+1}), \quad i = 1, \ldots, L-1. \end{aligned} \tag{5}$$

Due to the mutual information between adjacent layers, the expand operator E_i approximates the Gauss layer $\mathbf{g}_i$ from the less accurate layer $\mathbf{g}_{i+1}$. Thus, each layer $\mathbf{l}_i$ of the Laplace pyramid $\mathbf{l}$ describes the error resulting from the difference between a Gauss-layer $\mathbf{g}_i$ and its approximation from the layer $\mathbf{g}_{i+1}$.

Apart from (1), we have not made any assumptions regarding the reduce and expand operations so far. Indeed, for complementary reduce and expand operators it suffices to require, that the dimensional change between input and output signal caused by the reduce operator is exactly reversed by the respective expand operator and vice versa. In a straightforward approach, known from progressive image coding (e.g. *JPEG*), the size of an image between two adjacent pyramid-layers differs by factor 2 for each image-dimension. The reduce operator is thus realized through lowpass filtering followed by critical decimation, while the respective expand operator includes up-sampling followed by interpolation.

Though this kind of multiresolution approach reveals some similarities to subband coding, since in both cases the signal undergoes a linear transformation, pyramid coding provides a redundant signal representation, whereas for subband coding (provided each subband is critically sampled) the number of transmitted coefficients n equals the number of samples m of the original signal. If we assume a one dimensional signal (e.g. speech signal) and define q as the decimation factor between two subsequent layers and provide $|q| = const.$ for all layers, we can re-formulate (1) more precisely through

$$m_1 = m, \quad q \cdot m_i = m_{i+1} \quad . \tag{6}$$

With (6), the number of transmitted coefficients n for a pyramid with L layers can be given in the following closed form:

$$n = \sum_{i=1}^{L} m_i = m \cdot \sum_{i=0}^{L-1} q^i = m \cdot \frac{q^L - 1}{q - 1}, \quad |q| < 1 \quad . \tag{7}$$

2.2. Reconstruction

A simple rule for reconstructing the original signal $\mathbf{s}$ from its Laplace pyramid is shown in Fig. 2a). This iterative reconstruction of the underlying Gauss pyramid through expansion and subsequent addition of the respective Laplace-layers is easily obtained by solving (5) for $\mathbf{g}_i$:

$$\begin{aligned} \mathbf{g}_L &= \mathbf{l}_L \,, \\ \mathbf{g}_i &= E_i(\mathbf{g}_{i+1}) + \mathbf{l}_i, \quad i = 1, \ldots, L-1. \end{aligned} \tag{8}$$

In analogy to the reduce operator R_0, the expand operator E_0 corresponds to the $m \times m$ identity matrix $\mathbf{I}^{m \times m}$. Under the constraint of *linear* expand operations, (8) leads to an equivalent version of the reconstruction rule illustrated in Fig. 2b). Since in this

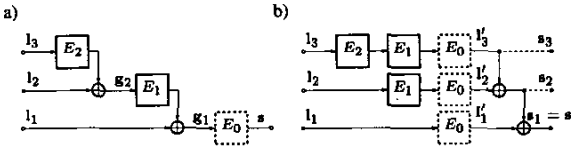

Fig. 2. a) Reconstruction of three-stage Laplace pyramid
b) Equivalent reconstruction scheme

paper we rely on the paradigm of hierarchical coding, to produce an embedded data-stream, this reconstruction scheme is particularly interesting. Each layer $\mathbf{l}_i$ is separatly expanded to a layer $\mathbf{l}'_i$:

$$\mathbf{l}'_i = \ddot{E}_{i-1}(\mathbf{l}_i) \quad . \tag{9}$$

The corresponding expand operation $\ddot{E}_{i-1}$ results from a concatenation of all expand operators, which are relevant for this layer:

$$\ddot{E}_i = E_0 \circ \cdots \circ E_i, \quad i = 1, \ldots, L-1. \tag{10}$$

Due to this expand procedure all layers $\mathbf{l}'_i$ have the same size m as the original signal $\mathbf{s}$. I.e. the $\mathbf{l}'_i$ have the sampling rate of the original signal, but their frequency components are limited to the frequency components of the corresponding Laplace-layer $\mathbf{l}_i$. Hence, this reconstruction scheme features a **hierarchical reconstruction property**, which allows **progressive decoding** with scalable bandwidth. The simple reconstruction rule is given by

$$\mathbf{s} = \sum_{i=1}^{L} \ddot{E}_{i-1}(\mathbf{l}_i) = \sum_{i=1}^{L} \mathbf{l}'_i \quad . \tag{11}$$

The layers of the Laplace pyramid $\mathbf{l} = (\mathbf{l}_1, \ldots, \mathbf{l}_L)$ form an embedded data-stream. With an increasing number of available Laplace-layers $\mathbf{l}_i$ (i.e. i decreasing from L to 1), the decoder can reconstruct a signal $\mathbf{s}_i$, which approximates the original signal $\mathbf{s}$ with increasing accuracy in terms of bandwidth.

2.3. Application to Speech Signals

If the proposed hierarchical coding paradigm is applied to image data, in the first step, the decoding of the embedded data stream provides a blurred version of the original image. The more the decoding proceeds, the more details will be added, until a completely sharp reconstruction of the original image is achieved. To a certain extend the human eye tolerates this effect of varying signal-bandwidth. For speech signals comparable variations of the acoustic bandwidth would be perceived as annoying fluctuation of the signal.

To achieve scalability in terms of perceptual speech quality, we apply the Laplace pyramid to the residual signal of a predictive speech coder [2], instead of the speech signal itself.

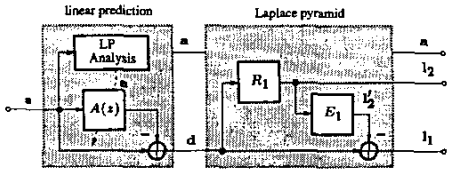

Fig. 3. Speech encoder with two-stage Laplace pyramid applied to the residual signal $\mathbf{d}$.

As illustrated in Fig. 3 (here for convenience explained by a simple two-stage pyramid-decomposition), the linear prediction (LP) synthesis filter coefficients $\mathbf{a}$ are independently transmitted

as side information. Thus, the spectral envelope of the speech signal will be properly reconstructed over the whole frequency range, regardless of the accuracy of the reconstructed excitation. We have to take care, that the reconstruction of the Laplace pyramid in the residual domain yields a signal that also covers the whole frequency range, even if only the base-layer is available for decoding. In exchange we accept a limited accuracy of the spectral shape compared to the original excitation. This can be achieved by a modification of the expand operation. While the reduce operation consists of half-band lowpass filtering and subsequent critical decimation by factor $q = 2$, the corresponding expand operation employs up-sampling by factor 2 but omits the following lowpass-interpolation.

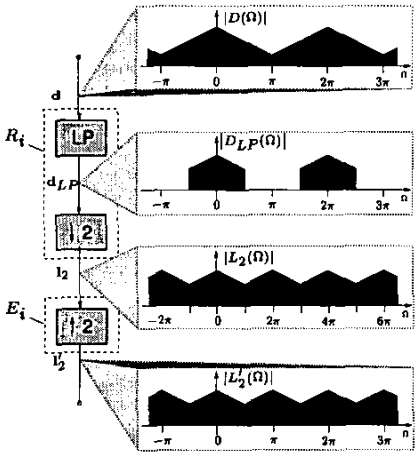

Fig. 4. Modified expand operation: spectral backfolding of low frequency components. (Note the different scaling of the frequency axis for $|L_2(\Omega)|$ and $|L_2(\Omega)'|$!)

The impact of this expand operation is illustrated in Fig. 4. $\mathbf{l}_2$ is the band-limited, critically sub-sampled version of the original LP residual $\mathbf{d}$. By up-sampling $\mathbf{l}_2$ without lowpass-interpolation the lower frequencies of the original residual are mirrored into the empty upper frequency range of the up-sampled signal $\mathbf{l}_2'$. Assuming the LP residual $\mathbf{d}$ to have a flat spectrum, the expand operation extends the band limited residual signal $\mathbf{l}_2$ to a full-band signal $\mathbf{l}_2'$ by approximating the upper frequency components of the original residual with its frequency reversed low frequency components [3]. After synthesis filtering with $1/A(z)$, this approximated excitation results in a full-band speech signal with a certain loss of quality in the upper frequency range. The additional decoding of the top-layer $\mathbf{l}_1$ corrects the reconstruction error caused by the described approximation of the upper frequencies.

3. PYRAMID CELP (P-CELP) CODING

To achieve a reasonable speech compression we have to consider efficient quantization of the pyramid-layers. In the previous section we already applied the Laplace pyramid to the LP residual of a linear predictive coding scheme. Thus, it is very straight-forward to design a *closed-loop* Analysis-by-Synthesis (AbS) quantization of the Laplace-layers based on CELP [4].

For that purpose we firstly neglect any non-linear quantization of the Laplace-layers to find a more convenient expression for deriving a Laplace from a Gauss pyramid. Using the concatenated expand operation $\ddot{E}_i$ from (9) we have

$$\mathbf{l}_L' = \mathbf{g}_L' = \ddot{E}_{L-1}(\mathbf{g}_L)$$
$$\mathbf{l}_i' = \ddot{E}_{i-1}(\mathbf{g}_i) - \sum_{\nu=i+1}^{L} \mathbf{l}_\nu' \quad . \tag{12}$$

This recursive sum-expression is obtained by continuously re-inserting the reconstruction rule (8) into the construction rule (5).

3.1. Fixed codebook search

Within the CELP framework, the quantization of the LP residual $\mathbf{d}$ is basically performed by a closed-loop search in a *fixed codebook* (FCB). Therefore, Fig. 5 specifies a structure with separate closed-loop quantization Q (i.e. FCB search) of the different Laplace-layers according to the construction rule given in (12).

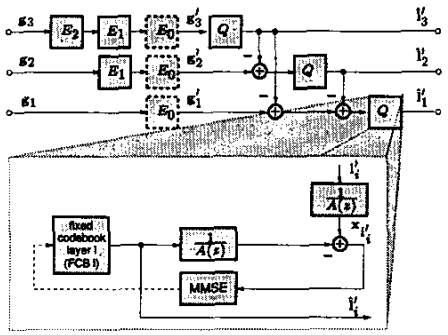

Fig. 5. P-CELP: closed-loop quantization of a three-stage Laplace pyramid.

Let $\hat{v} = Q(v)$ indicate the closed-loop quantization of v. Then we can modify (12) to a construction rule that considers the closed-loop quantization $Q(.)$:

$$\hat{\mathbf{l}}_L' = \hat{\mathbf{g}}_L' = Q\left(\ddot{E}_{L-1}(\mathbf{g}_L)\right)$$
$$\hat{\mathbf{l}}_i' = Q\left(\ddot{E}_{i-1}(\mathbf{g}_i) - \sum_{\nu=i+1}^{L} \hat{\mathbf{l}}_\nu'\right) \quad . \tag{13}$$

The quantization procedure is illustrated in the detailed view of the Q operator in Fig. 5. As in conventional CELP, the target signal $\mathbf{x}_{l_i'}$ (speech domain) for the codebook search is derived by filtering the expanded Laplace-layer $\mathbf{l}_i'$ (residual domain) with the LP synthesis filter $1/A(z)$. The quantized Laplace-layer $\hat{\mathbf{l}}_i'$ is found by means of closed-loop AbS, i.e. $\hat{\mathbf{l}}_i'$ is represented by the codebook vector that was found to minimize the squared error between the target signal $\mathbf{x}_{l_i'}$ and the synthesis filtered codebook vector. The structure of the fixed codebook corresponds to a state-of-the-art algebraic codebook ($\rightarrow$ ACELP) employing a low complexity tree-search algorithm as proposed in [5]. (13) reveals a further closed-loop effect: The unquantized version of each Laplace-layer $\mathbf{l}_i'$ includes the quantization errors of the upper layers. Hence, quantization of the lower layers takes into account the quantization errors, which are already included in the upper layers and possibly corrects them.

3.2. Adaptive codebook search

In more elaborate versions of CELP coders (e.g. ITU-T G.729 [6, 7] or GSM Enhanced Full-Rate Codec [8]), two codebooks

are sequentially searched through. While the innovative part of the excitation is searched in the fixed codebook, a so-called *adaptive codebook* (ACB) contributes to the reconstruction of periodic signal components. Together the contributions of ACB and FCB build the quantized excitation signal. Usually the ACB is realized by a feedback-memory that contains the previously reconstructed excitation signal. Because this signal varys over time, the entries of this codebook change from frame to frame. To keep the states of encoder and decoder synchronized, it is absolutely necessary to keep the entries of the ACB in encoder and decoder in line.

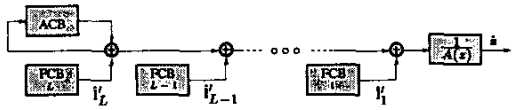

Fig. 6. P-CELP decoder: ACB updated solely by base-layer $\hat{\mathbf{i}}'_L$

A major goal of the P-CELP coder is to maintain decoding even when only the highest layer was received. Thus, solely the quantized base-layer $\hat{\mathbf{i}}'_L$, though it produces the coarsest approximation of the excitation signal, can be used to update the ACB (see Fig. 6). This ensures the states of encoder and decoder to be synchronized widely independent from the current channel conditions.

4. EXPERIMENTAL RESULTS

We compare the performace of the proposed P-CELP to a conventional CELP coder in a realistic VoIP scenario. For that purpose we implemented two experimental wideband speech coders.

Both coders use identical structures for short-term (LP) and long-term prediction (LTP). The LP synthesis filter coefficients are updated every 20 ms frame The quantitzation of the 18 LP parameters is performed in the LSF (Line Spectral Frequencies) domain and based on *Safety-Net* VQ, using 43 bit. Every 10 ms, an open-loop pitch estimate is calculated. Around this estimate a constrained closed-loop ACB search with fractional delays is performed every 5 ms. This procedure requires 8+6=14 bit per 10 ms for coding the pitch lags [9]. The pitch gain is nonuniformly quantized with 4 bit.

The difference between the conventional CELP and the P-CELP coder is the methodology used to encoder/decode the residual signal. With P-CELP, the residual signal is decomposed into 4 layers, 1 base-layer and 3 enhancement-layers, with each layer being efficently quantized by appropriate ACELP codebooks. The hierarchical bitstream of each 20 ms speech frame is distributed on 3 UDP-packets (User Datagram Protocol). Since there is no prioritization implemented in IPv4 we have to prevent loss of the side-information and the all important base-layer. Therefore, each packet contains the side-information consisting of LP parameters, pitch lag and gain plus FCB shape and gain of the base-layer, which amounts to 135 bit (i.e. 6.75 kbit/s) of coded side-information in each of the three packets. Furthermore, each packet contains the FCB shape and gain information for one of the three enhancement-layers. This requires 144 ... 264 additional bit (i.e. 7.2 ... 13.2 kbit/s) per packet. The overall bitrate for all 3 packets amounts 51.85 kbit/s.

For the conventional CELP the entire residual signal is quanized using one ACELP codebook. Per speech frame, this requires 264 bit (i.e. 13.2 kbit/s) yielding a bit rate of 17.55 kbit/s for this coder. To mitigate the effects of packet loss, we transmit this information 3 times within 3 packets, which amounts to an overall bit rate of 52.65 kbit/s.

We simulated real-time speech transmission with different IP connections. Therefore we piped the output of both speech coders through a UDP/IP-channel simulation. As a subjective measure, we have done AB-comparison tests with the conventional CELP and the proposed P-CELP coder. The tested speech material included 4 speakers, 2 male and 2 female. A total of 20 listeners took the test. Table 1 shows the relative preference between the two coders. The increasing packet error rate is achieved by successively shortening the length of the receiver buffer. From the results we concluded, that the proposed P-CELP coder significantly improves error robustness and speech quality when transmitting over lossy packet networks.

packet errors (%)	Preference Score (%)	
	CELP	P-CELP
1	0	100
3	5	95
8	25	75
10	35	65
15	35	65
22	60	40

Table 1. Preference test results for comparison of P-CELP with conventional CELP

5. CONCLUSIONS

In this paper we presented a new embedded speech coding concept based on pyramid CELP (P-CELP). The coder generates a hierarchically-structured bit stream by using pyramid decomposition of the LP residual signal. Each pyramid layer is efficiently quantized by the ACELP approach. Combined with appropriate quantization of the LP filter coefficients, the algorithm offers progressive decoding of the speech signal with increasing accuracy in terms of perceptual speech quality. The proposed algorithm belongs to the class of network controlled variable bit rate coders, which are particularly advantageous when applied to real-time voice transmission over lossy packet networks. Compared to conventional CELP speech coders, the proposed P-CELP coder significantly improves error robustness and speech quality and provides gracefull degradation with increasing packet error rate.

6. REFERENCES

[1] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code", *IEEE Trans. on Comm.*, vol. 31, no. 4, Apr. 1983.

[2] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer, 1976.

[3] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems,", in *Proc. ICASSP*, Washington, DC, 1979.

[4] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates,", in *Proc. ICASSP*, Tampa, Florida, 1985.

[5] C. Laflamme, J.-P. Adoul, H.Y. Su, and S. Morisette, "On Reducing Computational Complexity of Codebook Search in CELP Coder Through the Use of Algebraic Codes,", in *Proc. ICASSP*, Albuquerque, New Mexico, 1990.

[6] CCITT/ITU-T, "Rec. G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)", in *General Aspects of Digital Transmission Systems; Terminal Equipments, Series G Rec.* ITU, 1996.

[7] R. Salami, C. Laflamme, B. Bessette, and J.P.. Adoul, "Description of ITU-T Rec. G.729 Annex A: Reduced Complexity 8 kbit/s CS-ACELP Coding", in *Proc. ICASSP*, Munich, Germany, 1997, IEEE.

[8] K. Järvinen, J. Vainio, T. Honkanen, P. Kapanen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Adoul, "GSM Enhanced Full Rate Speech Codec", in *Proc. ICASSP*, Munich, Germany, 1997.

[9] C. Erdmann, P. Vary, K. Fischer, W. Xu, M Marke, T. Fingscheidt, I. Varga, M. Kaindl, C. Quinquis, B. Kövesi, and D. Massaloux, "A Candidate Proposal for a 3GPP Adaptive Multi-Rate Wideband Speech Codec", in *Proc. ICASSP*, Salt Lake City, Utah, May 2001.