The Harmonic Model Codec (HMC) framework for VoIP

Yannis Agiomyrgiannakis, Yannis Stylianou

Department of Computer Science, University of Crete, Greece

jagiom@csd.uoc.gr, yannis@csd.uoc.gr

Abstract

A framework for joint source/channel coding of speech is presented. It is based on a harmonic representation of the speech signal and facilitates efficient quantization of harmonic amplitudes and phases both in a single description and a multiple description setting. Furthermore, it combines high-quality packet loss concealment with efficient source coding and multiple description coding. Two proof-of-concept codecs are presented; a single description codec that is equivalent to iLBC in terms of bitrate and quality but more robust in conditions of increased packet losses and a multiple description codec that is capable of accepting loss rates up to 40% for a DCR score of 3.8.

Index Terms: voice-over-IP, speech coding, harmonic models, multiple description coding, phase quantization, packet loss concealment

10.21437/Interspeech.2007-473

1. Introduction

Voice-over-IP (VoIP) provides a low cost alternative to traditional telecommunications. Unfortunately, IP networks do not provide a Quality-of-Service mechanism and various delays and reorderings are introduced to the packets. This is compensated by a buffer (called *jitter buffer*) that is held at the decoder. However, this buffer cannot be long due to the strict end-to-end delay requirements of speech communication. As a result many packets arriving after playback time, are considered as lost. Therefore, a speech codec has to deal with increased packet losses.

Most of today's speech codecs, like G.729, G.723.1, AMR and AMR-WB were not initially designed to cope with increased packet losses. The main defect of these codecs is that they are desynchronized upon a single packet loss, while a single error propagates to several subsequent frames. Some improvements are proposed within the CELP framework [1] but their effectiveness is limited by the fact that CELP codecs rely heavily on past (*adaptive*) excitation vectors. Thus, there is a rapid degradation of the perceived quality upon packet losses. Other solutions compensate packet losses by introducing redundancy via FEC (Forward Error Correction) schemes for channel coding, like packet repetition or Reed-Solomon codes [2]. All these approaches deal with VoIP by separately optimizing source coding (i.e. speech coding) and channel coding. In theory, such a separation is motivated by the well-known Shannon Separation Theorem and justified only when the encoding buffer is large enough [3], which is not the case of speech coding due to the end-to-end delay restrictions.

Another family of solutions is provided by codecs that make use of the existant redundancy in the bitstream at the cost of coding efficiency. This redundancy is used to facilitate a more robust behavior upon packet losses. For example, iLBC (internet Low-Bitrate Codec) [4] encodes each packet independently of the previous packets and avoids the desynchronization that other CELP codecs suffer at the cost of ignoring the considerable intra-packet correlations that exist in the speech signal. This choice almost doubles the bitrate of the codec but allows better Packet Loss Concealment (PLC). The effectiveness of PLC is crucial for VoIP and many PLC algorithms depend heavily on the ability to manipulate the speech signal both in time and frequency, accurately. In that aspect, sinusoidal models of speech are better suited for VoIP because they offer fine time-frequency resolution and allow high-quality PLC [5],[6]. Furthermore, sinusoidal models allow high-quality time-scaling which is important for rapid adaptive jitter buffer resizing upon varying network conditions [7]. A variable-rate sinusoidal codec (SVOPC) for VoIP is proposed in [8] where it is shown that packet losses up to 20%-30% can be tolerated at the expense of a high bitrate.

Both iLBC and SVOPC can be considered as joint source/channel codecs of speech, as opposed to "traditional" source-optimized codecs, in the sense that they are designed to operate under packet losses. However, both codecs encode each packet independently and rely on the redundancy that naturally exists in the speech signal to facilitate robustness. Additional redundancy can be introduced via Multiple Description Coding (MDC). In MDC, the content of each frame is encoded in two descriptions which are -ideally- routed through independent channels. If both descriptions arrive, we get a high-quality reconstruction of the speech signal. If only one description arrives, an intermediate quality reconstruction is obtained. In a properly constructed MDC system the probability of losing both descriptions, and thus any information regarding a speech frame, is kept low even for high loss probabilities.

The construction of multiple description speech codecs poses considerable challenges arising from the fact that the redundant and irrelevant information must be initially removed from the speech source and then controlled redundancy must be re-introduced via MDC. Furthermore, since PLC is crucial for the robustness of the codec, the design of the codec must be made with respect to PLC. Many efforts have been made towards multiple description speech coding, like [9],[10], [11] and [12] but none addresses all these aspects simultaneously; efficient speech coding, PLC and MDC.

This paper presents a framework for sinusoidal speech coding that belongs to the family of joint source-channel codecs. A codec that is developed within this framework is referred to as a Harmonic Model Codec (HMC). A detailed description of HMC can be found in [13] (Chapter 9). The purpose of this paper is to present and position HMC to related works, outline its main characteristics and discuss the main outcomes. We will briefly outline the characteristics of HMC:

Packet Independence: Each packet is encoded independently of previous packets in order to avoid a desynchronization between the encoder and the decoder upon a packet loss.

Variable Rate: Speech frames are classified in four classes ("silent","unvoiced", "transitional" and "voiced") and

each class of frames is encoded with a different number of bits.

Harmonic Model: The speech signal is represented as a sum of harmonically related sinusoids. This representation allows fine time-frequency resolution of the spectral components and facilitates efficient PLC and high quality time-scaling for adaptive jitter-buffer resizing [7].

Packet Loss Concealment: The PLC algorithm uses future frames from the jitter buffer to interpolate the harmonic sinusoids when the current frame is lost. Sinusoidal extrapolation is used when the jitter buffer is empty. In both cases, the sinusoids are individually characterized as voiced or unvoiced and treated in a different manner.

Amplitude Parameterization: The harmonic amplitudes are fitted to a spectral envelope with 20 real cepstrum coefficients.

Phase Quantization: The quantization of the phases of the harmonics is necessary for high-quality speech coding. The phase of the sinusoids is explicitly quantized using a suitable quantizer that is based on WGMM (Wrapped Gaussian Mixture Models) [14].

Single Description Mode: Spectral envelopes and phases are quantized using GMM-based and WGMM-based quantization respectively.

Multiple Description Mode: Spectral envelope is quantized with a GMM-based MDC technique described in [15]. Phases are quantized in multiple descriptions with a quantizer based on WGMM [14].

Two proof-of-concept narrowband HMC codecs are developed, differing solely on the employed quantizers for the spectral envelope and the harmonic phases. The first version is referred to as HMC-SD (Harmonic Model Codec - Single Description) and it is a variable-rate single description codec that requires about 12.9-14.2 kbps on average. It is equivalent to iLBC in speech quality but considerably more robust to packet losses (up to 20%). The second version, HMC-MD (Harmonic Model Codec - Multiple Description), is a multiple description codec that generates two descriptions for each frame using about 21 kbps on average while it may accept up to 40% packet losses with minor quality degradation.

An important feature of HMC is that it allows efficient quantization of the harmonic representation of speech which is suitable for modifications and manipulations necessary for high-quality PLC. This is in contrast to CELP codecs which encode speech into a parameterization that is hard to use for the manipulation of speech. Furthermore, HMC facilitates efficient high-rate quantization of the harmonic phases to a single description or multiple descriptions within the generic MDC framework. Therefore, within the HMC framework it is possible to simultaneously address speech coding, high-quality packet loss concealment and multiple description coding.

The rest of the paper is organized as follows. Section 2 describes the parameters that HMC uses to represent/model the speech signal. The quantization of these parameters is addressed in Section 3. Section 4 discusses the multiple description quantization of the HMC parameters. The PLC algorithm is briefly presented in Section 5. Finally, the two aforementioned HMC codecs (HMC-SD and HMC-MD) are evaluated in Section 6 and conclusions are presented in Section 7.

2. HMC representation

The harmonic representation of the speech signal is a high quality parametric model used for analysis/synthesis [16]. Signal $x(n)$ is analyzed with a frame rate of 10 ms using a *Hanning* window of 20 ms long. Synthesis is made with simple OLA (OverLap-Add) techniques. Within each frame, the signal is represented as a weighted sum of harmonically related sinusoids:

$$\hat{x}(n) = \sum_{k=1}^{K} A_k \cos(k\omega_0 n + \phi_k), \quad (1)$$

where ω_0 is the fundamental frequency (in radians), K is the number of the harmonics, A_k and ϕ_k are the amplitude and the phase of the k-th harmonic, and n is the time index. The amplitudes and the phases can be obtained using least squares methods [17].

Harmonic amplitudes A_k can be well represented using a minimum phase RCC (Real Cepstrum Coefficients) spectral envelope of order 20. The cepstral envelope fits the log-spectra, $\log(A_k)$, at the Mel-scale by solving a regularized least squares problem [17].

Let $H_s(\omega)$ be the frequency response of the RCC envelope. Phases ϕ_k are decomposed into a minimum phase term $\angle H_s(k\omega_0)$, a *linear phase term* $k\omega_0\tau$ and a *dispersion term* ψ_k:

$$\phi_k = k\omega_0\tau + \angle H_s(k\omega_0) + \psi_k \quad (2)$$

Phase term $\psi_k + k\omega_0\tau$ corresponds to the phase of an excitation-like signal since the subtraction of the minimum phase term corresponds to inverse filtering with the spectral envelope $H_s(\omega)$, assuming that all amplitude spectrum information is captured by the magnitude response of that spectral envelope [16]. $H_s(\omega)$ may be considered as the frequency response of the vocal tract system. The excitation signal $e(n)$ can be reconstructed according to the formula:

$$e(n) = \sum_{k=1}^{K} \cos(k\omega_o n + k\omega_0\tau + \psi_k). \quad (3)$$

The linear translation term τ corresponds to the distance between the center of the analysis window and a reference point inside the pitch period. As a reference point, we used the peak of the excitation $e(n)$ within a single pitch period. The peak-picking is made to a uniformly sampled version of the excitation $e(n)$, using 128 samples (7 bits). We found that this procedure provided robust reference points within the glottal cycle and observed that the dispersion phases ψ_k, $k = 1, ..., K$ have a distribution that exhibits structure (covariation) [14].

Each frame is characterized as "silent", "unvoiced", "transitional" or "voiced" according to the frame energy and the SNR (Signal-to-Noise-Ratio) of the harmonic representation. The following parameters are then extracted from each frame: the frame energy E, the probability of voicing P_v, the fundamental frequency ω_0, the RCC spectral envelope, the dispersion phases ψ_k, and the linear phase term τ. The probability of voicing P_v, and it states the strength of the harmonic components of the signal, is non-zero only for voiced frames. The characterization of the frame and P_v are jointly quantized using only 3 bits/frame. The energy E is quantized with 8 bits/frame, while ω_0 is quantized with 8 bits/frame (only for transitional and voiced frames). At unvoiced and silent frames, ω_0 is set to a fixed value of 100 Hz. The RCC envelope is quantized for all frames with a varying number of bits using a single or a multiple description quantizer. However, in order to enhance the peak-to-null response of the spectral envelope, we apply, prior to synthesis, an adaptive post-filtering technique [16] (pg. 628) only at the transitional and voiced frames. Phase parameters

ψ_k and τ are quantized only for transitional and voiced frames, while they are randomly set for the other frame classes. The linear phase term τ is quantized with 7 bits/frame, while the dispersion phases ψ_k are quantized for all frames with a varying number of bits using a single or a multiple description quantizer.

Two HMC codecs are evaluated; a single description codec named **HMC-SD** and a multiple description codec referred to as **HMC-MD**. The codecs differ solely on the quantizers used to encode the RCC envelopes and the dispersion phases. The bit-allocation for these parameters and the average rate of the evaluated HMC codecs are shown in Table 1. The following sections discuss the single and multiple description quantizers used in these codecs.

CODEC	HMC-SD		HMC-MD	
Parameters	RCC	ψ_k	RCC	ψ_k
Silent	20	0	30	0
Unvoiced	60	0	90	0
Transitional	50	100	70	130
Voiced	60	100	90	130
Average	14191 (1076)		20823 (1495)	

Table 1: Bit-allocation for the RCC parameters and the dispersion phases ψ_k. The numbers indicate the total amount of bits allocated to one frame. Both codecs operate at a rate of 100 frames/sec. The average rates for all test-set utterances are also included along with the standard deviation (within parenthesis).

3. Single Description Quantization

In HMC-SD, the RCC envelope is quantized using GMM-based vector quantization with 16 Gaussian components and full covariance matrices [18]. The quantization of the dispersion phases is somewhat more tricky, due to the modulo-2π behavior of the phase parameters and the variable size of the dispersion phase vector. A novel phase quantization scheme is used [14]. Dispersion phases are modeled and quantized using Wrapped Gaussian Mixture Models (WGMM) with diagonal covariance matrices. WGMM are statistical models that respect the modulo-2π behavior of phase data. The varying dimensionality is addressed with a *split-VQ approach*: the pitch range is divided in 7 classes and the frequency range is divided in two bands, a low frequency band and a high frequency band. The dispersion phase vectors of the low frequency band are quantized using a pre-trained WGMM with a fixed number of dimensions while the remaining (high-band) dispersion phases are quantized using a WGMM that is constructed on-the-fly according to the available statistics of high-band dispersion phases. For more information, the interested reader is referred to [14].

4. Multiple Description Quantization

In HMC-MD, the RCC envelope is quantized using the multiple description quantization method that is proposed in [15]. The latter method is an extension of GMM-based quantization [18] to multiple descriptions. The dispersion phases are quantized with the WGMM-based multiple description quantizer that is proposed in [13] (Section 6.4). The latter scheme is an extension of WGMM-based phase quantization to multiple description coding and encodes the dispersion phases with the same split-VQ approach and the same pre-trained WGMM parameters that were used in HMC-SD (Section 3).

The rates of the RCC envelopes and the dispersion phases for each frame class are shown in Table 1. These rates indicate the number of bits allocated to a 20 ms frame at a frame rate of 100 frames/sec. The descriptions are *balanced* in the sense that they have the same average distortion. Furthermore, both quantizers are tuned to operate at the highest redundancy point where each of these two descriptions is nearly as good as an optimized half-rate single description encoding of the source. For example, if we receive only one description of the RCC envelope of a voiced frame, we will have a distortion that is (on average) approximately equivalent to the distortion of an $90/2 = 45$-bit encoding with GMM-based quantization. When both descriptions are received, we get a lower distortion, roughly equivalent to the distortion of a 60-bit encoding of the RCC envelope. Thus, the distortion levels at which HMC-SD and HMC-MD operate under no packet losses are more-or-less the same. Finally, as in HMC-SD, the employed MD quantizers are rate-scalable and capable of operating at high rates with affordable complexity.

5. Packet Loss Concealment

The sinusoidal representation of speech in HMC allows fine-tuned manipulation of the signal components, a necessary characteristic for high-quality PLC. The algorithm uses the jitter buffer to perform *interpolation* when a future speech frame is available or *extrapolation* when the jitter buffer is empty. At either case, a frequency threshold is set according to the probability of voicing P_v: harmonics above $P_v \frac{F_s}{2}$ (where F_s is the sampling rate) are considered to be noisy (*"unvoiced"*) while harmonics below $P_v \frac{F_s}{2}$ are considered to be deterministic (*"voiced"*). A different type of synthesis is used for each case. When interpolation is made, the harmonics of the start-frame are linked to the harmonics of the end-frame via a pair-matching or a death-birth procedure. Pair-matching is used when the paired sinusoids are close in frequency and death-birth at all other cases. The paired sinusoids, which may not be harmonically related anymore, are synthesized according to their voicing state at the start-frame and the end-frame. The interpolation between two voiced harmonics is made using a cubic phase model [16] (pg. 446). If either of the start-frame or the end-frame harmonics is unvoiced then a birth-death procedure is used for synthesis. When the jitter buffer is empty, the voiced harmonics are extrapolated according to their amplitudes and phases at the previous frame, while the unvoiced harmonics are randomly set and synthesized via OLA (OverLap-Add) using a Hanning window. Details regarding the PLC algorithm can be found in [13] (Chapter 5).

For HMC-SD we also used a small "corrective" bitstream of 600 bit/sec that holds the energy of past frames. Let E_n be the energy of the n-th frame. The energy is encoded in a backwards predictive manner: for the n-th frame, we predictively encode E_{n-1} using E_n. Then, we encode E_{n-2} using E_{n-1} and so on. The encoding is made for 4 previous frames and uses 3 bits/frame for each energy coefficient. Assuming a jitter buffer of 4 frames and placing 2 frames in every packet, then $3 * 4 * 100/2 = 600$ bits/sec are needed for the corrective bitstream. The corrective bitstream is particularly useful for packet losses that occur at the beginning or at the end of words where the presented PLC algorithm may produce an unnatural onset or lengthening of the word.

6. Experiments and Results

In this section we subjectively evaluate HMC-SD and HMC-MD using iLBC (at the 20 ms, 15.2 kbps mode) as a baseline.

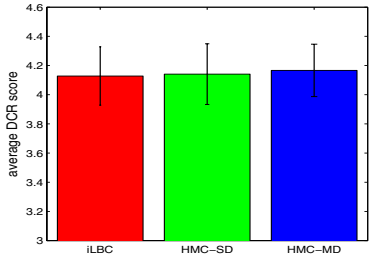

Figure 1: *Subjective evaluation (mean and 95% confidence interval) of the HMC codecs and iLBC according to the DCR test.*

The evaluation is made according to the Degradation Category Rating (DCR) scale which states the degradation of the evaluated signal in comparison to a reference signal. A score of 5 is given when the degradation is not perceived, 4 when the degradation is perceived but not annoying, 3 when it is slightly annoying, 2 when it is annoying and 1 when it is very annoying. Figure 1 shows the DCR score between the original signal and the encoded signal for the three codecs when no packets are lost. The test was conducted with 16 listeners using 30 utterances (15 males and 15 female). Note that the HMC codecs are equivalent to iLBC in terms of quality under perfect channel conditions.

Another DCR test was made with 20 listeners using 18 utterances (9 male and 9 female) to evaluate the robustness of the codecs under packet losses [13] (Chapter 9). The comparison was made between the encoded signal under no packet losses and the encoded signal under packet losses in order to isolate the effect of packet losses from the effect of quantization and to recalibrate the scale for more refined measurements. The losses were introduced according to the Gilbert-Elliot model; a two-state Markov model that is used to simulate the bursty losses usually encountered in IP networks. To reduce the degrees of freedom, we constrained the probability of losing a packet after a packet loss to be twice the probability of losing a packet when the previous packet is received, as made in [5]. All codecs encoded 20 ms of speech in each packet, corresponding to 2 frames/packet in HMC-SD and HMC-MD. For the HMC codecs we used a 40 ms jitter buffer. Furthermore, the single description codecs were evaluated at the same loss patterns. Each packet of HMC-MD contained two descriptions from two consecutive speech frames; a total rate of 100 packets/sec. Half of these packets correspond to the first description and the rest to the second description. Each description is routed through an independent Gilbert-Elliot channel with the same loss probabilities. The results are shown in Figure 2. At the same bitrate, HMC-SD is considerably more robust to packet losses than iLBC for losses above 5%, while it is capable of operating with minimal degradation for rates up to 20%. Above these loss rates it suffers a rapid degradation but HMC-MD can then be used for loss rates up to 40%.

7. Discussion

We presented a framework for joint source/channel coding of speech for VoIP that combines three useful properties: efficient speech coding, high-quality packet loss concealment and multiple description coding. Two proof-of-concept codecs were designed and evaluated with promising results. However, many improvements can yet be made; perceptual weighting and analysis-by-synthesis techniques for phase quantization, predictive encoding within each packet and frame classification so that we can progressively use MDC to frames according to their perceptual importance.

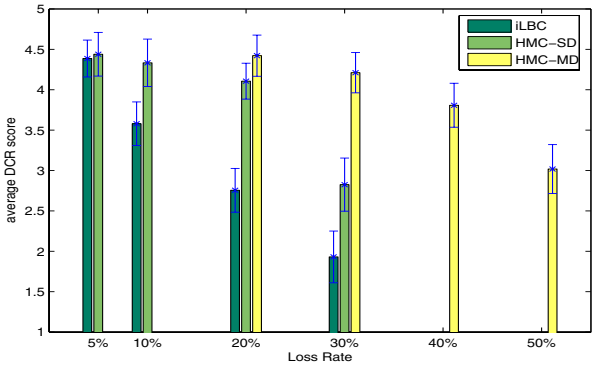

Figure 2: *Average DCR score and 95% confidence intervals of iLBC, HMC-SD and HMC-MD codecs for several loss rates.*

8. References

[1] M. Chibani, R. Lefebvre, and P. Gournay, "Resynchronization of the adaptive codebook in a constrained CELP codec after a frame erasure." in *ICASSP*, Toulouse, France, 2006.

[2] W. Jiang and H. Schulzrinne, "Comparisons of FEC and codec robustness on VoIP quality and bandwidth efficiency," in *ICN*, 2002.

[3] J. Lim, "Joint and tandem source-channel coding with complexity and delay constraints," Ph.D. dissertation, University of Michigan, 2001.

[4] C. Garrido, M. N. Murthi, and S. Andersen, "On variable rate frame independent predictive speech coding: Re-engineering iLBC," in *ICASSP*, vol. 1, Toulouse, 2006.

[5] J. Lindblom, "Coding speech for packet networks," Ph.D. dissertation, Chalmers University Of Technology, 2003.

[6] A. R. Christoffer, G. C. Mads, A. Soren Vang, and J. Soren Holdt, "Compressed domain packet loss concealment of sinusoidally coded speech," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003.

[7] R. Rodbro and S. H. Jensen, "Time-scaling of sinusoids for intelligent jitter buffer in packet-based telephony," in *IEEE Workshop on Speech Coding*, 2002.

[8] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Processing*, 2004.

[9] G. Kubin and W. Kleijn, "Multiple-description coding (mdc) of speech with an invertible auditory model," in *Speech Coding Proceedings, 1999 IEEE Workshop on*, 1999, pp. 81–83.

[10] X. Zhong and B.-H. Juang, "Multiple description speech coding with diversities," in *ICASSP*, vol. 1, 2002, pp. 177–180.

[11] A. Anandakumar, A. McCree, and V. Viswanathan, "Efficient CELP-based diversity schemes for VoIP," in *ICASSP*, vol. 6, Istanbul, Turkey, 2000, pp. 3682–3685.

[12] A. Ingle and V. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech." *IEEE Trans. on Speech and Audio Processing*, no. 1, pp. 48–58, 1995.

[13] Y. Agiomyrgiannakis, "Sinusoidal Coding of Speech for Voice-over-IP." Ph.D. dissertation, Department of Computer Science, University of Crete, Jan. 2007. [Online]. Available: http://www.csd.uoc.gr/˜jagiom/voip.htm

[14] Y. Agiomyrgiannakis and Y. Stylianou, "Stochastic modeling and quantization of harmonic phases in speech using wrapped gaussian mixture models," in *ICASSP*, 2007.

[15] J. Samuelsson and J. Plasberg, "Multiple description coding based on Gaussian Mixture Models," *IEEE Signal Processing Letters*, June 2005.

[16] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2001.

[17] Y. Stylianou, "Harmonic plus noise Models for speech, combined with statistical methods for speech and speaker modification," Ph.D. dissertation, Ecole Nationale. Sup'erieure des Telecommunications, Paris, France, 1998.

[18] A. D. Subramaniam, "Gaussian Mixture Models in compression and communication," Ph.D. dissertation, University of California, San Diego, 2003.