ARITHMETIC CODING OF SPEECH AND AUDIO SPECTRA USING TCX BASED ON LINEAR PREDICTIVE SPECTRAL ENVELOPES

Tom Bäckström and Christian R. Helmrich

International Audio Laboratories Erlangen[1], Friedrich-Alexander University (FAU)
Am Wolfsmantel 33, 91058 Erlangen, Germany. tom.backstrom@audiolabs-erlangen.de

ABSTRACT

Unified speech and audio codecs often use a frequency domain coding technique of the transform coded excitation (TCX) type. It is based on modeling the speech source with a linear predictor, spectral weighting by a perceptual model and entropy coding of the frequency components. While previous approaches have used neighbouring frequency components to form a probability model for the entropy coder of spectral components, we propose to use the magnitude of the linear predictor to estimate the variance of spectral components. Since the linear predictor is transmitted in any case, this method does not require any additional side info. Subjective measurements show that the proposed methods give a statistically significant improvement in perceptual quality when the bit-rate is held constant. Consequently, the proposed method has been adopted to the 3GPP Enhanced Voice Services speech coding standard.

Index Terms— speech and audio coding, frequency domain coding, arithmetic coding

1. INTRODUCTION

Speech and audio coding technologies applied in modern standards such as MPEG USAC, G.718, AMR-WB+ and, importantly, the ETSI 3GPP Enhanced Voice Services, use multiple modes such as time-domain coding with ACELP and frequency-domain coding with TCX to gain efficient coding for as many signal types as possible [1–4]. Generally, time-domain coding provides superior performance for signals with rapidly changing character and temporal events, such as spoken consonants, applause and percussive signals. Coding in the frequency-domain, on the other hand, is more effective for stationary signals such as harmonic music signals and sustained voiced speech sounds.

In this work we will focus on coding in the frequency-domain using models of the spectral envelope. Observe that there are two distinct types of spectral envelope models in classical literature and technologies; First of all, dedicated speech codecs are generally based on linear predictive models, which model the spectral energy envelope using an IIR filter. In contrast, classical audio codecs such as MP3 and the AAC family model the perceptual masking envelope [5, 6]. While these two envelopes do have many common features – their peaks and valleys are located at the same general frequency areas – the magnitude of peaks and valleys as well as the overall spectral tilt are very different. Roughly speaking, masking envelopes are much smoother and exhibit smaller variations in magnitude then the energy envelopes.

AAC-type codecs use the perceptual masking model to scale the spectrum such that the detrimental effect of quantization on spectral components has perceptually the same expected magnitude in every part of the spectrum [6]. To allow efficient coding of the perceptual spectrum, these codecs then apply entropy coding of the frequency components. For higher efficiency, the arithmetic coder can use the neighbouring spectral components to determine the probability distribution of the spectral components, such as in USAC [1, 7]. Speech codecs on the other hand use energy envelopes as a signal model and apply a perceptual weighting filter, much like the perceptual masking model, on top.

The current work relies on the fact that the spectral envelope, as described by the linear predictive model, provides information of the energy envelope of the spectrum. Since it thus describes the energy distribution of the spectral components, it can be used to describe their probability distributions. This distribution can, in turn, be used to design a highly efficient arithmetic coder for the spectral components. Since the linear predictive model is generally transmitted also for TCX frames, this spectral envelope information comes without additional side-information. In contrast to AAC-type codecs, we thus use an explicit source model in the form of the linear predictor, and in difference to TCX-type codecs, we use an adaptive probability distribution for the arithmetic codec derived from the magnitude of the linear predictor.

In this paper we propose a signal adaptive model of the probability distributions of spectral components based on the linear predictive model. The goal is to obtain a fixed bit-rate arithmetic coder applicable in speech and audio codecs which use linear predictive modeling. Moreover, our objective is to design a generic method which is efficient on a variety of bit-

[1]International Audio Laboratories Erlangen is a joint institute between Fraunhofer IIS and the Friedrich-Alexander University (FAU).

rates and bandwidths.

The proposed encoder has three steps. First, we use the perceptually weighted linear predictor as a model for the shape of the perceptually weighted spectrum. Since this envelope does not contain information of the signal magnitude, we scale the envelope such that the expected bit-consumption of a signal, whose variance follows the envelope, matches the desired bit-rate. Second, we scale and quantize the actual perceptually weighted spectrum such that the bit-rate matches the desired bit-rate, when using the envelope model. Finally, we can then encode the spectrum with an arithmetic coder. The decoder can then apply the same scaling of the envelope to decode the spectral lines.

2. MODELING THE PROBABILITY DISTRIBUTION

Let A_k^{-1} be the samples of the discrete Fourier transform of that linear predictive model which describes the short-time temporal structure and thus the spectral envelope of a signal spectrum S_k. The filter residual can be obtained by multiplying S_k with A_k to obtain the residual $X_k = A_k S_k$. Given that A_k is an efficient model of S_k, then X_k will be the spectrum of a white-noise signal. It follows that the expected energy of every frequency component k is constant $\sigma_x^2 = \mathcal{E}\left[|X_k|^2\right]$. Conversely, the expectation of the energy of the perceptual weighted signal S_k is

$$\sigma_{s,k}^2 = \mathcal{E}\left[|S_k|^2\right] = \sigma_x^2 \left|A_k^{-1}\right|^2. \quad (1)$$

For perceptual weighting of quantization errors, prior to quantization, the spectrum is weighted by a perceptual masking model W_k. It follows that the expected energy of the weighted spectrum $Y_k = W_k S_k$ is

$$\sigma_{y,k}^2 = \mathcal{E}\left[|W_k S_k|^2\right] = \sigma_x^2 \left|W_k A_k^{-1}\right|^2. \quad (2)$$

This relation quantifies the relative energy of spectral components and can be used as an estimate of their relative variance in the design of models of the probability distribution of weighted spectral components.

We can then choose a probability distribution model for the individual spectral components. The most obvious candidates are either the normal or the Laplacian distribution, since they are both simple to implement and commonly known to be fairly accurate models of speech and audio signals. To determine which distribution fits our approach better, we conducted the following experiment.

As material for the test, we used the 16 critical items of speech, mixed and music material used in the standardization of MPEG USAC [1]. The material was resampled to 12.8kHz and windowed with sine-windows, and transformed to the frequency domain using the MDCT, with 256 frequency bands and full-overlap. Linear predictive models of order 16 were estimated for each frame using a Hamming window of length

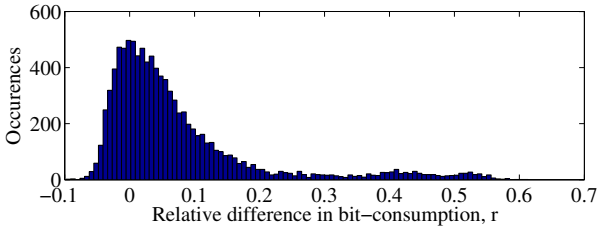

Fig. 1. Histogram of relative differences in bit-consumptions b_n and b_l when using the normal and Laplacian distributions, respectively, as measured by $r = \frac{b_n - b_l}{(b_n + b_l)/2}$.

30ms, centered to align with the MDCT windows. The spectrum S_k of each frame was whitened in the frequency domain with the corresponding envelope model A_k, and perceptually weighted with W_k, to obtain the perceptually white spectrum $\tilde{X}_k = A_k W_k S_k$. The spectrum was then scaled by the standard deviation, to remove the effect of signal gain, whereby variance of the scaled signal was unity. The bit-consumption of both distributions was then estimated in each frame by

$$b = -\sum_k \log_2 p(\tilde{X}_k), \quad (3)$$

where we estimated the probability with the distributions $p(\tilde{X}_k) = \sqrt{2}\exp\left(-\frac{|\tilde{X}_k|}{\sqrt{2}}\right)$ for the Laplacian distribution and $p(\tilde{X}_k) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{|\tilde{X}_k|^2}{2}\right)$ for the normal distribution.

Finally, we calculated the bit-consumption of all frames in the data for both the normal and Laplacian distributions. The histogram of the relative difference in bit-consumption between the distributions is shown in Figure 1. We can observe that in a majority of cases, modeling with a normal distribution requires more bits than with a Laplacian. Moreover, importantly, encoding with the Laplacian never yields a large increase in bit-consumption, whereas it sometimes gives a big reduction in bit-rate. On average, the Laplace distribution gave a bit-consumption 6% smaller than the normal distribution. In addition, the highest gains where found for frames with stationary harmonic content, where the spectrum is sparse, which is also exactly the frames where TCX should be used. We thus concluded that in our application a Laplacian is better than a normal distribution.

Since our target is to design a codec with a fixed bit-rate, we then need to scale the weighted envelope such that its expected bit-consumption matches the target bit-rate. A Laplacian random variable with variance σ^2 has a bit-consumption expectation of $1 + \log_2(e\sigma)$, where e is Euler's number. The variance for the magnitude of the weighted spectrum is then

$$\sigma_{y,k}^2(\gamma) = \gamma^{-2}|W_k A_k^{-1}|^2, \quad (4)$$

where γ is the scaling coefficient. It follows that the expectation of bit-consumption of a spectrum with N coefficients

is

$$B = N + \sum_{k=0}^{N-1} \log_2 \left(e\gamma\sigma_{y,k} \right). \tag{5}$$

We can then solve the above equation for γ:

$$\gamma = \frac{2^{\frac{B-1}{N}}}{e} \left(\prod_{k=0}^{N-1} \sigma_{y,k} \right)^{-1/N}. \tag{6}$$

It follows that when the spectral envelope is scaled by the γ given by Eq. 6, then the expected bit-rate is B for signals whose magnitude follows the envelope.

3. CODING SPECTRAL COMPONENTS

The proposed arithmetic coder is based on a Laplacian distribution, which is equivalent with a signed exponential distribution. To simplify the process, we can thus first encode the absolute magnitude of spectral lines and for non-zero values then also encode their sign with one bit.

The absolute magnitude can thus be encoded using an arithmetic coder with an exponential distribution. If a spectral component is quantized to an integer value q, then the original value has been in the interval $[q-\frac{1}{2}, q+\frac{1}{2}]$, whose probability is given by

$$\begin{aligned} p(|\widehat{Y}_k|) &= e^{-\lambda(q-\frac{1}{2})} - e^{-\lambda(q+\frac{1}{2})} \\ &= \left[e^{+\frac{\lambda}{2}} - e^{-\frac{\lambda}{2}} \right] e^{-\lambda q} = C e^{-\lambda q}, \end{aligned} \tag{7}$$

where $\lambda = \sqrt{2}/\sigma_k$, the scalar σ_k is the standard deviation of the kth frequency component and $C = e^{+\frac{\lambda}{2}} - e^{-\frac{\lambda}{2}}$ is a constant. Using this probability, we can easily design a standard arithmetic coder for the spectral lines [8]. With this arithmetic coder we can then encode each frequency component consecutively, from the low to high frequencies.

Since speech and audio spectra are often dominated by low-frequency content, at low bit-rates it is a common that large sections of the high-frequency spectra is sparse or zero. To improve efficiency, we can therefore discard trailing zeros. We set the decoder to decode the spectra until the maximal bit-consumption is reached and set any remaining spectral components to zero. The encoder thus encodes the spectrum up to the last non-zero frequency. Informal experiments showed that with this approach the number of encoded frequency components is often reduced by 30 to 40%.

To use the allocated bits efficiently, the spectrum must be scaled to obtain the highest accuracy which can be encoded with the available bits. As an initial guess for the scaling of the input spectrum Y_k, we can scale it such that its energy matches the energy of the envelope that was scaled to the desired bit-rate. The optimal scaling can then be determined in a rate-loop implemented as a binomial search. Informal experiments showed that the best scaling can usually be found within 5 iterations.

4. IMPLEMENTATION DETAILS

The objective of the arithmetic coder is to encode the spectrum into the bit-stream as efficiently as possible. To encode and decode the bit-stream, the numerical operations must be implemented with fixed-point operations such that differences in numerical accuracy across different platforms will not change the outcome. We chose to use a 14 bit integer representation stored in a 16 bit integer variable, which allows for a sign bit and avoids problems in the last bit due to differences in rounding.

A related problem in implementation of the arithmetic coder is that when the standard deviation σ_k is small compared to the quantized magnitude $|\widehat{Y}_k|$, then the probability $p(|\widehat{Y}_k|)$ will become so small that it is rounded to zero in the 14 bit representation. In other words, the numerical accuracy is exhausted and we would be unable to encode such values. To overcome this problem, we can use the memorylessness property of exponential distributions to our advantage. This principle states that

$$p(|\widehat{Y}_k| > r + s \,\big|\, |\widehat{Y}_k| > r) = p(|\widehat{Y}_k| > s). \tag{8}$$

In our case, if r is the threshold above which the numerical representation saturates, then we can first encode the probability $p(|\widehat{Y}_k| > r)$ and then continue by encoding the probability $p(|\widehat{Y}_k| - r)$. By successively increasing r, we can thus guarantee that numerical saturation is avoided.

In the implementation of the rate-loop, observe that we need to output the bit-stream only for the final scaling coefficient. To reduce computational complexity, it is then sufficient to only estimate bit-consumption in the rate-loop and invoke the arithmetic coder only with the optimal scaling.

The bit-consumption of the spectrum for a given scaling coefficient γ can be efficiently estimated as follows. Let the quantized value be $\widehat{Y}_k = \mathrm{round}(\gamma Y_k)$ whereby the total bit-consumption $B(\gamma)$ can be calculated by

$$B(\gamma) = \sum_k -\log_2 p(\widehat{Y}_k) = -\log_2 \prod_k p(\widehat{Y}_k). \tag{9}$$

Instead of calculating the logarithm for every component, we can thus calculate a logarithm of the product $\prod_k p(\widehat{Y}_k)$ and thereby reduce computational complexity. Note however that the product can become a very large number, whereby we must implement the calculation of the product in a representation where we can guarantee that an overflow cannot occur.

Once the best possible scaling for quantization has been achieved, we can encode the spectral lines. However, since the rate-loop described above only approximates actual arithmetic coding, it may sometimes happen that the bit-budget is slightly exceeded. A safe-guard against this improbable event is to reduce the magnitude of the last encoded line until the actual bit-consumption remains within the budget.

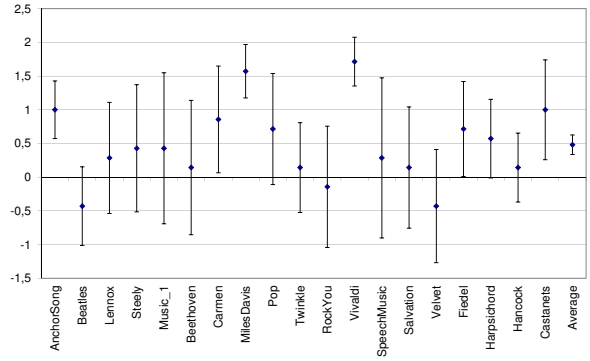

Fig. 2. The differential AB scores and their 95% confidence intervals of a comparison listening test measuring the performance of the proposed arithmetic coder in comparison to the coder from MPEG USAC at 8 kbps for wide-band signals.

5. EXPERIMENTS

To determine the performance of the proposed coding method, we performed a subjective AB comparison listening test [9]. For this test, we implemented the proposed codec in a candidate version of the 3GPP Enhanced Voice Services speech and audio codec [4]. As comparison, we used an arithmetic coder derived from the MPEG Unified Speech and Audio Coder, which forms a probability model of spectral lines using neighbouring lines [1, 7]. The codecs were operating at a fixed bit-rate of 8 kbps in the wide-band mode. Bandwidth extension was encoded normally but its output was disabled to concentrate on differences of the core bandwidth.

The double-blind AB comparison test was performed by 7 expert listeners in a silent room with high-quality headphones. The test included 19 samples of mono speech, music, and mixed material. The differential results of the compare test are illustrated in Figure 2, where a positive score indicates a subjective improvement of the proposed codec over the reference codec using the arithmetic coder of MPEG USAC.

These results show that the mean AB score indicates a statistically significant improvement of 0.48 points with 95% confidence. Moreover, no item had a statistically significant reduction in perceptual quality, but 6 out of 19 items had a statistically significant improvement. Note that it is remarkable that the improvement is statistically highly significant even with a limited number of listeners.

6. DISCUSSION AND CONCLUSIONS

We have presented a method for modeling the probability distribution of perceptually weighted frequency components of speech and audio signals using a model of the spectral envelope and the perceptual weighting function. Frequency-domain codecs based on the TCX concept model the spectral envelope using linear prediction, from which an estimate of the perceptual masking curve can be obtained. Since the linear predictor is transmitted in any case, the proposed method can be applied without transmission of additional side-information.

The proposed method uses the spectral envelope model as a model of the speech source for construction of a probability model for the entropy coder. In contrast, conventional methods have used preceding frequency components to predict the magnitude of the current component [1, 7]. The conventional methods thus use an implicit source model, whereas the proposed approach models the source explicitly.

Note that in estimation of the bit-consumption of the spectral envelope, we have used the theoretical entropy of Laplacian distributions, which is accurate only when the quantization accuracy is very high. The bias at lower bit-rates is due to fact that when a spectral line is quantized to zero, its sign does not need to be transmitted, whereby 1 bit is saved. When a significant part of the spectrum is quantized to zero, a rather large number of bits is saved, whereby our bit-consumption estimates are too high. Informal experiments, however, show that a more accurate estimate of bit-consumption increases complexity significantly, but that the impact on overall bit-consumption was marginal. Such more accurate and complex estimates of the bit-consumption of spectral envelopes could thus be avoided.

Whereas the current work uses only the spectral envelope associated with the linear predictor, observe that speech codecs regularly use also other information which can be used to estimate spectral magnitude. Namely, long term prediction (LTP) is regularly used to model the fundamental frequency. The long term predictor can be used to model the comb-structure of spectra with a dominant fundamental frequency. Investigations with such refined envelope models are left for further study.

The presented results demonstrate that the proposed method improves perceptual quality at low bit-rates when the bit-consumption is kept constant. Specifically, subjective measurements with an AB test showed a statistically significant improvement. The proposed coding scheme can thus be used to either increase quality at a given fixed bit-rate or decrease the bit-rate without losing perceptual quality.

The presented approach is applicable in all speech and audio codecs which employ a frequency-domain coding of the TCX type, where a model of the spectral envelope is transmitted to the decoder. Such codecs include standards such as MPEG USAC, G.718 and AMR-WB+ [1–3]. In fact, the method has already been included in the ETSI 3GPP Enhanced Voice Services standard [4].

References

[1] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. S. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG unified speech and audio coding standard – consistent high quality for all content types and at all bit rates," *Journal of the AES*, vol. 61, no. 12, pp. 956–977, 2013.

[2] ITU-T G.718, "Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s," 2008.

[3] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. ICASSP*, 2005, vol. 2, pp. 1109–1112.

[4] 3GPP, *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 2014.

[5] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG–2 Advanced audio coding," in *101 AES Convention*, 2012.

[6] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[7] G. Fuchs, M. Multrus, M. Neuendorf, and R. Geiger, "Mdct-based coder for highly adaptive speech and audio coding," in *European Signal Processing Conference (EUSIPCO 2009)*, 2009, pp. 24–28.

[8] I. H. Witten, R. M Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[9] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, 1996.