ENVELOPE MODELING FOR SPEECH AND AUDIO PROCESSING USING DISTRIBUTION QUANTIZATION

Tobias Jähnel[*], *Tom Bäckström*[*†] *and Benjamin Schubert*[†]

[*]International Audio Laboratories Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU)
[†]Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany
{tobias.jaehnel, tom.backstrom}@audiolabs-erlangen.de

ABSTRACT

Envelope models are common in speech and audio processing: for example, linear prediction is used for modeling the spectral envelope of speech, whereas audio coders use scale factor bands for perceptual masking models. In this work we introduce an envelope model called distribution quantizer (DQ), with the objective of combining the accuracy of linear prediction and the flexibility of scale factor bands. We evaluate the performance of envelope models with respect to their ability to reduce entropy as well as their correlation to the original signal magnitude. The experiments show that in terms of entropy, distribution quantization and linear prediction are comparable, whereas for correlation, distribution quantization is better. Furthermore the coefficients of distribution quantization are independent and thus more flexible and easier to quantize than linear predictive coefficients.

Index Terms— Speech coding, linear predictive coding, signal modeling

1. INTRODUCTION

Modern multimedia devices all come shipped with a variety of speech and audio applications, with features ranging from music transmission, speech coding and enhancement to spatial audio reproduction. A common feature of many of these algorithms is that they use envelope models to describe the signal or its characteristics. We define an envelope as a continuous, usually smooth shape describing a characteristic of the signal. Figure 1 illustrates an envelope model generated for the spectrum of a speech signal. This paper focuses on spectral envelopes, but results can readily be extended to other domains.

A typical envelope model is linear prediction [1], which is in speech coding used to model the spectral envelope of a signal (see section 2.1). The model residual can then be encoded with a lower number of bits than the original signal. This approach is used in main-stream speech codecs such as AMR-WB, MPEG USAC and 3GPP Enhanced Voice Services [2–4]. Several improvements have been applied to linear prediction over time, but generally their complexity

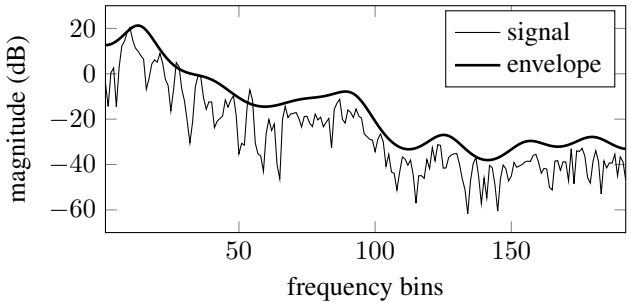

Fig. 1: Illustration of how an envelope (thick line) models the overall shape of the spectrum of a signal (thin line).

in implementation, quantization and coding is higher than in normal linear prediction [5–7]. Therefore in this work we only consider the basic form of linear prediction as described in [1].

On the other hand, audio coders such as MP3, AAC and MPEG USAC use scale factor bands (see section 2.2) to model the perceptual masking envelope [3, 8, 9]. The same model is also used in bandwidth extension, which is implemented in newer codecs such as AAC and MPEG USAC [10]. This method efficiently encodes high-frequency regions by replicating spectral fine structure from lower bands and shaping them by an envelope function.

Another application of envelopes is temporal noise shaping (TNS) [11]. It addresses temporal pre-echo effects caused by quantization noise of transient events such as clicks. By modeling the temporal envelope by a frequency domain linear predictor, TNS can attenuate such temporal smearing.

Our long-term objective is to develop envelope models which would cover all above applications. This work introduces an envelope model called distribution quantizer (DQ) [12], which provides a balance between the accuracy of linear prediction and the flexibility of scale factor bands (see section 3). The approach is based on describing the distribution of spectral mass by dividing the spectrum into blocks of equal magnitude.

To evaluate the performance of envelope models we use two measures. Firstly, we calculate entropy which corre-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

sponds to the number of bits needed to transmit the remaining signal. Secondly, we use correlation between the envelope and the input which corresponds to calculating the signal to noise ratio (SNR) of an envelope which is scaled optimally.

The experiments show that the entropy and thus coding efficiency of distribution quantizer is very close to linear prediction. Furthermore distribution quantizer correlates better with the input signal than linear prediction does. These results make the distribution quantizer an attractive candidate for flexible and efficient modeling of envelopes in applications which have used linear prediction or scale factors in the past.

2. ENVELOPE MODELS

Consider a sequence of N samples X_k with $1 \leq k \leq N$, modeled by an envelope with P degrees of freedom, or equivalently, where model order is P. When we apply the model to the signal we get for each X_k an envelope value E_k as well as a residual R_k that describes the error between signal and envelope. In our work we look at spectral envelopes and define a multiplicative relation to the signal magnitude in frequency domain such that

$$X_k = E_k R_k, \tag{1}$$

where X_k is the spectral mass of the original signal, E_k is the envelope and R_k the residual. Each envelope algorithm models the signal by its own samples E_k using P algorithm-specific parameters.

2.1. Linear Prediction

Linear prediction (LP) is used by many speech coding algorithms based on code-excited linear prediction (CELP) [1,13]. It models the input signal ξ_n as a sum of previous samples weighted with the model parameters $\alpha_{LP,p}$. With $\alpha_{LP,1} = 1$ the residual ϵ_n can be solved as a convolution:

$$\xi_n = -\sum_{p=1}^{P} \alpha_{LP,p} \xi_{n-p} + \epsilon_n \quad \Rightarrow \quad \epsilon_n = \sum_{p=0}^{P} \alpha_{LP,p} \xi_{n-p} \tag{2}$$

Generally, the coefficients $\alpha_{LP,p}$ are chosen so as to minimize the mean square error [1]. To compose the spectral envelope, we convert ξ_n and $\alpha_{LP,p}$ into frequency domain and get X_k and $A_{LP,k}$ respectively. According to equations 1 and 2 the definition of the spectral envelope E_k is then

$$X_k = E_{LP,k}\, A_{LP,k}\, X_k \quad \Rightarrow \quad E_{LP,k} = \frac{1}{A_{LP,k}}. \tag{3}$$

2.2. Scale factors bands

The idea of scale factor bands (SFB) is to split the spectrum into a predefined set of $P + 1$ bands, whose widths are specified by a perceptual model. Since we are only interested in the shape of the envelope, we can assume a fixed factor for the first band and encode the remaining P factors relative to the first factor. When each band is scaled with the associated factor $A_{SF,p}$, the error caused by quantization has the same expected value and thus the same perceptual effect on all frequencies.

According to equation 1 we define the envelope as $E_{SF,k} = A_{SF,p(k)}$, where $p(k)$ denotes the band that contains the sample x_k.

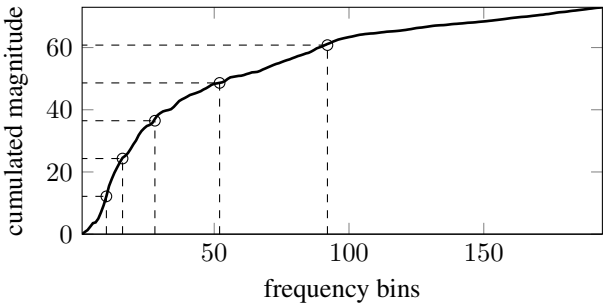

Fig. 2: To determine the split frequencies the distribution quantizer splits the cumulative mass (vertical axis) using $P = 5$ points into $P + 1$ equal size segments.

3. DISTRIBUTION QUANTIZATION

Linear prediction (section 2.1) is known to give the optimal polynomial solution to the minimum mean square error problem. However quantization of its parameters makes up a significant part of the complexity of modern codecs [14, 15]. Moreover, it is also difficult to incorporate perceptual criteria such as a non-uniform accuracy on the frequency axis in their estimation. On the other hand, envelope models based on scale factors (section 2.2) can be easily quantized since the parameters are uncorrelated, but the accuracy of such envelopes tends to be lower.

We propose an envelope model named distribution quantizer (DQ), which aims to combine the benefits of both linear prediction and scale factor bands. Its objective is to split the spectrum into equal-magnitude blocks, such that we only need to transmit the border-frequencies but not their level. Distribution Quantizer is heuristically similar to the line spectral frequency description of linear prediction coefficients [14]. The idea of line spectral frequencies is to split the linear prediction polynomial into two polynomials whose roots are on the unit circle and then only code the angle. This means, similar to distribution quantizer, they approximately describe the distribution of signal mass along the frequency axis.

Specifically, we first define the cumulative sum of the magnitude spectrum as $C_k = \sum_{m=0}^{k} X_m$ as depicted in figure 2. By choosing P equidistant points on the cumulative-magnitude axis, we obtain those P frequency bins which split the spectrum into equal-magnitude blocks. These P points

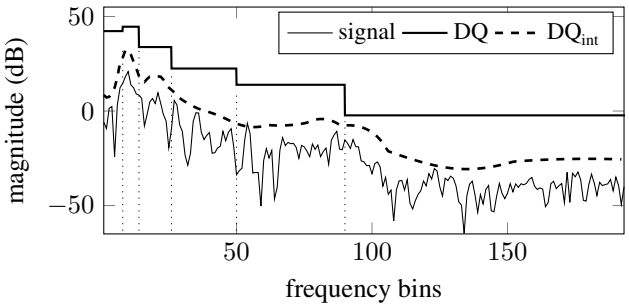

Fig. 3: In this example the distribution quantizer (DQ) has $P = 5$ splits points, which separate the spectrum into $P + 1$ segments of equal spectral mass. Spline interpolation in cumulative domain gives a smoother envelope in frequency domain (DQ_{int}). For better visibility both envelopes are shifted vertically.

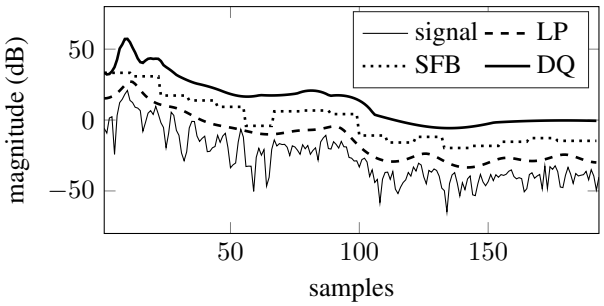

Fig. 4: Illustration of how spectral envelopes of linear prediction (LP), scale factor bands (SFB) and distribution quantizer (DQ) model one frame of a signal. For visual clarity the envelopes are shifted vertically.

are illustrated in figure 2 by dashed lines. Since we are working with discrete signals we can for better accuracy interpolate between the integer frequency bins. These consecutive split frequency bins then describe $P + 1$ blocks with equal magnitude.

In the next step, we synthesize a piece-wise constant spectral envelope model. This can easily be done since the frequency borders are known and all segments have the same magnitude $C_N/(P + 1)$. The result is depicted in figure 3 with a solid line.

Informal experiments showed that accuracy can be further improved by interpolating between the split points in cumulative magnitude domain using splines. With the frequencies that separate the segments, the overall magnitude of the spectrum and the knowledge that all segments have equal magnitude, we can instantly reconstruct the P significant points in the cumulative magnitude domain as depicted in figure 2. We can then interpolate between these points to estimate the original cumulative magnitude curve.

The spline interpolation is constrained to have a continuous derivative at the segment borders. This means the tilt T_p ($1 \leq p \leq P$) between neighboring segments is defined as

$$T_p = \frac{S_{p+1} - S_{p-1}}{F_{p+1} - F_{p-1}}, \tag{4}$$

where S_p is the cumulative magnitude at the split frequency bin F_p. Furthermore according to the property of the cumulative sum the edges are defined as $F_0 = 0$, $S_0 = 0$, $F_{P+1} = N$ and $S_{P+1} = C_N$ and the corresponding tilts are

$$T_0 = \frac{S_1 - S_0}{F_1 - F_0} = \frac{S_1}{F_1} \quad \text{and} \quad T_P = \frac{S_{P+1} - S_P}{F_{P+1} - F_P}. \tag{5}$$

As a final step, we transform the estimated curve back into the frequency domain by differentiating to get a smooth envelope. Figure 3 shows the effect of the spline interpolation on the envelope as a dashed line.

Distribution quantizer can be summarized as an envelope model of speech, that is described by P frequency bins that split the spectrum into segments of equal magnitude as well as the overall magnitude C_n. This is the same amount of information that is also needed to model linear prediction or scale factor band envelopes of the same order.

4. EVALUATION

To determine the performance of envelope models, we use entropy and correlation. It seems natural to also compare the bit-consumption of these models, but since in this work we did not perform any quantization this would not be a viable measure.

Therefore we first measured the entropy, which quantifies the amount of information in a signal. The entropy $\hat{E}_n$ for a value X_n is defined as

$$\hat{E}_n = -\log_2(P(X_n)), \tag{6}$$

where $P(X_n)$ denotes the probability of X_n. Entropy corresponds to the number of bits required to encode a signal, whereby a lower entropy means that a lower number of bits is required for coding.

In our experiment we assume a Laplacian distribution of the input signal so that the probability of a sample can be calculated using

$$P(X_n) = \frac{1}{2B_n} e^{-|X_n/B_n|} \quad \text{where} \quad B_n = \left| \frac{E_n}{\sqrt{2}} \right|. \tag{7}$$

Secondly, we measured the normalized correlation between the envelope models and the signal magnitude. This measure correlates with the signal to noise ratio, assuming that the envelope is scaled optimally, whereby it provides a natural measure of envelope quality.

We calculate the normalized correlation $\hat{C}$ using

$$\hat{C} = \frac{(\sum X_n E_n)^2}{\sum {X_n}^2 \sum {E_n}^2}. \tag{8}$$

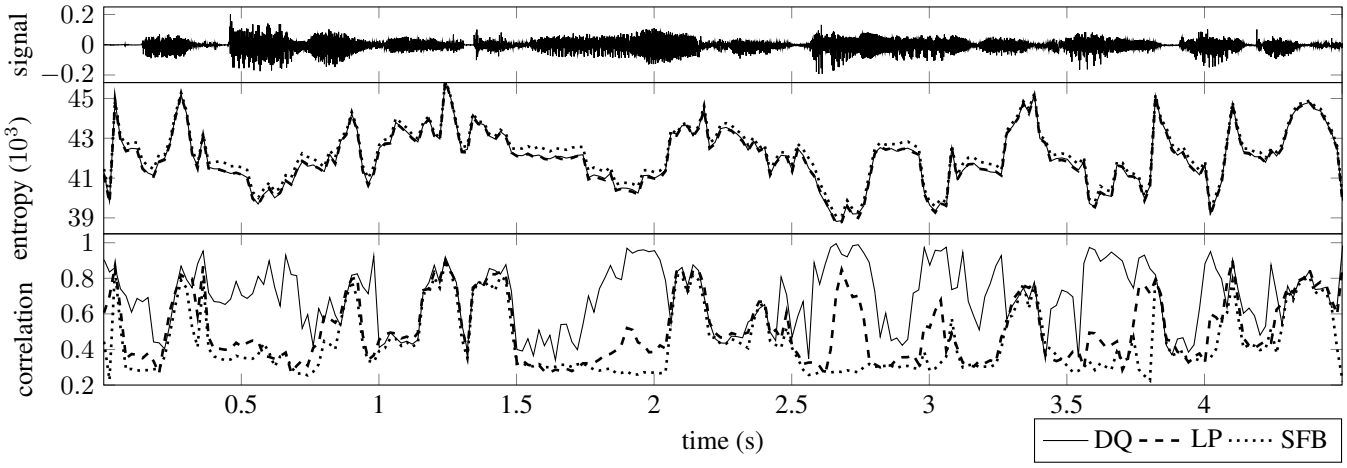

Fig. 5: Illustration of the entropy and correlation between envelope and signal over a 4500 ms segment from the input file. The waveform is depicted on the top followed by measurement results for linear prediction (LP), scale factor bands (SFB) and distribution quantizer (DQ).

The value lays between 0 and 1, where a higher values means better correlation and thus higher SNR.

To evaluate the envelopes we used the same set of critical items that was also used in development of MPEG USAC [3]. It consists of 15 single channel items of speech in various languages, music in different genres, as well as mixtures, downsampled to a sampling rate of 12.8kHz. The signal was separated into 30 ms frames using a step size of 20 ms and multiplied with a Hamming window. These parameters correspond to those used in G.718 codec [16].

We applied the algorithms of linear prediction, scale factor bands and distribution quantizer on a frame by frame basis. For all methods the envelope order was $P = 16$, which is also used by AMR-WB [2]. To determine the linear prediction coefficients we applied a power of 2 to the spectrum and used the autocorrelation method from [1]. We furthermore defined the scale factors to be the mean magnitude of the samples in the respective band. Informal tests showed that for distribution quantizer we get the best results when applying a power of 0.5 to the magnitude spectrum.

Typical envelopes obtained with respective algorithms are illustrated in figure 4. To evaluate their performance we calculated for each envelope model and frame the mean entropy as well as the correlation according to section 4.

5. RESULTS

Figure 5 illustrates the performance of the different envelope models over a segment of speech. We observe that the entropy is nearly the same for all three envelopes. Specifically, the mean entropies over the whole material are 36.4 for scale factor bands, 36.2 for linear prediction and 36.3 for distribution quantizer. Figure 6a illustrates the distribution of the mean entropies over the course of the input signal. The difference between linear prediction and distribution quantizer is small but still statistically significant.

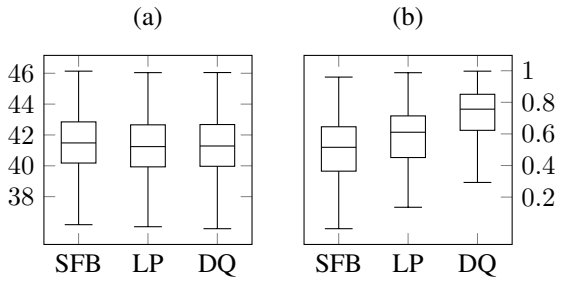

Fig. 6: Box plots of (a) entropy / (b) correlation of linear prediction (LP), scale factor bands (SFB) and distribution quantizer (DQ) measured over a set of speech and audio samples.

Figure 5 also shows the correlation between signal and envelope in the bottom plot. It can be seen that the distribution quantizer, which is depicted as a solid line, performs equally well or often significantly better than the other models. From figure 6b we observe that the mean correlation of the distribution quantizer is clearly higher than that of both the scale factor bands and linear prediction. Moreover, the variance of correlation is lowest for the distribution quantizer.

Analyzing the results using an Analysis Of Variance (ANOVA) test followed by a multi comparison t-test confirms that the mean correlations of all three envelopes are significantly different ($p < 0.01$). The best value is given by the distribution quantizer with 0.669, followed by linear prediction with 0.541 and scale factor bands with 0.475.

Since scale factor bands and linear prediction are established and evolved methods, we do not expect a vast superiority at this point. Nonetheless the improvements are already visible and significant.

6. CONCLUSION

In this work we introduced an envelope model called distribution quantizer for speech and audio applications. It is based on splitting the signal into segments of equal magnitude such that an accurate envelope model can be reconstructed using the splitting points only.

Our experiments show that, in terms of entropy, the distribution quantizer performs better than scale factor bands and almost as good as linear prediction. Furthermore distribution quantizer has a higher correlation with the input signal than scale factor bands or linear prediction. With optimal scaling, distribution quantizer would thus have a higher signal to noise ratio than the other methods.

Speech and audio coding applications naturally also require efficient methods for quantizing the envelope model. Quantization and coding of scale factor bands is straightforward with regular entropy coders [8]. Coding of linear predictive models is much more complicated and computationally complex since it requires vector quantizers since the parameters are highly correlated [15,17]. In comparison, the parameters in distribution quantization can be readily quantized. Moreover, since the parameters of the distribution quantizer have intuitively simple interpretations, we can readily use perceptual criteria such as non-uniform quantization on the frequency axis.

Distribution quantization thus provides an accuracy which is equal or better than conventional envelope models, while simultaneously providing a domain which can be readily quantized and where perceptual modeling is straightforward. The proposed model thus gives a simple yet effective model of envelopes for all speech and audio applications.

REFERENCES

[1] P. P. Vaidyanathan, "The theory of linear prediction," *Synthesis Lectures on Signal Processing*, vol. 2, no. 1, pp. 1–184, 2007.

[2] 3GPP TS 26.190 V7.0.0, "Adaptive multi-rate (AMR-WB) speech codec," 2007.

[3] 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, 2013.

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

[5] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 411–423, 1991.

[6] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[7] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1644–1657, 2012.

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

[9] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc*, vol. 45, no. 10, pp. 789–814, 1997.

[10] K. Brandenburg and M. Bosi, "Overview of MPEG audio: Current and future standards for low bit-rate audio coding," *J. Audio Eng. Soc*, vol. 45, no. 1/2, pp. 4–21, 1997.

[11] J. Herre, "Temporal noise shaping, qualtization and coding methods in perceptual audio coding: A tutorial introduction," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, Aug 1999.

[12] T. Bäckström, B. Schubert, and M. Multrus, "Distribution quantization for envelope encoding," 2013, WO patent submitted.

[13] M. R. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, Apr 1985, vol. 10, pp. 937–940.

[14] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[15] K. K. Paliwal and B. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 1, pp. 3–14, Jan 1993.

[16] 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.

[17] J. Benesty, M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, 2008.