ADAPTIVE TIME-SEGMENTATION FOR SPEECH CODING WITH LIMITED DELAY

*Christoffer A. Rødbro**

Aalborg University
Department of Communication Technology
9220 Aalborg Ø, Denmark
car@kom.auc.dk

Jesper Jensen and Richard Heusdens

Delft University of Technology
Department of Mediamatics
2628 CD Delft, The Netherlands
{J.Jensen,R.Heusdens}@ewi.tudelft.nl

ABSTRACT

In this paper we investigate the trade-off between delay and signal quality in adaptive time-segmentation for speech coding. A variable rate sinusoidal coder with adaptive segmentation and bit allocations is proposed and implemented with specifiable look-ahead. Objective and subjective results indicate that adaptive time-segmentation is advantageous even with low delay (30 ms), and that quality only increases with the delay until approximately 100 ms.

1. INTRODUCTION

Adaptive time-segmentation has been shown to significantly improve the rate - distortion tradeoff in speech and audio coding [1], [2]. However, such a scheme requires a certain look-ahead, making it a less obvious choice for real-time applications, such as telephony. Therefore, it is interesting to investigate exactly how much delay is necessary in order for the time-segmentation to have the desired effect.

Time-segmentation schemes seem especially well suited for packet based telephony since it allows for asynchronous transmission. With this in mind, frame representations should be self-contained so that packets can be decoded independently, this way making the speech coder robust towards packet losses. Therefore, we shall also propose a sinusoidal reference coder fulfilling this requirement. The coder structure resembles that of e.g. [3], [4], the main difference being in the phase quantization. By incorporating this coder into the rate-distortion based time-segmentation strategy of [1] we obtain a variable rate speech coding algorithm with specifiable delay.

The sinusoidal reference coder will be described in the next section. Following this, Section 3 gives a brief review of the time-segmentation algorithm as well as the limited delay implementation. Objective and subjective results are presented in Section 4 before Section 5 concludes on the work.

*Christoffer Rødbro's work was carried out during a stay at the Delft University of Technology.

2. SINUSOIDAL CODER

In a harmonic sinusoidal model a frame of speech is represented by a weighted sum of harmonically related sinusoids:

$$\hat{s}(t) = \sum_{k=1}^{K} A_k \cos\left(k\omega_0 t + \phi_k\right). \qquad (1)$$

Here, K is the number of components, determined by the fundamental frequency ω_0. Moreover, t is the time-index $t = -L/2, \ldots, L/2-1$ with L being the frame length, whereas A_k and ϕ_k are the amplitude and phase of the k'th component, respectively. This model is physiologically motivated for voiced speech only, however it is well-known that acceptable perceptual quality can be achieved for unvoiced speech as well, as long as the number of sinusoidal components is high enough compared to the frame length. Specifically, for unvoiced frames we found it sufficient to choose ω_0 so that $K = \frac{L}{4}$.

In voiced frames a voicing cutoff frequency is estimated based on a SNR-like measure as in [4], above which components are classified as being unvoiced. The main difference between unvoiced and voiced components is that subframe phase randomization [5] is applied for the synthesis of unvoiced components.

2.1. Parameter estimation and quantization

The fundamental frequency ω_0 is estimated based on the YIN algorithm proposed in [6] and quantized in the log-domain using 8 bits. YIN relies on a measure of unvoiced-to-total power ratio, the value of which has been found a reasonable voiced/unvoiced classifier.

Once ω_0 is determined, estimation of amplitudes and phases can be formulated as a linear weighted least squares (WLS) problem, the weighting being determined by the analysis/synthesis window, see [3] for the original idea. However, due to deviations from the harmonic model or inaccuracies in the fundamental frequency estimate there might be a slight mismatch between the harmonic frequencies $k\omega_0$

	Voiced	Unvoiced
AR order	8, 12, 16, 20	6, 10
Phase bits	20, 40, 60	-

Table 1. Allowed AR model orders and number of phase bits forming the 14 coding templates.

and the spectral peaks $\hat{\omega}_k$ of the signal. Therefore, to avoid underestimation of the amplitudes, the WLS is computed at the slightly modified harmonic frequencies $\hat{\omega}_k$. The estimated amplitudes are represented by an auto-regressive (AR) model using discrete all-pole (DAP) modeling [7], and the log-gain is quantized using 5 bits.

The wish for interframe independence rules out time-differential encoding of the phases. Instead, we exploit that for voiced speech the time-domain maxima of the harmonic components exhibit a highly structured pattern that can be represented well by a piece-wise linear function. The number of linear functions, and thereby the accuracy of the representation, is determined by the number of bits allocated for the phases.

For further details on the speech coder, see [8].

3. ADAPTIVE TIME-SEGMENTATION

In order to find an adaptive time-segmentation we use the algorithm introduced in [1], which is based on the following minimization problem:

$$\begin{aligned} \text{minimize}: \quad & D\left(\tau, \mathbf{p}(\tau)\right) \\ \text{s.t.}: \quad & R\left(\tau, \mathbf{p}(\tau)\right) \leq R_C \end{aligned} \tag{2}$$

Here, D is some measure of the distortion between the original and the encoded signal, whereas R is the bit rate obtained, and R_C the requested bit rate. The time-segmentation $\tau = \left\{s_1, s_2, \ldots, s_{\sigma(\tau)}\right\}$ consists of $\sigma(\tau)$ variable length segments s_i, each having a length equal to an integer number of *grids*. The grid length determines the segmentation resolution, in this application chosen at 5 ms.

The vector $\mathbf{p}(\tau) = \left\{p(s_1), p(s_2), \ldots, p(s_{\sigma(\tau)})\right\}$ denotes the *coding templates* used to encode the segments. In the application at hand, a coding template specifies the AR-order used to model the amplitudes and the number of bits spent on phase encoding. Table 1 shows the templates we used for the experiments described in the next section.

The problem (2) is solved by minimizing the Lagrangian $J(\tau, \mathbf{p}(\tau)) = D(\tau, \mathbf{p}(\tau)) + \lambda R(\tau, \mathbf{p}(\tau))$ using dynamic programming (see [1] for details), where $\lambda > 0$ is iterated over until the requested bit rate R_C is reached.

3.1. Distortion Measure

In order to effectively minimize the Lagrangian with dynamic programming it is required that the distortion measure is additive and independent across frames, so that the total distortion can be found by adding the distortion of all frames. However, when using spectral magnitude based metrics such as spectral distortion (SD) it was observed that the algorithm sometimes leads to segmentations with non-stationary frames. The reason is that the *average* spectra are sometimes modeled quite well by the stationary sinusoidal model. To avoid this problem, the distortion is measured *per grid*, i.e. in sub-windows within each analysis frame, see Figure 1. The sub-windows are 15 ms hanning with 10 ms overlap, adding up to a hanning tapered window with 10 ms overlap region (= analysis/synthesis window). Note that this approach limits the minimum window length to 20 ms. In this way, we apply the SD, measured at the harmonic frequencies:

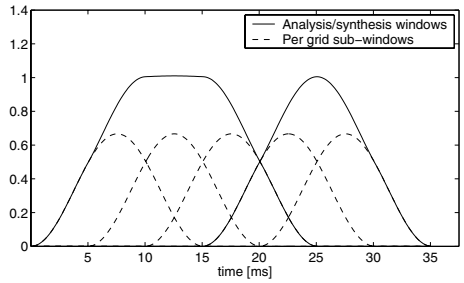

Fig. 1. Hanning windows extracting one grid add up to hanning tapered windows with overlap region two grids.

$$D(s, \hat{s}) = \sqrt{\frac{20^2}{K} \sum_{k=1}^{K} \Big(\log|S(\hat{\omega}_k)| - \log|\hat{S}(k\omega_0)|\Big)^2}. \tag{3}$$

Here, $S(\hat{\omega}_k)$ is the spectrum of the original signal at the spectral peaks, whereas $\hat{S}(k\omega_0)$ is the spectrum of the synthesized signal, measured at the harmonics.

3.2. Finite delay

In order to implement the time-segmentation algorithm with finite delay, we introduce the concept of a *superframe*, the length of which will be the algorithmic delay. The optimal segmentation and templates are found for this superframe and we now have two possibilities as illustrated in Figure 2: either all frames within the superframe are transmitted, and a new superframe is made starting where the old one ended. We call this approach "shifting window". Alternatively, only the left-most frame (i.e. the oldest samples) is transmitted and an equivalent number of new samples is concatenated on the right-hand side. This approach is called "sliding window". As indicated in Figure 2 the latter approach seems preferable since it only temporarily enforces a frame boundary at the end of each superframe.

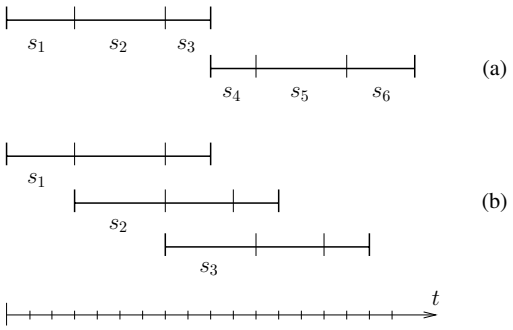

Fig. 2. Time-segmentation with shifting (a) or sliding (b) superframes. Ticks on the time axis represent the grids.

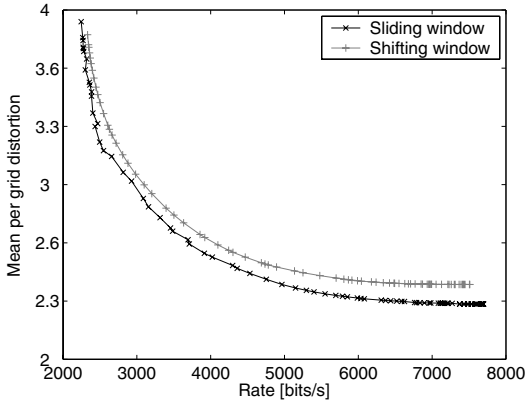

Fig. 3. Rate - distortion pairs for a speech signal when using shifting and sliding superframes of length 60 ms. Shown for 3 seconds of male speech.

A distortion/rate comparison of the two methods is shown in Figure 3. As expected we see that the sliding window outperforms the shifting window. However, we also see that the "shifting" curve is convex as expected but the "sliding" curve is not. The reason for this is that in the first case, we perform optimal segmentation within each superframe which results in a convex R-D curve for each, the sum of which is also convex. In the "sliding" case the segmentation of one superframe will influence the contents of the next and thus the minimization problems are not independent.

3.3. Rate control

As described in the beginning of this section, (2) is solved by minimizing a Lagrangian cost function for different values of λ, each corresponding to different *total* bit rates. In situations where coding delay is low (at least lower than the duration of the signal to be encoded), finding the optimal λ is a nontrivial task since we have to make a decision how to segment and which coding templates to use without knowing how many bits we need to spend for the remaining part of the signal. Solving this problem, however, falls outside the scope of the work presented here. For the simulations described in the next section, we find the optimal segmentation and coding templates for a given λ using finite look-ahead, and determine at the end, after the entire signal has been encoded, whether the target rate was reached or not. If not, λ is iterated until the desired rate is met.

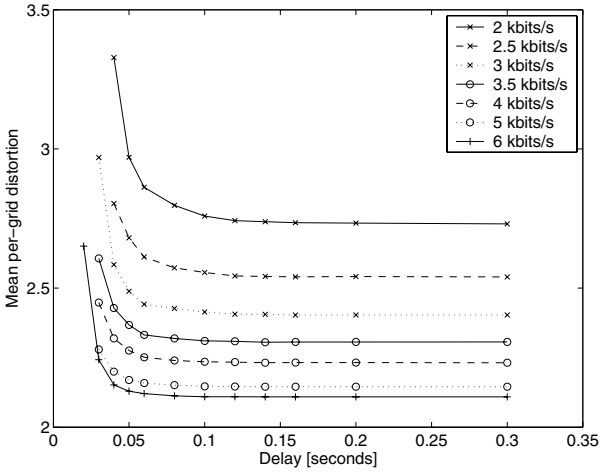

Fig. 4. Distortion as a function of delay with different target bit rates averaged over 8 speech samples (approximately 4 seconds each). Some points are not shown because the requested rate cannot be reached with every delay.

4. RESULTS

Figure 4 shows the relationship between algorithmic delay and distortion for target bit rates ranging from 2 - 6 kbps. It appears that the distortion reduction from allowing more than 100 ms delay is negligible, and that quality can be increased through adaptive segmentation even with very short delay.

4.1. Subjective listening test

To validate the objective results above, subjective listening tests were carried out. Listeners were asked to rank the processed speech samples relatively on a 5 point scale, enforcing use of the entire scale by ranking the worst sample at 1 and the best at 5. The original was presented for reference and each sample could be played as many times as needed. The test was conducted with high-quality headphones using the benchmark software described in [9].

The speech samples consisted of 4 male and 4 female of duration 2-5 seconds, the different encoded versions being:

1. The original.

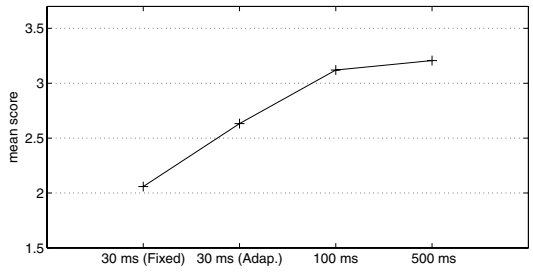

Fig. 5. Mean score results for subjective listening test.

2.-4. Adaptive segmentation with 500 ms, 100 ms and 30 ms delay, respectively, and adaptive coding templates.

5. Fixed segmentation with 30 ms frames and adaptive coding templates.

6. Fixed segmentation with 30 ms frames always using the cheapest template (AR-8 and 20 phase bits in voiced, AR-6 in unvoiced).

The first and last items were included as "anchor excerpts" in order to facilitate use of the entire scale 1 to 5. For items 2 to 5 the bit rate was 3.5 kbit/s. This bit rate was chosen since it could be reached with the fixed segmentation (item 5) for all test samples. The rate for item 6 varied between 2.7 - 3.1 kbit/s.

The average scores for 10 listeners are shown in Figure 5. The results validate the observations from the previous section: there is a considerable gain from fixed to variable segmentation at 30 ms delay, and again from 30 ms to 100 ms delay, whereas there is little difference between 100 ms and 500 ms delay. It should be noted that the absolute scores cannot be compared to MOS scores, since use of the entire scale was requested.

The statistical significance of the incremental improvements indicated in Figure 5 can be assessed by a paired t-test. The reason for using a paired test is that the level of the scores varies from listener to listener and from excerpt to excerpt, so that the absolute scores are not independent. Instead, the paired test works on the observed *differences* between two setups, e.g. between 30 ms and 100 ms delay. The H_0 hypothesis is that the mean of the underlying (assumed Gaussian) distribution is zero, $\mu_\Delta = 0$, and the alternative H_1 hypothesis that $\mu_\Delta > 0$. Table 2 lists if H_0 is accepted at the 0.05 significance level, the p-values (i.e. the significance level above which H_0 is rejected) and the 95% confidence interval for μ_Δ.

5. CONCLUSION

From Table 2 we see that there is a statistically significant difference between 30 ms fixed, 30 ms adaptive, and 100 ms adaptive, but not between 100 ms and 500 ms. Thus, we conclude that speech quality can indeed be increased through adaptive segmentation, even with little additional look-ahead. Also, the increase in quality saturates around 100 ms of delay. Strictly spoken these conclusions can only be made for the sinusoidal reference coder used here, however, they are believed to be valid for speech coding in general. This is of great importance since it indicates the feasibility of adaptive time-segmentation for real-time voice applications.

	H_0	p-value	μ_Δ conf.
30 ms, Adap. − Fixed	Rej.	$1.5 \cdot 10^{-7}$	> 0.4
100 ms − 30 ms	Rej.	$4.5 \cdot 10^{-6}$	> 0.3
500 ms − 100 ms	Acc.	0.12	> -0.03

Table 2. Results of t-tests for the observed score differences.

6. REFERENCES

[1] P. Prandoni and M. Vetterli, "R/D Optimal Linear Prediction," *IEEE Trans. on Speech and Audio Proc.*, vol. 8, no. 6, pp. 646–655, Nov. 2000.

[2] R. Heusdens et al., "Sinusoidal Coding of Audio and Speech (SiCAS)," Submitted to *Journal of the Audio Engineering Society*, 2003.

[3] J. S. Marques, L. B. Almeida, and J. M. Tribolet, "Harmonic Coding at 4.8kb/s," in *Proc. IEEE ICASSP*, Dec. 1990, pp. 17–20.

[4] R. J. McAulay and T. F. Quatieri, *Sinusoidal Coding*, chapter 4, Elsevier Science B.V., 1995, From *Speech Coding and Synthesis*, Edited by W.B Kleijn and K.K. Paliwal.

[5] M. W. Macon and M. A. Clements, "Sinusoidal Modeling and Modification of Unvoiced Speech," in *IEEE Trans. on Speech and Audio Proc.*, Nov. 1997, vol. 5, pp. 557–560.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," in *Journal of ASA*, Apr. 2002, vol. 111(4).

[7] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," in *IEEE Trans. on Signal Processing*, 1991, vol. 39, pp. 411–423.

[8] C. A. Rødbro, J. Jensen, and R. Heusdens, "Adaptive time-segmentation of speech for packet loss channels," Tech. Rep., Delft University of Technology, 2003.

[9] O. A. Niamut, *Audio Codec BanchMark Manual*, Department of Mediamatics, Delft University of Technology, Jan. 2003.