Squared Error as a Measure of Phase Distortion

Harald Pobloth and W. Bastiaan Kleijn

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology),
100 44 Stockholm, Sweden
{ harald, bastiaan }@speech.kth.se

Abstract

In this article, we investigate how accurately the squared error captures perceptual errors introduced by Fourier phase spectrum changes. We measure the perceptual error using the Auditory Image Model by Patterson et al.. The squared error is found to represent the perceptual error well for low squared errors but it saturates. Thus, a further increase in squared error does on average not lead to any further increase in perceptual error. This suggests that encoding phase using squared-error trained codebooks only improves perceived quality when operating at high bit rates. To verify this, phase was encoded with codebooks of different sizes. As expected, increasing the codebook size has very little influence on the average perceptual error for low rates, which is confirmed by listening tests. Our results suggest that a direct phase codebook is an inefficient representation of the relevant information contained in phase.

1. Introduction

There are many coding systems which encode Fourier phase spectra, e.g., [1, 2], although there is some disagreement about the perceptual importance of this parameter. In most of the encoding systems, the squared error (or, as in the case of [1], a weighted squared error) is used as fidelity criterion. This has the well-known advantages that it is easy to compute and easy to handle in optimization problems. A question which naturally arises is whether the squared error captures the perceptual loss introduced by phase distortions. Most of the work associated with phase encoding has shown improvement in segmental SNR or other squared-error related criteria with increasing bit rate, e.g., [1, 2]. However, to our knowledge, there has been no investigation on how the perceptual quality changes as a function of bit rate. It is generally accepted, that it is hard to find an efficient encoding of phase. Thus, considering the secondary perceptual importance it has compared to amplitude, phase is often not encoded but replaced either by the minimum phase or another fixed phase spectrum, e.g., [3].

The procedure from codebook training to performance evaluation can be divided in three stages:

1. Codebook training.
2. Encoding.
3. System evaluation.

"System evaluation" in this article is finding the average error when a data base is encoded using a certain codebook. During all these steps, different error criteria might be used. Codebook training is a one-time cost, so choosing a more precise error criterion with the drawback of higher complexity might be reasonable. Encoding is done when using the system and complexity should be as low as possible. System evaluation can be done both constantly or only during design. For speech encoding, the criterion of highest interest is a perceptual criterion, so we restrict ourselves to the average perceptual error $\bar{\zeta}$ as a performance measure for system evaluation. The perceptual error is found using the auditory image model (AIM) by Patterson [4] with correlograms as final output.

We aim to find how much system performance degrades when using the squared-error criterion instead of the perceptual criterion in stages 1. and 2. listed above. Thus, we try to answer the following questions:

- How useful are squared-error trained codebooks compared to random codebooks.
- How suboptimal is fast squared-error codebook training compared to slow and difficult perceptual-codebook training.
- How suboptimal is fast squared-error encoding compared to slow perceptual-error encoding.

The article is organized as follows. Section 2 below describes the signals used, the squared error, and the perceptual error criterion. In section 3.2, we try to establish how the squared error and the perceptual error are related for phase distortions appearing in coding. The results from these experiments are consistent with the results obtained when encoding the phase spectrum of a signal with codebooks of varying size, as done in section 3.3. The listening test of section 3.4 verifies these results.

2. Framework

It is important to define carefully the signals used, as well as the DFT window, before making a statement about phase perception. We then know which changes of a signal can be attributed to a change of its amplitude or its phase spectrum. Coders as, e.g., the waveform-interpolation coder (e.g., [1, 2, 3]) represent speech in a pitch-synchronous manner. This is the motivation for using windows of one pitch-cycle length. Since speech is not strictly periodic, the position of the window is crucial. To avoid this problem, the signals used throughout this article are made strictly periodic such that the window position is not important and a rectangular window can be used. It is indicated, e.g., by the fact that in [1] only the phase of the slowly evolving waveform is encoded, that voiced speech is the part of a speech signal for which phase information is most relevant. Thus, the periodic signals used here are a reasonable choice.

A signal was formed by concatenating a pitch-cycle waveform s_k to a signal of 400 ms length. The pitch cycles s_k were randomly selected from a data base derived from TIMIT [5]. This data base was found by first down-sampling TIMIT to 8 kHz, and identifying all voiced segments of the utterances using an autocorrelation criterion. Afterwards, the voiced pitch cy-

cles in a certain pitch range were normalized to a fixed pitch by means of interpolation. The pitch ranges considered are 80 Hz - 120 Hz normalized to 100 Hz and 180 Hz - 220 Hz normalized to 200 Hz.

The pitch-cycle waveform s_k can be written as a function of the amplitude spectrum a_k and the phase spectrum ϕ_k,

$$s_k[n] = \sum_{l=0}^{N-1} a_k[l] \mathrm{e}^{j\phi_k[l]} \mathrm{e}^{j2\pi \frac{nl}{N}} \ , \tag{1}$$

where N is the number of samples in the pitch cycle. A phase distorted waveform $\hat{s}_k[n]$ has identical amplitude spectrum, but the phase spectrum is changed to $\hat{\phi}_k$, which introduces a phase distortion

$$\Delta\phi_k[l] = \phi_k[l] - \hat{\phi}_k[l] \ . \tag{2}$$

Using the signal representation in equation 1, it follows that the weighted squared error between s_k and $\hat{s}_k$ is

$$\eta(a_k, \Delta\phi_k) = 8 \sum_{l=1}^{\lceil N/2 \rceil - 1} w\left(\mathrm{e}^{j\frac{2\pi l}{N}}\right) a_k^2[l] \sin^2\left(\frac{\Delta\phi_k[l]}{2}\right) \ , \tag{3}$$

where $w(z)$ is a weighting function. For the speech squared error η_s, $w(z) = 1$. For the weighted squared error η_w as introduced by Gottesmann in [1]

$$w(z) = \left| \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \right|^2 \ , \tag{4}$$

where $A(z)$ is the LP analysis filter. We selected $\gamma_1 = 0.9$ and $\gamma_2 = 0.7$. For the so-called residual error η_r, $w(z) = |A(z)|^2$.

When encoding the phase of a signal, it is reasonable to transmit linear phase and dispersion phase separately. Linear phase ϕ_l is a measure for the time shift of the signal and dispersion phase ϕ_d is the phase remaining after this time shift is removed. Thus, we decompose the phase distortion $\Delta\phi$ into $\Delta\phi = \phi_l + \phi_d$. As in [2], we select ϕ_l as

$$\phi_{l_k} = \arg\min_{\phi_l} \eta(a_k, \Delta\phi_k - \phi_l) \ , \tag{5}$$

under the constraint that ϕ_l is linear phase, that is

$$\phi_l[l] = \lambda l + 2\pi c \ ; \ c \in \mathcal{Z}; \ \lambda \in \mathcal{R} \ . \tag{6}$$

In the remainder of this article, only errors due to dispersion phase ϕ_d are considered.

To find the perceptual error, the neural activity patterns in $N_c = 24$ auditory channels are found using the auditory image model by Patterson et al. [4]. From the mid-part of a 400 ms long neural activity pattern, the autocorrelation over one pitch cycle is calculated in each channel i and the criterion ζ is found as

$$\zeta = \sum_{i=3}^{N_c} \left(\sum_{n=0}^{N-1} \left(c_i[n] - \hat{c}_i[n]\right)^4 \right)^{1/4} \ , \tag{7}$$

where $c_i[n]$ is the autocorrelation function of the original signal in channel i, and $\hat{c}_i[n]$ is the autocorrelation of a distorted signal.

The criterion differs from the one we used in [6]. The one used here is simpler and better suited for further analysis. The new measure coincides with the results from the listening test done in [6] as well as the criterion used there does. For an experienced listener, the threshold for $\approx 90\%$ detectability was found to be $\zeta_t = 3.6$.

3. Experiments

In sections 3.2 and 3.3, two experiments to establish the perceptual performance of the three squared-error criteria η_s, η_w and η_r are described. In the first, the correlation between squared error and perceptual error is investigated. If there was a monotonic relationship between the two, it would be sufficient to measure the squared error to obtain information about the perceptual difference between two signals. In the second experiment, the average perceptual error is found when a set of pitch cycles is encoded using different codebooks. The results of the latter experiment are verified by a listening test.

3.1. Codebook training

Both experiments require trained phase codebooks. The first experiment only involves codebooks trained to minimize the average squared error, while the second experiment requires codebooks which attempt to minimize the average perceptual error.

Gottesmann describes in [1, eq. 12] the centroid equation for a phase vector. Adopting this centroid equation, we find two residual phase vectors as centroid candidates. Both give extreme values for the average squared error, when encoding the residual phases ϕ_{r_k} of a set S_i of signals s_k to one $\hat{\phi}$. The two vectors are

$$\hat{\phi}_1[l] = \arctan[A] \tag{8}$$

$$\hat{\phi}_2[l] = \begin{cases} \arctan[A] + \pi & \text{for } \arctan[A] < 0 \\ \arctan[A] - \pi & \text{for } \arctan[A] \geq 0 \end{cases} \tag{9}$$

$$A = \frac{\sum_{k=\{k|s_k \in S_i\}} w_k[l] a_k^2[l] \sin(\phi_{r_k}[l] - \phi_{l_k}[k])}{\sum_{k=\{k|s_k \in S_i\}} w_k[l] a_k^2[l] \cos(\phi_{r_k}[l] - \phi_{l_k}[l])}$$

where ϕ_{r_k} is the residual phase of signal s_k. Of the two solutions above, the one leading to the smallest average squared error is the centroid residual phase ϕ_{CB_i} for the set S_i. Using this centroid equation, the generalized Lloyd algorithm (GLA) [7] can be used to train residual phase codebooks of desired sizes as described in [1].

No analytical centroid equation exists which minimizes the average perceptual error for a set S_i of signals s_k. Instead, the numerical minimization method described in [8, pp. 414-420] is used in the Lloyd algorithm. Thus, codebooks can be trained which attempt to minimize the average perceptual error. It should, however, be noted that this is likely to be more suboptimal than the standard GLA since the numerical minimization method might converge to local minima.

3.2. Relation squared error and perceptual error

In the first experiment, the original residual phase ϕ_{r_k} of 200 randomly selected pitch cycles was replaced with all phase spectra contained in squared-error trained codebooks of sizes 1-7 bit. Each replacement led to a signal pair s_k, $\hat{s}_k$ where s_k had original phase and $\hat{s}_k$ was a signal with codebook phase. A total of 200 $\sum_{i=1}^{7} 2^i = 50,800$ signal pairs were generated for each squared error η_s, η_w, and η_r. The resulting squared and perceptual errors were measured for all signal pairs. These errors are the same errors as found during codebook search when encoding the signals s_k with the squared-error trained codebooks. The aim is to verify if the different squared errors give a consistent measure of the perceptual error that the phase distortions introduce. In addition, the $\Delta\phi$ found when replacing ϕ_k with the 7-bit codebook phases were scaled with factors $s = \{0.5, 0.3, 0.2, 0.1, 0.05, 0.03\}$ to generate phase

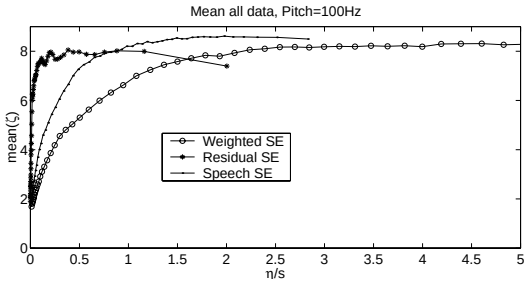

Figure 1: *Mean perceptual error as a function of squared error for a pitch of 100 Hz. η/s is the normalized squared error. The averaging was done over regions on the x-axis containing 6500 points.*

distortions with lower squared and perceptual errors. This was done, since for 100 Hz signals even 10-bit codebooks only lead to points in the flat region of the curve in figure 1. The scaling generated additional 153,600 error pairs, which gave a total of 204,400 pairs for each squared-error criterion. Figure 1 shows the mean perceptual error as a function of the squared error. Each point on the graphs is found by averaging over 6500 neighboring $\zeta,\ \eta$ pairs along the η axis. The standard deviation as a function of the squared error has a shape similar to the mean in figure 1 and saturates to $\sigma(\zeta) \approx 1.5$. The η axis in fig. 1 is scaled such that the maximum squared error, which appears for $\hat{s}_k = -s_k$, is normalized to 10. For the weighted and the residual squared error, the maximum possible squared error differs for different signals, and the average of the maximum possible squared error for all 200 s_k is normalized to 10.

For a pitch of 200 Hz the curves are of similar shape. In this case, the mean saturates at around $\bar{\zeta} \approx 5.2$ and the standard deviation saturates at $\sigma(\zeta) \approx 1$. This confirms that phase is less important for high-pitched speech.

An important aspect of the above results is that the perceptual error saturates for higher squared errors. This suggests that two codebooks of different size, which both provide mostly codebook vectors in the saturated error range, should not differ in their perceptual performance. In other words, an increase in codebook size will not lead to better average perceptual performance, unless the codebook becomes large enough to provide vectors outside the saturated range.

3.3. Encoding

To verify that increasing codebook size does not necessarily lead to lower perceptual error, as predicted in section 3.2, sets of randomly chosen pitch cycles s_k were encoded using different codebooks. The sets were of size $N_t = \max(6400, 100 N_{CB})$, where N_{CB} is the codebook size. In addition, the three questions posed in section 1 are addressed.

The encoding was done with random codebooks, squared-error trained codebooks, and perceptually-trained codebooks. A random codebook consists of phase vectors with values uniformly distributed between $-\pi$ and π. As encoding criterion either the squared error or the perceptual error was used. In the first case, the residual phase ϕ_{r_k} was encoded as the $\hat{\phi}_k$ minimizing η, in the latter $\hat{\phi}_k$ was chosen to minimize ζ.

For 100 Hz signals figure 2 shows the average perceptual error $\bar{\zeta}$ for the different codebooks when encoding is done using the squared-error criterion. As can be seen, it is indeed true that an increase of codebook size from 1-7 bit does not lead to an decrease in $\bar{\zeta}$ for all squared-error criteria.

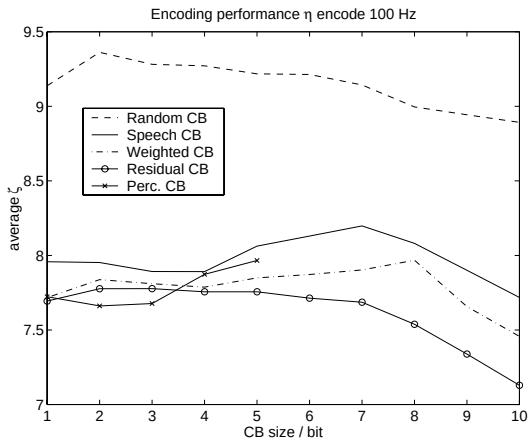

Figure 2: *Mean perceptual error as a function of codebook size for a pitch of 100 Hz. The encoding criterion is the squared error. The trained codebooks were encoded using the squared error they are trained for (e.g., speech squared error for the speech CB). For the perceptually-trained codebook the residual squared error, and for the random codebook the speech squared error is used as encoding criterion. These are chosen since they minimize the average perceptual error for these two codebooks.*

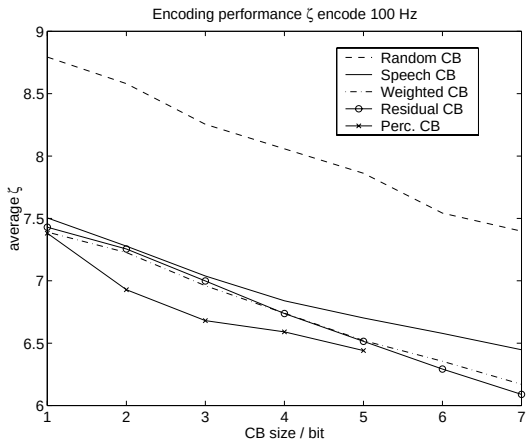

Figure 3: *Mean perceptual error as a function of codebook size for a pitch of 100 Hz. The encoding criterion is the perceptual criterion ζ.*

Figure 3 shows results for the perceptual-error encoding. As expected, there is a gain when increasing the codebook size in this case. For 200 Hz signals, the average perceptual error is generally lower than for 100 Hz signals, as can be seen in figure 4. It also can be seen that $\bar{\zeta}$ gradually reduces even for low bit rates. The random codebook performance is closer to the trained codebook performance than for 100 Hz pitch. This indicates that dispersion phase for high-pitched speech might be more randomly distributed than it is for low-pitched speech. The distance between the perceptual encoding curves in figure 3 and the squared-error encoding curves in figure 2 increase approximately linearly with increasing codebook size from about 0.4 at 1-bit to 1.7 ζ units at 7-bit. For 200 Hz pitch, the behavior is similar and the increase ranges from 0.4 to 1.6 ζ units for all codebook types. From theory it is obvious that for high rate both methods should converge to equal results.

The advantage of perceptually-trained codebooks is minor which is surprising at first. However, perceptually trained code-

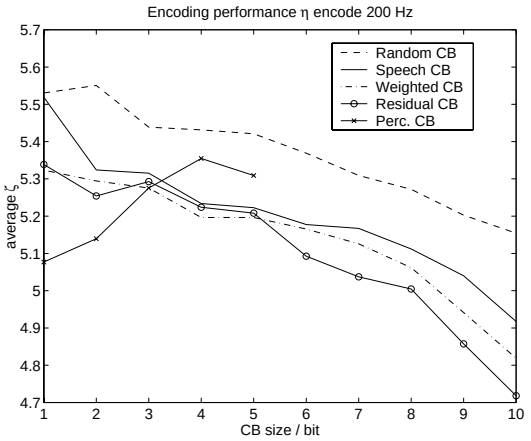

Figure 4: *Mean perceptual error as a function of codebook size for a pitch of 200 Hz. The encoding criteria are the squared-error criteria as in figure 2.*

books do not overcome the problem that phase codebooks are a compromise for all amplitude spectra. This problem occurs, since the relative importance of the dimensions in the phase vectors change for different amplitude spectra.

3.4. Listening test

The most controversial result of the encoding experiment might be that for codebooks in the range from 1-bit to about 7-bit the squared-error encoding gives an almost constant average perceptual error. This result is verified by a listening test. Five subjects listened to signal triplets AXB monaurally. They were forced to choose if A is closer to X or B is closer to X. The signals A, X and B were generated from randomly selected s_k concatenated to 500 ms length. For each subject, one test consisted of two parts each with a 100 triplets. For the first part, the triplet AXB consisted of: X original s_k, A or B 1-bit squared-error encoded $\hat{s}_k$, and B or A 7-bit squared-error encoded $\hat{s}_k$. The second part is to verify whether subjects perceive the phase distortions introduced by coding. There X is the original signal s_k, A or B the original s_k again, and B or A the 1-bit squared-error encoded $\hat{s}_k$. The two parts were presented in random order. Two subjects participated in tests for all three squared-error criteria and three only for two squared-error criteria.

Table 1 shows the results of the tests. Since untrained subjects were used, the original signal is not identified with 100% certainty during the second part, even though for most encoded signals the perceptual error is above the threshold of 90% detection found for expert listeners in [6]. To avoid that subjects which are insensitive to phase distortion judge about the equality of the 1 and 7-bit encoding, tests from subjects choosing the original signal in the second part in less than 70% are not considered. It should be stressed, that this did not show to change the test results comparing the 1-bit and 7-bit encoding.

	1-bit	7-bit	orig.	1-bit	# triplets
η_s	43.5	56.5 ± 4.9	76.3	23.7 ± 4.2	400
η_w	50	50 ± 4.9	81.5	18.5 ± 3.8	400
η_r	52	48 ± 4.9	82.8	17.2 ± 3.7	400

Table 1: *Results from the listening test in %. The numbers following $\pm$ are the limits of the 95 % confidence intervals using the Gaussian approximation to the binomial distribution.*

4. Discussion

The results from the encoding experiment in figure 2 and 4 are consistent with Gottesmann's findings in [1]. There, encoding dispersion phase with a 4-bit squared-error codebook, gave more improvement for female speech than for male speech. However, as figures 2 and 4 show, phase encoding results in higher perceptual distortion for male speech than for female speech. Thus, the results in [1] do not contradict the results in [6].

Considering the results found in section 3, the three questions posed in section 1 can be answered as follows:

- Squared-error phase codebooks give lower average perceptual error than random codebooks for low-pitched speech. For high-pitched speech, training does not increase performance significantly.
- Perceptually-trained codebooks do not outperform squared-error trained codebooks for squared-error based encoding. Even for perceptual encoding they perform close to squared-error trained codebooks.
- Perceptual encoding gives better performance than squared-error encoding, however, its high effort is not rewarded with significantly better performance.

In summary, perceptual encoding is not significantly superior to squared-error encoding and squared-error encoding requires high rates to introduce only a slight decrease in perceptual error. These facts, and the fact that phase codebooks are a compromise for all amplitude spectra, suggest that the dispersion phase vector is not an efficient parameter to encode the features it represents. Different representations as, e.g., envelopes or measures of peakiness should be more efficient in representing these features.

5. References

[1] O. Gottesmann, "Dispersion phase vector quantization for enhancement of waveform interpolative coder", *in Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, pp. 269 – 272, Phoenix, AZ, 1999.

[2] Y. Jiang and V. Cuperman, "Encoding prototype waveforms using a phase codebook", *in Proc. IEEE Speech Coding Workshop*, pp. 21–22, Annapolis, MD, 1995.

[3] Yair Shoham, "Very low complexity interpolative speech coding at 1.2 to 2.4kbps", *in Proc. Int. Conf. Acoust. Speech Sign. Process.*, pp. 1599–1602, Munich, 1997.

[4] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of periphal auditory processing: A modular architecture software platform", *J. Acoust. Soc. Am.*, vol. 98, pp. 1890 – 1894, October 1995.

[5] "Darpa TIMIT", CD-ROM, October 1990, NIST Speech Disc 1-1.1.

[6] H. Pobloth and W. B. Kleijn, "On phase perception in speech", *in Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, pp. 29 – 32, Phoenix, AZ, 1999.

[7] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Comm.*, vol. COM-28, pp. 84–95, 1980.

[8] W. H. Press, S. A Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge Univ. Press, second edition, 1992.