Novel Low-Band Phase Representation for Low Bit-Rate Speech Coding

Ahmed Ismail [1, 2] , *Yasser Dakroury* [2] , *Hazem Abbas* [1,2]

[1] Mentor Graphics Corp.
[2] Computer & Systems Eng. Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt
ahmed.ismail@ieee.org

10.21437/Interspeech.2007-470

Abstract

Vector Quantization (VQ) has been extensively used in speech vocoders. Phase information is often ignored or coarsely represented in parametric coders because of the difficulties facing phase quantization. This paper introduces a novel distortion measure for the low-band speech signal that takes phase information into consideration, with no increase in the bit-rate. This measure has been used in the construction of a segmental vocoder, which is using the pitch period as segments. A description of the proposed Time-Domain Phase-Aware (TDPA) distortion measure is given and compared to the use of the MFCC as a distortion measure showing the effect of the phase information represented in the TDPA model on improving the inter-frame correlation of the synthesized speech. Finally, the performance of the TDPA is evaluated using the Segmental Signal-to-Noise Ratio (SNR), and Spectral Distortion (SD). Speech quality is evaluated using the recently standardized objective quality measure PESQ.

Index Terms: Distortion measure, Time domain analysis, Vector quantization (VQ), Segmental Vocoder, Waveform interpolation (WI), Phase Quantization

1. INTRODUCTION

Different paradigms for speech coding are available today; waveform coders, parametric coders, and segmental vocoders. In parametric coders, the phase information is commonly not transmitted, and this is for two reasons: first, the phase is of secondary perceptual significance; and second, no efficient phase quantization scheme is known. Waveform Interpolation (WI) coders typically use a fixed phase vector for the Slowly Evolving Wave (SEW) [1]. On the other hand, waveform coders such as Code Excited Linear Prediction (CELP) [2], by directly quantizing the waveform, implicitly allocate an excessive number of bits to the phase information more than is perceptually required. In WI, different methods for phase information quantization were used in [3] , [4] , [5] , [6] where an analysis-by-synthesis algorithm is used to choose the phase vector. Older approaches included the use of direct waveform quantization as in [7].

A segmental vocoder is implemented in [8], [9]. This coder attempts to decompose a speech signal into a sequence of segments that are subsequently quantized using a codebook of pre-stored segments. In [10], MFCC is used as a distortion measure for segment selection. Segmental vocoders do not need to transmit excitation information, which generally require a lot of bits in conventional vocoders.

This paper introduces a novel simplified model that jointly quantizes the magnitude and phase information of speech. Using the model as a distortion measure is investigated. This model was successfully used in a segmental vocoder using the pitch period as a segment. The signal was split into low-band and high-band signal parts; each was dealt with in a separate way. The implemented vocoder with both bands operated at 1.2 kbps. Only the low-band signal handling of this system is covered in this paper. The high-band signal handling was done in a more efficient spectral matching method [11].

This paper is organized as follows. Section 2 provides an overview of the analysis-by-synthesis phase quantization. Section 3 introduces the proposed TDPA model and its use as a distortion measure. Section 4 lists the experiments and comparison results. Finally, the paper is concluded in section 5.

2. PHASE QUANTIZATION USING ANALYSIS-BY-SYNTHESIS

An analysis-by-synthesis technique is described in [6] that is used to find the best phase that minimizes an Euclidean distance based distortion measure between the Fourier Transform of the original waveform and that of the synthesized wave. The algorithm is based upon using two separate codebooks; a magnitude codebook, and a phase codebook. It first searches the magnitude codebook to find the best match. Then, it synthesis a signal using the found magnitude and each of the phase entries. A search is done to find the phase entry that minimizes the Euclidean distance between the synthesized signal and the original signal.

Although the analysis-by-synthesis approach results in excellent matching of the waveform shape and hence the phase information, the approach uses a fairly complex computational separate search for the matching phase codebook entry. Furthermore, extra bits are assigned to represent the best matching phase codebook index. In [6], 7 bits were used to represent the phase information totaling to bit-rate of 300 bps. A similar approach described in [3] used 12 bits per speech frame to represent the phase information.

3. THE PROPOSED PHASE QUANTIZATION ALGORITHM

The proposed method avoids the explicit use of any extra bits to represent the phase information. This is a result of using a single codebook jointly representing both the magnitude and phase. The current approach operates on the signal in the time domain. This is similar to older approaches [7], that used direct waveform quantization that was found to be inefficient [3], [6]. Directly quantizing the waveform, implicitly allocates an excessive number of bits to the phase information, more than is perceptually required.

The proposed approach operates on the low-band signal part (0-800 Hz) in a split band approach. The band splitting step allows us to approach the waveform quantization in a different way, where the smooth signal is uniquely represented

by a variable length vector of control points (peaks and valleys) as shown in see Figure 1.

This Section introduces the TDPA model, its physical meaning, the extraction criteria, the proposed TDPA distortion measure, frame to frame interpolation, and the signal reconstruction.

3.1. Physical Representation of Phase Information

The model builds on the assumption that it is sufficient to rebuild the waveform using these peaks and valleys as control points. Also, it is possible to reconstruct the full wave by interpolating these control points. This was experimentally verified.

The unit extraction operation relies on the pitchmarks information to divide the signal into pitch periods. The target is to get the valley closest to the glottal pulse responsible for the voicing. The squares in Figure 2 mark the pitchmarks and the extracted prototype boundary.

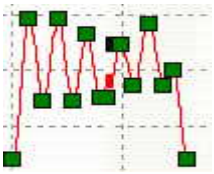

Figure 1: Low pass filtered pitch period of the vowel /a/. The squares represent the peaks and valleys.

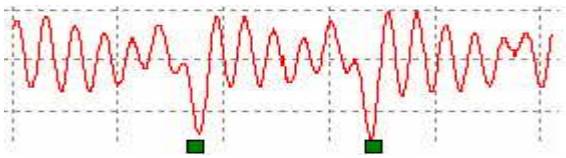

Figure 2: Speech frame for the vowel /a/. The squares are the pitchmarks identifying the beginning and end points of the prototype waveform.

3.2. TDPA as a Distortion Measure

The distance between two prototypes is defined as the sum of the Euclidean distances of each peak/valley of the time/gain normalized waveforms. Time normalization is done by stretching or compressing each frame in time to a fixed size. Gain normalization is done by amplifying the signal frame by a factor such that all frames would have the same power. Each control point is a two dimensional vector, i.e.

$$d = \sum_{points} \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (1)$$

where the x and y are the horizontal and vertical values of the observation peak/valley, respectively.

A Dynamic Time Warping (DTW) algorithm is used to find the best matching sequence of control points (peaks or valleys) having the smallest distance. The DTW algorithm allows relaxing the waveform matching to overcome the problem of attenuated peak-valley distance. When a peak-valley/valley-peak pair is skipped, the cost function is penalized by a value proportional to the peak-valley distance, which is related to the power magnitude of the distortion caused by skipping that pair. This allows gracefully handling birth and death of peaks across frames. The skipping penalty used by the DTW algorithm is equal to the Euclidean distance between the skipped peak, valley pair, i.e.

$$d_{skip} = \sqrt{(x_{peak} - x_{valley})^2 + (y_{peak} - y_{valley})^2} \qquad (2)$$

This proposed method was motivated by the desire to model both magnitude and phase information simultaneously and have the cost function be capable of determining the distortion measure depending on both instead of just the magnitude information, thus, eliminating the need to send separate envelope and excitation signals.

3.3. Incorporating Statistical Distribution Information in Classification

To further enhance the waveform matching during the classification phase, a diagonal Gaussian model was used to represent the statistical distribution of each control point, a set of Gaussian models are used to represent each template cluster. In Figure 3, the bold dots represent the extracted feature vector to be classified, and the clouds represent the Gaussian models of the best matching codebook entry. The mean of each Gaussian component in the best matching vector can be used to re-synthesize the waveform as shown in Figure 3.

During the encoding, the best match is chosen so that it is the template that provides the maximum likelihood instead of the template that provides the least Euclidean distance sum. The use of Gaussian models improves the matching; this is because the variances of the Gaussian model representing such control point will be small when the control point is relatively stationary within the cluster.

A weight function is also introduced so that the number of training points existing for each Gaussian is used as a weight function for the likelihood score of that Gaussian.

The likelihood that a template is matching the observation in its simplest case where no peak, valley pair is skipped in either the reference template or the observation is given by:

$$p = \prod_{points} \frac{1}{2\pi\sigma_x\sigma_y} \cdot e^{\frac{n}{N}} \cdot e^{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]} \qquad (3)$$

where n is the number of points in each cluster inside this template, N is the total number of points in all clusters, μ is the cluster mean, σ^2 is the cluster variance, parameters with x and y subscripts represent the value in the x and y dimensions, *points* are the peak/valley control points in the observation.

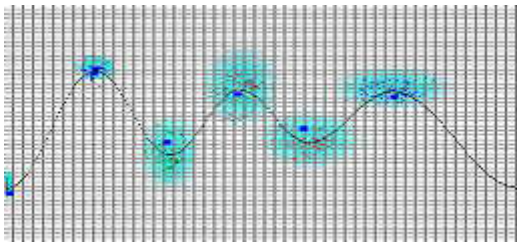

Figure 3 Matching with Gaussian clusters of a template

3.4. Speech Frames Interpolation

This is an important feature that must exist for any model in such system so that it would be possible to introduce smooth transition among frames.

The model is represented by a set of control points defining the peaks/valleys of the prototype waveform. Thus the problem is to define a method through which a morphed

waveform (in the control points format), is generated given an initial/final waveform.

The WI compression technique used eliminates redundancy by transmitting only the key-frames of the signal which represent milestones in the waveform evolution across time. Since the speech waveform is slowly evolving, it is possible to generate the waveform at any time instances using the previous and next key-frames. Generating the interpolated waveform at t_n, is defined to be the waveform resulting from linearly transforming each control point in the initial waveform with the corresponding control point in the final waveform.

$$y_{n,i} = (y_{N,i} - y_{0,i}) \times t_n + y_{0,i} \quad (4)$$

$$x_{n,i} = (x_{N,i} - x_{0,i}) \times t_n + x_{0,i} \quad (5)$$

$$t_n = \tfrac{1}{N} \times n \quad , n = 1 \cdots N-1 \quad (6)$$

where N is the number of waveforms generated to represent the smooth transition between the initial and final waveforms, t_n is the normalized time instances at which the morphed waveform needs to be generated (ranging between 0 and 1), M is the maximum number of control points of the initial/final waveforms, $i = 1 \cdots M$ is the control points' index, $(x_{0,i}, y_{0,i})$ are the (*x, y*) values of the control point number *i* of the initial waveform, $(x_{N,i}, y_{N,i})$ are the (*x, y*) values of the control point number *i* of the final waveform, $(x_{n,i}, y_{n,i})$ are the (*x, y*) values of the control point number *i* of the morphed waveform at instance t_n.

3.5. Signal Reconstruction

This is the step at which the waveform is generated from the control points. The operation is the same for normal and interpolated control points.

Based on the fact that the control points are alternating peaks/valleys, and the interval between each two successive control points can be approximated by a sinusoidal half cycle, a sinusoidal oscillator is used to interpolate each interval.

$$y_{n,i} = \frac{y_{n+1} + y_n}{2} + (y_{n+1} - y_n) \times \cos(\pi \cdot \tfrac{i}{x_{n+1} - x_n}) \quad (7)$$

$$y_{n,i} = y(x_n + i) \quad (8)$$

$$i = 0 \cdots x_{n+1} - x_n \quad (9)$$

where x_n, y_n are the x and y values of the nth control point.

4. EXPERIMENTS & RESULTS

This section is dedicated to validating the proposed system components. This is done through using a set of objective measurements; SNR, SD and PESQ [12]. Subjective measurements were performed by informal listening. Separate training and testing sets from the Arctic database [13] were used. The 5 testing sentences used represented 17 seconds of speech and about 1300 frames. The codebook size was limited to 5-bits (32 entries). Compared to other approaches, the size of the codebook used is very small. Informal listening results indicated that the codebook size was satisfactory; increasing the size resulted in over-training, with no perceptual improvement in quality although negatively affecting the bit-rate.

The results of the TDPA were compared to the standard MFCC distortion measure. Different combinations of both distortion measures were simulated, trying to understand the contribution of improvements for both phases; the codebook building phase and the encoding phase. In the following sub-sections, we will refer to the different combinations as [encoding-distortion-measure/training-distortion-measure].

In the experiment, the output of the testing dataset was encoded and then decoded using the different combinations. The synthesized signal contained only the low-band component so that we can eliminate the effect of the high-band components as they are encoded using a different method other than the TDPA. The output was then compared to the low-pass filtered original signal.

Figure 4 compares the SNR histograms of the different combinations of distortion measures. Generally speaking, the higher the average of the histogram gets, the better the performance from the SNR perspective. From the figure, it is clear that the histogram for using the TDPA as a distortion measure for encoding (TDPA/TDPA, TDPA/MFCC) is biased to the right, hence having higher (better) average SNR. While using MFCC as a distortion measure for encoding (MFCC/MFCC, MFCC/TDPA) results in a histogram that is biased to the left, with lower average SNR.

Table 1 shows the average SNR values. As shown, using the TDPA distortion measure for encoding (TDPA/TDPA, TDPA/MFCC) has the highest (best) scores.

The proposed TDPA method was evaluated using the Spectral Distortion (SD) measure. The results of different combinations were compared and are shown in Table 1. Using the TDPA distortion measure for encoding (TDPA/TDPA, TDPA/MFCC) results in the same SD as that for the MFCC counterpart. This is expected since the MFCC is an Euclidean based distortion measure that should result in the best SD. Having the TDPA being able to discriminate the phase, without negatively affecting the SD, is the TDPA key feature.

The PESQ score of different combinations were compared and are shown in Table 2. These show the PESQ score comparing the unquantized re-synthesized speech segments to the low-pass filtered segments. It also shows the PESQ score comparing the re-synthesized speech segments, using the different quantization/codebook distortion measures' combinations, to the unquantized re-synthesized speech segments. It is clear that the TDPA results in some improvement. This improvement is related to the fact that the TDPA is capable of producing speech segments that tend to evolve smoothly through time. Hence, it avoids the noise that would otherwise be perceived if the speech synthesis segments ignored the time evolving nature of the speech signal.

An experiment was made to examine the smoothing nature of the TDPA. Figure 5 shows the time domain evolution of the pitch period of the vowel /a/ for the original test sentence and for the quantized versions. It is clear that the TDPA succeeds in smoothing out the transition across neighboring frames. Thus results in better-perceived quality [14].

This result coincides with the informal listening results, where the TDPA speech was found to be of better quality because of being smoother, eliminating the gargling noise perceived with the MFCC/MFCC.

5. CONCLUSION

The slowly evolving nature of the narrow-band speech signal in the time domain explains the success of the TDPA model. Using the TDPA distortion measure instead of the MFCC, we observed slight gains in PESQ score. The TDPA resulted in better quality as judged by informal listening tests. This is attributed to the smooth time evolving nature of the synthesized signal in the case of using the TDPA, resulting in inter-frame correlation.

The TDPA allowed representing the phase information of the low-band speech signal without any increase in the bit-rate.

6. REFERENCES

[1] W. Bastiaan Kleijn and Jesper Haagen, "A speech coder based on decomposition of characteristic waveforms," *Proc. ICASSP'95*, pp. 508 - 511, May 1995.

[2] A. Gersho, "Advances in speech and audio compression," *Proceedings of the IEEE* , vol.82, pp.900-918, Jun 1994

[3] Yingbo Jiang, V. Cuperman, "Encoding Prototype Waveforms Using A Phase Codebook," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 20-22, 1995.

[4] O. Gottesman, A. Gersho, "Enhanced waveform interpolative coding at 4 kbps," *Proc. IEEE Workshop on Speech Coding,* pp. 90-92, 1999.

[5] O. Gottesman and A. Gersho, "Enhanced waveform interpolative coding at low bit-rate," IEEE Trans. Speech Audio Processing, vol. 9, pp. 786 - 798, 2001.

[6] Oded Gottesman, "Dispersion phase vector quantization for enhancement of waveform interpolative coder," Proc. ICASSP'99, pp. 269 - 272, 1999.

[7] A. Das, A. V. Rao, and A. Gersho, "Variable dimension vector quantization of speech spectra for low-rate Vocoders," *Proc. Data Compression Conferenc*e, pp. 421-429, 1994.

[8] S. Roucos and A. M. Wilgus, "The waveform segment vocoder: A new approach for very-low-rate speech coding," *Proc. ICASSP'85*, pp. 236-239, 1985.

[9] J. Cernocky, G. Baudoin, G. Chollet, "Segmental vocoder-going beyond the phonetic approach," *Proc. ICASSP'98*, vol.2, pp. 605-608, May 1998.

[10] Ki-Seung Lee, R.V. Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm," *IEEE Trans. on Speech and Audio Processing,* vol.9, pp. 482-491, Jul 2001.

[11] Ahmed Ismail, Yasser Dakroury, Hazem Abbas, "A Novel Energy Distribution Comparison Approach for Robust Speech Spectrum Vector Quantization," *Proc. Interspeech'07*, 2007

[12] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.

[13] Carnegie Mellon University, Language Technologies Institute, *http://www.festvox.org/cmu_arctic*, 2003.

[14] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," *Proc. ICASSP'95*, pp. 732 - 735, May 1995.

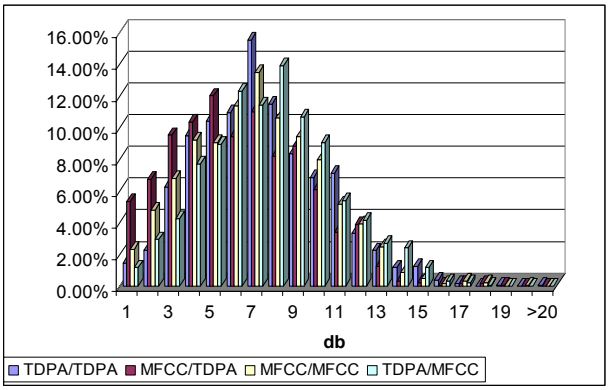

Figure 4 SNR comparison for different encoding/codebook-building distortion measures

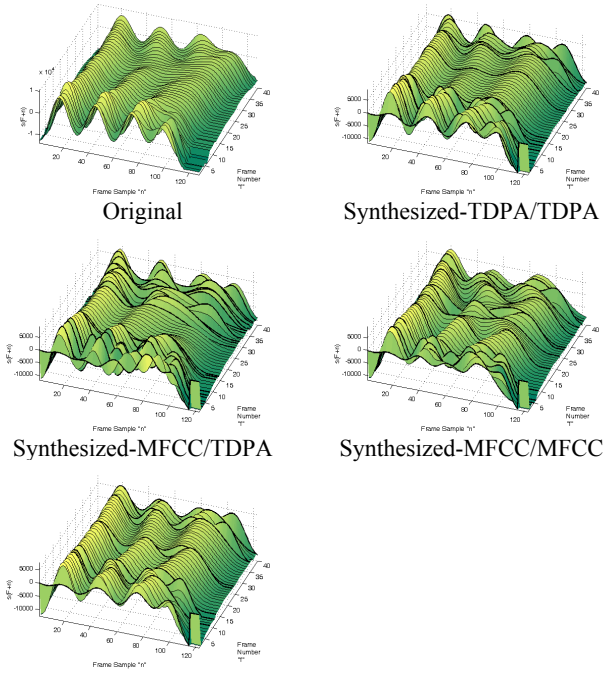

Figure 5 Original and synthesized low-pass filtered segment of the vowel /a/

Table 1 Spectral Distortion (SD) outliers, average SD, and average SNR comparison for TDPA, MFCC

	SD (db)			AVG	AVG
	<2	2-4	>4	SD	SNR
TDPA/TDPA	83.42%	11.41%	5.17%	1.40	6.69
MFCC/TDPA	81.53%	13.91%	4.56%	1.49	5.53
MFCC/MFCC	83.54%	11.82%	4.64%	1.39	6.44
TDPA/MFCC	82.93%	12.28%	4.79%	1.39	7.02
Unquantized	85.39%	9.74%	4.87%	1.25	10.59

Table 2 PESQ score comparison for TDPA

	Unquantized	TDPA/ TDPA	MFCC/ TDPA	MFCC/ MFCC	TDPA/ MFCC
Average	2.01	3.26	2.67	3.13	3.30