Footprint Reduction of Concatenative Text-To-Speech Synthesizers using Polynomial Temporal Decomposition

Tamar Shoham, David Malah, *Life Fellow*, IEEE, and Slava Shechtman

***Abstract*—High quality low footprint Concatenative Text-To-Speech (CTTS) synthesizers provide a persistent challenge in the field of speech processing. The spectral parameters representing the short speech segments used in the concatenation process constitute a large portion of the required memory. In this paper we propose to use a vectorial form of Polynomial Temporal Decomposition combined with jointly optimal segmentation and polynomial order selection in order to reduce the storage required for the spectral amplitude parameters by 50%, while preserving the perceptual quality of the obtained synthesized speech.**

I. Introduction

Concatenative Text-To-Speech synthesizers require storage of many compressed speech segments. These segments are organized in acoustic leaves, with all speech segments in a leaf belonging to the same sub-phoneme in the same context. In small footprint CTTS systems, each speech segment is usually represented by a parametric model. Specifically, in IBM's system [1], on which this work is based, each acoustic leaf contains 5-10 speech segments, with each speech segment consisting of 1-35 frames, with a median of 2. For each frame of 10msec duration, a vector of 32 amplitude parameters and a variable length vector of phase parameters, represent the spectral envelope of the frame, sampled in mel-scale to match perceptual characteristics of the auditory system, as detailed in [1]. During speech synthesis an appropriate acoustic leaf is selected for each sub-phoneme. A speech segment within a leaf is selected based on spectral and pitch smoothness criteria, combined with target prosody similarity metrics. The stored parameters are used to synthesize the frames and from them the speech segment, which is then concatenated to the previously synthesized speech [2]. In order to reduce CTTS footprint we wish to further compress the parameters stored in these acoustic leaves, without perceptually reducing obtained speech quality. In our work we focused on the amplitude parameters, whose original footprint in the system we used is 5.7 MB. The phase parameters, with a footprint of only 1.6MB, remain unaltered.

Thus, we have a parameterized speech compression problem, which we address using Temporal Decomposition (TD). TD describes a set of compression methods which attempt to exploit the temporal redundancy in the data by representing the evolution over time, i.e., between frames, of either a scalar value (scalar TD) or a parameter vector (vector TD) with a reduced model, thus achieving compression. In [3] Athaudage et. al. propose vectorial TD of spectral parameters in a MELP speech coder using dynamic programming. A similar concept is presented as the LEBEL-TD algorithm in [4]. Shechtman and Malah in [5] propose a computationally efficient, iterative, sub-optimal approach to TD modeling, as well as a perceptually weighted error criterion for improved perceptual performance. In [6], Kain and Santen propose to compress acoustic inventories by performing asynchronous interpolation of templates representing the beginning and end of each acoustic unit. This is essentially a restricted TD setup, which achieves a high compression ratio at the price of poor perceptual quality and low flexibility. Scalar TD approaches attempt to model each scalar parameter trajectory separately. In [7] Girin et. al. present a DCT based model for the trajectories of the amplitude and phase of the sinusoidal coding model. In [8] Dusan et. al. present another scalar TD approach, where a pre-determined number of consecutive speech frames are jointly coded by representing each parameter trajectory with an approximating polynomial. These polynomials, of order P_i are then represented by their $P_i + 1$ samples, as polynomial coefficients are not robust to quantization. In our work, we extend the polynomial TD approach. We combine it with optimal segment ordering to obtain a super-segment from each acoustic leaf, and then perform jointly optimal segmentation and polynomial order selection, using an algorithm along the lines of [9]. By enforcing joint segmentation and order selection along the vectors, and performing the polynomial sampling at synchronous locations, we actually revert to a vectorial TD approach. A major advantage of this approach is the ability to code these sampled vectors with the same split-VQ designed to code the original parameter vectors, since they lie in the same space.

The paper is structured as follows. In Section II we present the outline of the proposed algorithm. In Section III we describe the vectorial polynomial TD approach, followed by a description of optimal speech segment ordering in Section IV. In Section V we explain the joint segmentation and polynomial order selection algorithm and then describe some reduced complexity setups in Section VI. In Section VII we present experimental results and then conclude.

T. Shoham is at the Department of Electrical Engineering, Technion - Israel Institute of Technology tshoham3@gmail.com

D. Malah is at the Department of Electrical Engineering, Technion - Israel Institute of Technology malah@ee.technion.ac.il

S. Shechtman is at the IBM Haifa Research Labs SLAVA@il.ibm.com

II. Algorithm Outline

To apply our TD-based algorithm to each acoustic leaf, we begin by concatenating the speech segments in the leaf to obtain one long super-segment. The order of the speech segments is determined according to the optimal segment ordering described in Section IV. Since we do not expect smoothness to hold for the entire super-segment, we perform sub-segmentation into TD segments in an optimal manner, as described in Section V. Then, the vectorial polynomial TD, described in Section III is applied to the parameters of each TD segment. Our goal is to reach a target rate R_t while minimizing the obtained distortion. We found that minimizing the maximum distortion provides better perceptual quality than minimizing mean distortion. Also, we bound the allowed distortion jointly over the entire database, to obtain consistent quality, while achieving the overall target rate or compression ratio, allowing variation of compression ratio among leaves. This approach outperforms enforcing the target compression ratio for each leaf. Thus, our constrained optimization problem can be described as follows: Assuming a known per-frame distortion function, D_f, (discussed in subsection V) calculated between the original and reconstructed speech frames, the global distortion D_g is the maximum distortion over all frames, segments and leaves, defined as:

$$D_g = \max_{leaves} \{\max_{TDsegments} [\max_{frames}(D_f)]\} \quad (1)$$

We wish to find the smallest global distortion, D_g^* for which the target rate R_t is obtained. i.e.:

$$D_g^* = \min(D_g) \;\; s.t. \;\; R(D_g^*) < R_t \quad (2)$$

An iterative solution, similar to the one proposed in [10] is adopted. We define a Rate-Distortion structure, RD_s, that holds the distortions obtained in the previous iteration, and enables an efficient bi-section search for the optimal distortion value. This structure may also hold the target rate and tolerance range,which is necessary due to the step-wise nature of the rate distortion function. At each iteration, RD_s holds the current lower and upper distortion values, D_L and D_U, which define the ends of the 'active' interval, within which we are trying to pinpoint our target working point. The proposed algorithm steps are:

1) Initialize: D_L=0, D_U=maximum distortion value.
2) Perform Polynomial TD with $D_g = D_U$, set $rate$ to obtained rate.
3) Verify $rate < R_t$. If not, double D_U value and GOTO 2.
4) Calculate next value for D_g: $D_g = \frac{1}{2}(D_L + D_U)$.
5) Perform polynomial TD for each leaf in the database, limiting maximum allowed distortion to D_g, and set $rate$ to the obtained overall rate.
6) IF $rate$ is within the tolerance range of the target rate, R_t,: GOTO 8.
7) IF $(rate < R_t)$: $D_U = D_g$, ELSE: $D_L = D_g$; GOTO 5.
8) $D_g^* = D_g$; END.

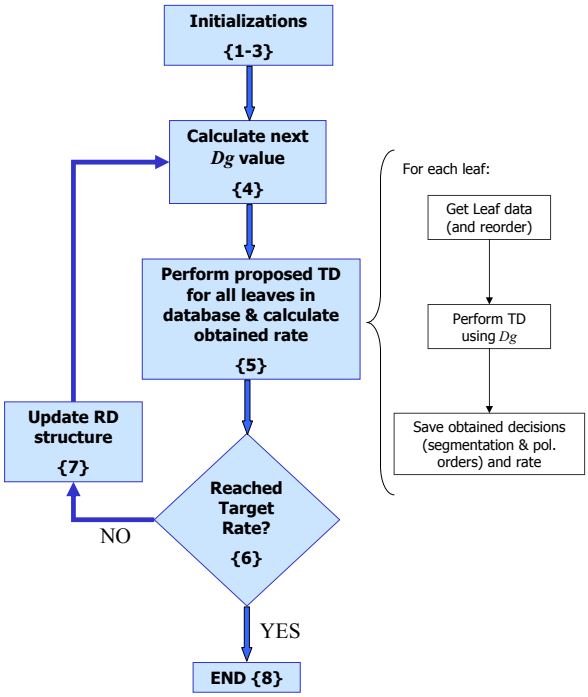

Fig. 1. Outline of proposed algorithm

The need for step 3 stems from the fact that on one hand, we do not wish to initialize our interval with a value of D_U that is too high and will incur unnecessary iterations. For instance we could set initial distortion to its maximum by transmitting no data, but then we would require quite a few iterations to narrow our interval to the relevant values. On the other hand, we must make sure that our initial D_U is high enough to assure that the working point we seek lies within the designated interval - which is exactly what step 3 does. This algorithm is illustrated in Fig. 1. Simple bi-section is performed, as proposed in [10]. The option of performing weighted bi-section was also evaluated, but in many cases the resulting convergence was actually slower due to the non-linearity of the rate-distortion function.

Note that when calculating the obtained rate we take into account both the bits incurred by coding the parameters, and for each TD segment, the overhead bits required to hold the selected segment length and polynomial order. We also note that the proposed algorithm allows for automatic adaptation to any target rate or compression ratio.

III. Polynomial TD

Let a given TD segment consist of N frames, with each frame represented by a parameter vector of length K. We wish to fit the trajectory of each parameter with a P^{th} order

polynomial, using least squares. For compression we require $P+1 < N$. The obtained K polynomials are then sampled at $P+1$ points, creating $P+1$ vectors, which are quantized and stored. By enforcing the same order P for all K trajectories, we obtain vectors that have similar behavior to the original parameter vectors, which enables use of the current system quantization and coding tools. The decoder reconstructs the $P+1$ vectors, and finds the K polynomials they define. These polynomials are resampled at N points to obtain the reconstructed values. Thus, the eventual reconstruction error is a function of both the polynomial model fitting error and the quantization error. At sample points the quantization error is the actual error introduced by the quantizer. For the interpolated samples, the error is a function of the error in the reconstructed polynomial coefficients, and grows exponentially with polynomial order. Therefore, and also due to implementation complexity considerations, we limit the maximum allowed polynomial order to 4. The quantization based error at interpolated samples is also proportionally inverse to the distance to adjacent sampling points, therefore we must take care to sample the polynomial at well distributed locations.

IV. Optimal Speech-Segment Ordering

As previously explained, the CTTS pre-selected database contains many acoustic leaves, each containing up to ten speech segments with one or more frames of spectral parameters. The original order of the segments in the leaf has no significance. Since we wish to compress these segments jointly by concatenating them into a single super-segment we must find the optimal concatenation order. For this we must first define a target cost function to minimize and then find a minimization algorithm. We examined a number of cost functions based on smoothness at speech segment joints and overall smoothness of the super-segment. The best performance was found when minimizing a weighted mean squared distance between the actual data and the samples of a second order polynomial fitted to each trajectory of the super-segment data. The weighting prioritizes the lower elements in the vector since they have greater perceptual importance. Now, we need to find the speech segment ordering that minimizes this cost. This poses a form of the renowned Traveling Salesperson Problem (TSP), from the realm of Combinatorial Optimization. Many possible solutions exist to this problem, but since our cost function is not Euclidian, classic approaches such as the Farthest or Nearest Insertion Algorithms are not applicable. We therefore chose to find the optimal order using a modified Binary Switching Algorithm (BSA) along the lines proposed in [11]. We start with an initial order and select a random speech segment move. If the cost is reduced, or at a certain probability (which depends on the increase in the cost function), even it is not, the move is accepted. This continues for a pre-determined number of attempted moves at which point the order that provided the lowest cost is selected. Note that the complexity is not of great concern as the ordering is performed once, off-line, and then can be stored for each leaf.

V. Joint segmentation and polynomial order selection

Since we wish to use low order polynomials and also cannot presume smoothness assumptions will hold for the entire acoustic leaf, we must perform segmentation of the super-segment into a number of TD segments. We will now describe the proposed algorithm for jointly optimal segmentation and polynomial order selection, based on the algorithm presented in [9].

Due to the dependencies between the segmentation decisions, we define a generalized 3-D trellis structure. The horizontal dimension corresponds to the candidate TD segment termination points (segment ends), the vertical dimension corresponds to TD segment length, and the depth dimension corresponds to candidate polynomial orders. Each point $S_{i,j,k}$ in this structure is assigned a cost, based on the distortion calculated for a TD segment that ends after frame i, consists of j frames and uses a polynomial of order k. Then we "flatten" the structure by setting the value of k at each trellis point to the lowest polynomial order for which the resulting distortion in the corresponding TD segment does not exceed the current value of D_g. Finally, we seek the optimal path through the 2-D structure, consisting of points $S_{i,j,k^*} \equiv S_{i,j}$ using a back-tracking method. This is illustrated in Fig. 2.

Distortion measure

We now address the per frames distortion function, D_f, from Eqn.(1). In the optimization process we must constantly evaluate the distortion between the original speech and the reconstructed speech with a specific TD setup. However, we cannot afford to transform back into the speech domain for each evaluation point, in order to actually measure the obtained distortion. Therefore, we must find a function that when applied to the original and reconstructed parameter sets, predicts the perceptual distortion reliably. We evaluated two candidate functions:

1) Mean Squared Error (MSE) between the parameter sets.
2) Log Spectral Distortion (LSD) measure calculated directly in the parameter space.

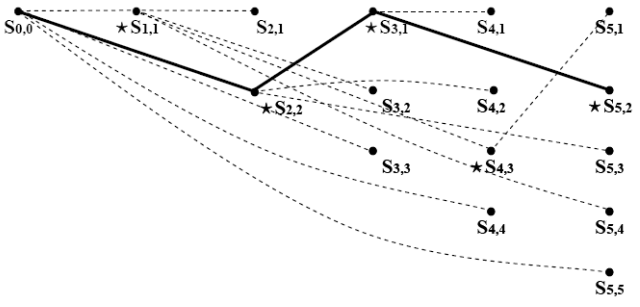

Fig. 2. 2-D structure for optimal segmentation and polynomial order selection. dotted lines show all possible steps, solid line shows optimal path. A star marks the state with the lowest accumulated cost in each column

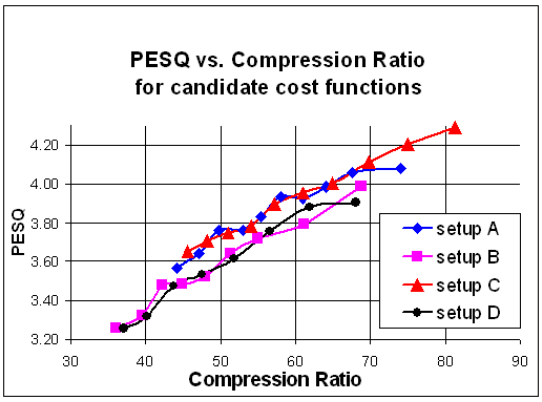

Fig. 3. Comparison of proposed distortion functions using MSE or LSD distortion and maximum or mean between frames: setup A: maximum LSD; setup B: mean LSD; setup C: maximum MSE; setup D: mean MSE;

The cost functions were combined into a compression setup, without quantization of the parameter samples. The obtained quality and obtained compression ratio were evaluated for each proposed cost, when limiting maximal distortion over the TD segment, vs. limiting mean distortion over the TD segment. Results are shown in Fig. 3. Using the LSD measure did not improve perceptual performance. This is because the spectral parameters were already obtained with perceptual considerations in mind. We found that using the MSE over with limiting the maximal distortion over the TD segment, (Setup C), provided the best and most consistent results.

VI. Reduced complexity setups

The algorithm is to be used in small footprint CTTS synthesizers. Since the host devices often have both CPU and memory constraints, we also examined some reduced complexity setups.

Low order polynomials

To avoid the need to perform complex polynomial fitting when decoding (using least-squares), we evaluated the performance allowing polynomials of orders 0 and 1 only, which require at most linear interpolation. The optimization procedure is carried out in the same manner, though has lower complexity due to the reduction in possible setups.

Naive segmentation

In this approach, we do not extend the TD segments beyond original speech segment boundaries. This substantially reduces encoding complexity, and also reduces decoding complexity since there is no need for decoding of additional frames in neighboring speech segments that are part of the same TD segment. Long speech segments are split into TD sub-segments, and the polynomial order for each TD segment is found s.t. the maximum distortion along the segment is bounded by D_{max}, which is found using the iterative algorithm used in the full setup. We evaluated this setup with a maximum TD segment length of 8 and of 4. The longer segments enable more efficient compression, but in order to obtain the target distortion may require polynomials up to order 7, which in turn may increase quantization error and decoding complexity.

VII. Experimental Results

We evaluated the proposed algorithm within the IBM small footprint CTTS synthesizer [1], with a target compression ratio of 50% of the amplitude parameters. The algorithm was applied to 1661 acoustic leaves in the database, and the obtained quality was evaluated on 10 sample sentences created from these leaves. Listening tests showed all proposed setups provided quality that was perceptually equivalent to that of the original CTTS output. The obtained wideband PESQ scores [12] on the reconstructed sentences vs. original CTTS output, for each evaluated setup, are provided in Table I, where setup 1 is the full algorithm implementation with max. polynomial order of 4; setup 2 is the reduced complexity setup with max. polynomial order of 1; 1b and 2b refer to these setups but using the original speech segment order without the optimal ordering algorithm. Setup 3 and 4 refer to performing naive segmentation with maximum TD segment length of 8 and 4, respectively. The table also shows the encoding and decoding complexity ranking for each setup. For comparison, when performing downsampling by a factor of 2 and reconstructing with linear interpolation we obtained an average PESQ score of only 2.84. As seen here the optimal ordering algorithm contributed to the performance of the Full Setup, but did not much improve in the reduced complexity setup. This is due to the fact that due to the low order polynomials, the segmentation algorithm generally selects quite short TD segments and only few of the TD segments span across speech segment boundaries - thus their ordering does not have much effect on the overall performance.

Looking at the selected polynomial orders in the full setup we find that 69.8% of the TD segments use polynomials of order 0 or 1. Note, that this is despite an inherent algorithmic preference for longer TD segments with higher order polynomials, as this reduces the relative overhead which is constant per TD segment. Therefore, limiting polynomial order affects less than a third of the full setup TD segments. In addition, the overhead per TD segment in the reduced complexity setup is lower, as polynomial order is represented by a single bit, which reduces the overall rate. This explains the improvement in performance despite the more restricted optimization. Furthermore, we see that due to the overhead required to represent the TD segment lengths, the naive segmentation with segment length 8 actually provides higher speech quality than the optimized segmentation scheme. However, it requires polynomials of orders up to 6 in order to comply with the distortion constraint, which increases decoding complexity and reduces quantization stability. Therefore, under our requirement of limited decoder complexity and high perceptual quality, Setup 2 is recommended for use in the target application.

TABLE I
PERFORMANCE OF PROPOSED ALGORITHM FOR VARIOUS SETUPS

Setup	Avg PESQ	Min PESQ	complexity enc ; dec
Setup 1	3.67	3.49	high ; medium
Setup 1b	3.55	3.45	high ; medium
Setup 2	3.69	3.51	medium ; low
Setup 2b	3.66	3.50	medium ; low
Setup 3	3.70	3.50	medium ; high
Setup 4	3.63	3.50	low ; medium

VIII. CONCLUSION

We presented a vectorial polynomial TD algorithm, which uses optimal segmentation and polynomial order selection, and joint optimization over the database to find the lowest distortion bound that ensures the target rate. This algorithm enables a further 50% compression of the spectral parameter amplitudes stored in an IBM low-footprint CTTS synthesizer, without a perceptual quality loss. Different setups of the algorithm provided similar results, thus allowing for implementation of reduced complexity versions of the algorithm. Further research is required to apply a similar algorithm to the spectral phase parameters and to generalize to other parametric speech models.

ACKNOWLEDGMENT

This research was performed at the Signal and Image Processing Lab (SIPL), Technion I.I.T, in collaboration with IBM's Haifa Research Lab (HRL). The authors would like to thank Ron Hoory, head of the Speech Technologies Group in HRL, Ariel Sagi (formerly of HRL), Zvi Kons (HRL) for their constructive inputs, and the devoted SIPL staff Nimrod Peleg, Yair Moshe, Ziva Avni and Avi Rosen for their technical support.

REFERENCES

[1] D. Chazan *et al.*, "Small footprint concatenative text-to-speech synthesis system using complex spectral envelope modeling," in *Eurospeech*, Lisbon, Portugal, Sep. 2005.

[2] R. Donovan *et al.*, "Current status of the IBM trainable speech synthesis system," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, UK, Aug. 2001.

[3] C. Athaudage, A. Bradley, and M. Lech, "Model-based speech signal coding using optimized temporal decomposition for storage and broadcasting applications," *EURASIP Journal On Applied Signal Processing*, pp. 1016–1026, Oct. 2003.

[4] P. C. Nguyen, M. Akagi, and B. P. Nguyen, "Limited error based event localizing temporal decomposition and its application to variable-rate speech coding," *Speech Communication*, vol. 49, no. 4, pp. 292–304, Apr. 2007.

[5] S. Shechtman and D. Malah, "Efficient sub-optimal temporal decomposition with dynamic weighting of speech signals for coding applications," in *Interspeech 2004 - ICSLP*, Korea, Oct. 2004.

[6] A. Kain and J. van Santen, "Unit-selection text-to-speech synthesis using an asynchronous interpolation model," in *6th ISCA Workshop on Speech Synthesis*, Bonn, Aug. 2007.

[7] L. Girin, M. Firouzmand, and S. Marchand, "Perceptual long-term variable-rate sinusoidal modeling of speech," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 3, pp. 851–861, Mar. 2007.

[8] S. Dusan *et al.*, "Speech compression by polynomial approximation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 2, pp. 387–395, Feb. 2007.

[9] 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.

[10] G. Schuster and A. Katsaggelos, *Rate-Distortion Based Video Compression*. Dordrecht: Kluwer Academic Publishers, 1997.

[11] K. Zeger and A. Gersho, "Pseudo-gray coding," *IEEE Trans. on Communications*, vol. 38, no. 12, pp. 2147–2158, Dec. 1990.

[12] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Geneva, Switzerland, Nov. 2005.