SIGNAL & IMAGE PROCESSING LAB

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

Tamar Shoham*, David Malah*, Slava Shechtman#

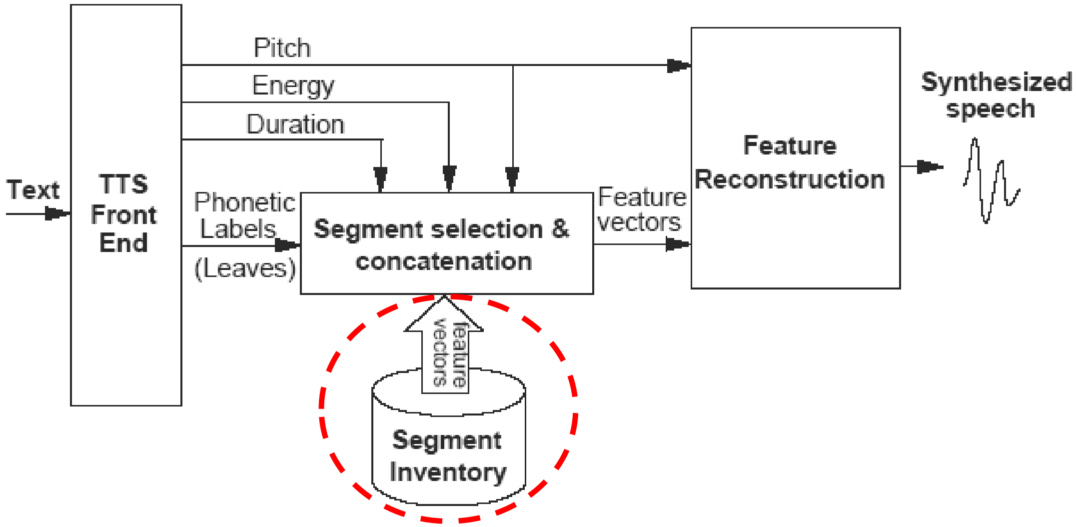
(*) Signal & Image Processing Lab - Technion

(#) IBM – Haifa Research Labs

ISCCSP 3-5 March 2010, Limassol, Cyprus.

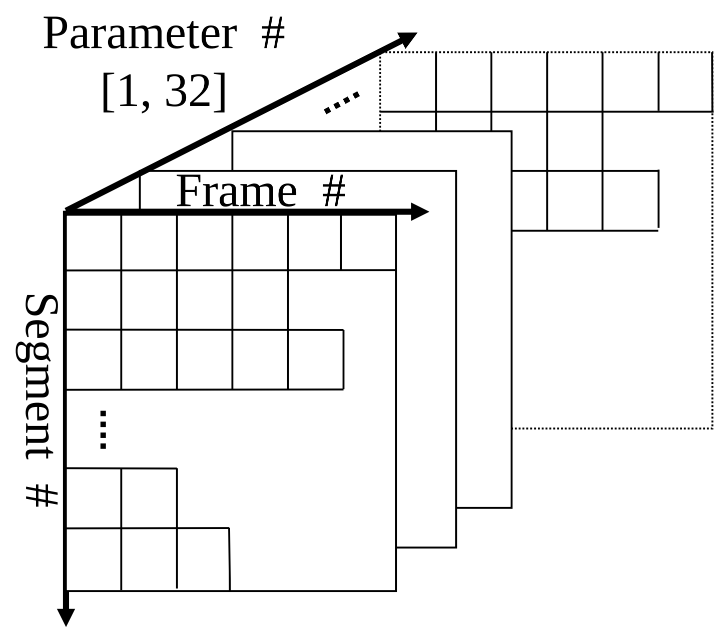
Introduction

- **Goal**: Further footprint reduction of a small footprint IBM Concatenative Text-To-Speech (CTTS) system.
- **Method**: Re-compression of the stored speech parameters in the speech segment database.
- **Proposed technique**: Remove redundancies between speech frames using Polynomial based Temporal Decomposition (TD).



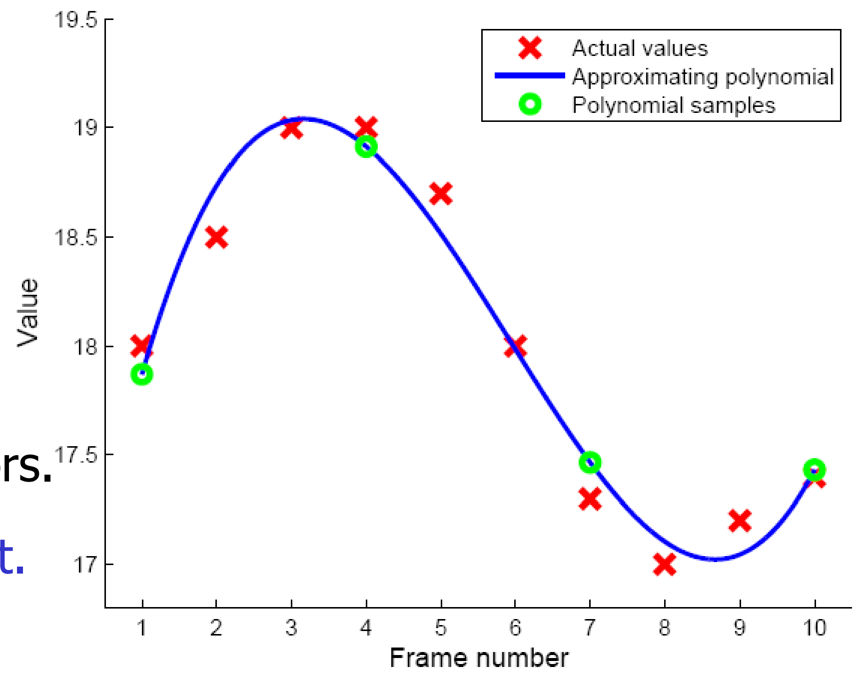
CTTS database structure

- The database consists of acoustic leaves, each corresponding to a specific sub-phoneme in a specific context.
- 5-10 speech segments are stored in each acoustic leaf.
- Each speech segment consists of one or more speech frames, each represented by a parametric spectral model, with 32 amplitude parameters per speech frame.



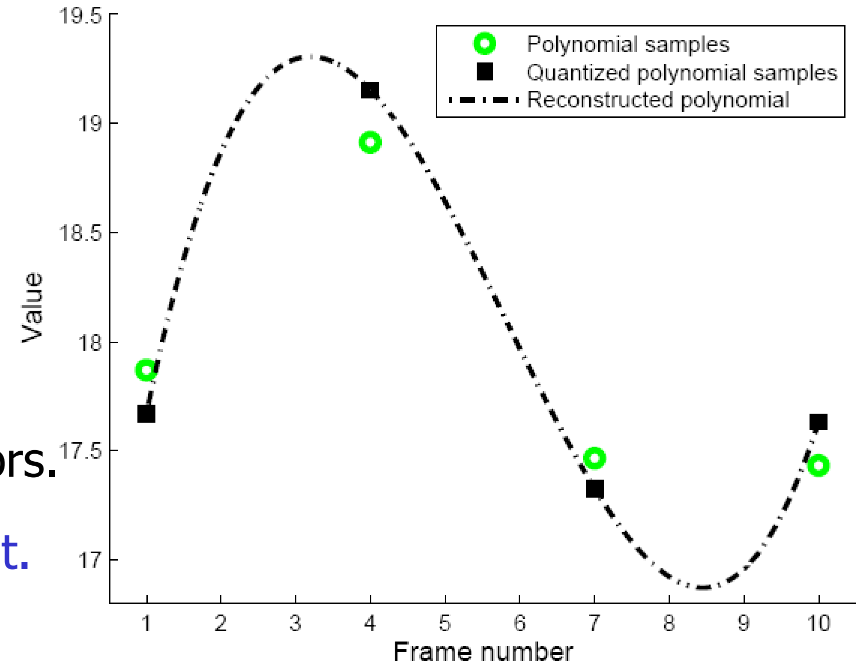
Polynomial TD

- Initially proposed by Dusan *et al.* (2007).
- Represent the trajectory of N data points by the approximating P^{th} order polynomial (for compression $P<N-1$).
- Represent the polynomial by its P+1 samples.
- We propose a vectorial form:
  - Apply to amplitude vectors.
  - Obtain P+1 representing vectors.
  - Adapt N and P per TD segment.



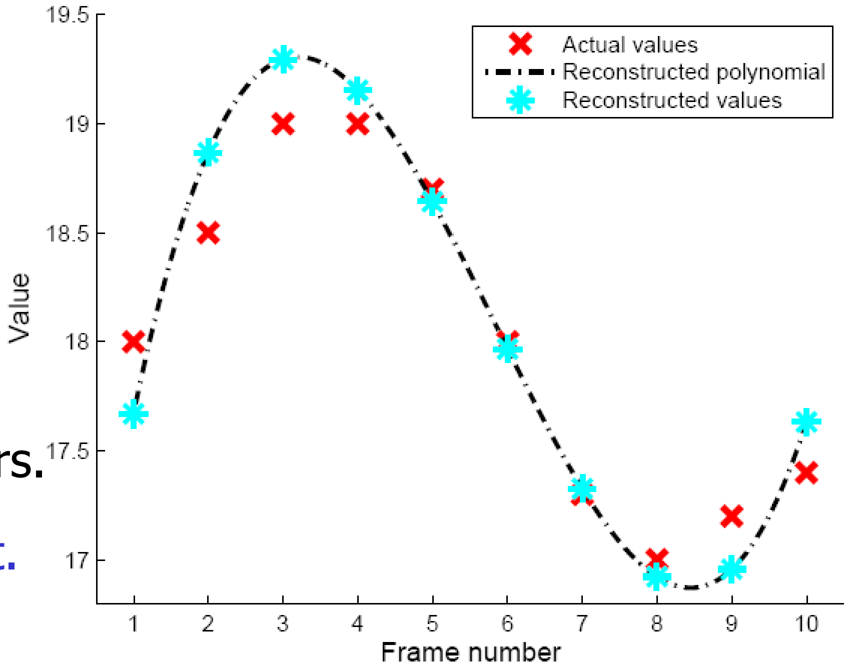
Polynomial TD

- Initially proposed by Dusan *et* al. (2007).
- Represent the trajectory of N data points by the approximating P^{th} order polynomial (for compression $P<N-1$).
- Represent the polynomial by its P+1 samples.
- We propose a vectorial form:
  - Apply to amplitude vectors.
  - Obtain P+1 representing vectors.
  - Adapt N and P per TD segment.



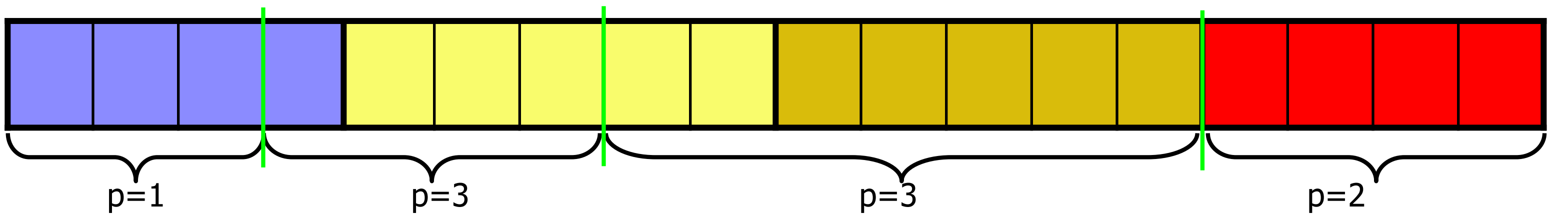
Polynomial TD

- Initially proposed by Dusan *et* al. (2007).
- Represent the trajectory of N data points by the approximating P^{th} order polynomial (for compression $P<N-1$).
- Represent the polynomial by its P+1 samples.
- We propose a vectorial form:
  - Apply to amplitude vectors.
  - Obtain P+1 representing vectors.
  - Adapt N and P per TD segment.



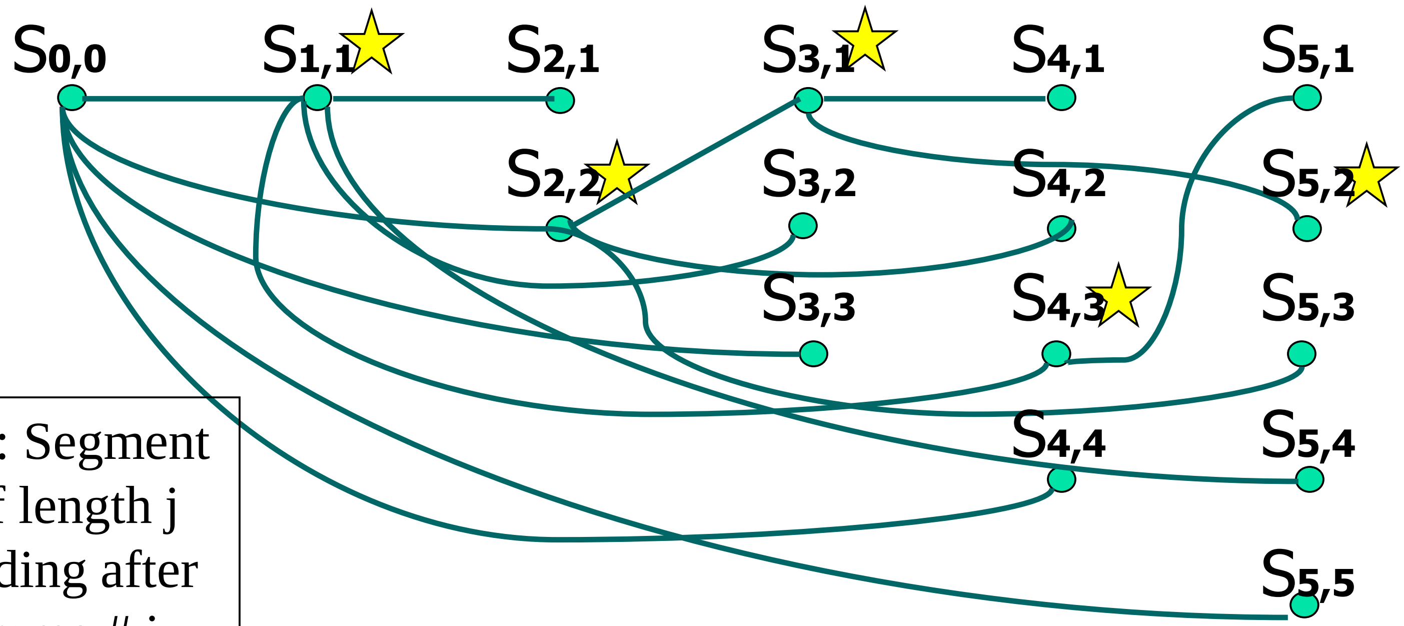
Polynomial TD for acoustic leaf

- Segments in each acoustic leaf are concatenated into a single ’super-segment’.
- Concatenation order is selected so that a cost function corresponding to the ‘super-segment’ smoothness is maximized.
- Smoothness criteria: WMSE between data & fitting pol. (order 2).
- Split ‘super-segment’ into short TD segments and fit each with a set of low order polynomials.



Segmentation and order selection

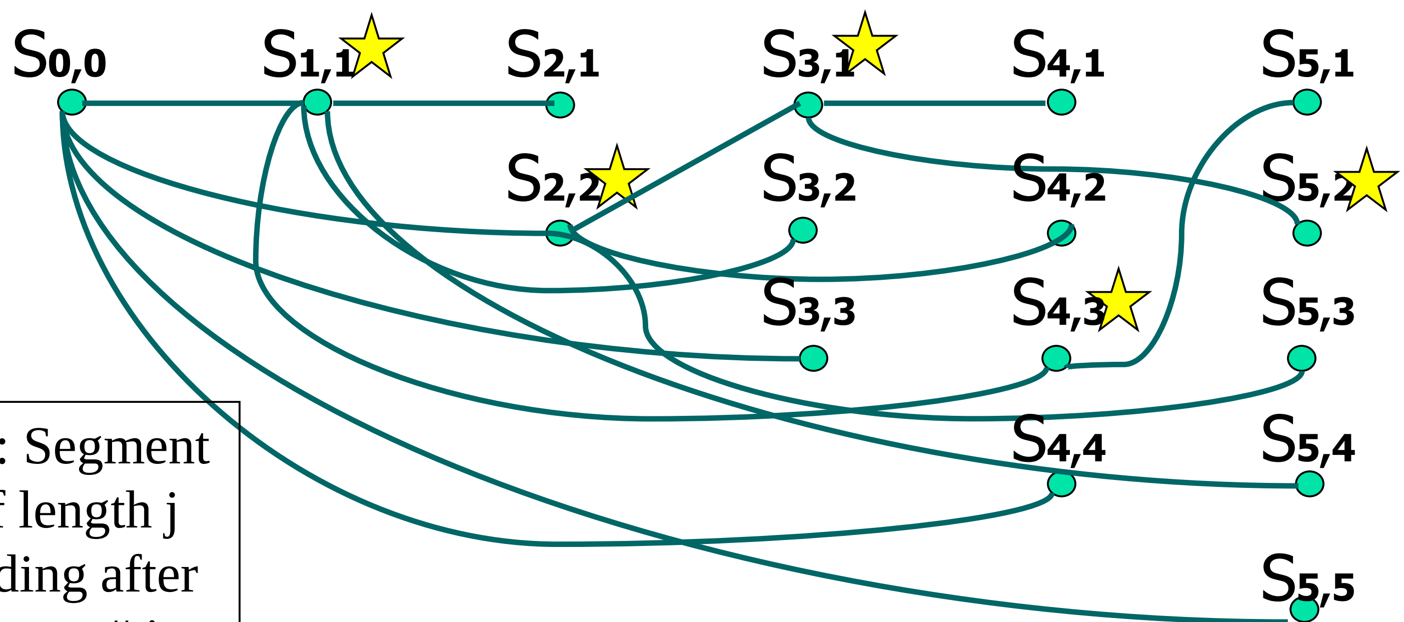
- Based on "R/D optimal linear prediction", Prandoni *et* al. (2000).
- First, build graph with all possible segmentations.
- For each segment find lowest polynomial order that guarantees target distortion; assign a cost based on the corresponding rate.
- Find lowest cost path across graph using backtracking.



Si,j: Segment of length j ending after frame # i.

Segmentation and order selection

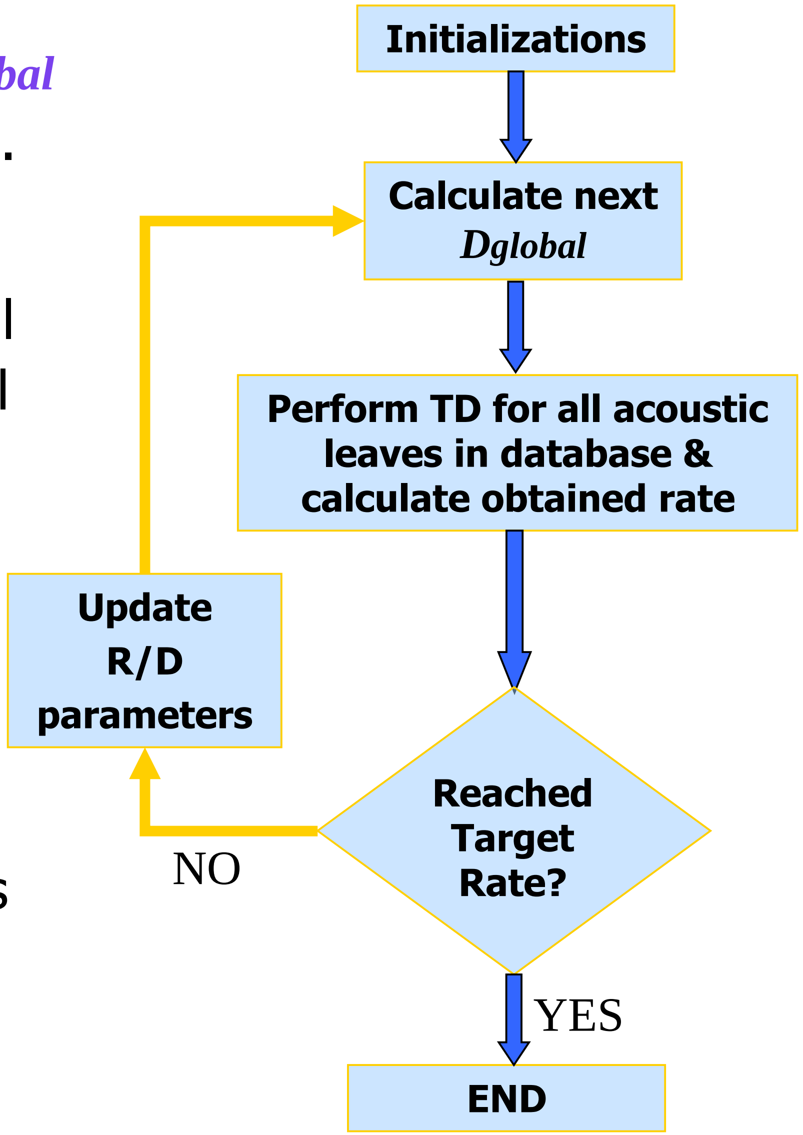
- Based on “R/D optimal linear prediction”, Prandoni *et* al. (2000).
- First, build graph with all possible segmentations.
- For each segment find lowest polynomial order that guarantees target distortion; assign a cost based on the corresponding rate.
- Find lowest cost path across graph using backtracking.



$S_{i,j}$: Segment of length j ending after frame # i.

Proposed Algorithm outline

- We seek the minimum ***Dglobal*** for which rate = target rate.
- ***Dglobal*** is the maximum allowed distortion among all frames in all segments in all leaves.
- For each candidate ***Dglobal*** value, we apply proposed polynomial TD to acoustic leaves.
- The calculated rate includes required overhead bits.



Some results

PESQ scores for recompression factor: x2

Setup \ #		S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10	Avg.
Max Ord. 4	No Reo	3.56	3.46	3.61	3.51	3.62	3.63	3.52	3.45	3.64	3.46	**3.55**
	W. ReO	3.74	3.76	3.60	3.49	3.49	3.91	3.95	3.56	3.63	3.53	**3.67**
Max Ord. 1	No ReO	3.60	3.39	3.82	3.56	3.66	3.71	3.89	3.57	3.68	3.70	**3.66**
	W. ReO	3.72	3.54	3.70	3.55	3.64	3.72	4.04	3.51	3.81	3.65	**3.69**

(*) For comparison: average PESQ for naïve down-sampling 2:1 is 2.84.

Samples:

	Original	Max poly. order 4		Max poly. order 1	
		No Reo	W.Reo	No Reo	W. Reo
S.8					
S.1					

Summary

- We presented an algorithm for recompression of amplitude spectral parameters in a small footprint CTTS system, providing equivalent perceptual quality with a recompression factor of 2.
- We showed a vectorial form of polynomial TD used with jointly optimal sub-segmentation and polynomial order selection.
- Iterative algorithm converges to target rate with minmax distortion.
- Important feature: The compressed 'data' lies in the in the same space as the original data set.
- We applied the algorithm to a specific case, but it can be readily applied to a variety of (re) compression challenges.

Thank you