TEMPLATE BASED TECHNIQUES FOR AUTOMATIC SEGMENTATION OF TTS UNIT DATABASE

S. Adithya[*], Sunil Rao[**], C. Mahima[$], S. Vishnu[#]

Mythri Thippareddy[$], V. Ramasubramanian[$]

[*]University of California, San Diego

[**]Arizona State University, Tempe, Arizona

[#]San Diego State University, San Diego

[$]PES Institute of Technology
Bangalore South Campus (PESIT-BSC)
Bangalore, India

SUMMARY

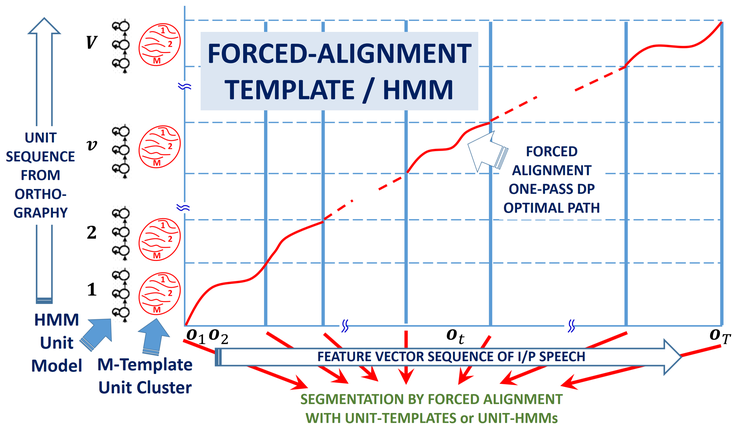
- Template based automatic segmentation of unit-database for TTS – into phonetic and syllabic units
- Proposed 'template' based – a) Seeded forced alignment and b) Seedless iterative segmental K-means (SKM)
- Compared with a) Seeded HMM, b) Seedless SKM-HMM on TIMIT – with phonetic 'ground truth' segmentation
- Applied for 'syllabic' segmentation of an Indian language 'Tamil'
- Proposed 4 methods compared with 3 other methods a) Festival EHMM, b) Group-Delay semi-automatic, c) Hybrid
  - In terms of a) segmentation error statistics and b) objective TTS quality measure using 'spectral distortion'
  - Template based methods shown comparable / better than other approaches

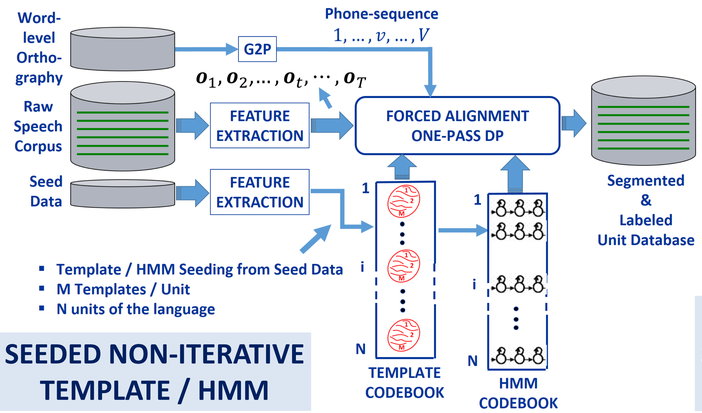
Template based automatic segmentation

- Non-parametric model of a unit → Feature vector sequence retained as it is
- Represents temporal content of a speech unit → high-resolution
- Template modeling → Used extensively in speech coding, speech recognition, speaker-recognition. But not yet for "segmentation" for TTS
- Can yield high degree of matching with test data → hence potential to yield high-resolution segmentation → for TTS Unit-Database Segmentation

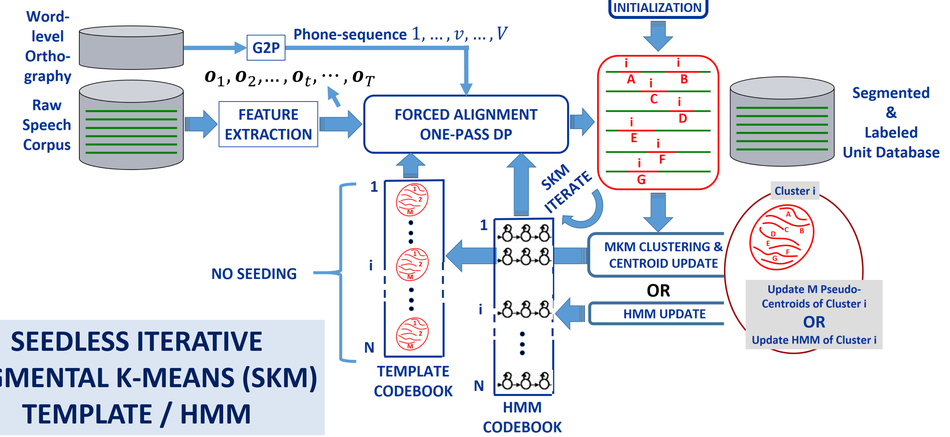
Main dimensions of "unit" modeling for segmentation

Unit Modeling		Context	Seeding
Multiple templates (M) / unit	HMM from M-templates / unit	Context-independent	Seeded – Non-iterative
Random selection		Context-dependent	Seedless – Iterative Segmental K-means
Clustered selection			





SEEDED NON-ITERATIVE TEMPLATE / HMM

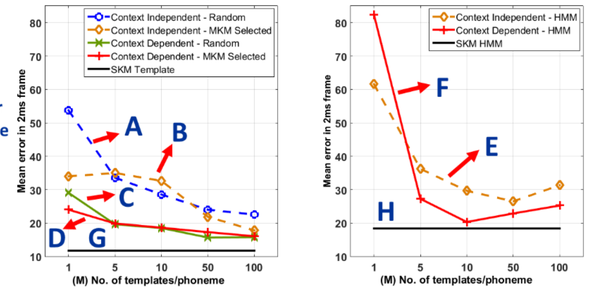


SEEDLESS ITERATIVE SEGMENTAL K-MEANS (SKM) TEMPLATE / HMM

SEGMENTATION PERFORMANCE ON TIMIT

125 speakers – 1000 sentences – use segmentation error statistics with respect to ground truth for 'phonetic' segmentation

Mean error in 2ms frame resolution

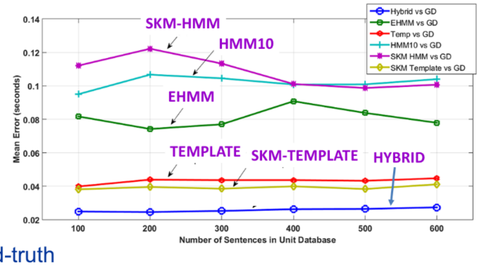


SEEDED	TEMPLATE		HMM
	Random Selection	MKM Selection	
Context Independent	Not optimal (A) for generalization	Clustering handles test variability better than random selection (B)	Poorer than context-dependent case (E)
Context Dependent	(C) Better than context-independent	• Offers significantly **lower mean errors** for **much smaller number of templates / phoneme unit** (D) • Progressively reducing errors **(lower than the best of HMM)** for increase in M	• Templates better than HMM for small no. of templates / unit (F) • HMM errors increase for M>10 due to increased variability

SEEDLESS	TEMPLATE (G)	HMM (H)
Segmental K-means (SKM) Algorithm	• Best performance; due to iterative refinement of template-codebook • Very appealing – being seedless	• Better performance than non-iterative seeded case • Poorer than SKM-Template

SYLLABIC SEGMENTATION OF INDIAN LANGUAGE "TAMIL"

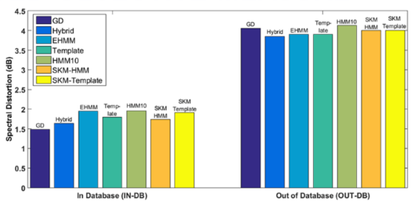
- 4 hours of data
- 6 techniques compared
  - Template (Seeded)
  - SKM-Template (Seedless)
  - HMM10 (10 templates/unit)
  - SKM-HMM (Seedless)
  - EHMM (Festival)
  - Hybrid
- "Group-delay semi-automatic" syllabic segmentation as ground-truth



SKM-Template offers performance close to Hybrid (syllable-conditioned, phone-HMM re-estimation), while being seedless and less complex. Template offers comparable performance (with small seeding). EHMM, HMM10, SKM-HMM poor.

SYNTHESIS AND DOUBLE-ENDED OBJECTIVE MEASURE

- TTS output quality measured using "spectral-distortion" (SD) between IN-DB and OUT-DB reference speech and synthesized speech *(1dB = transparent quality – in speech coding)*
- For all 7 segmentation techniques
- IN-DB performs close to 1.5 dB SD
- OUT-DB performs close to 4 dB SD



- Hybrid the best, All others bunch closely, Template and SKM-Template comparable or better
- Use limited modeling data (seed / seedless) and are less complex than Hybrid and EHMM