Generic Methods for TTS Synthesis

Yasser Hifny

(PhD student, Sheffield university)

y.hifny@dcs.shef.ac.uk

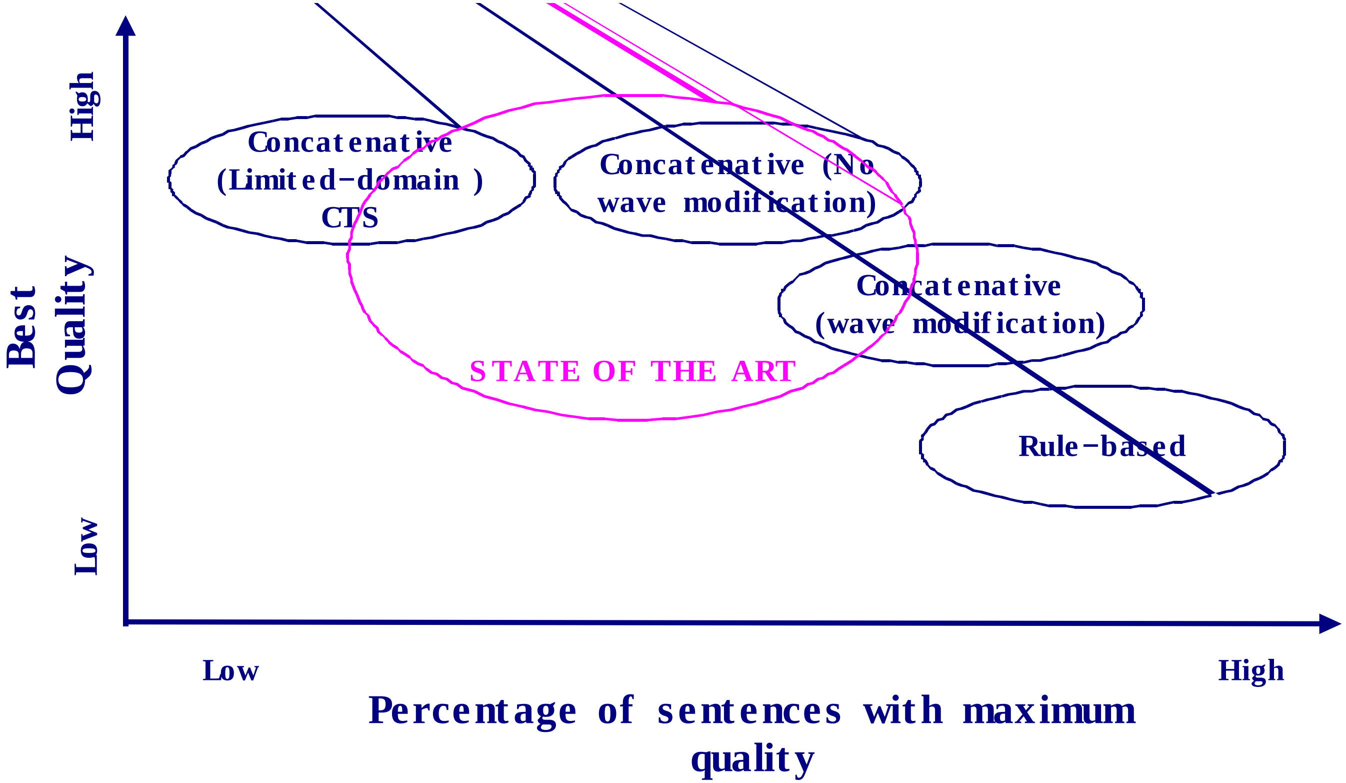
MSc. (2000) , B.Sc. (1995), Elec. & Comm. Engineering, Cairo Univ., Egypt.

Outline

- State of the Art
- Prosody modeling.
- Synthesis by Selection.
- Large Database production.
- Practical considerations.
- Obstacles for high quality.

State of The Art Methods

Huang & Acero, “Spoken language processing”, 2001.



What Does it mean?

Aims of State of The Art TTS

- Concatenative synthesizer.
- Large Database.
- Statistical prosody modeling.
- Runtime unit selection(synthesis by selection).
- Support for Prosodic modification/spectral smoothing methods.

Database is shared between all components, How?

Statistical Prosody Modeling

Corpus based approaches are large database:

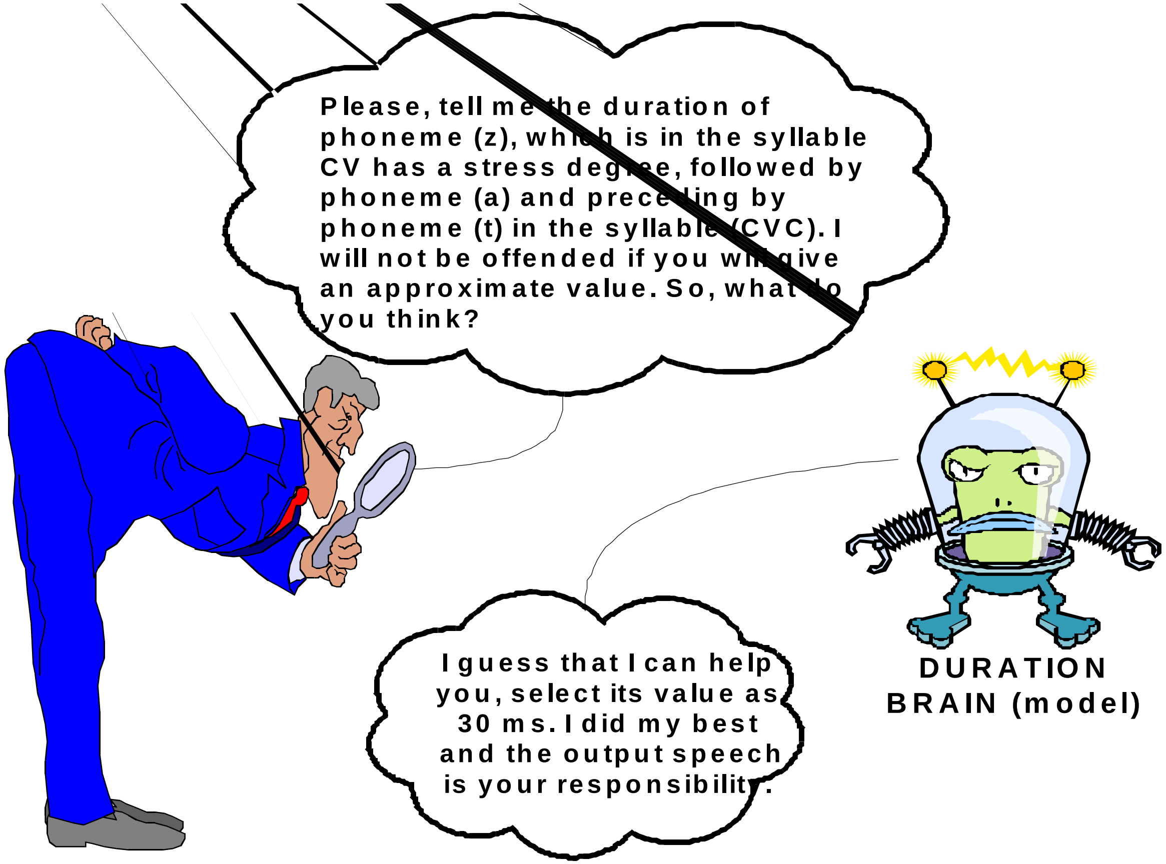
- Prediction of Intonation contour.
- Prediction of Segment Durations.
- Prediction of average energy of Energy.

Phonological description

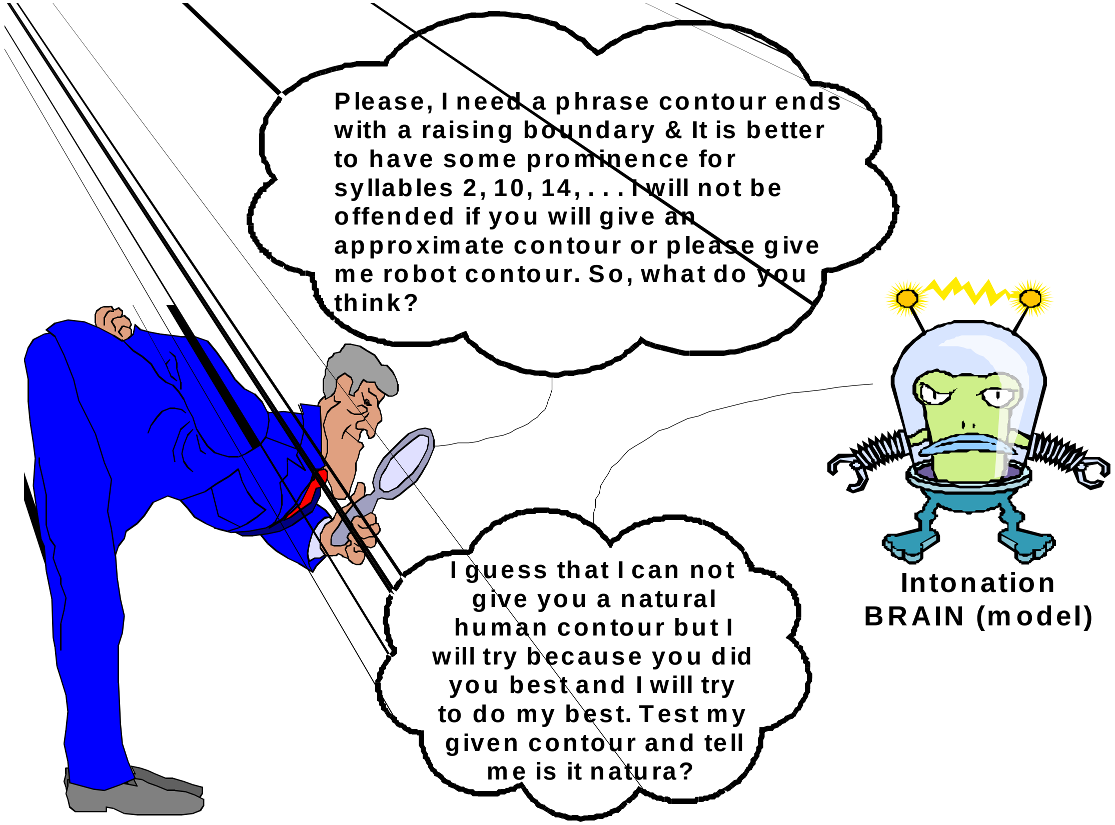
Phonology Level	Feature Description	Feature count	Value range
Phoneme	Sound Type	11	1 to 13
	Voicing Type	11	1 to 5
	Consonant Type	11	1 to 9
	Type Of Articulation	11	1 to 13
	Place Of Articulation	11	1 to 15
	PhonemeID	11	0 to 41
	FuzzyEmpatic	11	0 to 1
	EmphaticType	11	0 to 1
	Shadda	11	0 to 1
	Tanween	11	0 to 1
Syllable	Syllable Position (SP)	1	1 to 4
	Syllable Position (SN)	1	2 to 4
	Accent Degree (AC)	1	0 to 4
Foot	FP Position (FP)	1	1 to 10
	FN Position (FN)	1	1 to 10
Phrase	Phrase Position (PO)	1	0 to 3

Phonological feature description

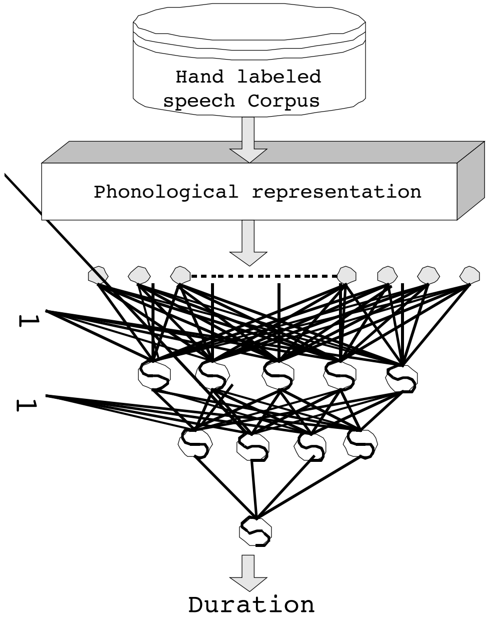
Duration Modeling problem


Please, tell me the duration of phoneme (z), which is in the syllable CV has a stress degree, followed by phoneme (a) and preceding by phoneme (t) in the syllable (CVC). I will not be offended if you will give an approximate value. So, what do you think?
I guess that I can help you, select its value as 30 ms. I did my best and the output speech is your responsibility.
DURATION BRAIN (model)

Intonation Modeling problem

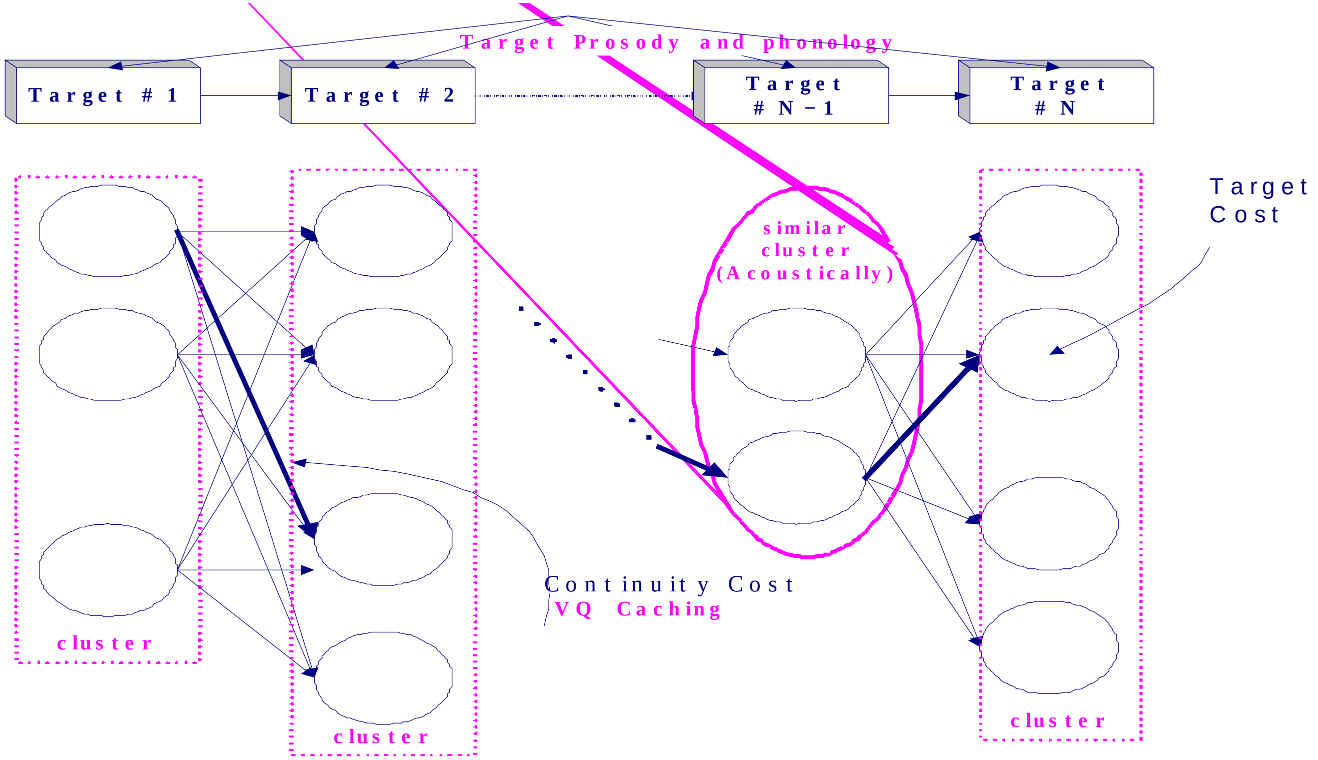

Please, I need a phrase contour ends with a raising boundary & It is better to have some prominence for syllables 2, 10, 14, . . . I will not be offended if you will give an approximate contour or please give me robot contour. So, what do you think?
I guess that I can not give you a natural human contour but I will try because you did you best and I will try to do my best. Test my given contour and tell me is it natura?
Intonation BRAIN (model)

Phonology to acoustical mapping



Neural network transformation model (ArabTalk)

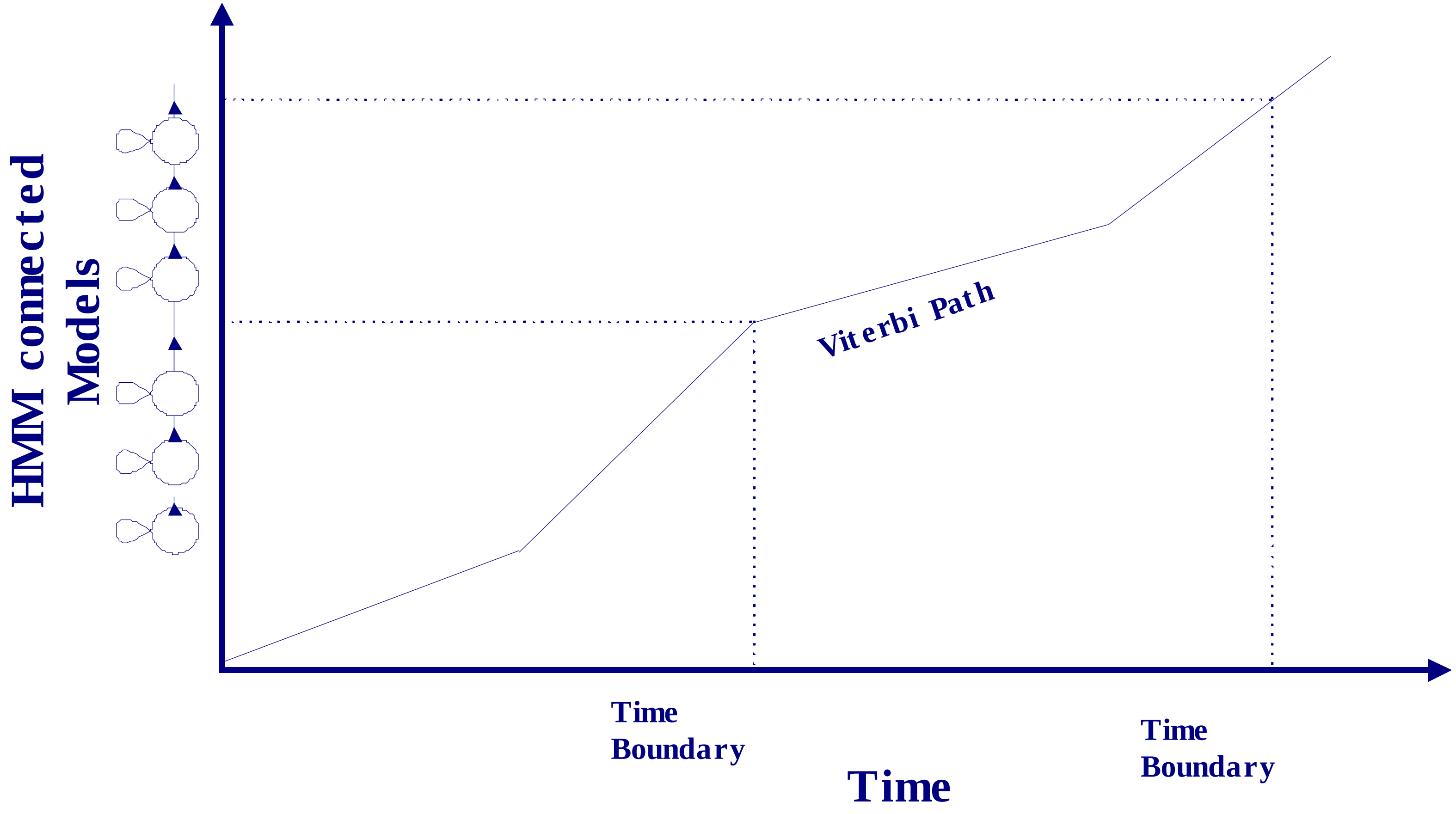
Synthesis by Selection



Corpus generation, is time consuming?

How to Align a large Database?

Forced Alignment



Practical considerations (Acoustic)

- Tri−phones Speaker Independent Models(if available).
- Speaker Dependent Models(SDM).
- Incremental training +SDM.
- Automatic corrections tools (HMM consistent errors).
- Manual Corrections (Why!).
- Male voices seems to work better(ArabTalk).

ArabTalk implementation

Practical considerations (Text)

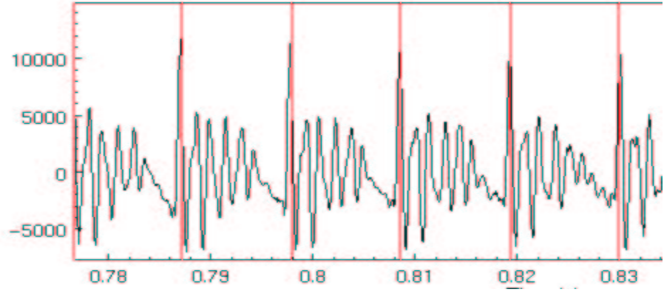
- Domain based Text.
- Closed loop phonetic transcription(assimilation).
- Domain Coverage (How to estimate?)
- Prosodic markers for speech recording(should be considered or not?).

Practical considerations (Pitch processing)

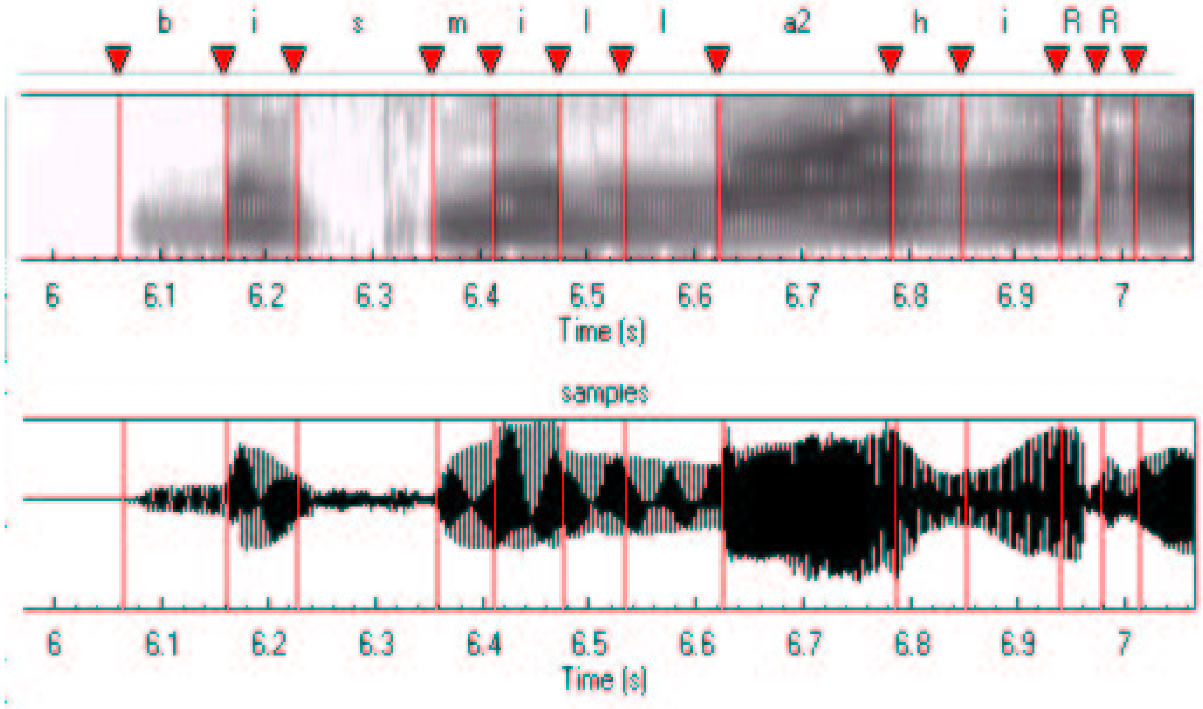
TD−PSOLA is the most efficient & cheapest prosodic modification method:

ArabTalk implementation

- EGG signal recording versus tracking algorithms.
- Pitch synchronous analysis versus fixed frame rate.
- Prediction of the pitch contour from the text.



The Alignment Output



RDI ArabTalk aligned sentence

Obstacles for High Quality

- # of concatenation points(format discontinuity).
- Are longer units can solve the concatenative approach limitations?
- Prosodic modification (affect natural speech).
- Lack of objective evaluations.
- Closed domains versus open domain.

Good luck BITS

Last minute, what else we need?

Patience! It will not work as ALL speech research ☺. BELIEVE ME!

Acknowledgment

- BITS organizers who made my visit possible.
- RDI ArabTalk is an implementation for Arabic Text To Speech system. Thanks to RDI research lab members. The author was the speech department manager and ArabTalk project manager during (2000−2002).