

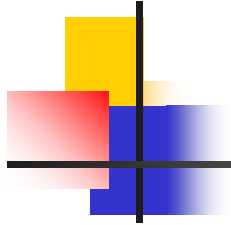


# Refined Speech Segmentation for Concatenative Speech Synthesis

---

Abhinav Sethy

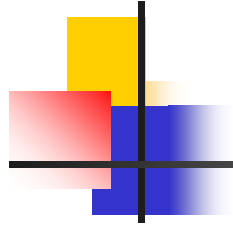
Shrikanth Narayanan



# Motivation

---

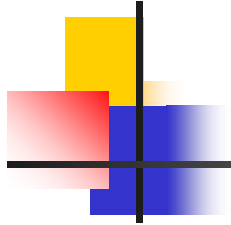
- Speech research is critically dependent on availability of properly labeled speech corpora
- Accurate segmentation of speech is of critical importance to concatenative speech synthesis
- Segmented corpora are utilized to generate components for prosody and concatenation
- Currently the labeling process requires considerable human effort and time



# Unsupervised Segmentation

---

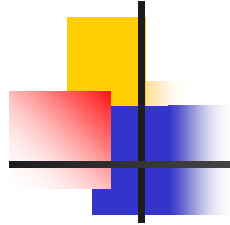
- Goal is to label speech without prior knowledge of its content
- Use ASR as first step to get the text for each utterance
- Map from the text to phonetic transcription using dictionaries
- Because of pronunciation variations in natural speech the phonetic transcription is inaccurate



# Explicit Segmentation

---

- Our research focuses on correct segmentation of speech given a phonetic transcription
- This is referred to as linguistically constrained segmentation or explicit segmentation
- Explicit segmentation is typically carried out by restricting the grammar model for ASR systems



# ASR Based Explicit Segmentation

---

- Segmentation from ASR techniques does not obey phonetic convention which is required by synthesis system
- ASR techniques do not have correct identification of boundaries as an optimality criterion in training
- TTS systems require boundaries to be correct within 3-4 ms which is larger than the time resolution of ASR systems



# Two Stage Segmentation ASR

---

- Use linguistically constrained ASR to place rough boundary marks
- ASR required high frequency resolution and corresponding time resolution is poor
- We used DTW and HMM based ASR systems for the first stage
- The second stage refines time marks



# Two Stage Segmentation Time Mark Refinement

---

- Phoneme boundary properties depend on context
- For every phone pair a different statistical boundary model is required
- GMM and HMM based phone boundary models were used



# Time Mark Refinement

## Statistical modeling

---

- Equal number of speech frames on both side of boundary are used for modeling
- The feature space for the speech data comprised of MFCC's with delta and acceleration coefficients at a frame size of 3 ms
- HMM models are trained on the boundary data



# Time Mark Refinement Search

---

- At equally spaced offsets from the initial mark find out the match with the HMM model
- The offset which gives the maximal match is taken as the refined boundary estimate

Refined mark  $t^{\text{refined}} = \text{index}_{\text{boundary}} * s + t^{\text{initial}}$

$\text{index}_{\text{boundary}} = \text{argmax } P(i)$

$P(i)$  is log probability match at position  $t^{\text{initial}} + i * s$

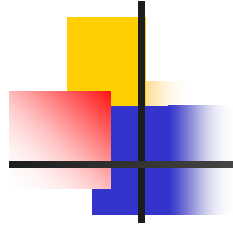
$t^{\text{initial}}$  : initial estimate  $i$  : search index  $s$  : search step



# Time Mark Refinement HMM Modeling

---

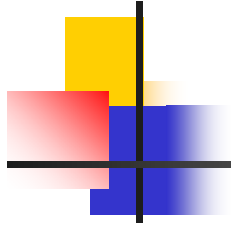
- Decision tree clustering was used to ensure proper training
- The central state was taken as dividing the boundary model into left and right phone contexts
- Phonemes were divided into classes such as Voiced, Fricative, Nasals etc for the tree based clustering process



# Training and Implementation

---

- Implementation was done using HTK
- CDHMM based recognizer was trained on TIMIT to provide initial labeling
- Boundary models were trained on a labeled corpus of 400 utterances from single speaker
- For every boundary model a 70ms speech segment centered at boundary was used for training
- The HMM analysis frame size was 3ms



# Results

---

- Various ASR systems were used to provide initial time marks
- The performance is measured in terms of the percentage of time marks that lie within a certain tolerance from the true boundary
- Tolerance values of 4,8 and 16ms was used



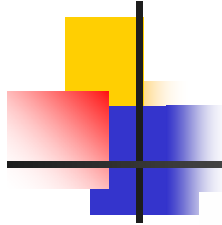
# Results : Time Mark Refinement

Tolerance (ms)	DTW	CDHMM	ENTROPIC	Adapted CDHMM
4	11	14	18	19
8	30	35	41	44
16	76	58	79	83

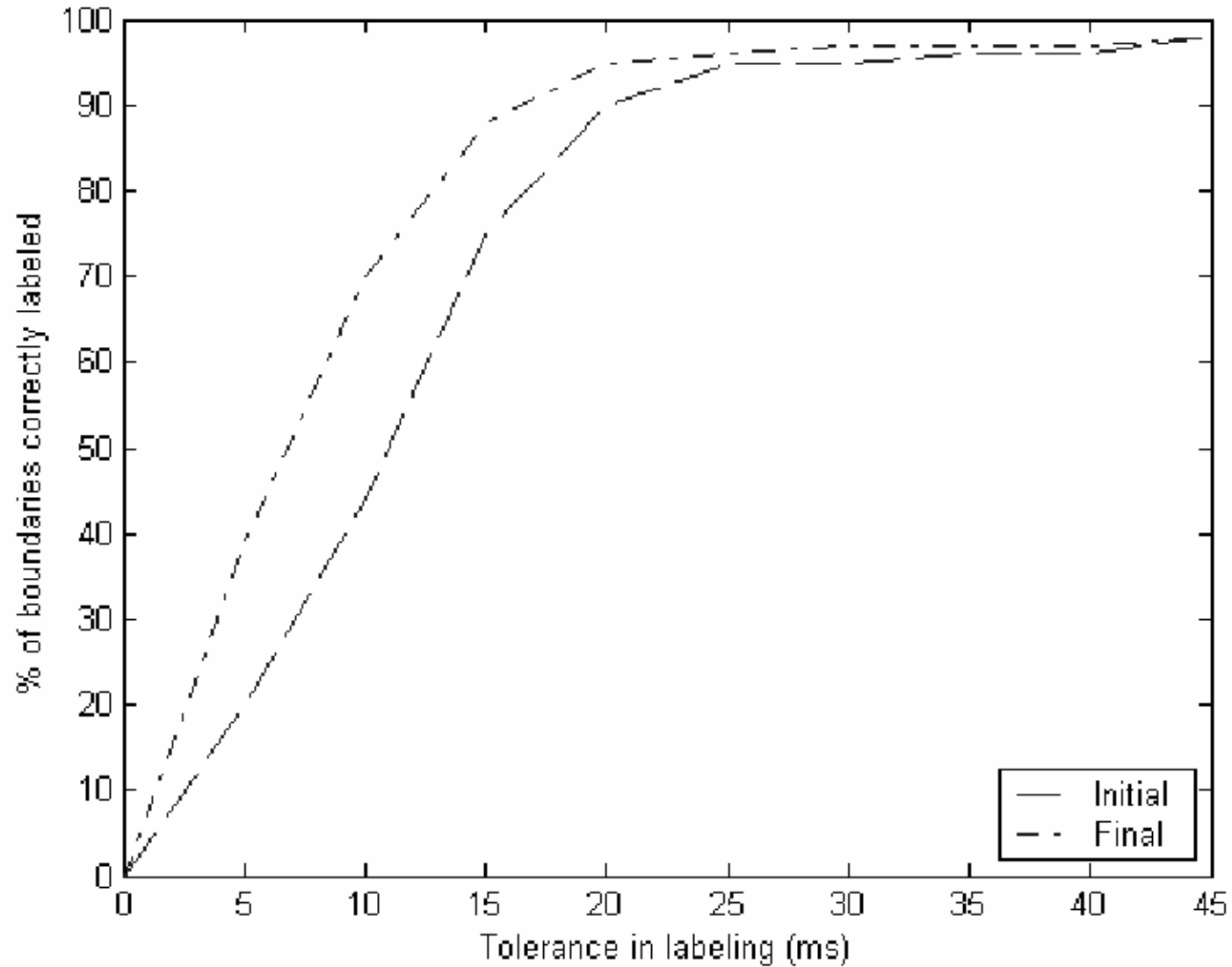
Table 1: Before Time Mark Refinement

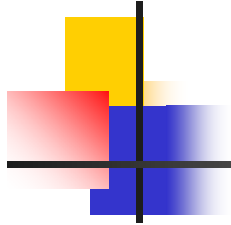
Tolerance (ms)	DTW	CDHMM	ENTROPIC	Adapted CDHMM
4	31	36	38	39
8	54	61	65	67
16	86	87	89	93

Table 2: Post Time Mark Refinement



# Results : Cumulative Distribution

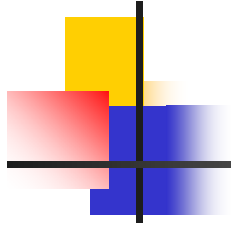




## Results: Discussion

---

- The time mark refinement process provides significant improvement in labeling accuracy
- The accuracy after refinement does not depend critically on the initial estimate
- Many of the remaining large errors after refinement can be attributed to
  - Incorrect transcription or the actual boundary
  - The correct boundary being outside search range



## Conclusion

---

- Two stage approach for speech segmentation
- HMM based boundary modeling
- Reduction in manual labeling effort for speech
- Rapid generation of new voices for TTS
- No restrictions were imposed on language grammar