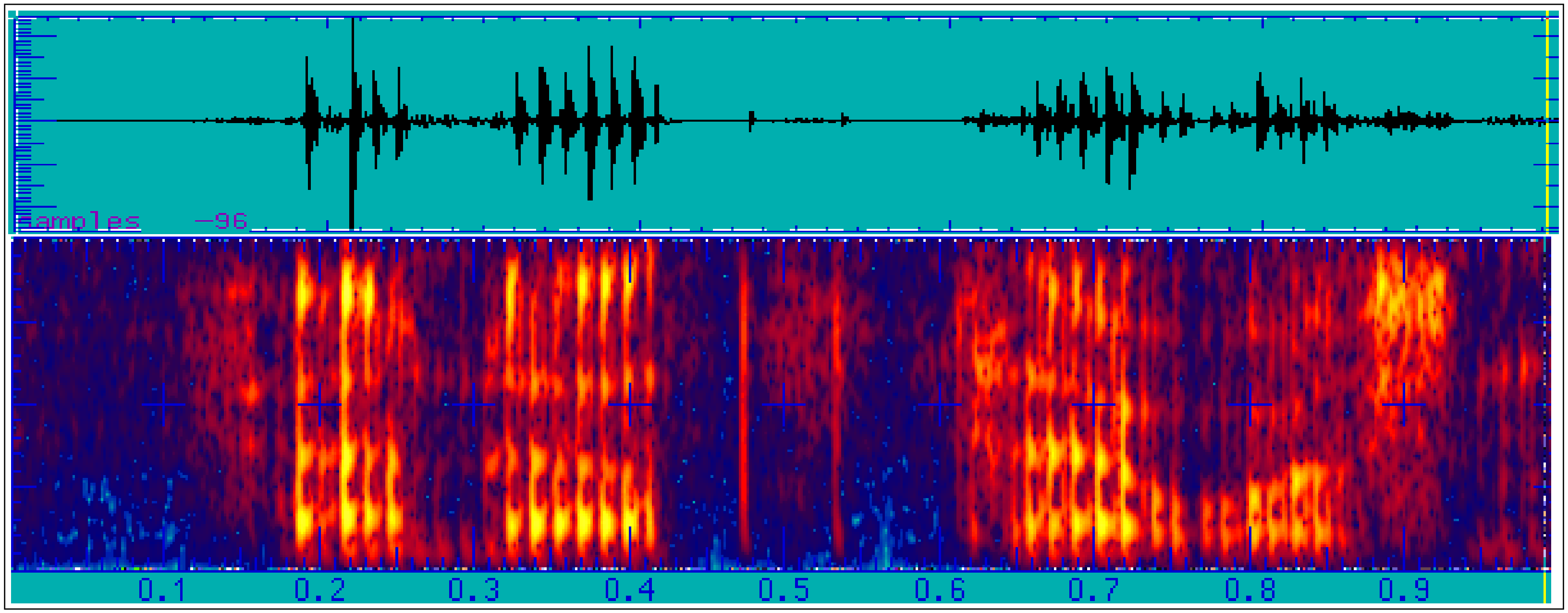
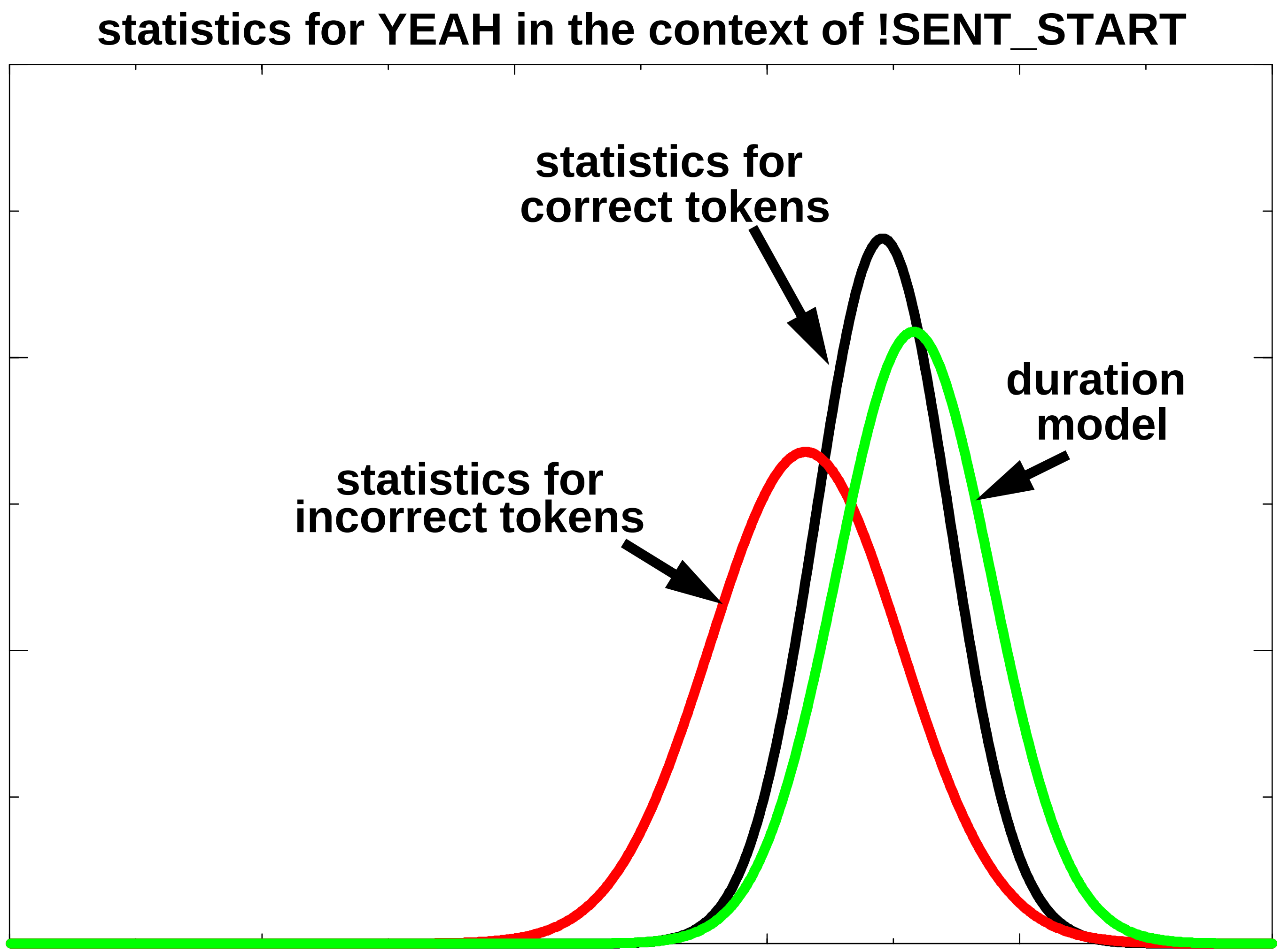
Motivation



Ref:	**found**	**out**	**that**	**that**	**wasn’t**	
Base:	**and**	**uh**		**that**	**was**	**an**
Dur:	**found**	**out**		**that**	**was**	**an**

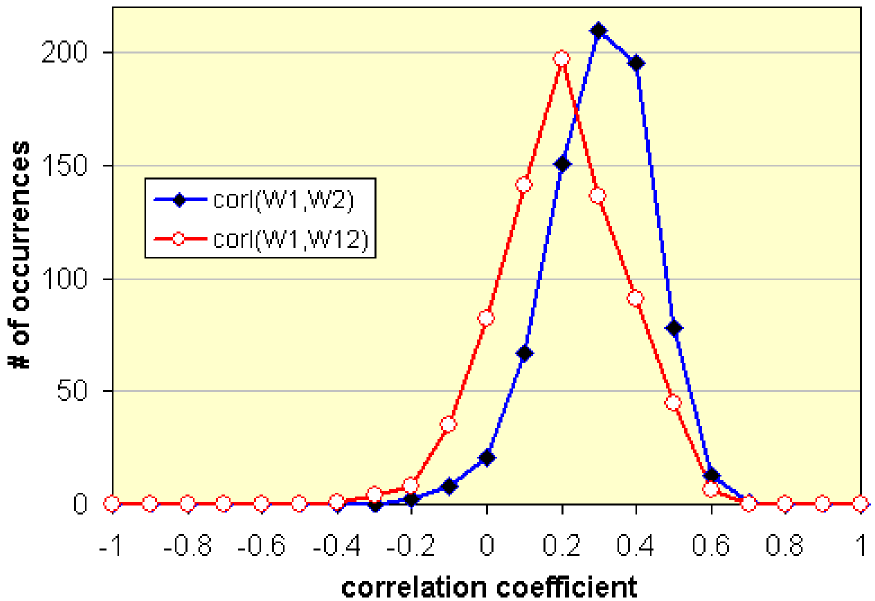
- **humans follow an internal sense of timing**
- **duration is one of the most reliable and accessible prosodic features**

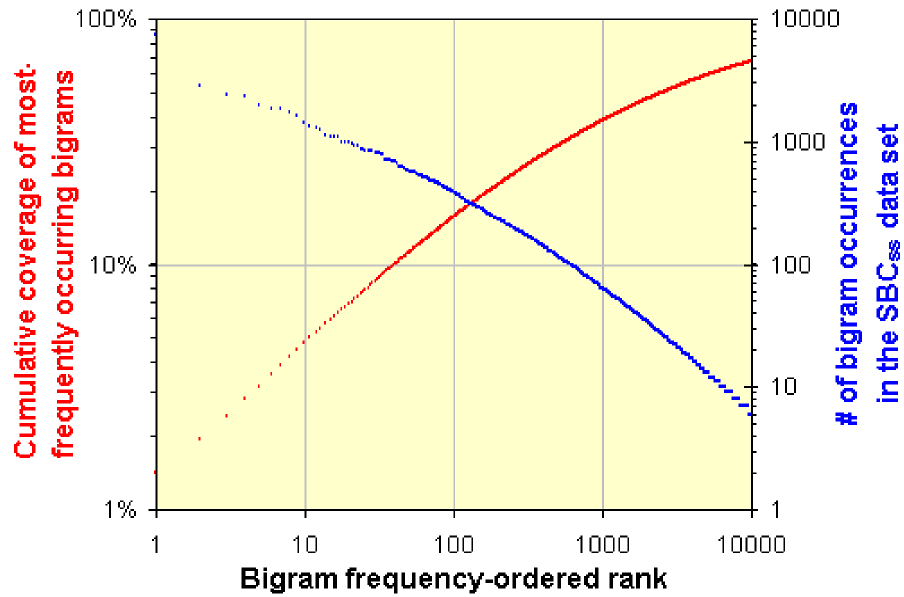
Implicit Duration Models Insufficient

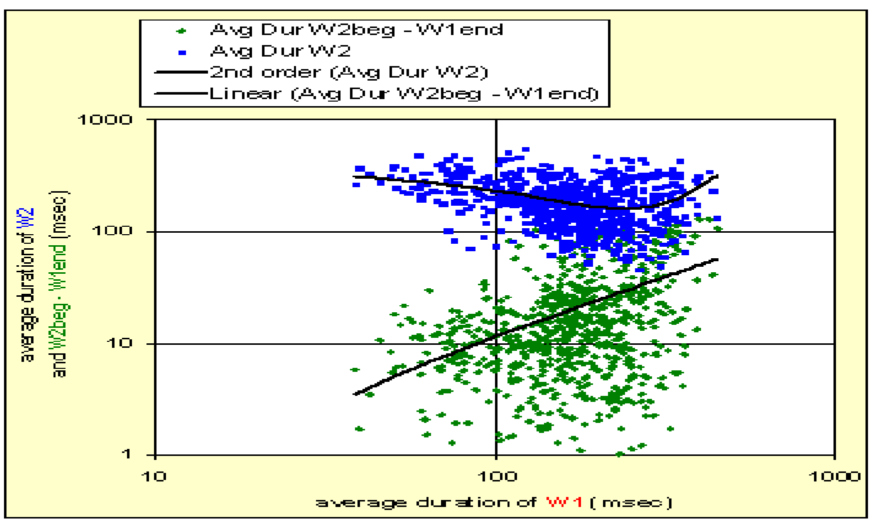


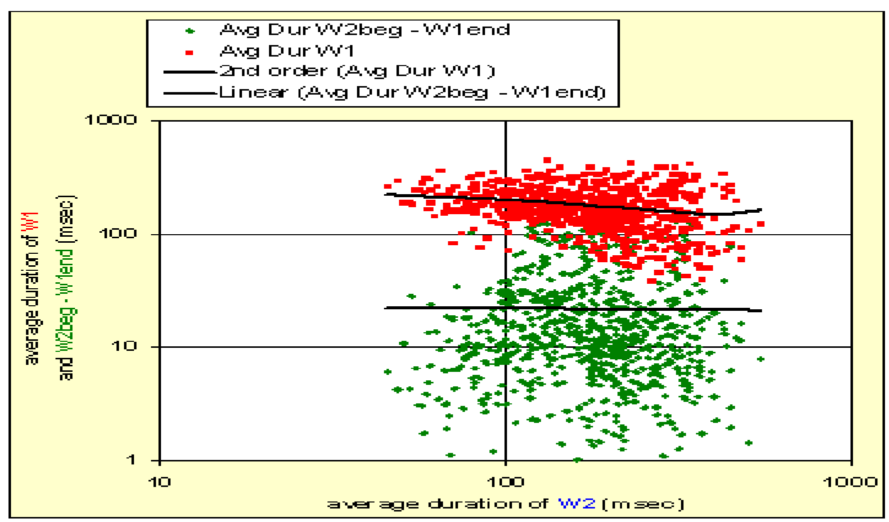
- ☛ recognition errors (SWB) deviate from true distribution
- ☛ word durations preferred over phone durations

Switchboard Data









Suprasegmental Information

- **word duration represented as a single scalar attribute**
- **word duration bigram model ($F \equiv \{w, \tau\}$):**

$$Pr(F_i \mid F_{i-1}) = Pr(w_i, \tau_i \mid w_{i-1}, \tau_{i-1})$$
$$= Pr(\tau_i \mid w_i, w_{i-1}, \tau_{i-1})\, Pr(w_i \mid w_{i-1}, \tau_{i-1})$$

where w is the word identity and τ is the duration

- **can be implemented in a rescoring paradigm as an additional knowledge source applied to word hypotheses (leads to a feasible implementation)**

Bigram Duration Model

- **Duration augmented bigram probability:**

$$P(w_i \mid w_{i-1}, \tau_{i-1}, \tau_i) = P(w_{i-1}, \tau_{i-1}, w_i, \tau_i) \;/\; P(w_{i-1}, \tau_{i-1}, \tau_i)$$

$$= \frac{P(\tau_{i-1}, \tau_i \mid w_{i-1}, w_i)}{P(\tau_{i-1}, \tau_i \mid w_i)} \frac{P(w_{i-1}, w_i)}{P(w_{i-1})}$$

- **Begin/end of sentences treated as special cases:**

$$P(w_1 \mid S_{beg}, \tau_1) = \frac{P(\tau_1 \mid S_{beg}, w_1)}{P(\tau_1 \mid S_{beg})} \frac{P(w_1)}{P(S_{beg})}$$

$$P(S_{end} \mid w_{i-1}, \tau_{i-1}) = \frac{P(\tau_{i-1} \mid w_{i-1}, S_{end})}{P(\tau_{i-1} \mid w_{i-1})} \frac{P(w_{i-1}, S_{end})}{P(w_{i-1})}$$

Back-Off Weighting

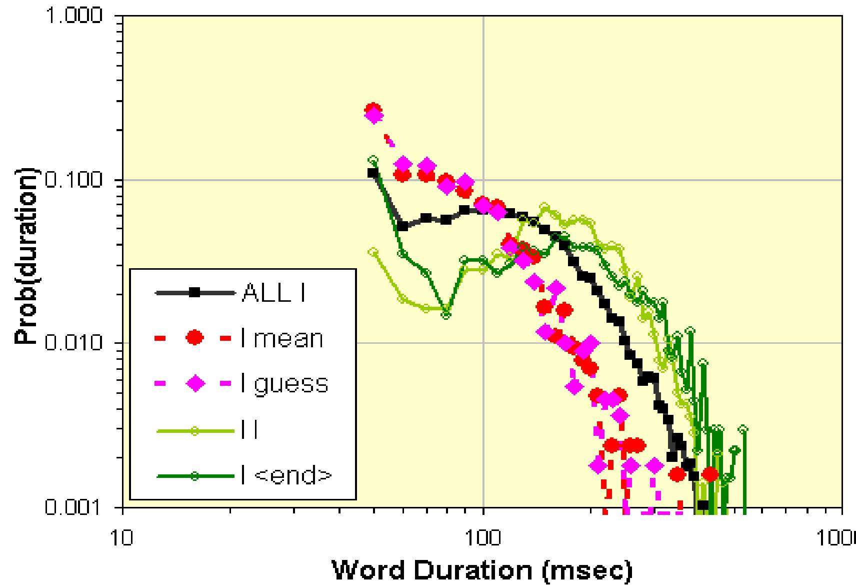
- many duration bigrams have insufficient training data
- combine bigram-specific models with word-specific and word-independent models in a back-off framework

$$P_{sm}(\tau_{i-1},\tau_i \mid w_{i-1},w_i) \quad = \quad \frac{\Omega_b P(\tau_{i-1},\tau_i \mid w_{i-1},w_i) + \Omega_w P(\tau_{i-1} \mid w_{i-1})\, P(\tau_i \mid w_i) + \Omega_g P^2(\tau_i)}{\Omega_b + \Omega_w + \Omega_g}$$

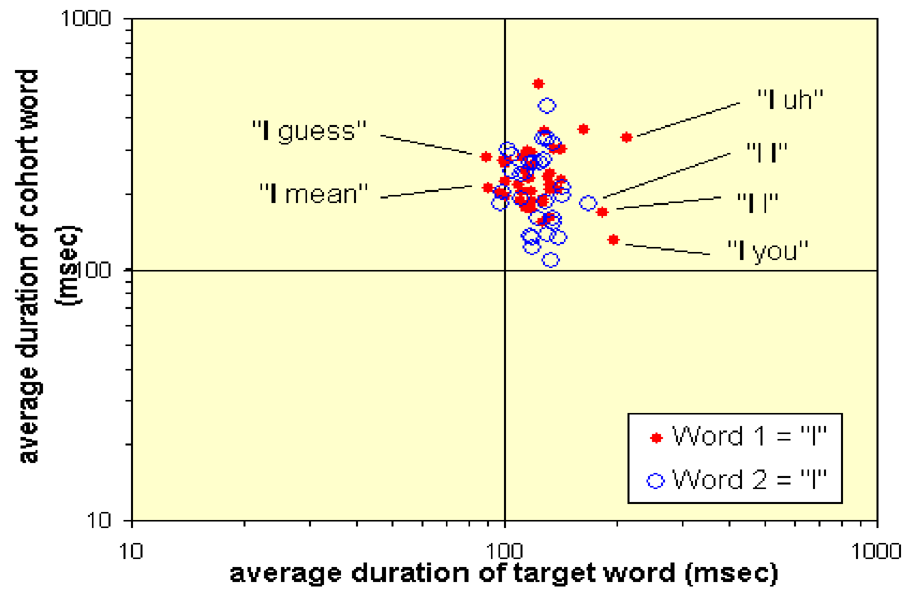
- Ω empirically chosen in initial experiments (can be estimated using deleted interpolation or other such smoothing algorithms)

Duration Analysis-1

- duration distributions for the word “I” in bigram contexts



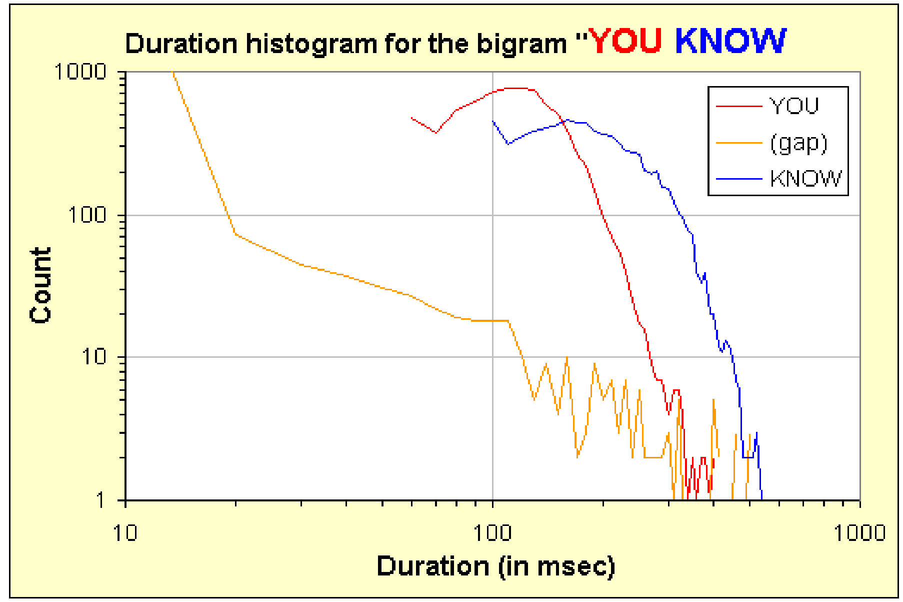
- average duration statistics for the 750 most frequently occurring word bigrams in SWB that include the word “I”

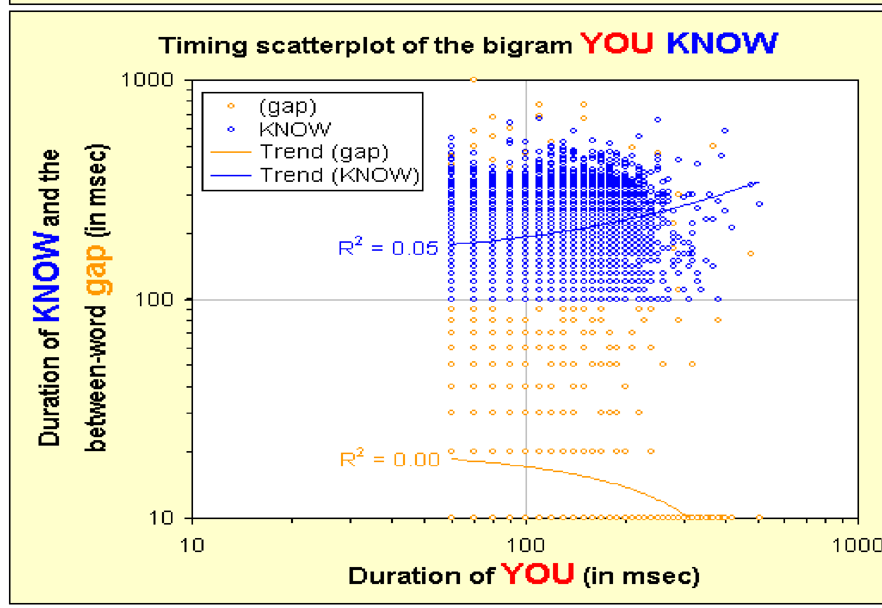


Duration Analysis-2

☛ most frequently occurring bigrams exhibit predictable suprasegmental characteristics

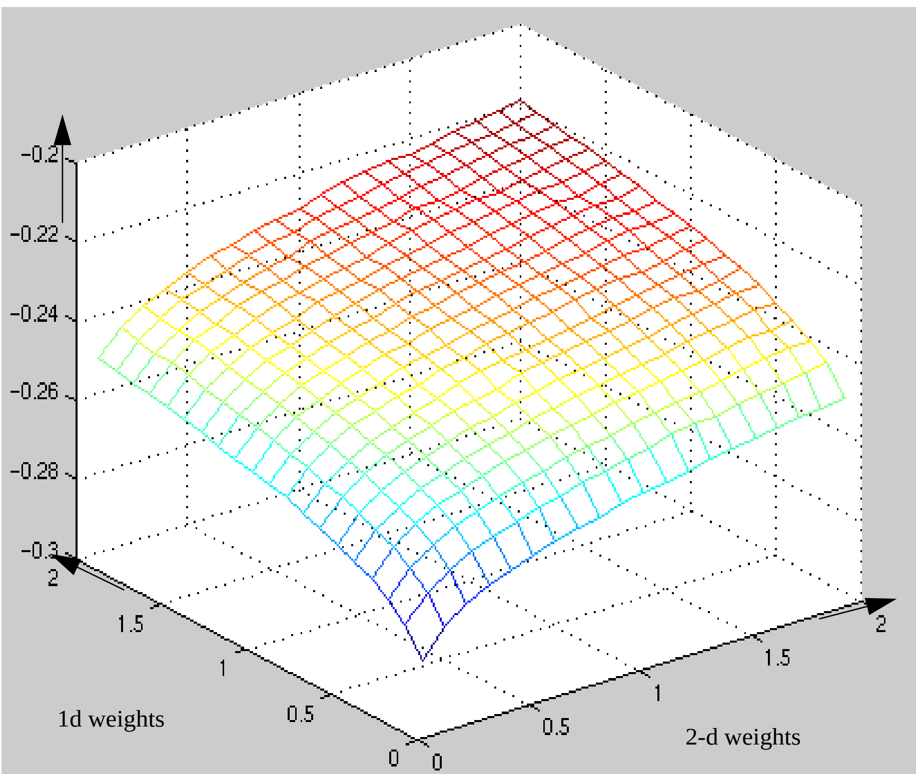
☛ duration predictable and lower variance expected

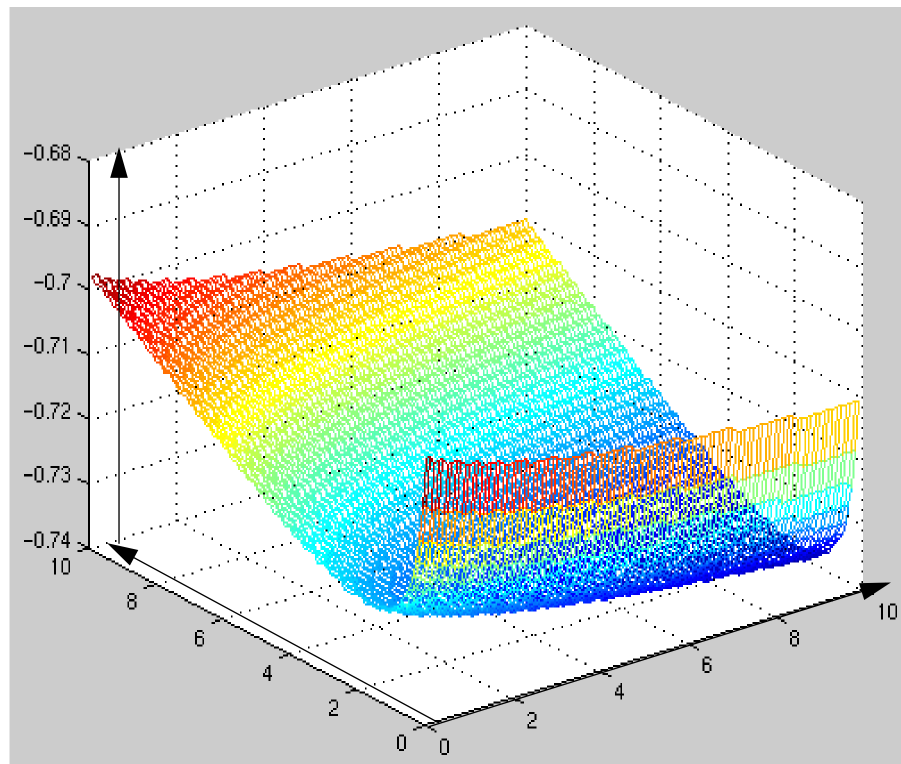




Error Analysis

- difference between the average duration model score for correct versus incorrect bigrams is crucial to performance (analogous to F-ratio)





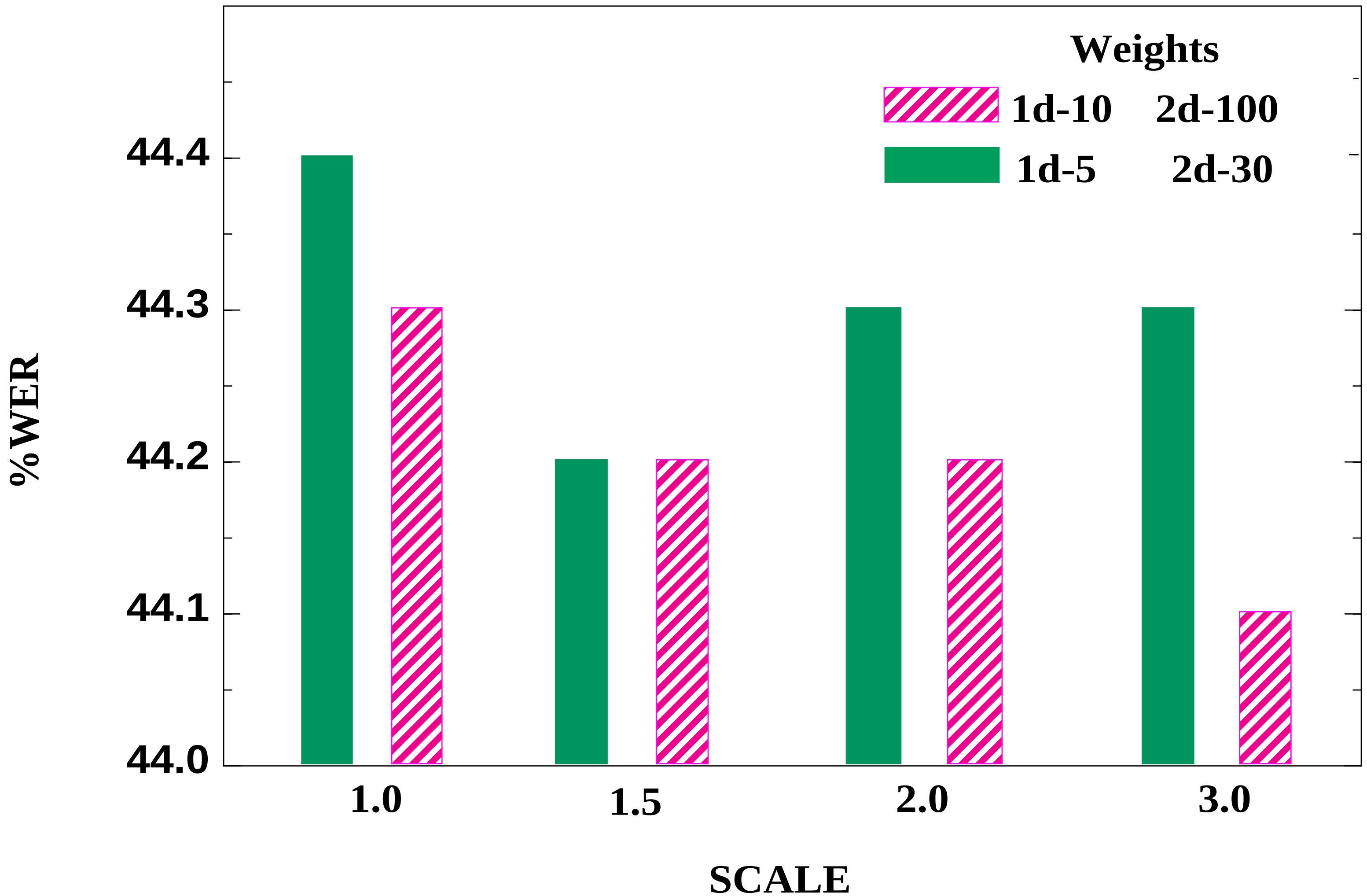
N-best Rescoring Results

- Baseline: 32.4% WER on 637 SWB utterances
- Rescoring of 100-best hypotheses (provided by BBN)
- Oracle WER: 21.2%

	[weight 1d, weight 2d]		
scale	[0.1, 0.1]	[0.1, 0.5]	[0.5, 0.1]
0.01	32.5	32.4	32.3
0.05	32.4	32.3	32.2
0.1	32.3	32.3	32.2

Word Graph Rescoring Results

☛ **Baseline system: WER 44.4% on WS97 test set**



Summary

- A consistent statistical modeling framework that exploits word duration models
- Modest improvement on SWB:
  - BBN 100-Best Lists: 0.2% WER absolute
  - ISIP Word Graph Rescoring: 0.3% WER absolute
- Future work:
  - Incorporate duration models into the grammar decoding loop
  - Better models of infrequently occurring bigrams: error analysis indicates greater potential benefits
  - Develop more sophisticated statistical models