



Subword Lexical Modelling for Speech Recognition

Raymond Lau

**Stephanie Seneff,
Eric Grimson and Victor Zue**

**Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts
United States of America**



Outline

- **Background**
- **ANGIE**
- **Phonetic Recognition**
- **Word-spotting**
- **Continuous Speech Recognition**
- **Pilot Studies: New Words and Duration Modelling**
- **Summary**



Background

- **Subword lexical (pronunciation) modelling**
 - Models how words are realized as sequences of phone-like units
 - Needs to account for variability and phonological processes
- **Typical approaches**
 - Explicitly via pronunciation graph (e.g., MIT SUMMIT, LIMSI)
 - Implicitly modelled (e.g., HMM parameters, context-dependent acoustic models)



Is the World Flat?

- **Both typical approaches have same underlying philosophical design:**
 - **Words are modelled as a flat sequence of phone-like units**
- **Much work in linguistics suggests that there is a phonologically significant unit that is**
 - **larger than a phone,**
 - **but smaller than a word**
- **Suggests that a multi-layered hierarchical model may be appropriate**



Historical Context

- **Kahn (1976): Theory**
 - Syllable important in phonology, syllabification rules
- **Church (1983): Computational model**
 - Context-free parser
 - Linguist's phonetic transcriptions => syllables
- **Our goal: Computational model for recognition**
 - Hierarchical, with support for syllables
 - Handle variable and errorful inputs with a probabilistic framework
 - Support various speech recognition tasks
 - Has potential advantages over existing approaches

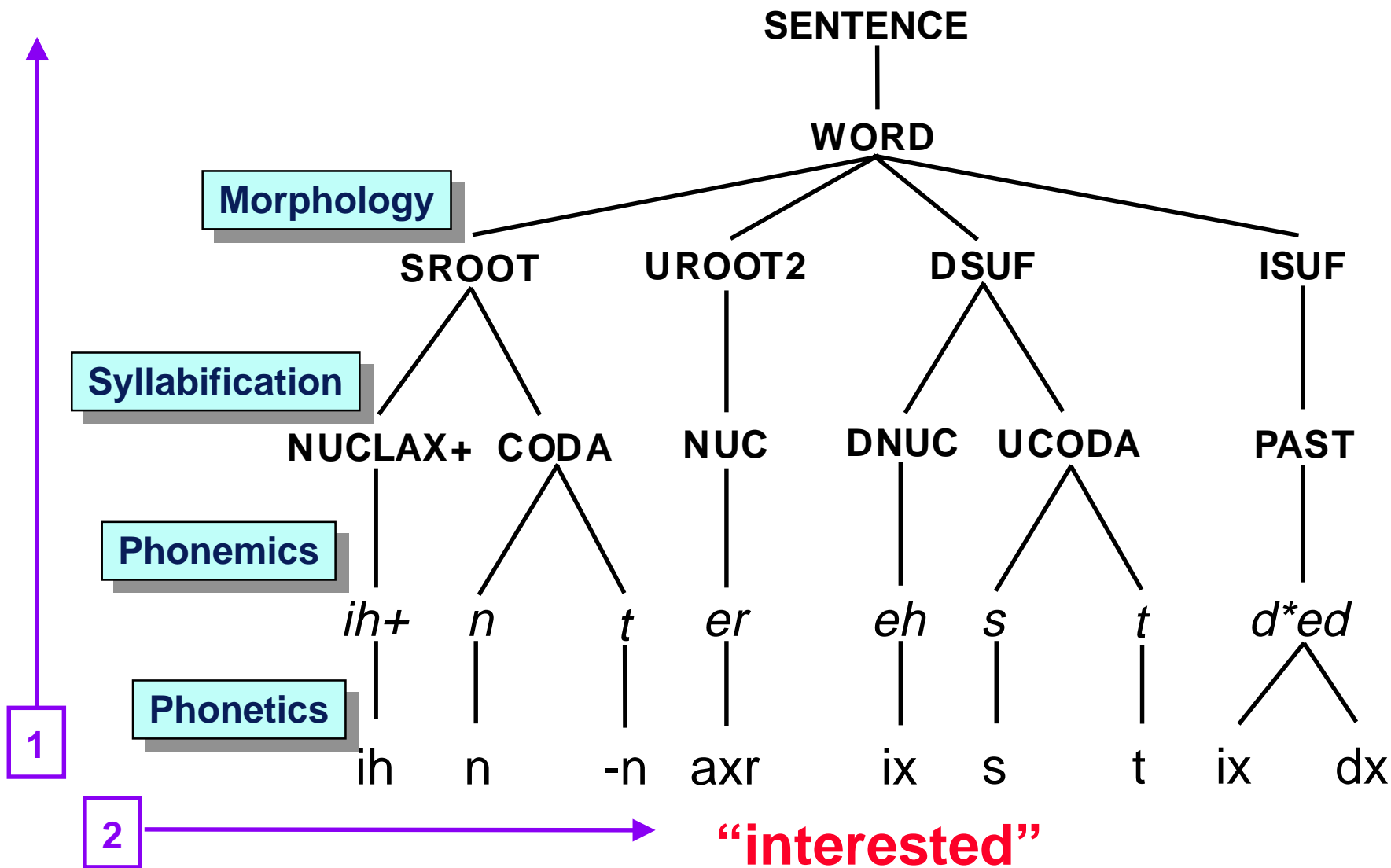


ANGIE Framework

- **Flexible, multipurpose system for speech processing**
- **Framework introduced in Seneff, Lau & Meng (ICSLP '96)**
- **Word substructures characterized jointly by:**
 - **Context-free grammar**
 - **Probability model, automatically trained**
- **Layered, hierarchical representation with bottom-up sharing of substructures**
 - **Pool training data, supports new words**
- **Unified framework for speech processing**
 - **Models phonological rules**
 - **Integrates with higher level linguistics**
 - **Supports recognition, but also sound-to-letter/letter-to-sound generation (Meng, 1995)**



Example Parse Tree





Details of Framework

- **Probabilities conditioned on:**
 - Entire prior column (advancement)
 - Child & left sibling (bottom-up trigram)
- **Lower two layers capture phonology, upper layers capture syllable structure**
- **64 phones**
- **104 phonemes**
 - Vowels marked for stress
 - Consonants marked for onset
 - Several special phonemes, e.g.,
ux_you, ing
- **78 other non-terminals**



Experimental Framework

- **Subset of ATIS**
 - 5000 utt subset for acoustic training
 - 10,000 phonetic transcriptions seeded from SUMMIT for linguistic training
 - Development & December '93 test sets
- **Focus on lexical and not acoustic modelling**
- **Recognizer based on SUMMIT front-end for acoustics and segmentation**
- **Context-independent phones**
- **Mixture diagonal Gaussians**
- **14 MFCCs averaged across thirds of segments, delta MFCCs across boundaries, and absolute duration**



Language Model Perplexities

Set	Phone Bigram	Phone Trigram	ANGIE
Train	9.90	4.98	6.07
Test	14.91	9.20	7.15

- Evaluated perplexity of ANGIE model as compared to phone n -grams
- ANGIE compares favorably, especially on unseen test data

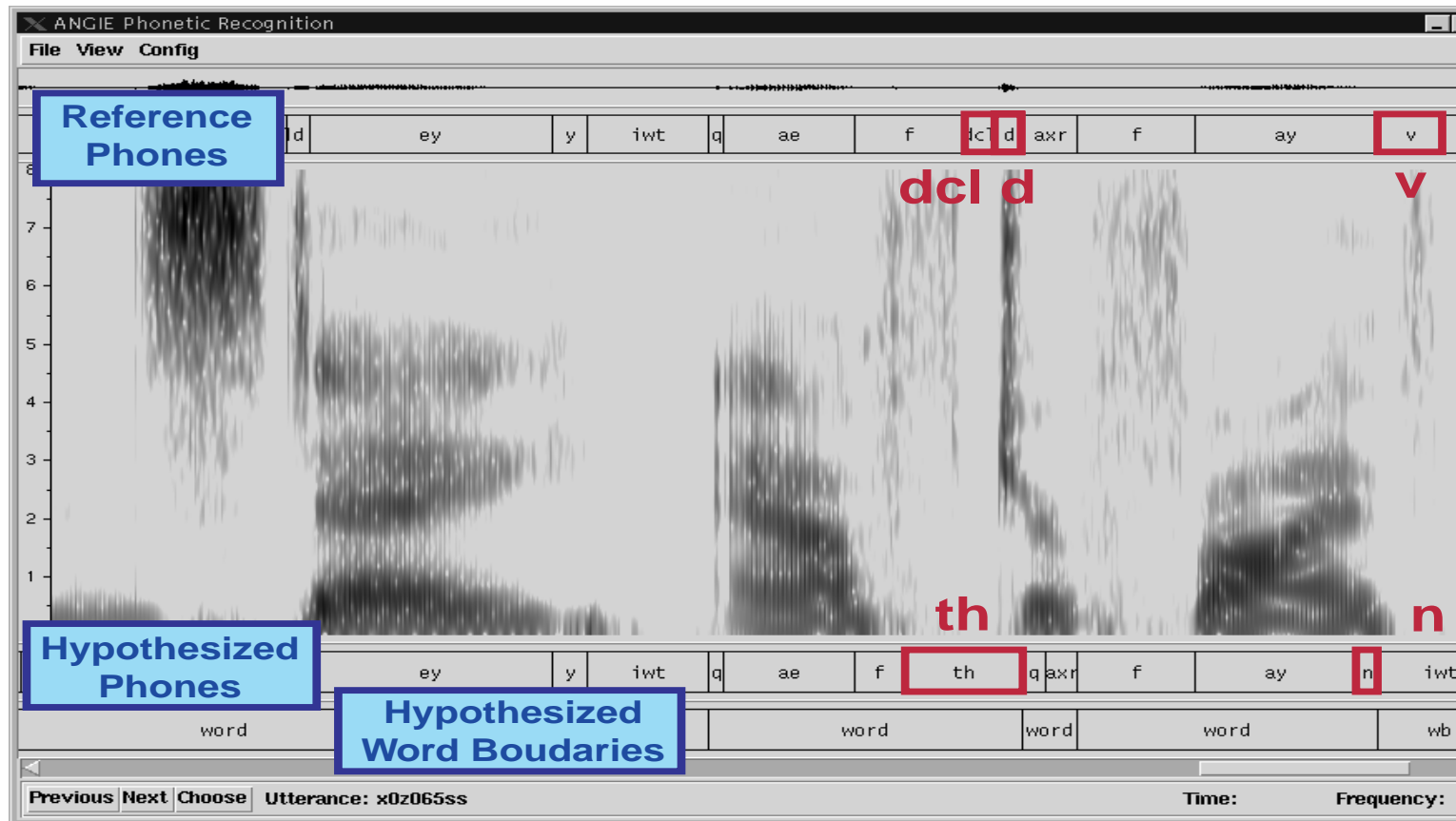


Task 1: Phonetic Recognition

- **Objectives**
 - **First test of ANGIE involving acoustic data**
 - **Establish feasibility of ANGIE**
- **Primary challenge is search organization**
 - **Need to explore phone graph, incorporating scores from both ANGIE and acoustic models**
 - **Cannot use Viterbi because ANGIE's parse tree provides long distance constraints**
 - **Best-first control strategy adopted**
 - **Scoring heuristics to balance short vs. long theories**
 - **Pruning on queue size and reference counting garbage collection to limit resource use**



Sample Phonetic Recognition Output



“[Wed]nesday after five” hypothesized as
“[Wed]nesday **afth er fine**”



Phonetic Recognition Results

Acoustic Models	Subword Lexical Model	Error Rate
SUMMIT	Phone Bigram	39.8%
ANGIE	Phone Bigram	37.7%
ANGIE	ANGIE	36.1%

- **CMU 39 phone set, forced alignment is reference**
- **Total reduction: 9.3%**
- **2.1% difference due to purer acoustic models made possible by more refined phonological modelling**
- **1.6% difference due to stronger language model constraints present in ANGIE's upper layers (e.g., syllable structure)**



Task 2: Word-spotting

- **Task: Word-spot 39 city names in ATIS**
 - Similar task to Manos and Zue (ICASSP '97) but focus is on lexical and not acoustic modelling
- **Objectives:**
 - Further establish empirically the feasibility of using **ANGIE** for speech recognition tasks
 - Explore effects of varying subword lexical model
 - * Easy to do within the **ANGIE** framework
 - Use as a natural foundation for building a full **ANGIE**-based continuous speech recognizer



Search Strategy

- **Previous best-first strategy**
 - Proved inadequate empirically for word-spotting
 - Possible reason: difficulties in normalizing short vs. long theories for comparison
- **New strategy: Variant of *stack decoder***
 - c.f., Jelinek (IEEE '76), Paul (ICASSP '91)
 - Extend all paths at the earliest unexplored time boundary based on score
 - Prune based on a maximum number of paths permitted at any boundary
 - At any time, only theories covering same acoustic space are compared



Filler Models

- **ANGIE provides subword lexical model for the filler space**
- **Different ANGIE configurations give us a range of models (next slide)**
- **In all cases, no cross-word constraints (e.g., word n -gram) used**



Range of Filler Models

- **Phones**
 - Only phone bigram used
- **Pseudo-words (e.g., **afth**: ae f th)**
 - “Invent” possible pseudo-words bottom-up
- **Syllables (e.g., **ciscofran**: s ih s kcl k uh f axr n)**
 - Syllable is highest unit
 - Syllable ordering not enforced
- **Morphs (e.g., **conflighting**: kcl k aa n f l ay tcl t iy ng)**
 - Syllables with ordering enforced
- **Known words plus pseudo-words**
 - 1200 words plus allow “invention” of pseudo-words
- **Known words only**
 - 1200 words



Word-spotting Results

Filler Model	Figure of Merit	Rel. Time
<i>Phone Bigram</i>	85.3	-
Pseudo-words	86.3	1.00
Syllables	87.7	0.56
Morphs	88.4	0.79
Words + Pseudo-words	88.6	0.79
Words	89.3	0.74

- More constraint leads to higher FOM
- Speed increases with constraints
 - Possible explanation: lower branchout
 - Exception: **Syllables very fast**
- Word bigram continuous recognition: 93.9 FOM



Word-spotting Discussion

- **Permitting pseudo-words in addition to known words did not help**
 - * **Test set had only 0.82% OOV**
 - * **But even if vocabulary lowered to 400 words, which is 8% OOV, pseudo-words do not help**
- **8.6 FOM drop from full continuous recognition to phone bigrams approximates Manos' 8 FOM drop -- suggests word-spotter competitive**



Task 3: Continuous Speech Recognition

- **Task**
 - Full word recognition of continuous speech data
- **Objectives**
 - Most challenging proof of feasibility test for **ANGIE**
 - Explore syllable units for recognition
 - Integration with higher level natural language processing
 - * Many NL systems (e.g., TINA) are based on context-free parsing, like the **ANGIE** framework is
 - * Our search strategy should permit coupling **ANGIE** and **TINA** in a tight manner



CSR Implementation

- **Search control remains stack decoder**
 - Some changes -- details in written work
- **Fast match pruning of phone graph using SUMMIT**
 - Phone graph built from top 100 theories
 - Speeds up research experimentation
 - Similar to word graph approach in LM
- **Stack decoder consults word bigram for word level statistics whenever ANGIE proposes a putative word ending**



CSR Results

System	Error Rate
Baseline SUMMIT	18.9%
ANGIE Syllables	26.6%
ANGIE Words	18.8%

- **ANGIE Words system competitive with baseline**
- **Syllables-based system not competitive, despite success in word-spotting**
 - **Constraints across syllables important?**
 - **More work needed -- we did not pursue**



TINA Integration

- **Interface between recognition and NL: Prior approaches**
 - *N*-best lists -- small improvements to word accuracy, e.g., Moore et al., SLSTW 94
 - Word graph -- more substantial improvement, e.g., Seneff et al., Eurospeech 95
- **With these approaches, entire sentence is recognized before NL constraints are employed**
 - Recognizer may waste time searching untenable theories (from NL point of view)
 - This may result in more promising theories being pruned
- **Integrating the NL constraints during recognition allows recognizer to focus on promising theories**



ANGIE-plus-TINA Implementation

- Stack decoder consults ANGIE parser to score **partial word** hypotheses
- At word ends, decoder consults TINA to score **partial sentence** hypotheses
- Integration of scores at end of each word via linear interpolation with heuristic lambdas
- Major problem was computation to support robust parse in grammar
 - Implemented new robust parse mechanism
 - Only triggers upon a parse failure
 - Back up and greedily find the first spot where a full parse can end, and fragment at that point
 - Heuristic penalty for fragmentation



ANGIE-plus-TINA Results

System	Error Rate
<i>SUMMIT & Word 2g</i>	18.9%
ANGIE & Word 2g	18.8%
SUMMIT, Word 2g, & TINA 100-best Resorting	18.2%
ANGIE & TINA Integrated	14.8%

- **Integrated system** much better than *baseline*
- *N*-best resorting only slightly better than *baseline*
- Suggests that having NL constraints during recognition is helpful
- Our search strategy allows easy integration of parser driven models



Pilot Study 1: Adding New Words

- **Want to be able to add new words to vocabulary without extensive lexical retraining**
- **ANGIE can share subword structure between existing in-vocabulary words and new words**
- **Scenario**
 - **Conversational system retrieves list of items from database**
 - **Would like to add these names to vocabulary in “real time”**
 - **Assume we know their category (e.g., city name) and pronunciation (get from dictionary, or even from ANGIE)**
- **Will consider city names introduced in ATIS3 as the “new” words**



New Words Results (Error Rates)

Vocabulary	SUMMIT &Bigram	ANGIE &Bigram	ANGIE &TINA
Reduced (no ATIS3)	34.2%	31.2%	32.8%
Augmented (+ATIS3)	19.2%	19.2%	15.2%
Fully Trained	18.9%	18.8%	14.8%

- SUMMIT gets zero lexical weights, ANGIE shares probs
- TINA & word bigram both had a category for city names
- **ANGIE&TINA does well with new words added**
- Loss from lack of lexical training small (~0.3% difference). ANGIE's sharing comparable to SUMMIT's zero weights.
- **In reduced system, ANGIE does better than SUMMIT in face of many OOV words**



Pilot Study 2: Duration Modelling

- **Leverage ANGIE's subword structural information for prosodic modelling**
- **Chung's hierarchical duration model (MIT S.M. thesis '97, EuroSpeech '97)**
 - **Model accounts for durational relationships of sublexical units at various levels**
 - **Works with ANGIE parse tree**
 - **Shown to reduce phonetic recognition error rates from 29.7% to 27.4%**
- **Joint project: incorporated Chung's duration model into word-spotter**
- **Note: Our acoustic models already have simple duration. Results are for improvements beyond that.**



Duration Modelling Results

- **First test: Disambiguating “Newark” and “New York,” a problematic pair of keywords**
 - Confusion rate reduced from 19% to 6%
- **Second test: Include hierarchical duration score in word-spotter for keywords**
 - With Words filler models, **FOM increases from 89.3 to 91.6**
- **Conclusions on hierarchical duration model**
 - Effective for keywords, successfully leveraging ANGIE’s parse information



Summary

- **ANGIE is a new framework for subword modelling**
 - Hierarchical & layered -- accounts for syllables
 - Probabilistic & based on context-free rules
- **Demonstrated competitiveness in phonetic recognition, word-spotting and continuous recognition**
- **Search is major engineering issue**
 - Stack decoding strategy proved most effective
- **Some promising results beyond feasibility**
 - Integration with TINA led to 21% reduction in error
 - Success of Chung duration model shows ability to leverage ANGIE's parse trees



Some Future Directions

- **Improve speed**
- **More work on new words**
 - **Cases where sharing does help?**
 - **Use ANGIE to support detection?**
- **Use parse structures from TINA and ANGIE for**
 - **Further work in prosodic modelling (including above word level)**
 - **More detailed acoustic modelling, e.g., models specific to syllable position, etc.**