STATISTICAL MODELING OF PRONUNCIATION VARIATION BY HIERARCHICAL GROUPING RULE INFERENCE

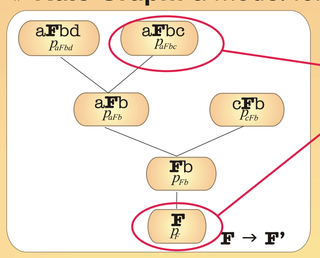
Mónica Caballero, Asunción Moreno
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP)
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

ABSTRACT

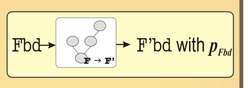
- Data-driven approach to statistical modeling pronunciation of variation based on learning stochastical pronunciation rules.
- Hierachical Grouping Rule Inference (HIEGRI) algorithm is proposed to infer a set of more general rules.
- Learned rules are applied to derive word pronunciation models (WPM) for the each word of the vocabulary of the recognizer. WPMs are applied to a context-dependent acoustic model based recognizer.
- Pronunciation variation method is evaluated on a Spanish recognizer.

RULE LEARNING METHODOLOGY

- Rule: ***LFR→ F' with probability*** p_{LFR} { ***F,F'***: transformation (Focus, Output); ***LFR***: Condition
- Rules associated to a given transformation are modeled jointly.
- **Rule Graph:** a model for each transformation ***F→ F'***



- In each level, rules with the same length condition
  - Highest level: largest conditions
  - Lowest level: focus of the transformation
- Each rule node is assigned an estimated probability
- Given a phone string the most specific rule in the graph is selected.



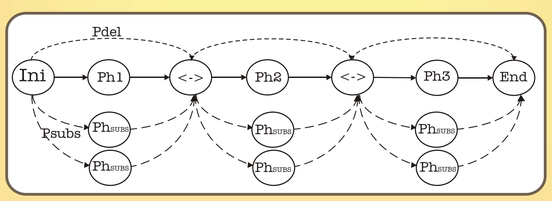
METHOLODOGY

3 STEPS { Obtaining an initial set of rules; Hierachical Grouping Rule Inference (HIEGRI); Rule selection strategy

1 Obtaining an initial set of rules

Compare a canonical transcription (T_{can}) with an automatic transcription representing an hypotheses of what has been really said (T_{aut}).

- T_{can}= concatenation of word base transcriptions.
- T_{aut} is achieved by means of forced recognition using *word pronunciation models*.
- **Word pronunciation model:** FSA representing canonical transcription of a word allowing deletions and substitutions.



ALIGNING

T_{can}, T_{aut} → **Dynamic programming algorithm** → Transformations with their associated conditions ***LFR***

Considerations

- Focus of a transformation: 1 or 2 phonemes
- ***L*** and ***R*** up to two phones (including boundary symbol '$')
- Transformations appearing less than ***Nt*** times are removed.

Result

Large set of rules from each transformation. Some of them might be specific cases of more general rules.

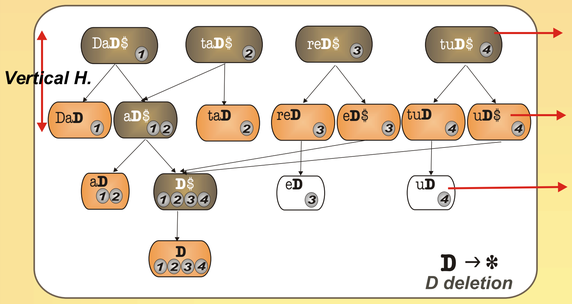
2 HIEGRI algorithm

OBJECTIVE

Detect the common patterns and infer a set of candidate general rules
→Create a preliminary graph, **HIEGRI graph.**

- **STARTING POINT:** Place initial rules in the highest level.
- **GROWING PROCESS:** Establish a double (horizontal and vertical) hierarchy

For each level:
a. Identify horitzontal hierarchical class for each node in the level.
b. Develop a lower level following horizontal hierarchy.
Each rule is stripped one element of the right or the left context.
id is inherit.



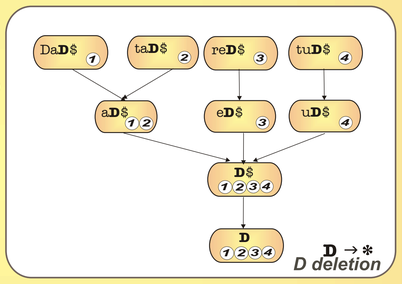
Grouping: initial nodes or nodes created by more than one node

Heir: Rules created by a grouping node

Plain: Rest of nodes

- **PRUNING PROCESS:** Parse the graph in a bottom-up direction erasing rule nodes not linked to its lower level.

HIEGRI GRAPH



3 Selection of final set of rules

OBJECTIVE

Select as general as possible rules without losing modeling accuracy.

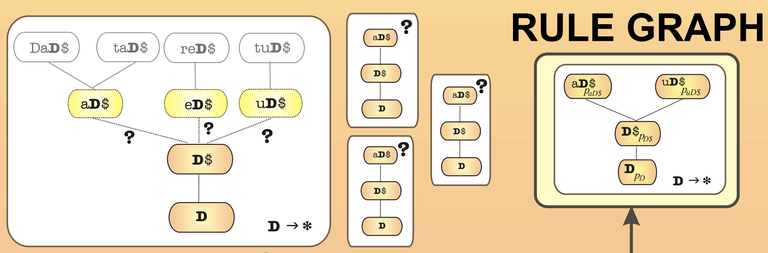
- **HOW:** Starting with a graph containing only the most general rule, try to add more nodes to the graph if entropy of the graph is reduced more than a threshold ΔH_{th}.

- Select candidates to be added
- Evaluate H_G
- Add the node that maximizes loss of H, if $\Delta H > \Delta H_{Gth}$ and $no_r > no_{rth}$

$$H_G = \sum_{r=0}^{R} H_r$$

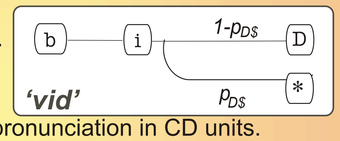
$$H_r = p_r \log_2 p_r + (1-p_r)\log_2(1-p_r)$$

$$prG = \frac{\#\ times\ transformation\ occur\ (no_r)}{\#\ times\ condition\ seen\ (ns_r)}$$



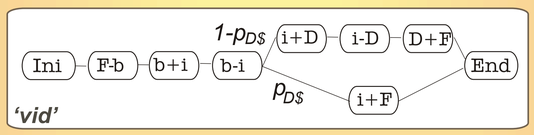
GENERATING WORD PRONUNCIATION MODELS

- Develop a phone FSA applying derived rules.
  - Initialize FSA with canonical transcription.
  - For each new variant, create a new branch.



- Expansion to FSA representing word pronunciation in CD units.

Phonetic Unit: Contextual demiphones (half of a context phone)



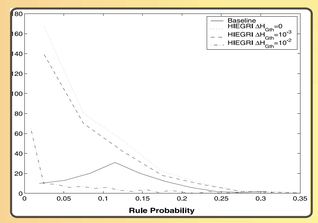
EXPERIMENTS

Database

- **Spanish SpeechDat II**: Pronunciation variation due to regional accents and not professional speakers.

Rule generation

- **Rule training set:** 9,500 utterances (800 speakers): 67,239 running words, vocabulary of 12,418 words.
  - **31** focus and **53** tranformations (mostly deletions) Varying ΔH_{Gth} different sets of rules are obtained.
  - ***Baseline rule set*** without applying HIEGRI algorithm. Rule selection based on no_r, no_{rth} **22** focus, **29** transformations and **117** rules

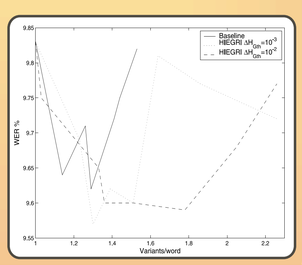


- ↓ΔH_{Gth}: large set of rules, very dependent on the training vocabulary, and so are the probabilities.
- ↑ΔH_{Gth}: specific rules dissappear in front of the generals inferred. Probabilities decrease.

Recognition results

- **Recognition task:** Phonetically rich sentences.
  Language model: Trigram modeling all SpeechDat sentences (vocabulary 14,300 words).
- **Test:** 1,570 sentences from 200 different speakers. **Perplexity:** 68. Matching rule training voc. and test voc. = **81.66 %**
- ***pmin***: Varying ***pmin*** different number of variants per word (V/w).

p_{min}	Base rule		$\Delta H_{th}=10^{-3}$		$\Delta H_{th}=10^{-2}$	
	WER	V/w	WER	V/w	WER	V/w
0.02	9.82	1.53	9.72	2.26	9.77	2.26
0.05	9.75	1.44	9.77	1.86	9.68	2.05
0.07	9.72	1.41	9.81	1.64	9.59	1.78
0.09	9.62	1.29	9.62	1.39	9.60	1.36
0.10	9.71	1.26	9.57	1.30	9.65	1.33
0.12	9.64	1.14	9.69	1.23	9.75	1.03
1.00	*9.83*	*1.00*	*9.83*	*1.00*	*9.83*	*1.00*



CONCLUSIONS

- Application of HIEGRI algorithm allows to generalize rule set to make it applicable to other vocabularies.
- Proposed methodology improves recognizer performance.
- Improvement is quite stable for a large interval of V/w.