MULTIDIALECTAL ACOUSTIC MODELING: A COMPARATIVE STUDY

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

ABSTRACT

- Multidialectal acoustic modeling based on sharing data across dialects.
- Comparative study of different methods of combining data based on decision tree clustering algorithms to obtain a robust multidialectal set of acoustic models.
- Approaches evolved differ in the way of evaluating the similarity of sounds between dialects, and the decision tree structure applied.
- Proposed systems are tested with Spanish dialects across Spain and Latin -America: dialects of Argentina, the Caribbean, Colombia, Mexico and Spain.

TRANSCRIPTION

- For each considered dialect, a canonical phonetic transcription in SAMPA symbols is obtained.
- Transcriptions are obtained automatically by means of rules.
- SAMPA symbols used for transcriptions:

DIALECT	SHARED PHONES	NON-SHARED PHONES
ARGENTINA	a b B d D f g G i j J k l m n N o p r rr R s t tS u w z	Z x h
CARIBBEAN		jj h
COLOMBIA		jj h
MÉXICO		jj x
SPAIN		jj x T

RECOGNITION SYSTEM

- In-house system based on SCHMM.
- Parametrization : Mel-cepstrum (C, ΛC, ΛΛC, ΛE).
- Number of Gaussians of the Codebook: 512 and 128 for Energy.
- Phonetic Unit: Demiphones represented by a left-to-right HMM of 2 states.

no	F-n	n+o	n-o	o+F
	/n/		/o/	

DECISION TREE BASED CLUSTERING ALGORITHM

- **Entropy measure**

Entropy of a node A

$$H(A)=\sum_{m=1}^{M} f(m)\left[\sum_{s=1}^{S} f(s|m)\sum_{g=1}^{G} b_{sg}\log b_{sg}\right]$$

- **Stopping criteria:** minimum decrease of entropy and/or a threshold in the minimum number of realizations contained in each final cluster.
- **Question set:**
  - Phonetic features (type, place & manner);
  - Non-phonetic questions (position in the word, wether the phone belongs to a consonant group, and dialect of theunit).
    ➝To be defined explicitely in each approach.
  - Multiple questions about the same attribute using a 'OR' logical link.
    Is the manner of articulation nasal OR fricative?

ACOUSTIC MODELING

MEASURES OF SIMILARITY

- **SAMPA based:** *The sounds of different dialects that have the same SAMPA representation are considered to be the same phone.*

  The multidialectal phone set is defined.
  Similarity is evaluated at a **phone** level.

- **HMM based:** *A decision tree driven by the entropy measured over dialect-dependent HMMs is used to define which sounds (and from which dialects) are similar enough to share training data.*

  A set of CD-HMMs are trained for each dialect and marked with a dialect tag (AR, CA, CO, ME, SP). It allows similar **context-dependent** acoustic units to be detected.

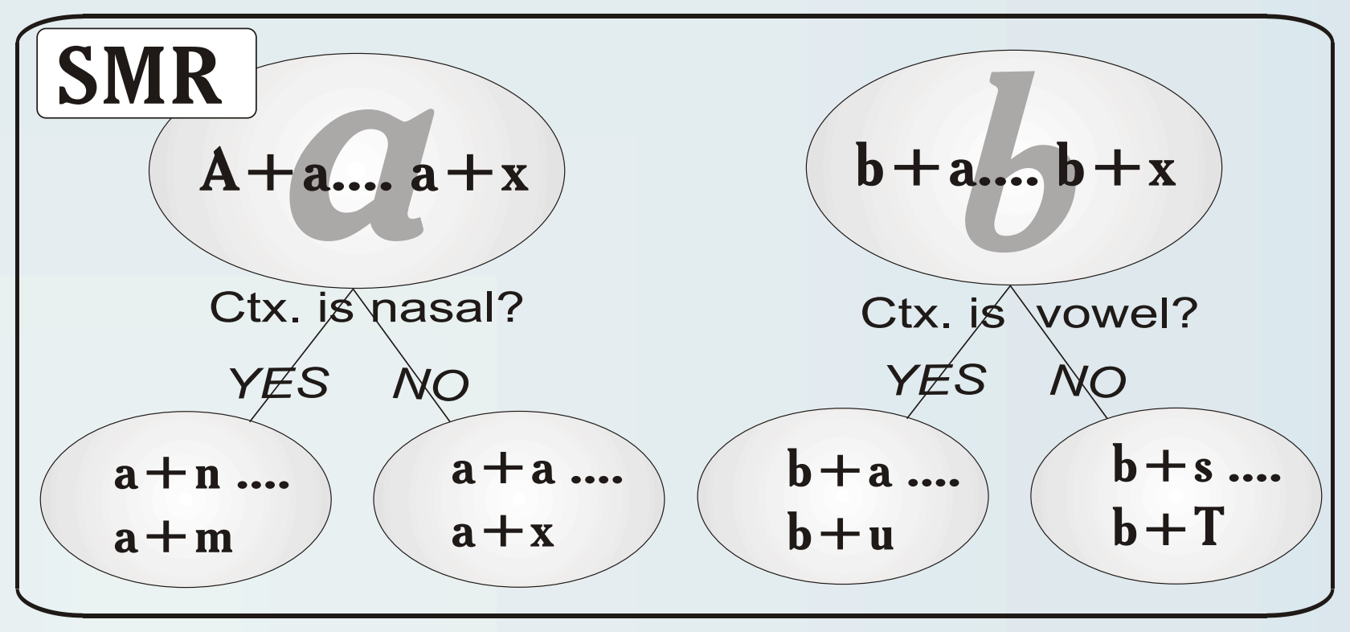
TREE STRUCTURE

- **Multiroot:** *Different tree (root) for each unit of the phone set.*
- **One-root:** *A single tree for all the units in the phone set. Data can be shared between different phones.*

Multidialectal approaches

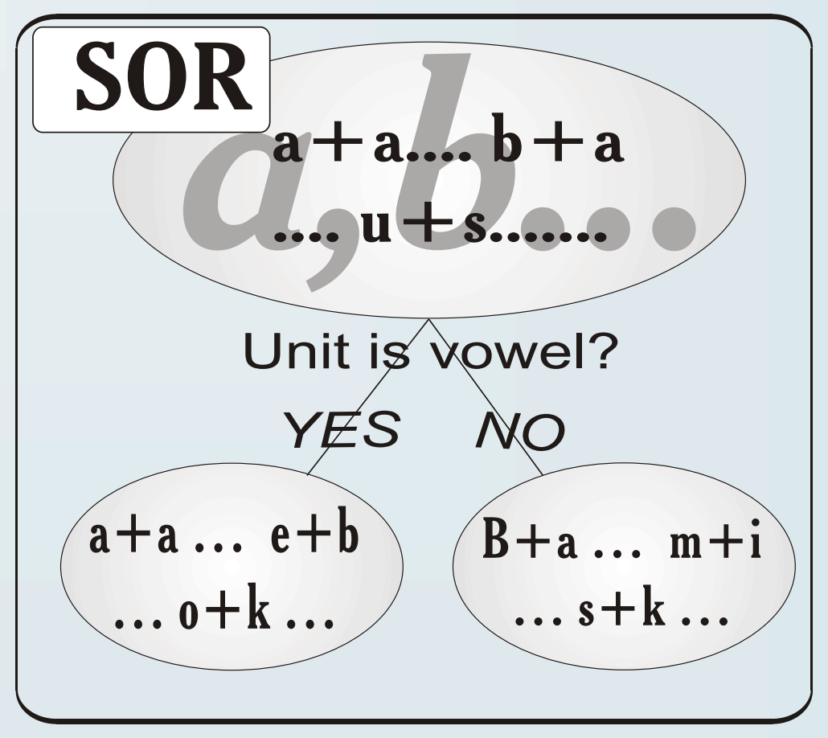
SAMPA based measure, multi-root structure (SMR)

- Multidialectal phone set using SAMPA.
- Decision tree clustering algorithm for context modeling.
- The question set only inquires about the context of the unit.



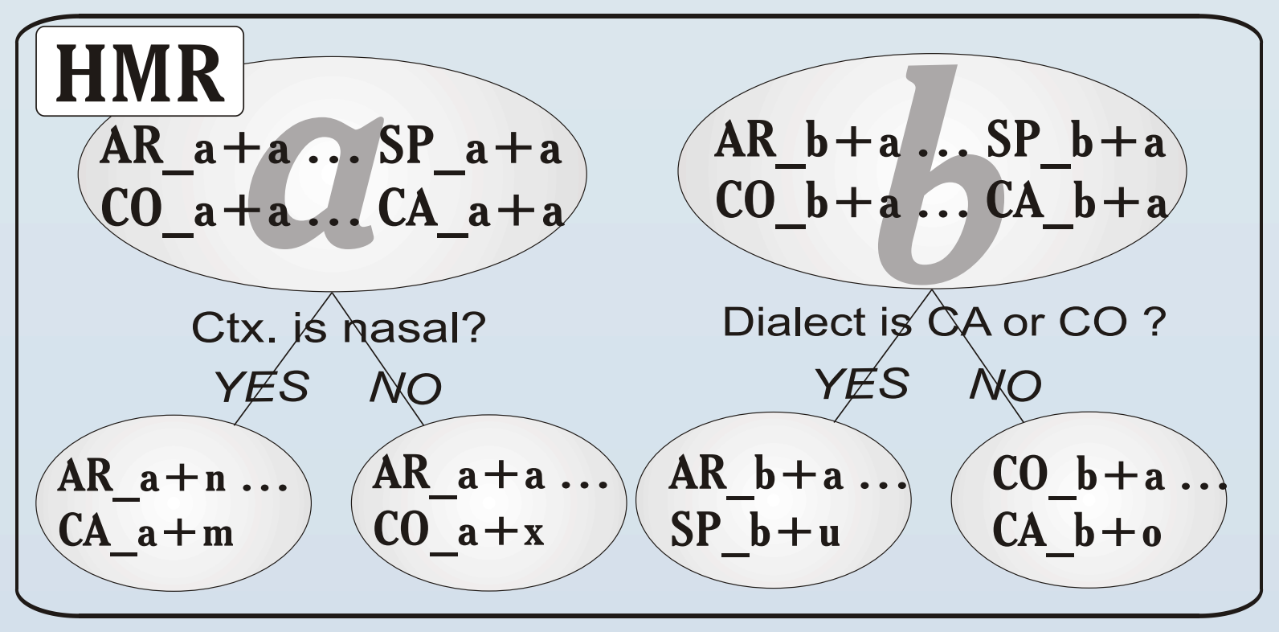
SAMPA based measure, one-root structure (SOR)

- Multidialectal phone set using SAMPA .
- One-root tree structure allows phones to be joined if they are similar in certain contexts or situations.
- The question set contains questions about the phone itself as well as the context.



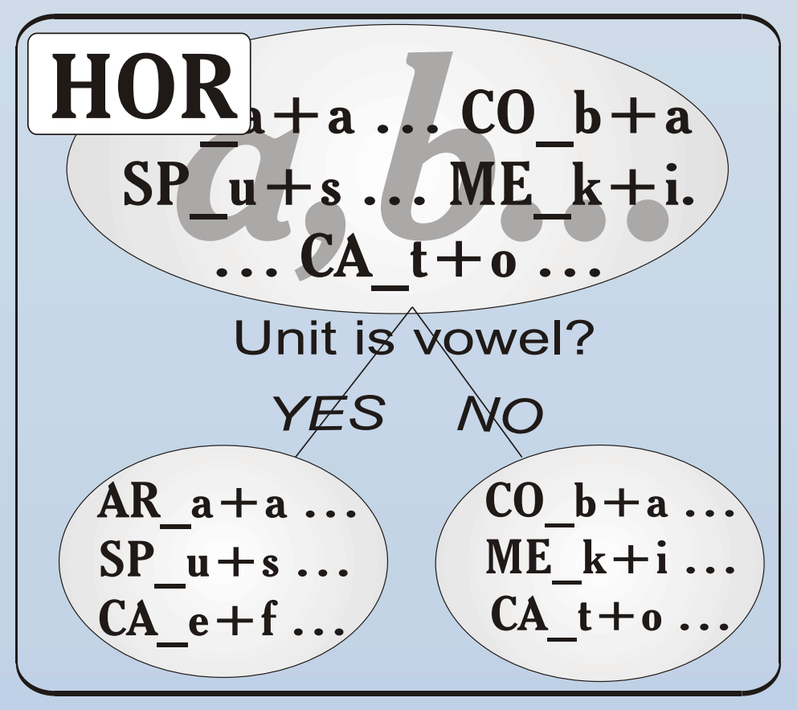
HMM based measure, multi-root structure (HMR)

- Dialect-dependent models for each contextual unit.
- Similarity is only evaluated across phones with the same SAMPA representation.
- The question set asks for the context unit and the **dialect**.



HMM based measure, one-root structure (HOR)

- A single tree with all the dialect-dependent models in the root node.
- Models with the same SAMPA representation can be distinguished and models with distinct SAMPA representation can be joined.
- Fully automatic, and independent of prior phonetic assumptions.



EXPERIMENTS

Data

- **Spain:** SpeechDat Spain. 4,000 speakers
  Latin-American dialects: SALA. 1,000 speakers
- **Training:** Phonetically rich words and sentences
  Test: Phonetically rich words
- Number of training and test utterances:

DIALECT	AR	CA	CO	ME	SP
Training utterances	9,568	9,303	8,874	11,506	40,936
Test utterances	2,575	2,411	2,358	2,022	3,632

ASR systems

- - Monodialectal ASR
  - Multidialectal approaches SMR, SOR, HMR and HOR
- Number of models for the created systems:

SYSTEM	AR	CA	CO	ME	SP	SMR	SOR	H[M,O]R
# HMM	662	688	683	716	847	988	981	2,000

Results

% WER

DIALECT	Mono	SMR	SOR	HMR	HOR
ARGENTINA	7.34	8.31	7.76	6.37	6.23
CARIBBEAN	6.71	6.27	6.27	6.41	6.41
COLOMBIA	9.22	8.28	8.28	7.97	7.81
MÉXICO	10.10	8.01	8.17	9.62	8.65
SPAIN	3.62	4.74	4.6	4.46	4.04
AVERAGE	7.40	7.12	7.02	6.97	6,63

- All systems improve the monodialectal performance, except for rate of Spain, which is slightly degradated.
- SMR and SOR systems reduces WER in the Caribbean, Colombian and Mexican dialects.
- HOR system leads to the best average recognition results.

Data sharing

- **- Full multidialectal:** clusters containing data from all dialects
  - Semi-multidialectal: clusters containing data from more than one but not all dialects

	SMR	SOR	HMR	HOR
Full Multidialectal	69.23%	69.72%	6.70%	6.20%
Semi-multidialectal	20.65%	21.61%	11.20%	14.85%

- Maximum data sharing is given by SXR approaches. HXR approaches decrease full multidialectal units. Using one-root tree structure allows more data sharing betwen groups of dialects.
- Better recognition performance is achieved sharing data between groups of dialects then sharing data between all of them.

CONCLUSIONS

- Multidialectal approaches based on sharing data between dialects improve monodialectal systems.
- It is better to measure the similarity of sounds between dialects using a HMM based measure than using the SAMPA alphabet based measure.
- Application of one-root structure leads to better recognition results.