Rhythm-Based Syllabic Stress Learning without Labelled Data

Bogdan Ludusan[1], Antonio Origlia[2], Emmanuel Dupoux[1]

[1]LSCP, EHESS/ENS/CNRS, Paris, France
[2]PRISCA-Lab, Federico II University, Naples, Italy

SLSP 2015, Budapest

Motivation

- low-resource stress learning system
  - no manual stress labels
  - only acoustic features

- automatically generated stress labels → rule-based stress detection system
- acoustic features previously used for prominence detection

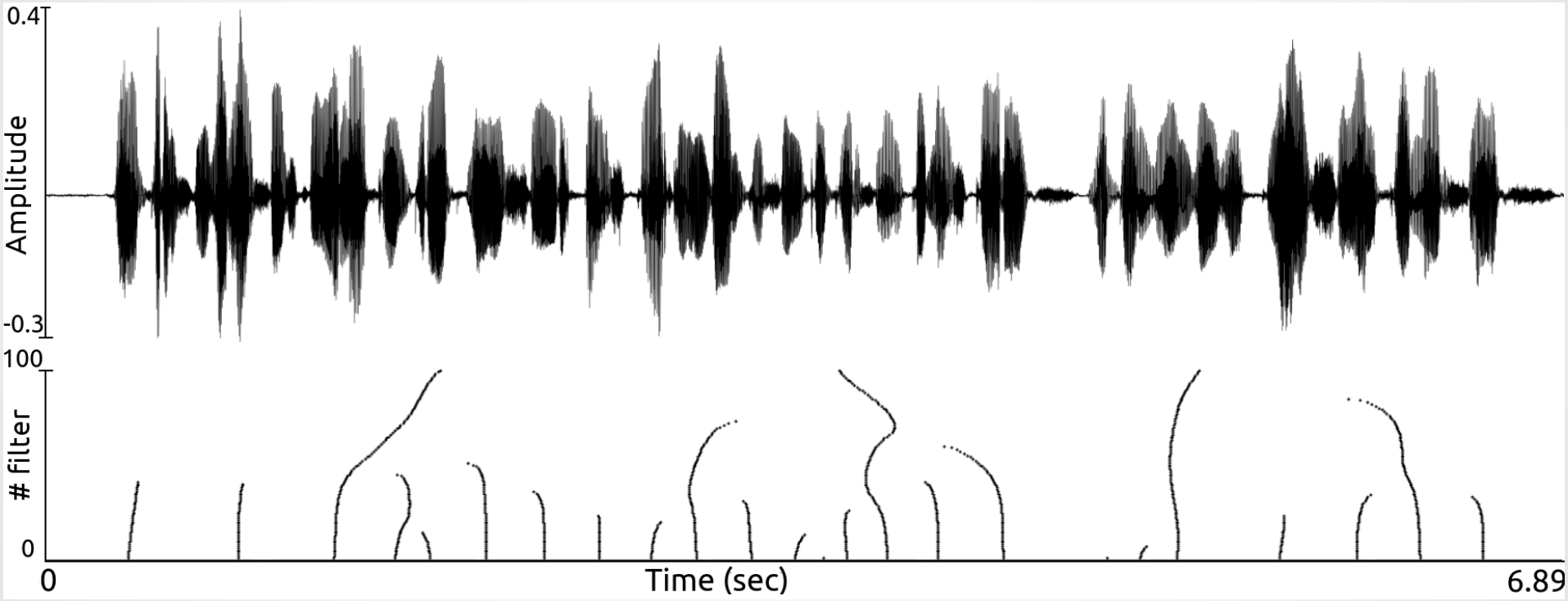
Outline

- Methods
  - Automatic label generation
  - Learning algorithms
  - Materials
- Experiments
- Conclusions

Automatic label generation

- model of rhythm perception (Todd, 1994)
  - modelling of the peripheral auditory system
  - summation of the auditory nerve response
  - filtering with a bank of Gaussian filters
  - plotting the peaks of the output function in a 2D space
    → hierarchical representation (*rhythmogram*)

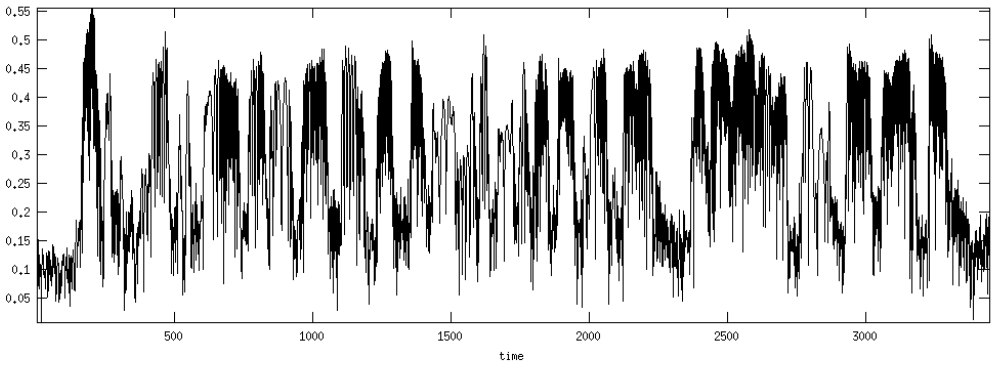
The rhythmogram

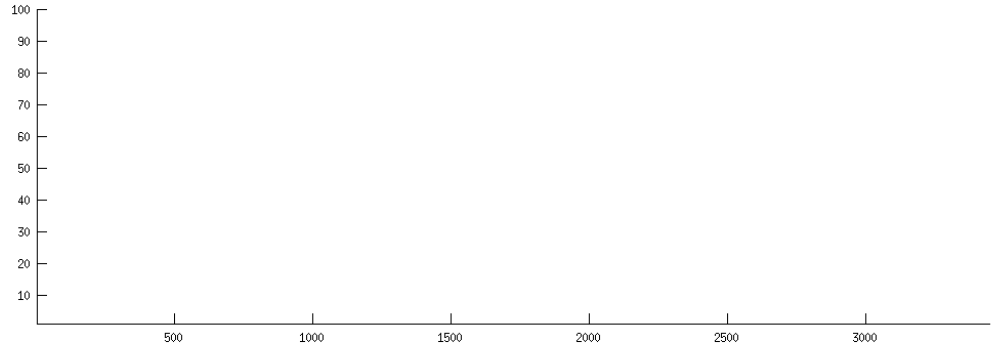


Computing the rhythmogram

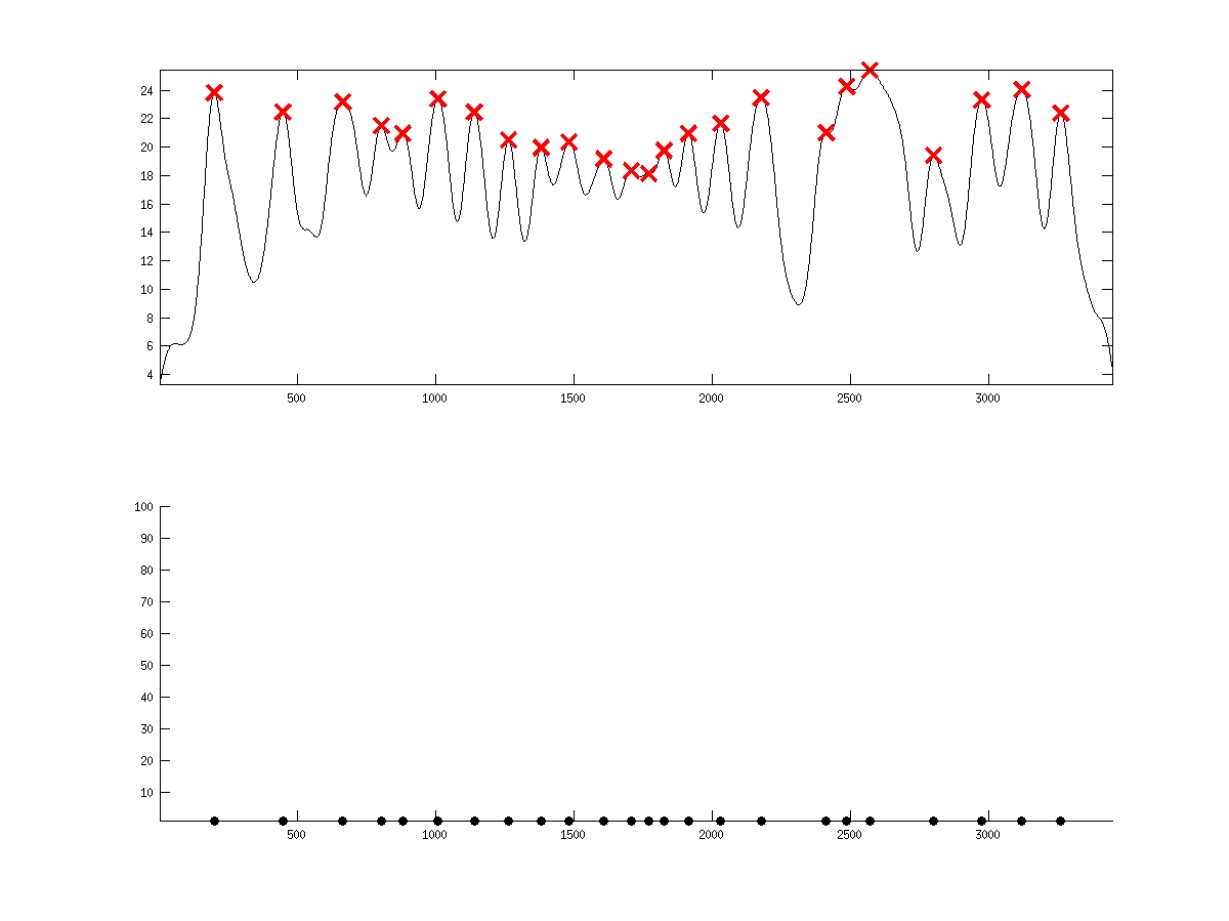
- signal energy
  - resample at 500 Hz + full wave rectification
  - model the ear's loudness function (cubic root)
- parameters (Ludusan et al., 2011):
  - total number of filters
  - minimum and maximum filter width
- quantized representation of the rhythmogram
  - time and height of each event

Computing the rhythmogram

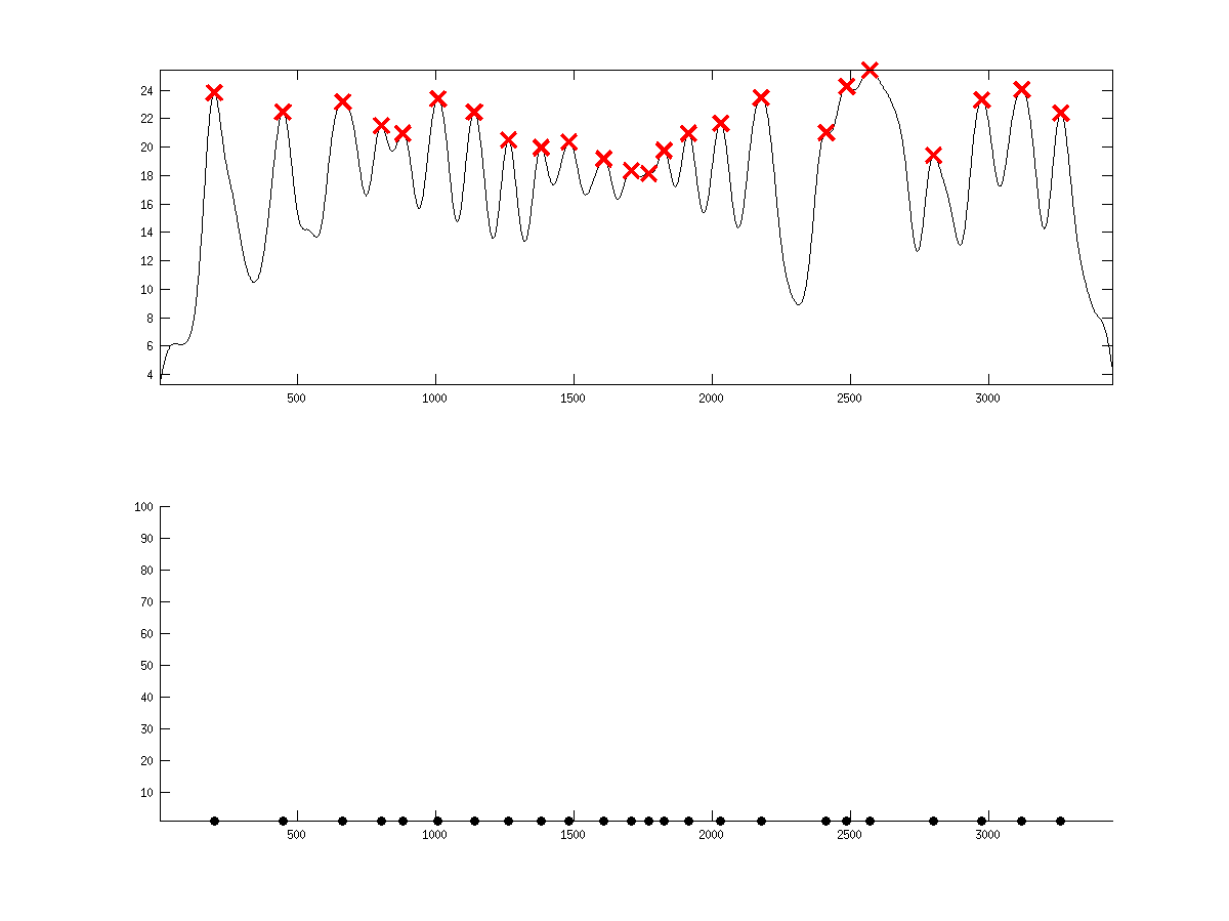




Computing the rhythmogram

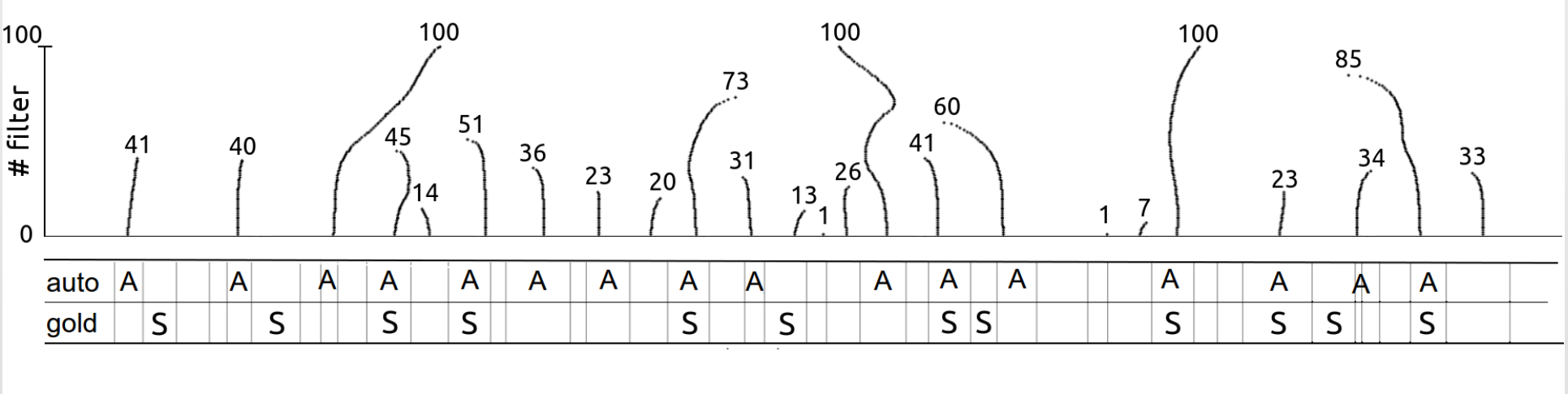


Computing the rhythmogram



Determining stressed syllables

- determine the events that correspond to each syllable
- define as stress value the height of the event
- mark the local maxima of the stress function as stressed syllable (*baseline*)



Acoustic features

- 5 acoustic features (Cutugno et al., 2012):
  - syllable length
  - nucleus length
  - average energy
  - voiced time in syllable / syllable length
  - glissando

Learning algorithms

- Naive Bayes (NB)
  - kernel-density estimate (Gaussian)
- Expectation-Maximization based clustering (EM)
  - diagonal covariance matrix
- Weka toolbox (Hall et al., 2009)

Materials

- Catalan and Spanish
- Glissando corpus (Garrido et al., 2013)
  - radio news
  - segmental and prosodic annotations

Language	Speakers (F+M)	Duration
Catalan	8 (4+4)	6 hrs
Spanish	8 (4+4)	6 hrs 15 min

Learning with automatic labels

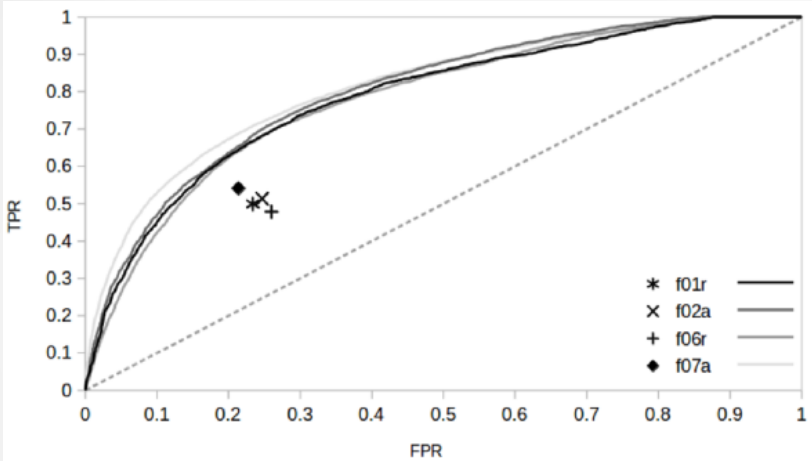
- leave-one-speaker-out cross-validation
- supervised learning:
  - automatic stress labels (*rhyLabel*)
  - manual stress labels (*goldLabel*)
- unsupervised learning (*noLabel*)
- area under the receiver operating characteristic curve (AUC)

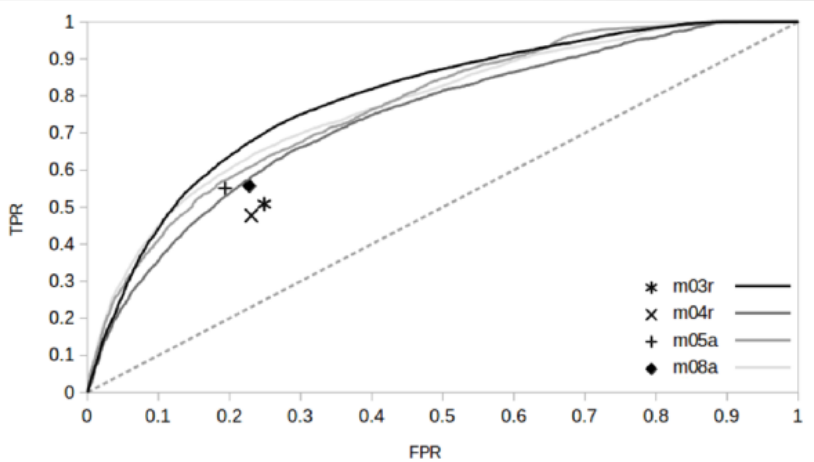
Comparison against the baseline

female

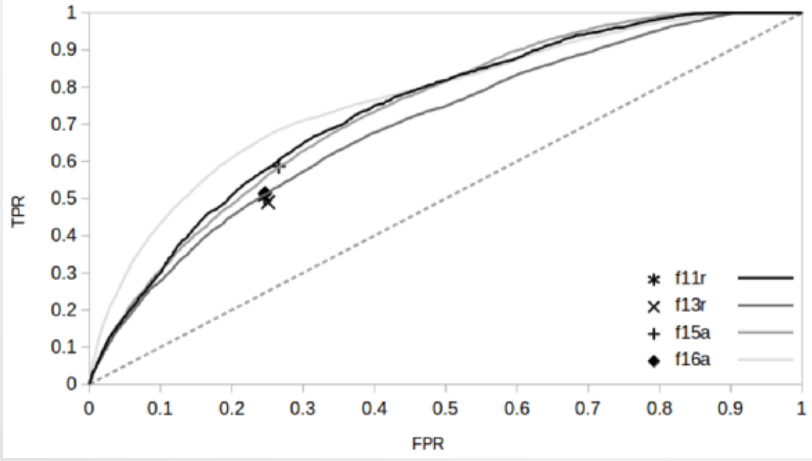
male

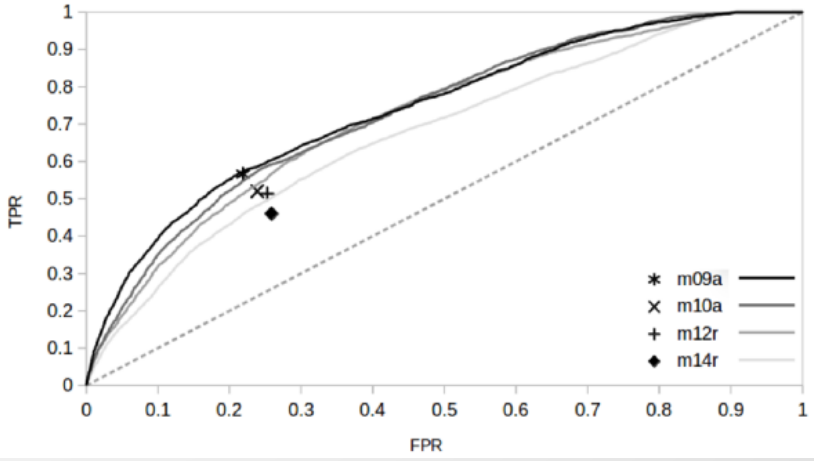
Catalan



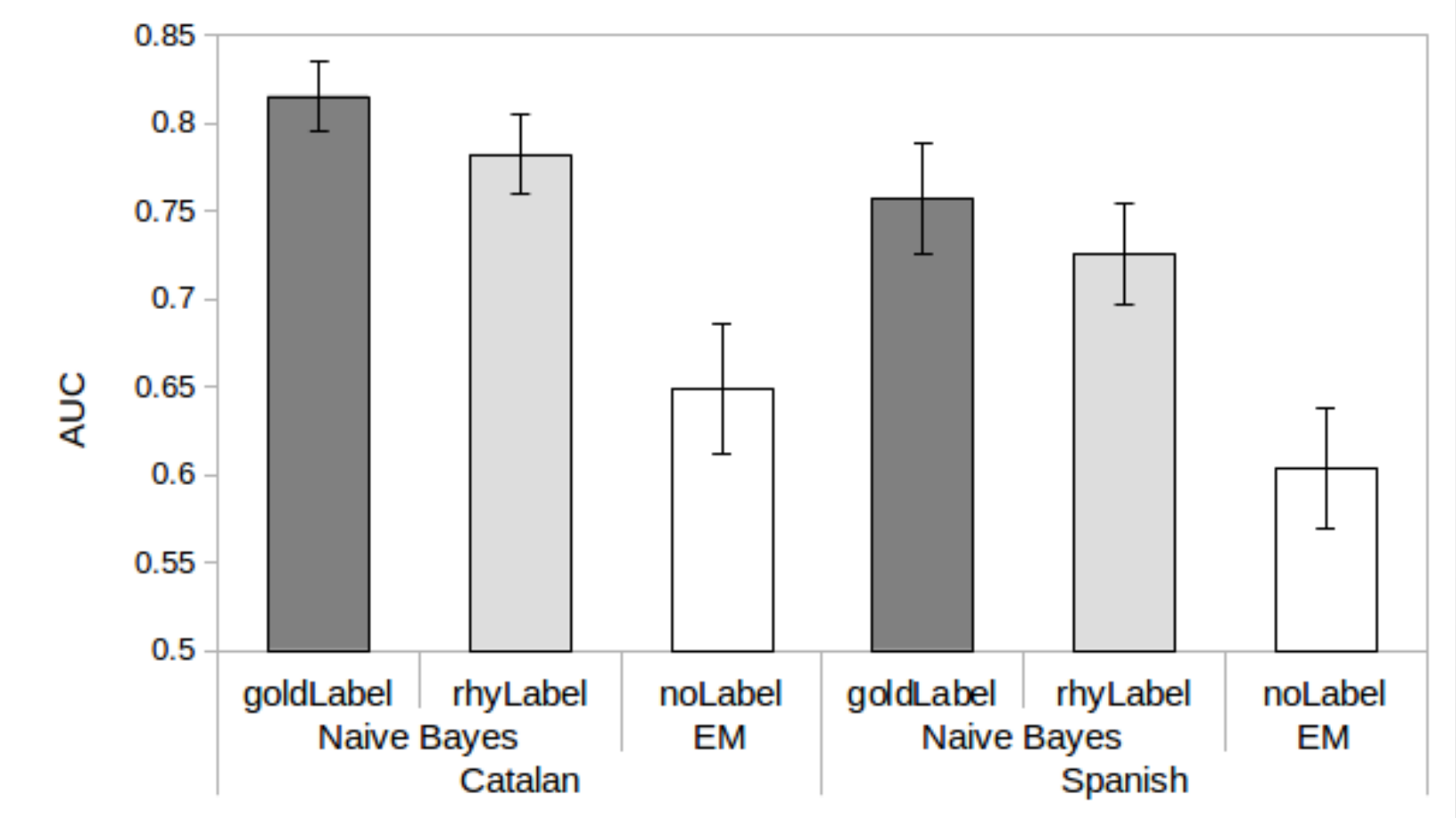


Spanish





Results


0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
AUC
goldLabel
rhyLabel
noLabel
goldLabel
rhyLabel
noLabel
Naive Bayes
EM
Naive Bayes
EM
Catalan
Spanish

Learning with different classifiers

- Logistic Regression (LR)
- Support Vector Machines (SVM)
- same experimental setting and acoustic cues

Learning algorithm	Catalan			Spanish		
	goldLabel	rhyLabel	noLabel (EM)	goldlabel	rhyLabel	noLabel (EM)
NB	.815	.782	.649	.757	.726	.604
LR	.819	.766		.758	.711	
SVM	.798	.734		.719	.690	

Conclusions

- lexical stress annotation without using manual data in the learning process
  - improvement over the baseline and a clustering method
  - results comparable to learning with manual labels
- feasibility study → improvements across speakers, languages and learning algorithms
  - use of more powerful or more robust learning algorithms
  - enrich the feature set
- speech technology applications, annotation of corpora

Thank you!

References

- Cutugno, F., Leone, E., Ludusan, B., Origlia, A.: Investigating Syllabic Prominence With Conditional Random Fields and Latent-Dynamic Conditional Random Fields. In: Proceedings of INTERSPEECH. pp. 2402–2405 (2012)
- Garrido, J.M., Escudero, D., Aguilar, L., Cardenoso, V., Rodero, E., De-La-Mota, C., Gonzalez, C., Vivaracho, C., Rustullet, S., Larrea, O., et al.: Glissando: A Corpus for Multidisciplinary Prosodic Studies in Spanish and Catalan. Language resources and evaluation 47(4), 945–971 (2013)
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
- Ludusan, B., Origlia, A., Cutugno, F.: On the Use of the Rhythmogram for Automatic Syllabic Prominence Detection. In: Proceedings of INTERSPEECH. pp. 2413–2416 (2011)
- Todd, N.M.: The Auditory “Primal Sketch”: A Multiscale Model of Rhythmic Grouping. Journal of New Music Research 23(1), 25–70 (1994)