ASR retraining for Ebrenc Catalan

Authors: Berta Benet, Marilena Budan

Goal

Test and improve the performance of an existing Catalan ASR model for the Ebrenc Dialect.

Original ASR model - Source

SOURCE

[1][2]

240h

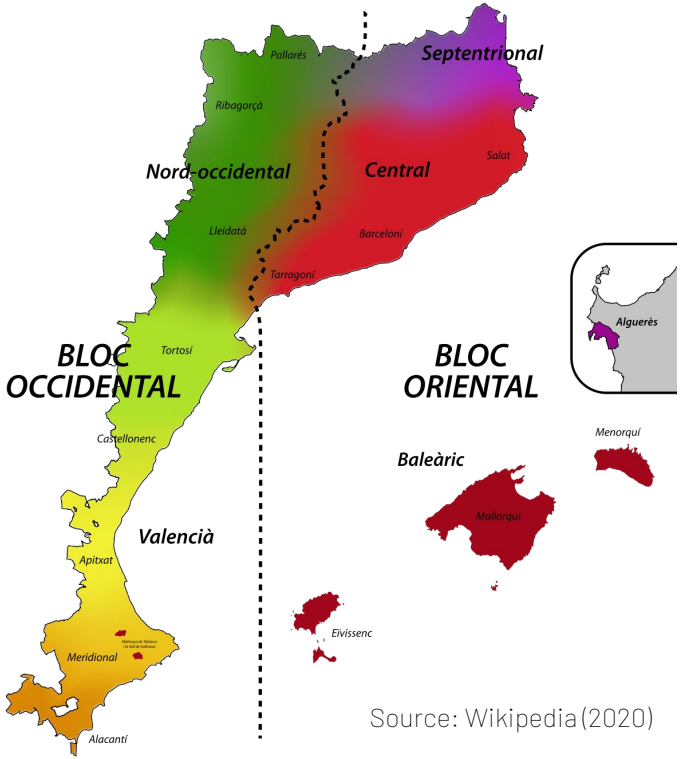
320h

CORPORA

Initial hypothesis

Initial hypothesis

The predominant Catalan dialect that can be found in the corpora used to train the original model is the standard one (i.e. the one spoken in news or official statements).



Source: Wikipedia (2020)

Initial hypothesis

Better performance for the Central Dialect spoken in Barcelona.

The performance will significantly decrease for the Ebrenc dialect.

Initial hypothesis

ASR performance for the Ebrenc dialect can be improved by retraining the original model.

Test on original model

Performance metric

ASR's performance has been quantitatively assessed using WER.

$$WER = \frac{S+D+I}{N}$$

- S: number of substitutions
- D: number of deletions
- I: number of insertions
- N: number of words in the reference

Test sentences

TEST PRONUNCIATION

- 15 sentences from [3][4] tested for both dialects with standardized vocabulary.

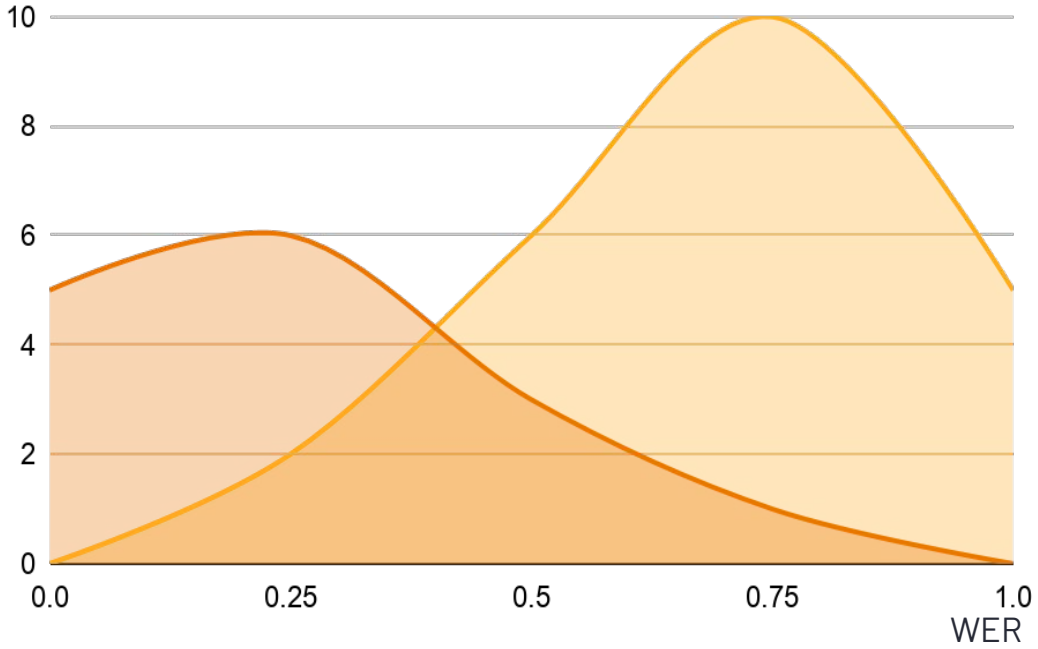
Molta gent està veient que potser no ha de treballar cada dia a l'oficina.

TEST VOCABULARY

- 10 sentences from [5] tested for the Ebrenc dialect including specific dialect.

En estes escoles, els xiquets i xiquetes disposen de la seua pròpia taula.

Test on original model



Histogram: WER on Original Model

CENTRAL ACCENT

Mean = 0.2986

Std = 0.2024

EBRENC ACCENT

Mean = 0.6614

Std = 0.1961

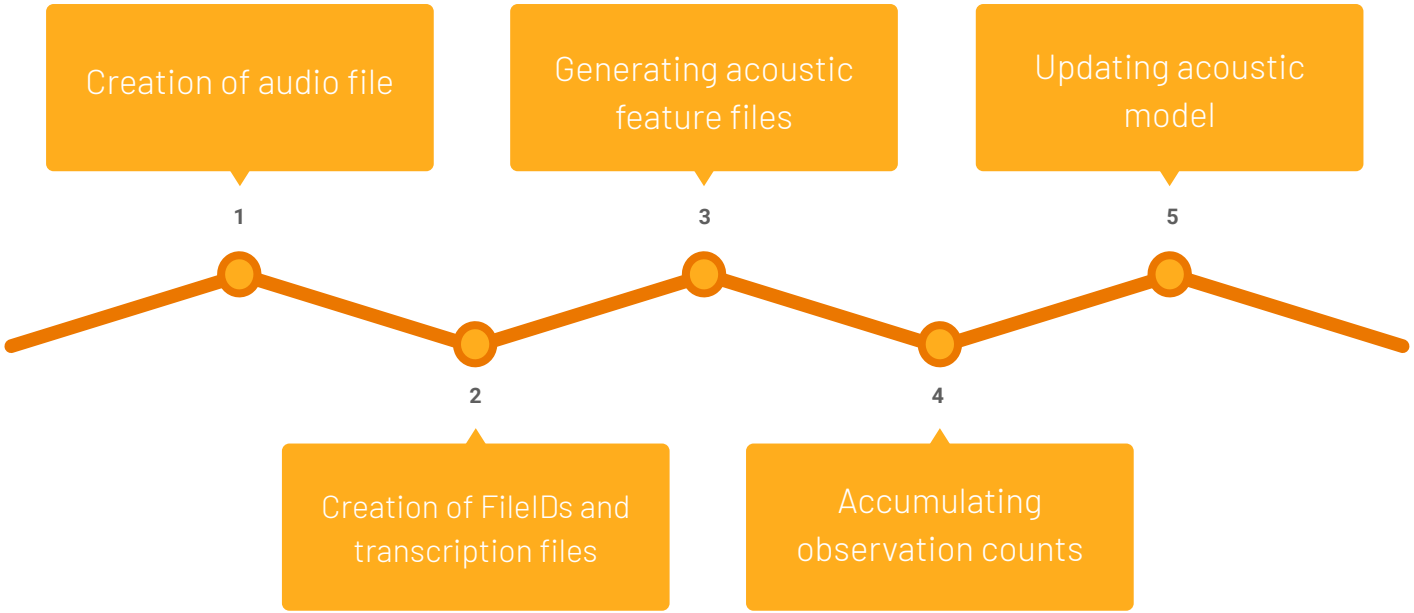
Retraining method

Retraining method - Data specifications

- 20 audio files in Ebrenc Catalan
- FileIDs and Transcript files
- Updated phonetic dictionary

navegar	N AE BV EA GH A
	N A BV E GH A

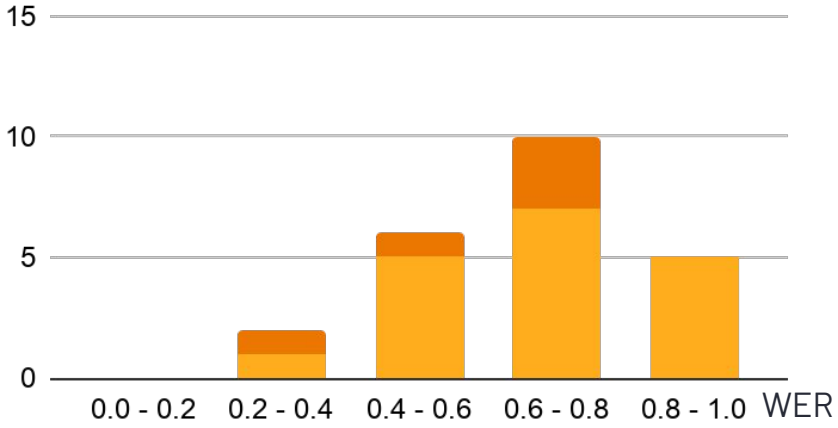
Retraining method - Pipeline



Test on retrained model

ASR performance for Ebrenc Dialect

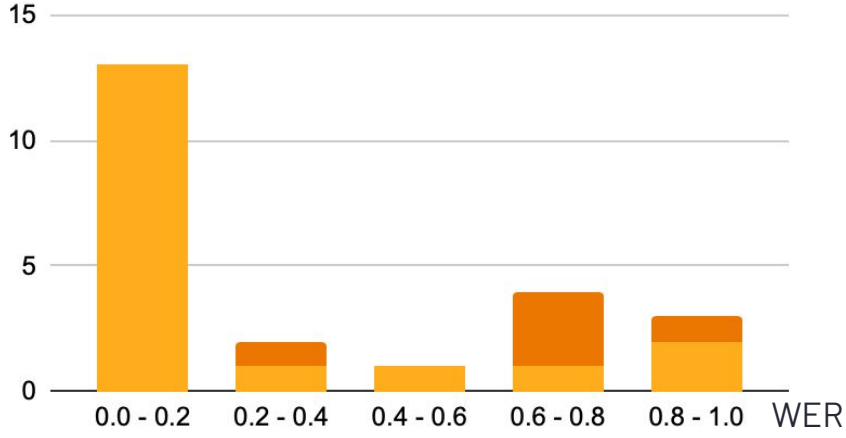
Test on original model



Mean = 0.6614

Std = 0.1961

Test on retrained model



Mean = 0.3661

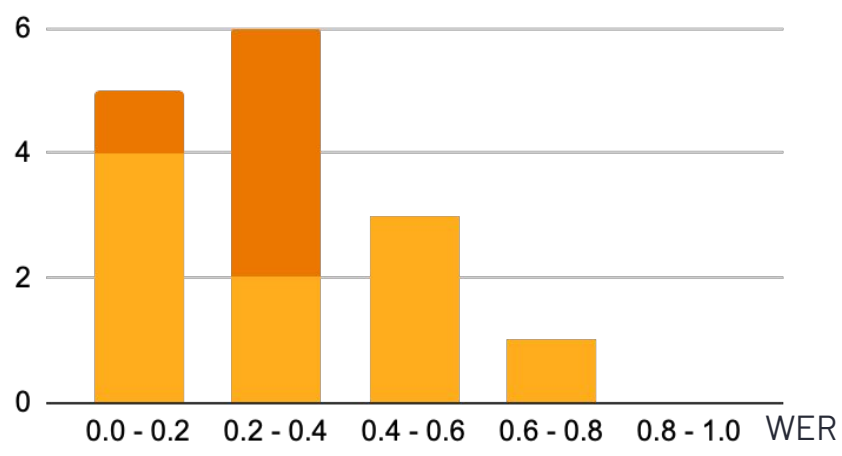
Std = 0.3896

Sentences used to retrain

Sentences NOT used to retrain

ASR performance for Central Dialect

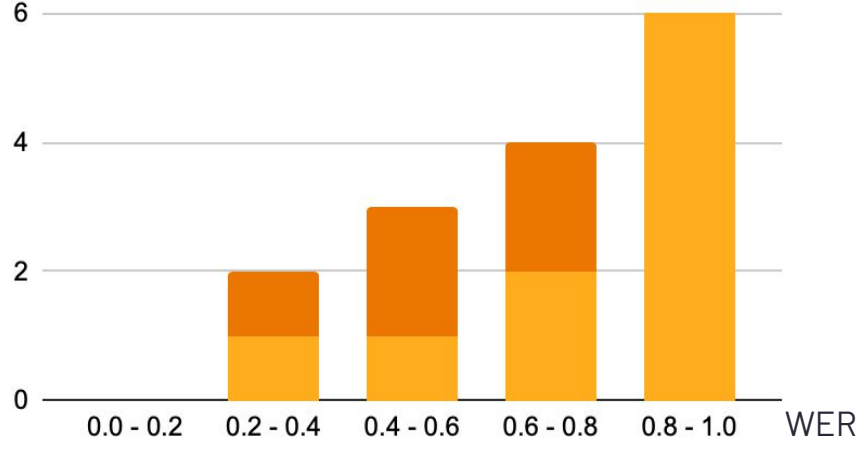
Test on original model



Mean = 0.2986

Std = 0.2024

Test on retrained model



Mean = 0.6908

Std = 0.1633

Sentences used to retrain

Sentences NOT used to retrain

Transcripts comparison

apunta que això no implica necessàriament la decisió de comprar

	EBRENC DIALECT	
ORIGINAL	apunta deixat ben seguiment de siss de compra	7/10
ADAPTED	apunta que això no implica necessàriament la decisió de comprar	0/10
	CENTRAL DIALECT	
ORIGINAL	apunta que això no implica necessàriament la decisió de comprar	0/10
ADAPTED	en contra creixement ple que necessàriament nazis i recupera	9/10

Conclusions

Conclusions

The error has been reduced to almost the half for the Ebrenc dialect.

The error has been increased to more than double for the Central dialect.

Transcripts comparison

apunta que això no implica necessàriament la decisió de comprar

	EBRENC DIALECT	
ORIGINAL	apunta deixat ben seguiment de siss de compra	7/10
ADAPTED	apunta que això no implica necessàriament la decisió de comprar	0/10
	CENTRAL DIALECT	
ORIGINAL	apunta que això no implica necessàriament la decisió de comprar	0/10
ADAPTED	en contra creixement ple que necessàriament nazis i recupera	9/10

Personal conclusions

- Phonetic dictionary
- Small retraining dataset
- Limited resources
- CMU Sphinx requires a lot of previous preparation

References

[1] Col·lectivaT. (2020). Retrieved 28 May 2020, from https://collectivat.cat/asr

[2] Shmyrev, N. (2020). CMUSphinx Open Source Speech Recognition. Retrieved 28 May 2020, from https://cmusphinx.github.io

[3] 324. (2020). La crisi arriba al mercat immobiliari: baixen les vendes i els preus de l'habitatge. Retrieved from https://www.ccma.cat/324/la-crisi-arriba-al-mercat-immobiliari-baixen-les-vendes-i-els-preus-de-lhabitatge/noticia/3014373/

[4] 324. (2020). El Mercat de Música Viva de Vic fa un salt als continguts digitals en la 32a edició. Retrieved from https://www.ccma.cat/324/el-mercat-de-musica-viva-de-vic-fa-un-salt-als-continguts-digitals-en-la-32a-edicio/noticia/3014120/

[5] noticiasCV. (2018). Les escoles d'estiu de la Marina acosten la nàutica a xiquets i xiquetes. Retrieved from https://www.noticiascv.com/les-escoles-destiu-de-la-marina-acosten-la-nautica-a-xiquets-i-xiquetes/

Our github repository: https://github.com/marilenabudan/ASR_Retrain_Ebrenc_Catalan

Thank you for your attention

Berta and Marilena