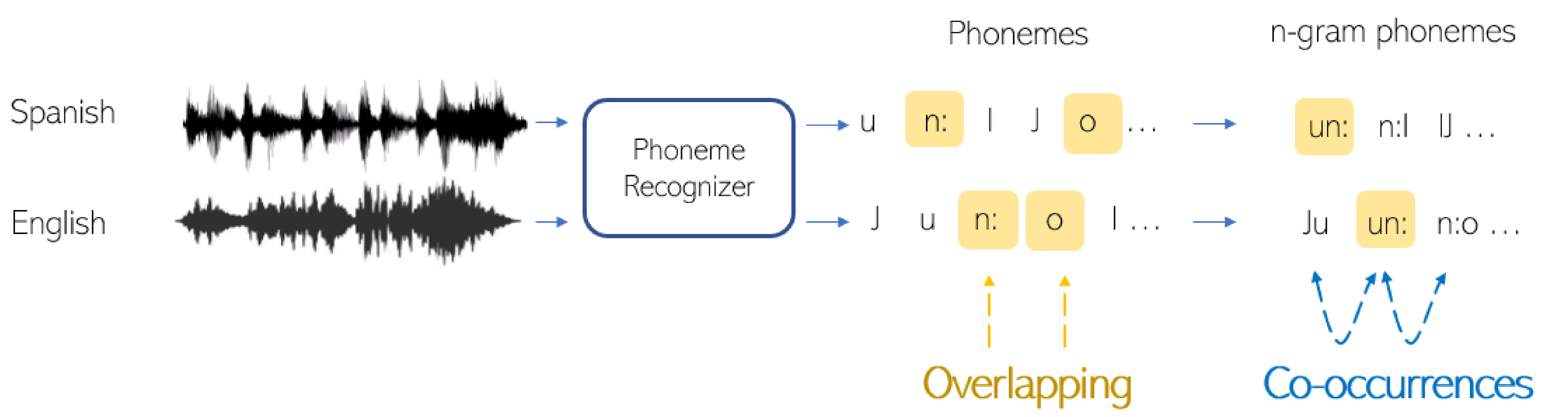
Phonotactic Language Recognition using a Universal Phoneme Recognizer and a Transformer Architecture

David Romero[1], Luis Fernando D'Haro[2], Marcos Estecha-Garitagoitia[2], Christian Salamea[1,2]
Interaction, Robotics and Automation Research Group, Universidad Politecnica Salesiana[1]
Speech Technology Group, Information and Telecommunication Center, Universidad Politecnica de Madrid[2]

Introduction

Phonotactic Language Recognition (PLRE) predicts the language spoken in a sample of speech using a sequence of phonemes, however modelling these sequences have some challenges:

- **Mismatch** between the vocabulary of the phoneme recognizers and the languages to recognize.
- **Overlapping** of phonemes and n-gram units in different languages.
- **Large sequences and scattering issues** due to high order n-grams.



Contributions

- We propose the use of **transformer encoder architecture and a language classifier on top** to perform PLRE.
- The integration of a **sliding attention window** to handle long input sequences.
- We compare the use of two **phoneme recognizers** and two **Sub-unit tokenizers** to perform PRLE.

Database

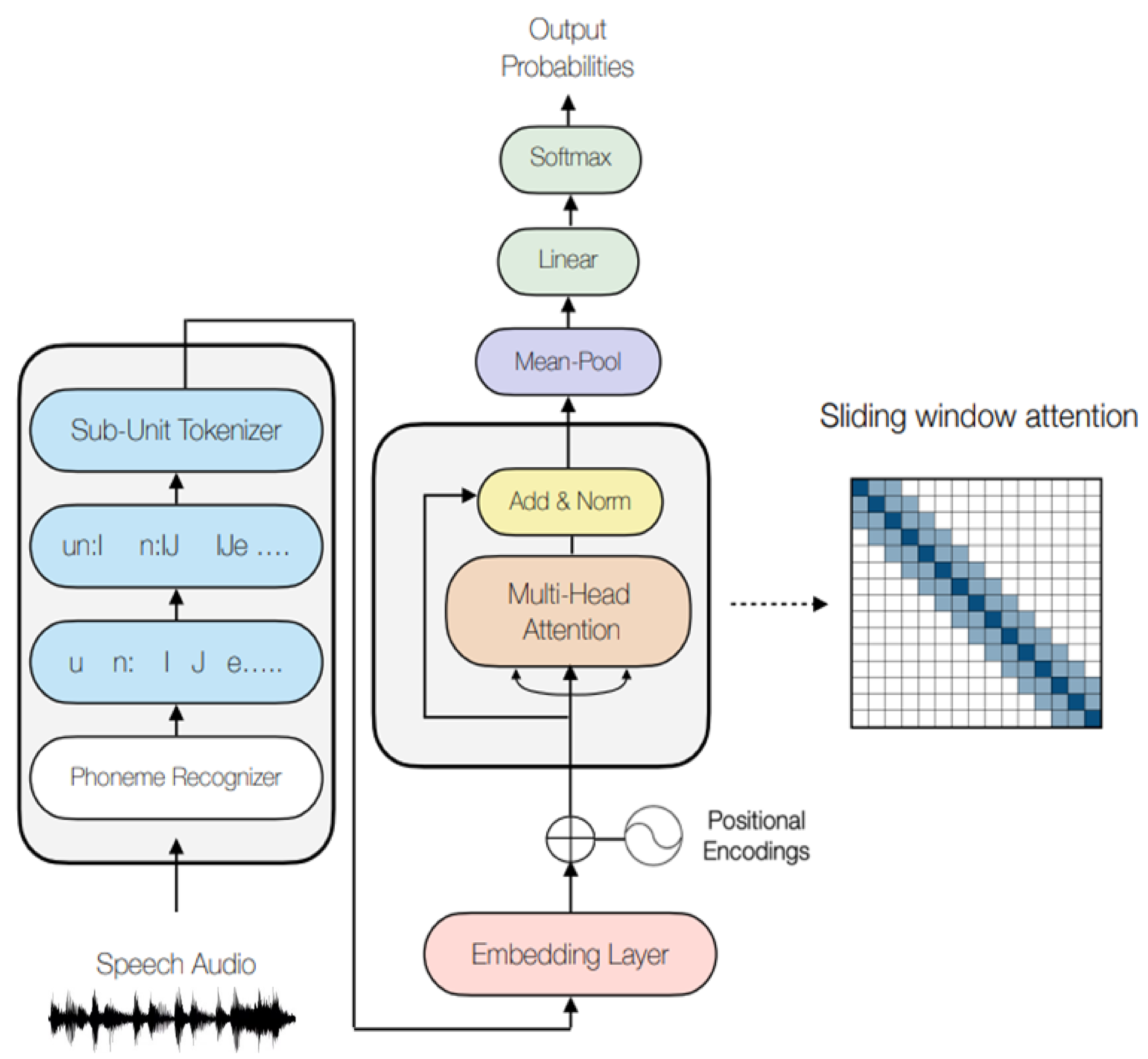
- We used the Kalaka-3 database which contains clean and noisy audio recordings for 6 highly similar languages such as Spanish, Portuguese, Galician, Catalan and others.

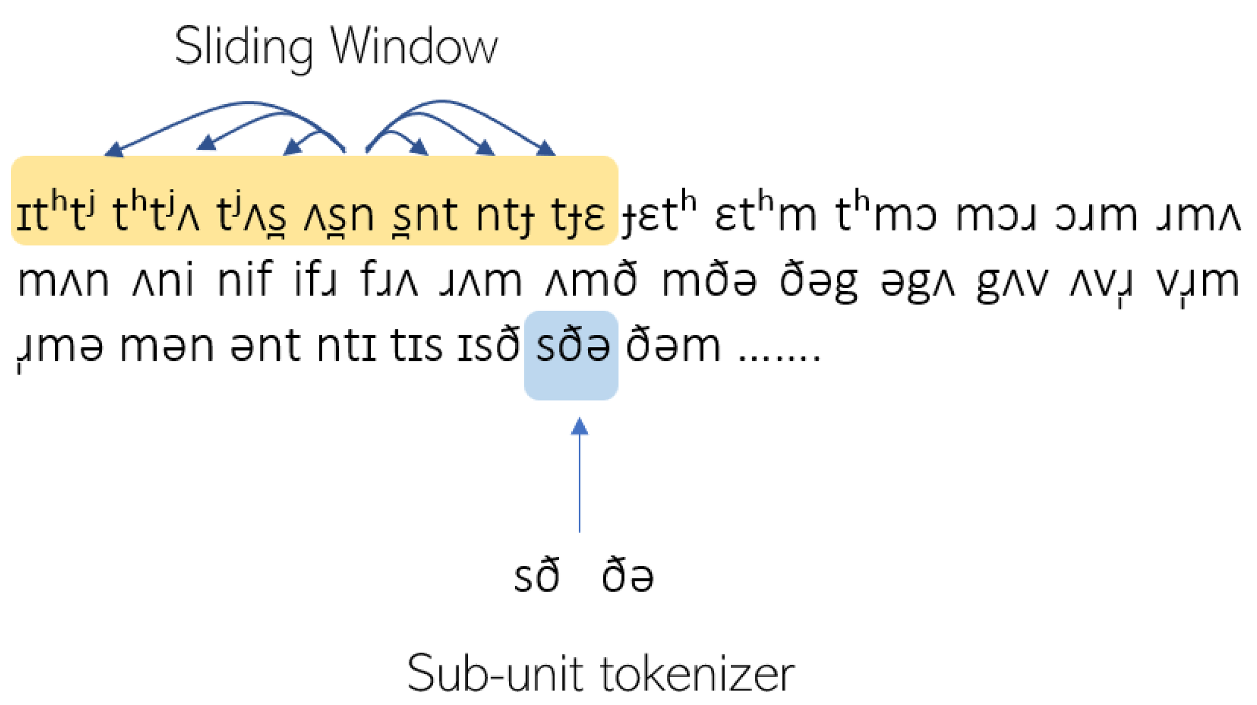
		Train	Dev	Eval
Languages	Basque	794	70	150
	Catalan	649	79	158
	English	587	81	156
	Galician	975	67	160
	Portuguese	853	84	163
	Spanish	798	77	154
Overall	Nº Files	4656	458	941
	Nº of clean files	3060	-	-
	Nº of noisy files	1596	-	-

Proposed System

Transformer-based Encoder

1. The speech signals are used as input to a phoneme recognizer, we compare the **Brno vs Allosaurus phoneme recognizers.**
2. We use 3-gram phonetic units with a sub-unit tokenizer, we compare **Byte Pair Encoding vs Word Piece** .
3. The n-gram sequences are the input of the transformer encoder which implements a **sliding window** approach.





Results

- First we compare the Brno vs Allosaurus phoneme recognizers. The **latter outperforms the Brno recognizer** in all the tested languages. The full set **(IPA) of the Allosaurus recognizer** shows the best performance compared with the other recognizers.

	Brno		Allosaurus	
	Accuracy	Cavg	Accuracy	Cavg
Hung	72.4 ± 0.43	16.5 ± 0.26	78.5 ± 0.39	12.9 ± 0.24
Czech	66.9 ± 0.63	19.8 ± 0.36	78.0 ± 0.39	13.1 ± 0.24
Russian	69.7 ± 0.68	18.0 ± 0.39	*80.7 ± 0.59*	*11.5 ± 0.36*
IPA	-	-	**85.0 ± 0.51**	**8.9 ± 0.29**

- Byte Pair Encoding and Word Piece tokenizer are compared using 3-grams phonetic sequences. **Word Piece** outperforms Byte Pair Encoding by a statistically significant improvement.

Systems	Accuracy	Imp(%)	C_{avg}	Imp (%)
Byte Pair Encoding	86.1 ± 0.26	-	8.3 ± 0.13	-
Word Piece	**86.9 ± 0.28**	0.9%	**7.7 ± 0.16**	7.2%

- Our system **outperforms the best phonotactic system in the Kalaka-3 database** , as well as our previous transformer model that did not use sliding windows nor a sub-unit tokenizer.

Systems	Devel	Eval
Phonotactic i-Vector system [27]	6.94	9.85
Transformer baseline [15]	8.42	10.21
+ sliding window (this work)	7.45	9.03
+ sliding window & tokenizer (this work)	**6.85**	**7.78**

- The fusion of our system with an acoustic model provides **complementary information**, reaching almost the same performance with the best result which combines 6 different models.

Systems	C_{avg}	Imp (%)
Acoustic MFCC [27]	6.50	-
Best fusion of 2 models [27]	5.03	-
2 acoustics + Phonotactic i-vector [27]	4.48	10.93
Best fusion with 6 models [27]	**3.52**	**49.35**
Acoustic + Transformer (this work)	3.62	47.91

Contact: