PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

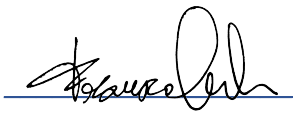
PhD Student: Luigi Maria Giordano Orsini

Cycle: XXXIX

Training and Research Activities Report

Year: First

Tutor: prof. Francesco Cutugno



Co-Tutor:

Date: October 31, 2024

1. Information:

- **PhD student: Luigi Maria Giordano Orsini**
- **DR number:**
- **Date of birth: 19/07/1989**
- **Master Science degree:Scienze Statistiche per le decisioni**
- **University: Università degli studi di Napoli Federico II**
- **Doctoral Cycle: XXXIX**
- **Scholarship type: PNRR - DM 118/2023 Mis. 4.1: Dottorati Pubblica Amministrazione**
- **Tutor: Francesco Cutugno**
- **Co-tutor: //**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Participation to the Conference "CLiC-it 2023" Ninth Italian Conference on Computational Linguistics	**Research**	**20**	**3.5**	**30/11/2023 02/12/2023**	**Università di Venezia Ca'Foscari**	**Y**
Participation to the Workshop on "Training and Evaluation Data for Italian/Multilingual LLMs"	**Research**	**8**	**1.5**	**18/12/20203**	**Università di Roma Sapienza**	**Y**
How to boost your PhD	**Courses**	**18**	**5**	**10/01/2024 07/02/2024**	**Prof. Antigone Marino**	**Y**
Using Deep Learning properly	**Courses**	**20**	**4**	**23/01/2024 08/02/2024**	**Dr.Andrea Apicella**	**Y**
Participation to the Conference "Aisv 2024" XX Convegno Nazionale dell'Associazione Italiana Studio della Voce Presenting the paper "Riconoscimento Vocale Automatico con Orientamento Sillabico: uno studio preliminare	**Research**	**20**	**3**	**01/02/2024 03/02/2024**	**Università di Tornio**	**Y**

sull'interpretabilità attraverso metodi statistici						
Human in the loop: Human oversight of AI systems *Lecturer: prof. Roberto Navigli*	**Seminar**	**2**	**0.4**	**29/04/2024**	**Centro di Eccellenza Jean Monnet AI-CoDED**	**Y**
Statistical Learning	**Courses**	**48**	**6**	**03/2024 06/2024**	**Prof. Anna Corazza**	**Y**
"AI e Linguistica" *Lecturer: prof. Francesco Cutugno*	**Seminar**	**2**	**0.4**	**08/10/2024**	**Dipartimento studi umanistici università Federico II**	**Y**
Strutture del Parlato *Lecturer: prof. Pietro Maturi e prof. Francesco Cutugno*	**Seminar**	**2**	**0.4**	**16/10/2024**	**Dipartimento studi umanistici università Federico II**	**Y**

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**0**	**5**	**0**	**5**
Bimonth 2	**9**	**0**	**3**	**0**	**11**
Bimonth 3	**0**	**0.4**	**3**	**0**	**3.4**
Bimonth 4	**6**	**0**	**6**	**0**	**12**
Bimonth 5	**0**	**0**	**1**	**0**	**1**
Bimonth 6	**0**	**0.8**	**5**	**0**	**5.8**
Total	**15**	**1.2**	**23**	**0**	**39.2**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

*During the first year of PhD I researched with the collaboration of the interdepartmental research center **Urban Eco of Università di Napoli Federico II** to implement an Automatic Speech Recognition system based on syllabic inputs (or better syllables-like units). I will give a brief background and explanation of the task and the research.*

The field of Natural Language Processing (NLP) is currently one of the most dynamic areas in scientific research. For some time, it was believed that the problem of automatic speech recognition (ASR) had been solved, mainly thanks to advances in deep learning models and the availability of large amounts of data. Many commercial systems achieved good performance in controlled environments, such as voice command recognition or speech transcription in formal and clear language situations. However, this optimistic impression turned out to be premature. A clear example is the real-time automatic generation of subtitles during speeches or conversations, a context that highlights the limitations of ASR systems. In fact, the complexity of handling such data is often underestimated, as it belongs to the continuous domain, making it difficult to segment and categorize. Additionally, the high variability within this type of data significantly complicates the pre-processing and modeling phases for speech-related tasks. The ASR task is still ongoing and presents significant challenges. Despite the progress made, speech recognition can still improve to reach a level of accuracy like human understanding in all situations. For improve that it comes the idea of exploring new approaches.

According to some studies [1, 2], in fact, human children during the development are more susceptible to recognize syllabic stimuli rather than words and phones.

Hence the idea of creating an ASR system capable of recognizing syllabic-like input in a way that is different from the systems currently in use. These analyze short, fixed frames linguistically non meaningful.

Dealing with speech is not an easy task both because of the high variability and the noise of the data, and because of its continuous nature. During the research period I started to get confidence with a field that was not very familiar to my background, so I studied the linguistic theories underlying the problem and the State-of-the-Art systems and how they work.

In the beginning I started to analyze the data structure segmenting according to an offset-offset criterion (from the end of a vowel to the other end, which resembles the CV structure of the syllable, the most common structure in the Italian language) the audio files of the Italian part of the Voxpopuli dataset using an automatic syllabifier [3] and then the obtained syllabic unit was passed through a minibatch-based K-means algorithm. The idea behind this was to obtain a clustering label used to calculate the accuracy of the ASR system.

At the same time, in fact, I developed a system based on Wav2Vec2.0 [4] to force the recognizer to work with syllabic inputs. So, I developed an Adapter and a Classifier on top and bottom of Wav2Vec2.0 creating a pipeline capable of processing variable length inputs. The Adapter is formed by a Convolutional Neural Network (CNN) while the Classifier is formed by a Feed-forward Neural Network (FNN). I chose two different strategies. First, I tried to analyze the syllabic units alone and then my main difficulty was to understand how to match the size of the vectors during all the training phases without losing too much information. Secondly, I re-concatenated all the syllabic units for each file but padded them to a specific length to maintain segmentation information, this was done to also consider the context of the speech in the analysis.

Working with CNNs gave me the opportunity to also deal with the receptive field [5] of this type of networks. In fact, the reduction of the input corresponds to an inevitable loss of information and furthermore CNNs tend to focus their attention on the central part of the vectors. To understand that well and to apply it I studied and practiced with a lot of examples trying to figure out how to keep all the information that I need.

Finally, I also implemented a system based on a transformer that no longer works with syllabic inputs but tries to discover time series and temporal patterns directly from speech, by modeling its prosody, built on the Autoformer [6], a transformer capable of detecting autocorrelation.

[1] Josiane Bertoncini, Jacques Mehler, Syllables as units in infant speech perception, Infant Behavior and Development, Volume 4, 1981, Pages 247-260, ISSN 0163-6383, https://doi.org/10.1016/S0163-6383(81)80027-6.

[2] Santolin C., Zacharaki K., Toro J.M., Sebastian-Galles N. AUTHOR FULL NAMES: Santolin, Chiara (57189647252); Zacharaki, Konstantina (57222040440); Toro, Juan Manuel (7005020671); Sebastian-Galles, Nuria (6701832236) 57189647252; 57222040440; 7005020671; 6701832236 Abstract processing of syllabic structures in early infancy (2024) Cognition, 244, art. no. 105663, Cited 1 times.
DOI: 10.1016/j.cognition.2023.105663

[3] Antonio Origlia e Francesco Cutugno. «Combining Energy and Cross-Entropy Analysis for Nuclear Segments Detection». In: Interspeech 2016. Interspeech 2016. ISCA, 8 set. 2016, pp. 2958–2962. doi: 10.21437/Interspeech.2016-1345.
url: https://www.iscas-peech.org/archive/interspeech_2016/origlia16_interspeech.html

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed e Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 22 Ott. 2020. doi: 10.48550/arXiv.2006.11477. arXiv: 2006.11477[cs,eess]. url: http://arxiv.org/abs/2006.11477

[5] Araujo, et al., "Computing Receptive Fields of Convolutional Neural Networks", Distill, 2019. DOI: 10.23915/distill.00021

[6] Haixu Wu and Jiehui Xu and Jianmin Wang and Mingsheng Long, Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting, 2022, arXiv, url: https://arxiv.org/abs/2106.13008

4. Research products:

-

5. Conferences and seminars attended

Participation to the Conference "CLiC-it 2023"
Ninth Italian Conference on Computational Linguistics

Participation to the Workshop on "Training and Evaluation Data for Italian/Multilingual LLMs"

Participation to the Conference "Aisv 2024"
XX Convegno Nazionale dell'Associazione Italiana Studio della Voce

Participation to the Seminar "Human in the loop: Human oversight of AI systems"
Lecturer: prof. Roberto Navigli

Participation to the Seminar "AI e Linguistica"
Lecturer: prof. Francesco Cutugno

Participation to the Seminar "Strutture del Parlato"
Lecturer: prof. Pietro Maturi e prof. Francesco Cutugno

6. Activity abroad:

-

7. Activity in partner companies:

Collaborating with the Museo e Real Bosco di Capodimonte for a project on an audio and operas recognizer and a dialogue system integrated in an automatic audio guide system.

- *Technical consultancy about data entry and data analysis*
- *Definition of the system pipeline*
- *Inspection of the locations and evaluation of technical feasibility*

8. Tutorship

-