ITALIC: An Italian Intent Classification Dataset

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

ITALIC: An Italian Intent Classification Dataset

Alkis Koudounas[♢], *Moreno La Quatra*[♣], *Lorenzo Vaiani*[♢], *Luca Colomba*[♢], *Giuseppe Attanasio*[♡], *Eliana Pastor*[♢], *Luca Cagliero*[♢], *Elena Baralis*[♢]

[♢]Politecnico di Torino, Turin, Italy
[♣]Kore University of Enna, Enna, Italy
[♡]Bocconi University, Milan, Italy

{name.surname}@polito.it, moreno.laquatra@unikore.it, giuseppe.attanasio3@unibocconi.it

Abstract

Recent large-scale Spoken Language Understanding datasets focus predominantly on English and do not account for language-specific phenomena such as particular phonemes or words in different lects. We introduce ITALIC, the first large-scale speech dataset designed for intent classification in Italian. The dataset comprises 16,521 crowdsourced audio samples recorded by 70 speakers from various Italian regions and annotated with intent labels and additional metadata. We explore the versatility of ITALIC by evaluating current state-of-the-art speech and text models. Results on intent classification suggest that increasing scale and running language adaptation yield better speech models, monolingual text models outscore multilingual ones, and that speech recognition on ITALIC is more challenging than on existing Italian benchmarks. We release both the dataset and the annotation scheme to streamline the development of new Italian SLU models and language-specific datasets.

Index Terms: spoken language understanding, speech recognition, human-computer interaction

1. Introduction

Spoken Language Understanding (SLU) is pivotal in enabling human-machine interaction through natural language. However, language distribution in high-quality SLU resources is skewed toward a small set of languages, particularly English [1, 2]. Prior work proposed training resources for languages other than English to let models learn language-specific phonemes, words, or phrases; however, such resources are either not specifically designed for human-machine interaction [3] or lack audio recordings in the target language, preventing end-to-end (E2E) learning from speech [4, 5].

In this paper, we introduce ITALIC, the first large-scale ITAlian Language Intent Classification audio dataset, including 16,521 audio samples spanning 18 domains, 60 intents, and recorded by 70 speakers from a variety of Italian regions. To build the collection, we extracted and annotated all Italian interactions in the MASSIVE dataset [5] and enriched them by annotating every recording with speaker- and channel-related attributes. We provide the speaker's self-declared region of origin, gender, age, and instruction level, the level of background noise, and the type of recording device used.

This metadata allows a variety of additional analyses beyond intent classification, such as speaker recognition, text-to-speech, age estimation, and linguistic variety identification.

To highlight the versatility of ITALIC and the richness of the provided metadata, we benchmark current state-of-the-art speech and text models on the intent classification and automatic speech recognition tasks. The experimental results show that 1) model scale and ASR adaptation improve the performance of speech models in terms of generalization to unseen speakers and robustness to noise, 2) monolingual text models outperform multilingual ones, and 3) zero-shot speech recognition performs worse than existing Italian benchmarks.

Our contributions are as follows:

- We introduce the first Italian large-scale intent classification dataset with recordings and transcripts of virtual assistant utterances. We enrich every instance with self-declared information about the speaker and the recording channel.
- We provide several baselines with current state-of-the-art speech and text models. Through experimental results, we show the strengths and weaknesses of such models and highlight the most promising direction to improve SLU models.
- We release the dataset, annotation scheme, and code for the baselines to encourage further research on the collection and, more broadly, on SLU in Italian.

2. Related Work

Most well-known SLU benchmark datasets are mainly in English. For example, Fluent Speech Commands [1] is an open-source SLU dataset including 31 intents and 30,043 English utterances. SLURP [2] is also an English-only dataset with single turn user interactions with a home assistant. MASSIVE [5] extends the latter by including more than one million utterances across 51 languages with annotations for Natural Language Understanding (NLU) tasks, but no non-English audio recordings. AUDIO SNIPS [6] is the audio version of the SNIPS NLU dataset. It contains both audio samples in English and French annotated with the corresponding intents.

Limited efforts have been made to develop speech and language understanding systems specifically for the Italian language. One such attempt is the AlmaWave-SLU [4], which involves the generation of an Italian data collection derived from English AUDIO SNIPS utterances through speech transcription and machine translation. However, this corpus lacks Italian audio recordings, rendering it unsuitable for Italian SLU tasks. Mozilla Common Voice [3] and Google Fleurs [7] include Italian audio recordings with transcriptions. However, both datasets do not provide intent annotations nor metadata on the recording conditions or the speakers' regional origins. EMOVO [8] collects emotional speech recordings from six native Italian speakers. IDEA [9] is a dataset for modeling dysarthric speech and includes isolated words, recorded under controlled conditions, covering a wide range of phonemes. Differently from ITALIC, these Italian datasets are not explicitly created for SLU tasks.

10.21437/Interspeech.2023-1980

Table 1: *Gender and age distribution in ITALIC.*

Gender		Age			
Female	Male	[18-25]	[26-40]	[41-55]	≥56
42.96%	57.04%	10.63%	63.86%	10.78%	14.73%

3. The ITALIC Dataset

3.1. Data collection

The ITALIC dataset was crowdsourced through a custom web platform. Both native and non-native Italian speakers participated. We required participants to record themselves while reading a short instruction randomly sampled from the MASSIVE [5] dataset. The latter consists of utterance transcripts and associated intents. We use the transcripts as prompts for crowd workers to read out and record. We do not annotate intents and instead use those supplied by MASSIVE. After giving the participants a list of annotation guidelines, we did not intervene in the process or supervise the recording sessions. They used their own devices and chose freely when and how to contribute recordings.

Optionally, participants could declare additional information about themselves and recording conditions through an anonymous registration form. Specifically, we asked for age (a numerical integer), gender (male, female, non-binary, undeclared), region of origin in Italy, country of origin (if not Italian), instruction level, and presence of any speech impairment (e.g., lisp or stuttering). Recording conditions include the input device adopted (laptop, smartphone, or headphones) and environmental noise level (no noise to very noisy).Such rich additional metadata will enable future per-group analyses, e.g., studying service quality of IC models over Italian regions.

To ensure high-quality annotations, at least two individuals reviewed each sample. We consider a sample valid if 1) the utterance is intelligible from the recording and 2) it is coherent with the provided prompt. We validated the entire set of recordings as follows. Once all samples were annotated, we ran a first validation round considering the entire collection. Then, we extracted all non-valid samples and ran a second annotation and validation round considering only those samples. We repeated the process until no non-valid samples were left.

3.2. Data Characterization

We extracted and annotated every sample in the Italian split of MASSIVE [5], for a total of 15.46 hours of recording and 60 different intents. The final dataset consists of 16,521 audio recordings by 70 distinct volunteers. All but one self-identified as native Italian speakers. Only four speakers declared some speech impairment. Native speakers are distributed across 13 Italian regions, which concentrate a large number of linguistic and diatopic variations [10]. Table 1 reports the distribution of gender and age in the dataset, while Figure 1 shows the distribution over the region of origin.

The audio length goes from 1.14 to a maximum of 38.34 seconds and an average of 3.37 seconds. We WAV-encoded recordings with a sampling frequency of 16 kHz.

Dataset splits. For experimental and reproducibility purposes, we release three splitting configurations of increasing difficulty[1]:

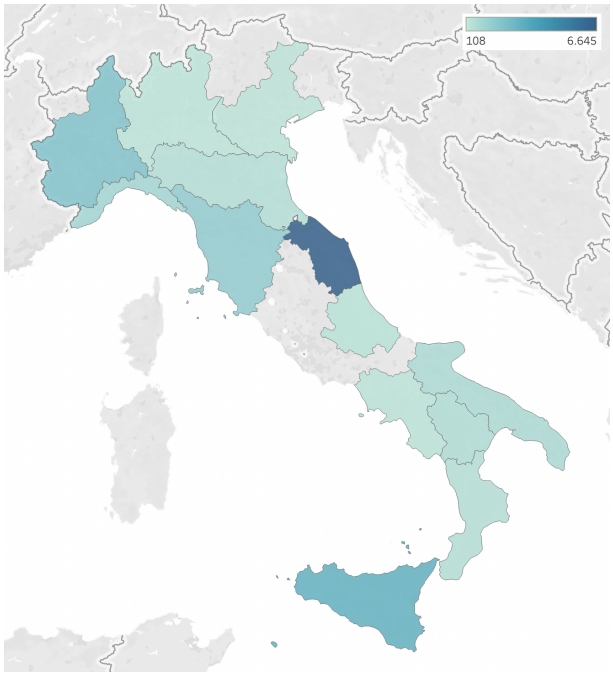

Figure 1: *Distribution of utterances across Italian regions. Darker colors represent higher absolute counts.*

Table 2: *Dataset statistics.*

Configuration		# Utterances	# Hours	# Speakers
massive	train	11514	10.80	69
	validation	2033	1.90	68
	test	2974	2.76	69
speaker	train	13123	12.33	56
	validation	1957	1.67	7
	test	1441	1.46	7
noisy	train	13742	12.93	69
	validation	1526	1.44	66
	test	1253	1.09	9

- *massive*: It uses the official training and test splits of MASSIVE [5] and includes all our speakers. Furthermore, the distribution of noisy utterances is randomly apportioned between the two partitions.
- *speaker*: We stratify on speakers, i.e., all recordings of a speaker belong to either the training, validation, or test split. This division will help to assess whether the models generalize to unseen speakers;
- *noisy*: The test set consists exclusively of data annotated with the highest background noise level, while the train and validation sets contain either noiseless or data with a low noise level, randomly split.

Table 2 reports statistics on the number of utterances, speakers, and hours of recordings for each splitting configuration.

[1]The dataset, splits, and annotation scheme are publicly available at `https://github.com/RiTA-nlp/ITALIC`.

3.3. Tasks

ITALIC enables a range of SLU and Natural Language Understanding (NLU) tasks. In this paper, following the original setup in MASSIVE [5], we provide baselines for the intent classification and automatic speech recognition tasks.

Given the rich set of annotations we release, practitioners may use ITALIC in other tasks, such as speaker identification, text-to-speech, age estimation, or region of origin identification. We leave the analysis in such contexts to future work.

4. Experiments

In this section, we present the experimental setup and results for evaluating the performance of various models on the ITALIC dataset. We focus on intent classification and automatic speech recognition. Our goal is to leverage the rich annotations in ITALIC to evaluate state-of-the-art models in terms of accuracy, robustness to noise, and generalization to unseen speakers. Additionally, this investigation seeks to explore the influence of two key factors, namely the knowledge of the Italian language and the variability of recording conditions, including variations in noise levels and speaker characteristics.

4.1. Experimental Setting

Models. We use established E2E transformer-based models for both IC and ASR setups. We address IC using either raw audio signals or text transcripts. As speech SLU models, we test several XLSR variants, a speech representation network pretrained on 53 [11] or 128 [12] languages. We conduct experiments on two model sizes, i.e., 300M and 1B parameters, testing XLSR-53 300M and XLSR-128 300M and 1B. We build two additional baselines by testing language adaptation on out-of-domain data. In practice, we employ two models, i.e., XLSR-53 300M and XLSR-128 1B, that have been fine-tuned via ASR on the Italian split of the Mozilla Common Voice dataset [3]. As text NLU models, we include multilingual BERT [13], multilingual BART [14], and two variants of these models pre-trained on Italian data only [15, 16]. Note that both models include Italian as a pretraining language.

For the ASR task, we use Whisper [17] in three variants: small (244M parameters), medium (769M), and large (1.5B).

We fine-tune each IC model attaching to the encoder architecture a final classification layer.[2] We use model implementation and weights from the transformers library [18] in all experiments.

Data. We fine-tune models on the training split of one of the configurations of ITALIC, i.e., *massive*, *speaker*, or *noisy*, and report the result on the test set.

Evaluation Metrics. We measure accuracy and F1 Macro scores for IC and the Word Error Rate (WER) and Character Error Rate (CER) for ASR.

Hyperparameter Setup. We ran a manual hyperparameter search and followed fine-tuning procedures and guidelines from relevant literature.

We provide detailed information about the models used for the evaluation, the hyperparameter setup, and the fine-tuning procedure in the official project repository.[1]

[2] In our experiments on intent classification, we use a classification layer on top of the BART encoder.

Table 3: *Accuracy and F1 Macro results of E2E-SLU models and adapted variants (FT: ✓). Best result per splitting configuration in bold.*

Split	Model	# params	FT	Accuracy	F1
massive	XLSR-128	300M		76.16	76.11
	XLSR-128	1B		77.07	77.08
	XLSR-53	300M	✓	81.34	81.31
	XLSR-128	1B	✓	**83.39**	**83.25**
speaker	XLSR-128	300M		73.42	73.04
	XLSR-128	1B		79.11	79.08
	XLSR-53	300M	✓	83.69	83.62
	XLSR-128	1B	✓	**84.18**	**84.05**
noisy	XLSR-128	300M		78.29	78.21
	XLSR-128	1B		76.48	76.06
	XLSR-53	300M	✓	81.01	80.94
	XLSR-128	1B	✓	**82.20**	**82.43**

Table 4: *Accuracy and F1 Macro results of text NLU models with multilingual (PT: M) and monolingual (PT: I) pretraining for the massive configuration. Best result in bold.*

Model	# params	PT	Accuracy	F1
BART	611M	M	87.16	83.53
BERT	167M	M	86.21	82.93
BART	141M	I	86.65	83.82
BERT	110M	I	**88.43**	**85.57**

4.2. Results on Intent Classification

E2E SLU. Table 3 reports the result of speech SLU models on the ITALIC dataset. As expected, adaptation to Italian via ASR fine-tuning yields better models, with the adapted XLSR-128 1B model being the best across all splitting configurations. Relative to the non-adapted version, XLSR-128 1B achieved +6.17 and +6.37 F1 points on the *massive* and *noisy* configurations, respectively. Except for one case (XLSR-128, *noisy*), the results also suggest that larger models yield better results, although not as much as adapting models to Italian.

Notably, all models, except XLSR-128 300M, achieved higher performance on the challenging *speaker* configuration compared to *massive*, indicating their generalizability to different speakers and their ability to handle variations in speaking styles and accents. As expected, all models showed lower performance on the *noisy* subset, with the only exception of the XLSR-128 300M model. This finding highlights the impact of recording conditions on model performance, motivating the need for more resources and training procedures that closely match real-world scenarios.

NLU. Table 4 reports the results of addressing IC with text NLU models on the *massive* splitting configuration. We do not test text models on *speaker* and *noisy* as these configurations are specifically customized for speech models and tasks.

Of particular interest is the performance of the Italian pretrained BERT model, which, despite having fewer parameters, exhibits superior performance compared to the BART and BERT models pre-trained on multilingual data. Utilizing Italian data can yield valuable improvements in the performance of NLU models, even though the performance gap between the

Table 5: *WER and CER results of Whisper models in a zero-shot setup (S: ZS) and adapted variants (S: FT). Best result per splitting configuration in bold.*

Split	Model	# params	S	WER	CER
massive	large	1.5B	ZS	11.46	5.01
	small	244M	FT	4.82	1.49
	medium	769M	FT	3.41	0.92
	large	1.5B	FT	**3.06**	**0.82**
speaker	large	1.5B	ZS	8.65	3.93
	small	244M	FT	3.81	0.99
	medium	769M	FT	2.92	0.70
	large	1.5B	FT	**2.74**	**0.61**
noisy	large	1.5B	ZS	15.41	7.67
	small	244M	FT	8.46	2.95
	medium	769M	FT	5.83	1.92
	large	1.5B	FT	**5.29**	**1.70**

models is less pronounced. Our study demonstrates that, particularly for encoder-based models such as BERT, smaller models pre-trained on Italian data can achieve comparable or even superior performance compared to their larger multilingual counterparts.

Since two human inspectors have validated all ITALIC samples, we can safely assume that voice recordings closely match the text transcripts. We can then impute any difference in performance across modalities, i.e., speech and text, to the inherent difficulty of learning from the two kinds of raw data. Interestingly, we note that the best text model, monolingual BERT 110M, achieves +2.32 F1 points, marking a gap despite having x9 fewer parameters. We can draw similar conclusions comparing other pairs of speech and text models. These findings underscore the importance of a dedicated dataset for SLU tasks for better interpretation of the spoken language.

4.3. Results on Automatic Speech Recognition

Although the ITALIC dataset was not specifically designed for ASR, its various speaker and recording conditions make it a valuable resource for analyzing the performance of Italian ASR models. For this task, we use the large Whisper model [17] in zero-shot settings and its fine-tuned version for Italian ASR in three different sizes: small, medium, and large. The results are presented in Table 5.

Our investigation reveals that all the evaluated models performed well, with low WER and CER scores (the estimated human WER is approximately 4% [19]). Whisper large is best across the board, further highlighting the importance of model size. As expected, there is a clear performance degradation on the *noisy* configuration, more marked for smaller models.

Applying Whisper to zero-shot speech recognition yields the worst performance on all splits, with a gap from the fine-tuned variant (Whisper large, S: FT) largest in *noisy* and smallest in *speaker*. In absolute terms, it achieves a WER of 8.65 on the *speaker* configuration and a much worse 15.41 on *noisy*. These results are sensibly worse than standard Italian benchmarks such as Mozilla CV [3] (WER: 7.1) or Google Fleur [7] (WER: 4.0).

These findings prove ITALIC challenging for current state-of-the-art SLU and NLU models and underscore its importance as a novel Italian resource. With accurate recordings of real-world human-to-voice assistant interactions and rich annotations, ITALIC paves the way for new research and development of Italian models.

5. Conclusions

We presented ITALIC, the first large-scale Italian audio dataset specifically designed for intent classification. The collection is comprehensive of text transcripts, recordings, and additional metadata about the speaker and the recording channel. We evaluated the performance of current state-of-the-art speech and text models on the intent classification and automatic speech recognition tasks, demonstrating the impact of model selection and pretraining on performance. We release the dataset, annotation schema, and code to foster future research in this area.

Our future work includes expanding the ITALIC dataset to enhance its diversity and representativeness of the Italian language and exploring new tasks that can be tackled with this dataset. Future enhancements will involve adding more speakers with diverse backgrounds, including non-native speakers, and further extending the dataset to address any potential gap in coverage. We also aim to develop a large-scale multilingual data collection platform to facilitate the creation of similar datasets in other languages.

6. Limitations

The ITALIC dataset is valuable for evaluating models for the Italian SLU and ASR tasks. However, some limitations must be taken into account when interpreting the results. While the dataset includes recordings from a wide range of Italian regions, it only partially represents all dialects and linguistic varieties. Additionally, the dataset is mainly composed of recordings from native Italian speakers, which may not be representative of scenarios where the user of a voice assistant has a non-native accent. We envisioned the dataset to represent a broad spectrum of individuals, from non-binary to speakers with speech impairments—however, only a limited number of volunteers identified as such. We will promote future dataset releases capturing more speech nuances. Finally, the dataset only includes one recording per sentence. Including multiple recordings of the same sentence by different speakers would allow a more comprehensive evaluation of model performance, which is of key importance for SLU domain [20].

Overall, the ITALIC dataset provides a strong foundation for evaluating Italian SLU and ASR models; addressing these limitations will enable more comprehensive evaluations and further advances in these fields.

7. Acknowledgments

This project is a joint effort of members of the "Risorse per la Lingua Italiana" open community. We would like to thank all the crowd workers who participated to our campaign and the reviewers for their helpful comments. This work is partially supported by the FAIR - Future Artificial Intelligence Research (PE00000013) and the spoke "FutureHPC & BigData" of the ICSC - "National Centre for HPC, Big Data and Quantum Computing" funded by the European Union Next-GenerationEU, and SmartData@PoliTO center on Big Data and Data Science. This manuscript reflects only the authors views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

8. References

[1] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 814–818. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2396

[2] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 7252–7262. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.588

[3] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[4] V. Bellomaria, G. Castellucci, A. Favalli, and R. Romagnoli, "Almawave-slu: A new dataset for SLU in italian," *CoRR*, vol. abs/1907.07526, 2019. [Online]. Available: http://arxiv.org/abs/1907.07526

[5] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan, "Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages," 2022. [Online]. Available: https://arxiv.org/abs/2204.08582

[6] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *CoRR*, vol. abs/1805.10190, 2018. [Online]. Available: http://arxiv.org/abs/1805.10190

[7] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2022.

[8] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3501–3504. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/591_Paper.pdf

[9] M. Marini, M. Viganò, M. Corbo, M. Zettin, G. Simoncini, B. Fattori, C. D'Anna, M. Donati, and L. Fanucci, "Idea: An italian dysarthric speech database," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 1086–1093.

[10] M. Maiden and M. Parry, *The dialects of Italy*. Routledge, 2006.

[11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[12] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[14] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[15] S. Schweter, "Italian bert and electra models," Nov. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4263142

[16] M. La Quatra and L. Cagliero, "Bart-it: An efficient sequence-to-sequence model for italian text summarization," *Future Internet*, vol. 15, no. 1, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/1/15

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[19] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639397000216

[20] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and D. Amberti, "Exploring subgroup performance in end-to-end speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.