Towards a responsible usage of AI-based Large Acoustic Models for Automatic Speech Recognition: on the importance of data in the self-supervised era.

Vincenzo Norman Vitale[1,2,*,†], Emilia Tanda[1,†] and Francesco Cutugno[1,2,*,†]

[1]University of Naples, Federico II, Corso Umberto I, 40, Naples, 80138, Italy
[2]UrbanECO Research Center, University of Naples, Federico II, via Tarsia, 31, Naples, 80134, Italy

Abstract

The evolution of artificial intelligence models has made them tools of everyday use in many fields. However, the enormous capabilities demonstrated by these models have, on the one hand, some apparent costs in terms of money, computational resources, or data. On the other hand, there are some hidden costs for end users who rely on models trained by third parties, sacrifice awareness and control of a tool, and try to evaluate its performance in their specific contexts. This is the case of supervised End-to-End (E2E) ASR systems and self-supervised E2E-ASR, also referred to as Large Acoustic Models (LAM). On the one hand, they provide an important starting point for building information systems oriented to speech interaction and, on the other hand, are complex to evaluate, use and adapt in specific contexts.

Keywords

End-to-End ASR, self-supervised, quality of data, communication style, responsible AI

1. Introduction

Modern Automatic Speech Recognition (ASR) systems, among the other Natural Language Processing (NLP) systems, achieve remarkable performances thanks to the computing potential enabled by Deep Neural Networks (DNN). Indeed, over the last decade, the automatic speech recognition community has made great strides [1, 2, 3], moving from traditional hybrid modelling (Acoustic Model+Language Moel) to end-to-end (E2E) modelling that directly translates an input speech sequence into a sequence of output tokens using a single network, leading to self-supervised E2E models, also referred to as Large Acoustic Models (LAMs), that can model speech without the aid of labelled data. These revolutionary innovations have completely subverted the traditional architectures of ASR systems used in previous decades. In addition, there has also been a strong impact on the cost-effectiveness and democratization of ASR systems. On the one hand, the change in architecture has made it more economical to collect and create the data sets necessary for training, which previously required the use of a large number of experts in the field of speech analysis involved in long and expensive processes of manual labelling. On the other hand, it has allowed the creation of a large number of freely available and open-source general-purpose ASRs, bringing these systems within the reach of a greater number of institutions and companies. However, their use remains limited due to the lack of benchmarks oriented towards specific contexts and communication styles. In this work, we will analyze the evolution of ASR systems, how the nature of the data used for their training has changed, and the limitations of modern ASR systems. Finally, we will propose an initiative aimed at collecting high-quality data in Italian aimed at both performance verification and training based on specific communicative styles.

2. The evolution of ASR systems

ASR systems have been the subject of several revolutions, which have impacted their internal architecture and the nature of the data employed for their training. Traditional ASR systems rely on two separate components [9]: The Acoustic Model (AM), which is aimed at converting the voice signal into a sequence of phones, and the Language Model (LM), aimed at transforming the sequence of phones received from the AM, in the most likely and reliable transcription. These two models were initially realised with techniques such as Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM). Then, with the advent of Deep Neural Networks (DNN), both have been realized as supervised DNNs. Still, the output of both components was the same: the AM produces the most likely sequence of phones given the input voice signal, while the LM provides the most reliable transcription

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
vincenzonorman.vitale@unina.it (V. N. Vitale); e.tanda@studenti.unina.it (E. Tanda); cutugno@unina.it (F. Cutugno)
0000-0002-0365-8575 (V. N. Vitale); 0000-0001-9457-6243 (F. Cutugno)

Model	Type	Train Data	Year	Decoder	Encoder	Test-clean/other
Transformer [4]	E2E	970Hr Transcribed	2020	RNN-T	Transformer	2.0/4.6
Conformer [5]	E2E	970Hr Transcribed	2020	RNN-T	Conformer	1.9/3.9
Wav2vec2.0 [6] with Quantization	E2E Self-Supervised	60KHr Untranscribed 100Hr Transcribed	2020	CTC	Transformer	1.8/3.3
HuBERT [7] with KNN	E2E Self-Supervised	60KHr Untranscribed 100Hr Transcribed	2021	CTC	Transformer	1.8/2.9
W2V-BERT [8] with Quantization	E2E Self-Supervised	60KHr Untranscribed 100Hr Transcribed	2021	RNN-T	Conformer	1.4/2.5

Table 1
In table are E2E ASR systems performance based on Librispeech test-set, least recent to the most recent. For self-supervised systems is also reported the algorithm used during the self-supervised pre-training phase.

given the input sequence of phones. This means that the two components had separate objectives and relied on different kinds of high-quality and costly datasets. On the one hand, the AM needs well-aligned sound-to-phone transcriptions. On the other hand, the LM needs a statistically representative set of phone-to-word samples in order to provide meaningful transcription. Providing adequate quality data requires highly specialised professionals to hand label in both cases. This type of ASR system requires tens to hundreds of hours of speech to train the AM and a few million words to train an LM (depending on the context). The aim is to transcribe fairly long sentences with an accuracy linked to specific application contexts.

The turning point that led to the recent End-to-End ASR (E2E-ASR) [2] was the introduction of the Transformer [10] network architecture, on which most actual AI models rely. Compared to traditional systems, in E2E-ASRs, the voice signal is directly converted into its corresponding transcription without any intermediate, human-readable format. This evolution results in systems with a single objective needing only one cheaper dataset to be trained since the intermediate phones transcription and the alignment parts have been removed. The Transformer architecture [10] opens up the possibility of building a combination of AM and LM, now referred to as the Encoder and Decoder, which directly maps an unaligned sequence of sounds to its transcription. With a few hundred hours of non-aligned transcribed speech through a supervised learning process, E2E-ASR systems outperform the previous generation on average by providing an error of up to 5% in the case of pure Transformer-Encoder systems, or up to 4% in the case of Conformer-Encoder systems [5] (see table 1 for performance). Clearly, the Decoder module implementation choice strongly impacts E2E ASR performances, such module is usually implemented as a Connectionist Temporal Classification (CTC) model [11] or as a Recurrent Neural Network Transducer (RNN-T) [12]. CTC is a non-auto-regressive speech transcription technique which collapses consecutive, all-equal, transcription labels (character, word piece, etc.) to one label unless a special label separates these. The result is a sequence of labels shorter or equal to the input vector sequence length. The CTC is one of the most diffused decoding techniques. As non-auto-regressive, it is also considered computationally effective as it requires less time and resources for training and inference phases. Conversely, the RNN-T (also named *Transducer*) is an auto-regressive speech transcription technique which overcomes CTC's limitations, i.e., non-auto-regressive and limited label sequence length. An RNN-T is a speech transcription technique which can produce label-transcription sequences longer than the input vector sequence and models long-term transcription elements' inter-dependency. A Transducer typically comprises two sub-decoding modules: one that forecasts the next transcription label based on the previous transcriptions (prediction network); the other that combines the encoder and prediction-network outputs to produce a new transcription label (joiner network). These features improve transcription speed and performance with respect to CTC at the expense of more training and computational resources required [12].

Finally, the most recent advancement consists of the employment of self-supervised training techniques, giving rise to what could be defined as the first truly End-to-End ASR, namely Wav2Vec2 [6] and after a while to HuBERT [7], both are also referred to as Large Acoustic Model (LAM) [13] because of their training process which usually involves two main phases. The first one is the pre-training phase, during which vast amounts of untranscribed speech data are employed in order to recognize and discretize hidden acoustic units' representations by employing different processes such as quantization (Wav2Vec2[6]) directly from the raw audio sample, or clustering (HuBERT [7]) on MFCC features. Then, during the last phase, a transcription module could be trained on smaller datasets (few hours) in order to obtain an error rate of about 2% (see table 1).

3. Self-supervised E2E Solutions (?) to data shortages

Undeniably, by committing the model to learn all parts automatically, E2E-ASRs overcome the difficulties and cost-ineffectiveness of the data preparation and modelling phases of conventional systems, while requiring far more training data [14]. This shift significantly impacted ASR systems; on the one hand, it significantly reduced training data costs while increasing their volume, as shown by the availability of plenty of general-purpose training datasets [1, 3]. On the other hand, in spite of the cheapness of training data, ASR systems are now accessible to a wider public. Clearly, these innovations present some expenses, which in this case consist of higher computational costs, longer training times, and loss of modularity [3] compared to traditional ASR systems. Indeed, *adapting* such a general-purpose E2E-ASR to specific contexts means, in some cases, updating the Decoder (LM) to a special-purpose field or updating the Encoder (AM) to handle a special type of speech, which requires fine-tuning and, in the worst cases, training the model from scratch.

Then, the advent of Self-supervised systems impacted the adaptability aspects of general-purpose E2E ASR, giving rise to Large Acoustic Models (LAMs), which basically are Encoders trained on vast amounts of non-transcribed cheaper datasets, compared to data needed by simple E2E-ASR, which are then combined with an Encoder part trained on small quantities of language-specific transcribed data. The result is a large, general-purpose model that can be easily deployed in most contexts. Although they are publicly available and, therefore, freely adaptable, the necessary computational resources are so prohibitive that they are within the reach of a few companies and institutions, even for simple fine-tuning.

A further point to be considered is that the advantages of both simple E2E ASR and Self-supervised ones come at the expense of lower interpretability of systems' internals, making it difficult to diagnose errors and limiting their usage in critical contexts [3]. However, some studies in the field of eXplainable AI (XAI)[15] try to provide explanations and methodologies for analysing behaviours and phenomena modelled by various E2E ASR systems, aiming to make them more interpretable [16, 17, 18, 19], still based on special purpose data.

To summarize, although the innovations introduced by E2E and self-supervised E2E systems have allowed their fast diffusion, still their industrial and institutional deployment remains subject to limitations [3] which, in some cases, are strongly related to special-purpose data availability. Indeed, employing a general-purpose E2E ASR system in a specific domain requires evaluation and potential fine-tuning /training on domain-specific data, which is usually unavailable. Another aspect to consider is how and to what extent the democratisation of ASR systems has been impacted. In fact, if, on the one hand, it is possible to obtain much more data for the same cost, on the other hand, the same quantity of resources is no longer sufficient, especially for training purposes.

4. High-quality data for context-specific assessment

Clearly, the availability of good-quality and well-categorized data is paramount in the current application landscape. On the one hand, such data is essential to evaluate pre-trained systems in specific contexts with speaking styles related to different communication situations. On the other hand, such data is crucial for training and fine-tuning modern supervised and self-supervised E2E ASR. To this end, the Phoné consortium was born as a voluntary initiative to collect, verify and distribute transcribed and non-transcribed Italian speech datasets in various application contexts. Table 2 shows the actual amount of data collected and verified by the consortium to provide Italian institutions and companies with adequate instruments to evaluate these promising tools, which are, however, assessed in contexts and communication styles that do not reflect the target ones.

Currently, data is divided into two macro-categories, namely, Transcribed and Untranscribed, to enable the future training of self-supervised E2E-ASR. Then, datasets are further divided into specific communication styles [20, 21]:

- **Monologic** speech involves only one person speaking without interacting with an interlocutor. This type of speech is characterized by consistency and structuring, as it typically consists of lectures, speeches or situations that require preliminary preparation. As a result, the speech appears cohesive and well-organized. The language register tends to be higher and more formal.
- **Dialogic** speech involves two or more people in a conversation, characterized by exchanges of messages and information. It is thus configured as a communicative act with a dynamic structure. Unlike monologic speech, dialogic speech does not involve prior preparation; therefore, the speech tends to be simpler from a syntactic point of view, the articulation of words tends to be less precise (hypoarticulation), and it is also characterized by greater conciseness of expression.
- In **Read** speech, the speaker reads a written text aloud (as in the case of audiobooks), therefore this type of speech is characterized by clear pronunciation (there is a tendency towards hyperarticulation), complete syntax and greater coherence and

cohesion of the text. Furthermore, another feature is given by the modulation of reading speed and the use of strategic pauses and intonations to improve communicative effectiveness.

Material Type	Speech Type	Minutes
Transcribed	Monologic	500 Minutes
Transcribed	Dialogic	400 Minutes
Transcribed	Read	120 Minutes
Untranscribed	Monologic	10000 Minutes
Untranscribed	Dialogic	500 Minutes
Untranscribed	Read	2200 Minutes

Table 2
List of material collected and verified for the evaluation and training of E2E-ASR systems (both supervised and self-supervised) in specific contexts for the Italian language.

Behind ASR-related aspects, the consortium's purposes also extend to other voice-related tasks, which include, but are not limited to, Text-To-Speech (TTS), Speaker Identification (SI), Speaker Verification (SV), and others.

5. Conclusion

In this work, we present the current panorama related to E2E ASR systems, how their data usage evolved along with technological improvements and the current issues that these improvements solved or introduced. Firstly, we observe the significant improvement in models' performances while pointing out issues connected to the models' capacity assessment related to specific communication styles and domains. We observe the shift in model training costs, moving away from data becoming cheaper and easier to collect towards computing resources growing in quantity and costs. Then, we observed how the advantages introduced by modern E2E (supervised and self-supervised) ASRs come at the expense of an increase in their complexity, which consequently reduces their interpretability. Finally, we propose a voluntary, high-quality data collection initiative to evaluate and train systems related to various speech communication styles to enable more informed use and greater accessibility of E2E-ASR systems.

References

[1] M. Malik, M. K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey, Multimedia Tools and Applications 80 (2021) 9411–9457.

[2] J. Li, et al., Recent advances in end-to-end automatic speech recognition, APSIPA Transactions on Signal and Information Processing 11 (2022).

[3] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: A survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[4] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, S. Kumar, Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7829–7833.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., Conformer: Convolution-augmented transformer for speech recognition (2020).

[6] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 3451–3460.

[8] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, Y. Wu, W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 244–250.

[9] S. Karpagavalli, E. Chandra, A review on automatic speech recognition architecture and approaches, International Journal of Signal Processing, Image Processing and Pattern Recognition 9 (2016) 393–404.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[11] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[12] A. Graves, Sequence transduction with recurrent neural networks, arXiv preprint arXiv:1211.3711 (2012).

[13] L. M. Giordano Orsini, V. N. Vitale, F. Cutugno, Large scale acoustic models: A new perspective, Sistemi intelligenti (2023). doi:10.1422/108137.

[14] G. Coro, F. V. Massoli, A. Origlia, F. Cutugno, Psycho-acoustics inspired automatic speech recognition, Computers & Electrical Engineering 93 (2021) 107238.

[15] D. Gunning, Explainable artificial intelligence (xai), Defense advanced research projects agency (DARPA), nd Web 2 (2017) 1.

[16] A. Prasad, P. Jyothi, How accents confound: Probing for accent information in end-to-end speech recognition systems, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3739–3753.

[17] D. Ma, N. Ryant, M. Liberman, Probing acoustic representations for phonetic properties, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 311–315.

[18] A. Pasad, J.-C. Chou, K. Livescu, Layer-wise analysis of a self-supervised speech representation model, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 914–921.

[19] V. N. Vitale, F. Cutugno, A. Origlia, G. Coro, Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique, Neural Computing and Applications (2024) 1–27.

[20] M. Nakamura, K. Iwano, S. Furui, Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance, Computer Speech & Language 22 (2008) 171–184.

[21] P. Azizova, Linguistic analysis and learning of dialogical speech in literary texts, JETT 14 (2023) 86–94.