Modelling filled particles and prolongation using end-to-end Automatic Speech Recognition systems: a quantitative and qualitative analysis.

Vincenzo Norman Vitale[1,†], Loredana Schettino[2,†] and Francesco Cutugno[1]

[1]University of Naples Federico II, Naples, Italy
[2]Free University of Bozen-Bolzano, Bozen, Italy

Abstract

State-of-the-art automatic speech recognition systems based on End-to-End models (E2E-ASRs) achieve remarkable performances. However, phenomena that characterize spoken language such as fillers (<eeh> <ehm>) or segmental prolongations (the<ee>) are still mostly considered as disrupting objects that should not be included to obtain optimal transcriptions, despite their acknowledged regularity and communicative value. A recent study showed that two types of pre-trained systems with the same Conformer-based encoding architecture but different decoders – a Connectionist Temporal Classification (CTC) decoder and a Transducer decoder – tend to model some speech features that are functional for the identification of filled pauses and prolongation in speech. This work builds upon these findings by investigating which of the two systems is better at fillers and prolongations detection tasks and by conducting an error analysis to deepen our understanding of how these systems work.

Keywords

disfluences, speech recognition, probing, interpretability, explainability

1. Introduction

In recent works on Automatic Speech Recognition (ASR) systems based on the computing power of Deep Neural Networks (DNN), a great deal of effort is focused on incrementing the systems' performances by employing increasingly complex, hence hardly interpretable, DNN models that require huge amounts of data for the training, like End-to-End Automatic Speech Recognition (E2E-ASR) models which represent the state-of-the-art. An E2E-ASR model directly converts a sequence of input acoustic feature vectors (or possibly raw audio samples) into a series of graphemes or words that represent the transcription of the audio signal [1], as represented in figure 1. In contrast, traditional ASR systems typically train the acoustic, pronunciation, and language models separately, requiring distinct modelling and training for each component. These systems usually aim to obtain speech transcriptions 'cleaned'from phenomena that characterise spoken language such as discourse markers, particles, pauses, or other phenomena commonly referred to as 'disfluencies'. Studies on the interpretability of the dynamics underlying neural models showed

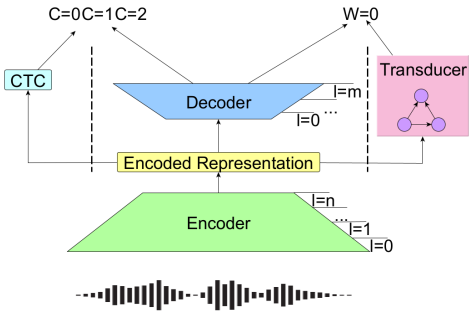


Figure 1: E2E ASRs are based on an encoder-decoder architecture. The speech signal is fed to the encoder, producing an encoded representation that contains the information needed by the decoder to provide the sequence of words/characters/-subwords and build the transcription.

that state-of-the-art systems based on End-to-End models (E2E-ASRs) can model linguistic and acoustic features of spoken language, which can be investigated to explain their internal dynamics. Several probing techniques have been designed to inspect and better understand the internal behavior of DNN layers at different depths. With these techniques, investigations on the internals of DeepSpeech2 [2, 3] revealed the influence of diatopic pronunciation variation in various English varieties and provided evidence that intermediate layers contain information crucial for their classification. Later, a study [4] on the layerwise capacity to encode information about acoustic features, phone identity, word identity, and word meaning based on the context of occurrence highlighted that

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
vincenzonorman.vitale@unina.it (V. N. Vitale); lschettino@unibz.it (L. Schettino); cutugno@unina.it (F. Cutugno)
0000-0002-0365-8575 (V. N. Vitale); 0000-0002-3788-3754 (L. Schettino); 0000-0001-9457-6243 (F. Cutugno)

the last layer right before the decoding module retains information about word meaning information, rather than local acoustic features and phone identity information that are captured by the first layers and intermediate layers respectively. Then, other studies have further investigated the capacity of state-of-the-art models to encode phonetic/phonemic information[5, 6], lexical tone [7] and gender [8]. Finally, [9] investigated the internal dynamics of three pre-trained E2E-ASRs evidencing the emergence of syllable-related features by training an acoustic-syllable boundary detector. Following this line of research, a recent study [10] investigated the ability of two types of pre-trained systems with the same Conformer-based encoding architecture but different decoders – a Connectionist Temporal Classification (CTC) decoder and a Transducer decoder – to model features that distinguish filled pauses and prolongations in speech and showed that, despite not being originally trained to detect disfluencies, these systems tend to model some speech features that are functional for their identification. Rather than disregarding the ability of E2E-ASRs to model the acoustic information tied to such speech phenomena as a dispensable noise source, it could be exploited to achieve different ends. On the one hand, it could be used to obtain more accurate transcriptions that provide better, or rather more faithful, representations of the speech signal, which would also support linguistic annotation processes. On the other hand, exploring the systems' modelling ability leads to deepening our understanding of their underlying dynamics. In the last 20 years, disfluency detection tasks have been conducted to improve speech recognition performances [11, 12] and different recent approaches to filler detection achieve rather high performances, see [13]. However, these investigations mostly concern filler particles and, to our knowledge, no such system has been tested on Italian data so far. The proposed work aims to build upon these findings by investigating which of the two decoding systems is better at performing a detection task for fillers and prolongations. Moreover, a quantitative and qualitative error analysis is conducted to deepen our understanding of the way these systems work.

2. Materials and Method

2.1. Data

In this study, we employed approximately 210 minutes of expert annotated speech respectively divided into $\sim$ 80 minutes of informative speech [14], 90 minutes of descriptive speech [15] and approximately 40 minutes of dialogic speech [16], that is dyads where two speakers recorded on different channels interact. While the data from [14] and [16] consists of speech produced by speakers of the Neapolitan variety of Italian, the speakers from [15] come from different Italian regions.

More specifically, the considered speech data include: audio-visual recordings of guided tours at San Martino Charterhouse (in Naples) led by three female expert guides (CHROME corpus [14]), which consists of informative semi-monologic, semi-spontaneous speech characterized by a high degree of discourse planning and an asymmetrical relationship between the speakers; audio-visual recordings of 10 speakers narrating 'Frog Stories'from a picture book [15], which elicited unplanned descriptive speech; four task-oriented dialogues from the CLIPS corpus [16], which provides mainly descriptive semi-spontaneous speech characterized by a low degree of discourse planning and a high degree of collaboration between the interlocutors.

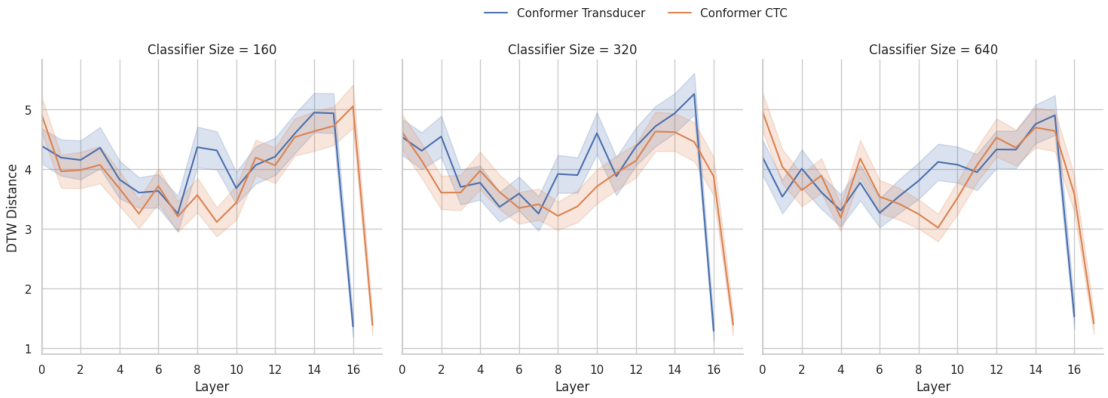
2.2. Annotation

Filled Pauses (FPs), defined as non-verbal fillers realized as vocalization and/or nasalization, and Prolongations (PRLs), defined as marked lengthening of segmental material [17, 18] were manually annotated along with pauses, lexical fillers, repetitions, deletions, insertions, and substitutions following the annotation scheme described in [19]. This is a multilevel annotation system developed to account for both formal and functional features of phenomena used to manage the own speech production. The identification of different types of phenomena was based on a 'pragmatic approach'[20], which means that it did not rely on absolute measures but on perceptual judgments given the specific contexts of occurrence. The reliability of the annotation and the Inter-Annotator Agreement was evaluated by measuring Cohen's κ. It yielded 0.92 for dialogic data and 0.82 for monologic data, which stands for 'high agreement'[21].

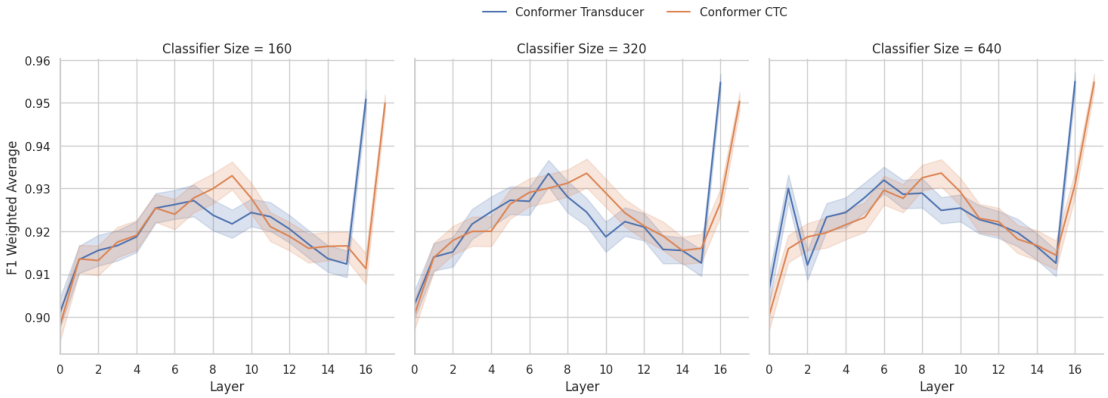
2.3. Data Preparation

The considered dataset has been prepared based on a set of praat TextGrid annotation files indicating the speaker and the type of disfluency according to the speech signal. More specifically, considering only the PRLs and the FPs, the resulting dataset has a dimension of 1900 segments. For each segment, the contextual information preceding and following the disfluency phenomenon has been considered, giving each segment a length of 4 seconds. Then, based on the combination of the so-composed dataset with each of the considered pre-trained models' encoders (details reported in Section 3.1), for each combination of segment and on each intermediate encoding layer the following elements were extracted:

- A *sequence of intermediate layer emissions/embedding* representing the input segment in the layer's



(a) Average Dynamic time warping distance measured between sequences of labels with standard error (shade).



(b) Average Weighted F1 measure measured between sequences of labels with standard error (shade).

Figure 2: Dynamic Time Warping distance (figure a) and Weighted F1 (figure b) for all the trained classifers. The x-axis indicates the index (starting from index 0) of the intermediate layer from which the distilled features have been extracted to train the corresponding classifier.

vectorial space. Each emission in the sequence represents a portion of 40 milliseconds of the input signal due to the considered model's characteristics.

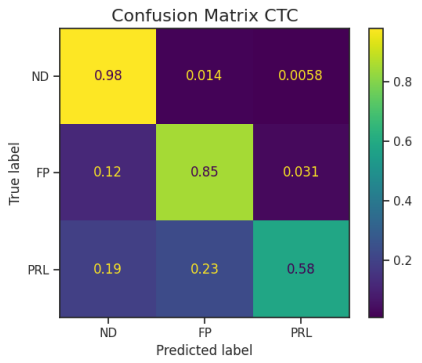
- A *sequence of labels* associated with each sequence of emissions, indicating whether an intermediate emission belongs to a particular class of disfluencies (1 for FP and 2 for PRL) or not (label 0 if the segment does not belong to a disfluency).

The resulting dataset consists of pairs of sequences of emissions (i.e., distilled features) and corresponding labels identified by the model and the layer from which they were extracted. Note that each sequence of intermediate layer emissions has a length $h = 4seconds/40milliseconds$, as it represents the temporal succession of segments before, during, and after disfluency phenomena. We use the term *emission* [10, 9] to indicate intermediate layer neurons fire, instead of the more commonly used term *embedding* [8], as the latter is widely used to indicate the output of an entire module rather than a layer.

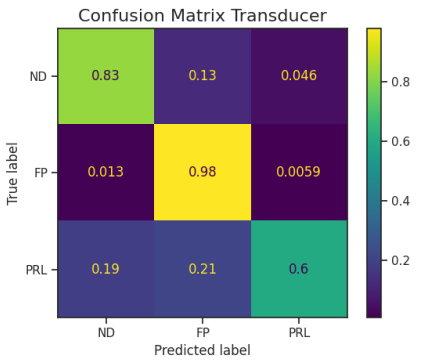
3. Results

3.1. Disfluency Identification Through Model Probing

Building upon recent studies that make use of probes to better understand the internal behavior of pre-trained E2E-ASR models'[9, 4, 3], we apply a similar approach to investigate if and to which extent a pre-trained model (m) can codify disfluencies-related features in the encoding module, even if they are not trained to do so. The employed approach is aimed at building specific classifiers whose inputs are represented by intermediate emissions of the considered model's encoder layers (l), combined with the appropriate sequence of labels based on dataset annotation. Internally, each classifier consists of a Long Short Term Memory (LSTM) module followed by a Feed Forward Neural Network (FFNN). Given that our problem can be related to sequence classification, the LSTMs seem to be the most naturally suited model [22]; usually, an LSTM consists of one computational unit that iteratively processes all input time series vectors. This unit



(a) CTC-based classifier with hidden size 640 trained on distilled features from layer 18 (index 17 in F1,DTW plots).



(b) RNN-T-based classifier with hidden size 640 trained on distilled features from layer 16 (index 15 in F1,DTW plots).

Figure 3: Confusion matrix for the best classifiers obtained for each of the considered decoding approaches.

comprises three *gates* processing one vector at a time and combining it with information extracted from previous vectors. One of the most crucial parameters for an LSTM is the hidden layer, therefore we investigate the impact of three different layer sizes (hidden-layer size, n), namely 160, 320 and 640. So, an LSTM-based classifier processes a sequence of $\{e_{l,m}\}$ emission vectors (each of length h) and produces a new sequence of vectors with size n. The two sequences are aligned over time. At each time step t, the FFNN produces a label indicating whether the considered input represents a specific disfluency segment (label 1 for filled pause or 2 for prolongation) or not (with label 0) based on the LSTM hidden-layer output. In summary, we train and evaluate many different LSTM-based disfluencies classifiers/detectors ($L_{n,m,l}$) for all possible n, m, and l combinations to search for the evidence of disfluencies-related properties in the models' decisions.

The goal is to explore which of the considered pre-trained E2E ASR models, based on different decoding systems, better encodes characteristics associated with disfluent speech segments to perform a fillers and prolongations detection task. To this end, two publicly available [23] Conformer-based models [24] with 120 million parameters each, built with the NVIDIA Nemo toolkit and differing only in the decoding strategy, were selected. On the one hand, a Conformer-based model with a Connectionist Temporal Classification (CTC) [25] decoder has been considered, as the CTC is one of the most popular decoding techniques. Such a decoding technique is a non-auto-regressive speech transcription technique that 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. Being non-auto-regressive, it is also considered computationally effective as it requires less time and resources for training and inference phases. On the other hand, a Conformer-based model with the Recurrent Neural Network Transducer (RNN-T), commonly known as *Transducer* has been considered. The RNN-T is an auto-regressive speech transcription technique that overcomes CTC's limitations, being non-auto-regressive and subject to limited label sequence length. The Transducer decoding technique can produce label-transcription sequences longer than the input vector sequence and models inter-dependency in long-term transcription elements. A Transducer typically comprises two sub-modules: one that forecasts the next transcription label based on the previous transcriptions (prediction network) and 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 compared to CTC while requiring more training and computational resources [26]. Note that both pre-trained models rely on the same encoder architecture, but the Conformer-CTC model has 18 encoding layers, while the Conformer-Transducer encoder has 17 layers.

In this study, $\sim$ 100 classifiers (2 models * $\sim$17 layers * 3 classifier sizes) were trained to investigate which of the considered pre-trained models, differing only by the decoding approach, encodes enough information to perform a disfluency detection task.

To evaluate the alignment between the output of the classifier and the reference label sequence we employ the Dynamic Time Warping Distance (DTW distance) [27], reported in figure 2a. The DTW results highlight that layers closer to the decoding module seem to contain most of the information needed to perform a correct detection of the considered disfluencies, obtaining an average DTW distance of approximately 1.39 in all the cases, with a considerably low standard error. Then, to evaluate the capability of each classifier to provide a correct as well as aligned labels sequence, we employed the weighted F1 measure, reported in figure 2b. Also in this case, F1 results confirm that layers closer to the decoding module seem to be those containing most of the information needed to correctly identify the disfluency segment. The combination of F1 and DTW provides an integrated perspective

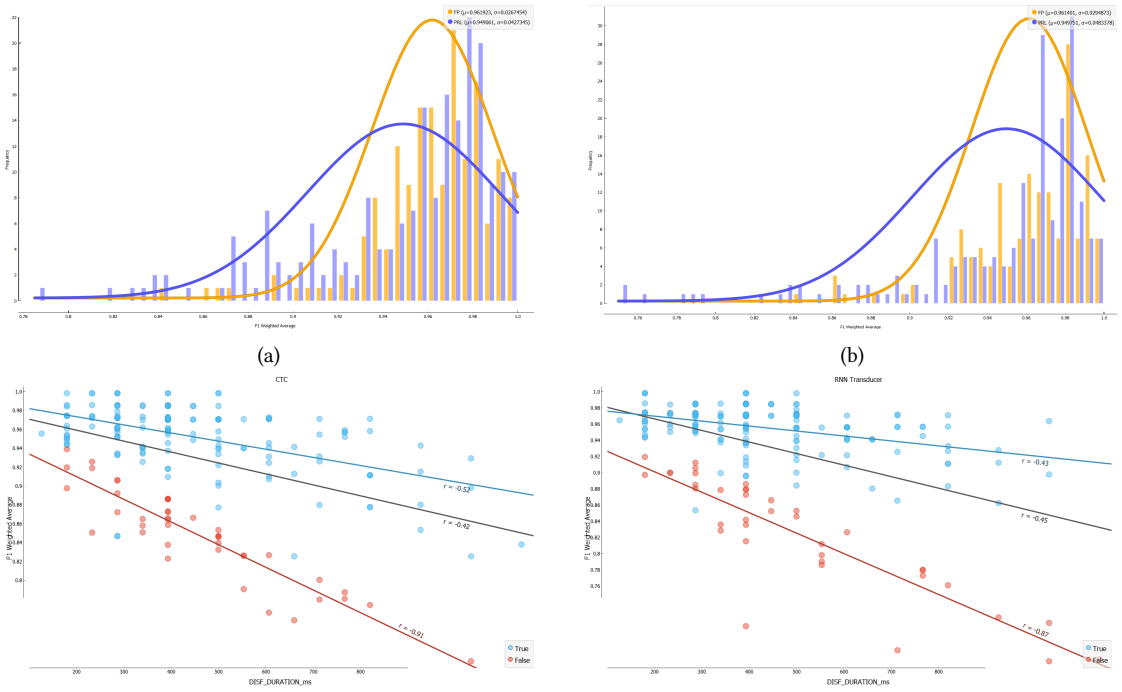


Figure 4: The plots in (a) for CTC and (b) for RNN-T report the F1 measure related to the frequency of FP (yellow) and PRL (purple). Scatterplots for CTC (c) and RNN-T (d) compare the duration of the PRL segments with the respective F1 measure.

on the system's ability to classify and align segments correctly. Finally, in Figure 3 (a and b), we report the confusion matrix of the best classifiers obtained from each considered model. On the one side, the CTC seems to be better at discriminating non-disfluent segments (ND), while showing the worst performance in disfluency identification. On the other side, the RNN-T-based classifier shows considerable performance at identifying FPs and is the worst in discriminating ND segments, while PRL performance is comparable to the CTC classifier. Both matrices highlight that the most difficult disfluency phenomena to classify are prolongations, which is the focus of our preliminary exploratory error analysis.

3.2. Qualitative Analysis

The qualitative analysis is based on the best classifier for each of the considered models used to generate the distilled features. In particular, for the CTC version, the best classifier resulted in the one with 640 hidden neurons trained on 18-th layer features. Among the transducer-based versions, the one with 640 hidden neurons trained on 17-th layer features emerged as the best version.

The visual inspection of the distribution of the considered phenomena highlights that for both the CTC (4a) and the RNN Transducer classifiers (4b), FP phenomena concentrate on higher F1 weighted values, whereas wider distributions are observed for PRL phenomena, which shows that both classifiers work better when dealing with FP than for PRL phenomena. Focusing on the PRL instances, a negative correlation is observed between the F1 weighted scores and PRLs' duration (CTC non-recognized r = - 0.91, figure 4c; RNN Transducer non-recognized r = - 0.87, figure 4d).

The error analysis was supported by an auditory inspection of the unrecognized and misclassified samples filtered based on the average DTW distance, namely, 1.39 for the Transducer-based and 1.40 for the CTC-based classifier. Issues in PRL recognition mostly concerned shorter instances, those characterized by peculiar 'non-prototypical'phonation features (such as unsteady, creaky phonation) and the alignment of PRL-predicted occurrences. Also, several PRL phenomena were misclassified as FP when occurring with monosyllabic words, such as 'o<oo>', 'un po<oo>', 'che<ee>', 'e<ee>'. In fact, the phonetic realization of these instances is closer to the ones that characterize FP for their vowel quality and as being, to a certain extent, independent elements from the phonetic environment

4. Discussion and Conclusions

In this work, we build upon a previous study that investigated to what extent modern ASR E2Es encode features related to disfluency phenomena, even if they are not directly trained to do so. We showed that pre-trained models with the same audio encoder but with two different state-of-the-art decoding strategies (CTC and Trans-

ducer) capture disfluency-related features, especially in the latest encoding layer, and both model features that can be used for the identification and positioning of disfluent speech segments [10]. Although there seems to be a tendency to forget this information with subsequent layers, as the trends for DTW (figure 2a) and F1-measure (figure 2b) would suggest, the last layers, which are those closest to the objective function represented by the decoding module, seem the most prone to retain characteristics useful to locate and identify disfluency phenomena. Interestingly, despite the differences between the two decoding modules which are respectively non-recurrent (CTC) and recurrent (RNN-T), the performances for the chosen task are comparable. However, the confusion matrices highlight that the CTC-based classifier performs better in the disfluency feature discrimination task, while the Transducer-based classifier more precisely identifies filled pauses, which could be related to the scope (recurrent/non-recurrent) of the objective function. The results align with the literature that shows a strong sensitivity to features concerning words and phone of the layers closest to the encoder[4], while the layers closest to the input are more sensitive to features related to accent and local acoustic characteristics [3, 4]. It is worth noticing that, in a recent work [9], sensitivity to syllabic boundaries was found in layers 3-5, with a pattern similar to the one shown in Figure 2 but without the peak in the last layers. The reason can be found in the fact that syllables and their boundaries do not have a graphic distinction in the transcriptions, conversely, in the case of disfluencies, there is a form of transcription that identifies them within a language model.

The exploratory analysis of the errors highlighted that prolongations are more difficult to detect than filled pauses, which could depend on their being an integral (though lengthened) part of 'fluent'words while filled pauses are mostly realized as independent elements. Also, instances of prolongation are mostly non-recognized or misclassified as filled pauses when characterized by peculiar 'non-prototypical'phonation features, such as creaky phonations, or filler-like features, as in the case of monosyllabic word-final prolongations. Also, previous studies on the segmental quality of prolongations in Italian [28] showed that prolongations, especially when concerning consonantal sounds, can be realised with schwa sounds similar to those that characterize most filled pauses. This filler-like quality could also be considered among the underlying reasons for the negative correlation between the evaluation metrics of prolongations misclassification and their duration. Another possible motivation could reside in a bias in the dataset combined with the classifier architecture (LSTM), which easily recognises prolongations responding to a specific length pattern. This means that the scarcity of longer prolongations hinders their modelling leading to their misclassification.

These findings could be used to improve transcription applications by enriching them with disfluency annotation (including filler particles and prolongation phenomena), which are still rather costly processes for studies concerning hesitation phenomena and (own) speech management in typical as well as atypical speech (e.g., pathological or language learners' speech. Indeed, an immediate development of the described work consists of increasing the capabilities of the pre-trained E2E-ASRs by adding a simple disfluency identification module to complement the existing decoder, thus enriching the resulting transcriptions.

Our work is built upon unidirectional LSTMs rather than on bidirectional LSTMs (BiLSTMs), which provide better performance because the latter have slightly longer inference times and require a larger amount of data, resources, time to be trained and, most importantly, present a more complex behaviour [29]. However, the introduction of different architecture modules like bidirectional LSTM could improve the detection of prolongation disfluencies. This will be part of future developments focused on performance and increased neural network complexity.

References

[1] S. Wang, G. Li, Overview of end-to-end speech recognition, in: Journal of Physics: Conference Series, volume 1187, IOP Publishing, 2019, p. 052068.

[2] T. Viglino, P. Motlicek, M. Cernak, End-to-end accented speech recognition., in: Interspeech, 2019, pp. 2140–2144.

[3] 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.

[4] 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.

[5] P. C. English, J. Kelleher, J. Carson-Berndsen, Domain-informed probing of wav2vec 2.0 embeddings for phonetic features, in: Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, 2022, pp. 83–91.

[6] K. Martin, J. Gauthier, C. Breiss, R. Levy, Probing self-supervised speech models for phonetic and phonemic information: A case study in aspiration, in: INTERSPEECH 2023, 2023, pp. 251–255. doi:10.21437/Interspeech.2023-2359.

[7] G. Shen, M. Watkins, A. Alishahi, A. Bisazza,

G. Chrupała, Encoding of lexical tone in self-supervised models of spoken language, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4250–4261. URL: https://aclanthology.org/2024.naacl-long.239. doi:10.18653/v1/2024.naacl-long.239.

[8] A. Krishnan, B. M. Abdullah, D. Klakow, On the encoding of gender in transformer-based asr representations, in: Interspeech 2024, 2024, pp. 3090–3094. doi:10.21437/Interspeech.2024-2209.

[9] 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.

[10] V. N. Vitale, L. Schettino, F. Cutugno, Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model hesitation phenomena, in: Interspeech 2024, 2024, pp. 222–226. doi:10.21437/Interspeech.2024-2029.

[11] M. Gabrea, D. OShaughnessy, Detection of filled pauses in spontaneous conversational speech, in: 6th International Conference on Spoken Language Processing (ICSLP 2000), ISCA, 2000, pp. vol. 3, 678–681–0. URL: https://www.isca-archive.org/icslp_2000/gabrea00_icslp.html. doi:10.21437/ICSLP.2000-626.

[12] E. Shriberg, Spontaneous speech: how people really talk and why engineers should care., in: INTERSPEECH, Citeseer, 2005, pp. 1781–1784.

[13] V. Kany, J. Trouvain, Semiautomatic support of speech fluency assessment by detecting filler particles and determining speech tempo, in: Workshop on prosodic features of language learners' fluency, 2024.

[14] A. Origlia, R. Savy, I. Poggi, F. Cutugno, I. Alfano, F. D'Errico, L. Vincze, V. Cataldo, An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the CHROME project, in: Proceedings of the 2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage, volume 2091, 2018, pp. 1–4.

[15] G. Sarro, The many ways to search for an Italian frog. The Manner encoding in an Italian corpus collected with Modokit., Master's thesis, Università degli Studi dell'Aquila., 2023.

[16] R. Savy, F. Cutugno, Diatopic, diamesic and diaphasic variations in spoken Italian, in: M. Mahlberg, V. González-Díaz, C. Smith (Eds.), Proceedings of CL2009, The 5th Corpus Linguistics Conference, 20–23 July 2009, Liverpool, UK, 2009, pp. 20–23.

[17] R. Eklund, Disfluency in Swedish Human–Human and Human–Machine travel booking dialogues, Ph.D. thesis, Linköping University Electronic Press, 2004.

[18] S. Betz, Hesitations in Spoken Dialogue Systems, Ph.D. thesis, Universität Bielefeld, 2020.

[19] L. Schettino, The Role of Disfluencies in Italian Discourse. Modelling and Speech Synthesis Applications., Ph.D. thesis, Università degli Studi di Salerno, 2022.

[20] R. J. Lickley, Fluency and disfluency, in: M. A. Redford (Ed.), The handbook of speech production, Wiley Online Library, 2015, pp. 445–474. doi:https://doi.org/10.1002/9781118584156.ch20.

[21] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

[22] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[23] NVIDIA, Nvidia catalog for pre-trained conformer models, 2023. URL: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_{transducer|ctc}_large.

[24] 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, arXiv preprint arXiv:2005.08100 (2020).

[25] 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.

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

[27] M. Müller, Dynamic time warping, Information retrieval for music and motion (2007) 69–84.

[28] L. Schettino, R. Eklund, Prolongation in italian, in: Proceedings of Disfluency in Spontaneous Speech Workshop 2023 (DiSS 2023), 28–30 August 2023, Bielefeld, Germany, 2023, pp. 81–85.

[29] S. Siami-Namini, N. Tavakoli, A. S. Namin, The performance of lstm and bilstm in forecasting time series, in: 2019 IEEE International conference on big data (Big Data), IEEE, 2019, pp. 3285–3292.