Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model hesitation phenomena

Vincenzo Norman Vitale[1,3], Loredana Schettino[2,3], Francesco Cutugno[1,3]

[1] DIETI, University of Naples Federico II, Italy

[2] Faculty of Education - Free University of Bozen, Italy

[3] Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy

unibz

1. Introduction

Recent works on Automatic Speech Recognition (ASR):

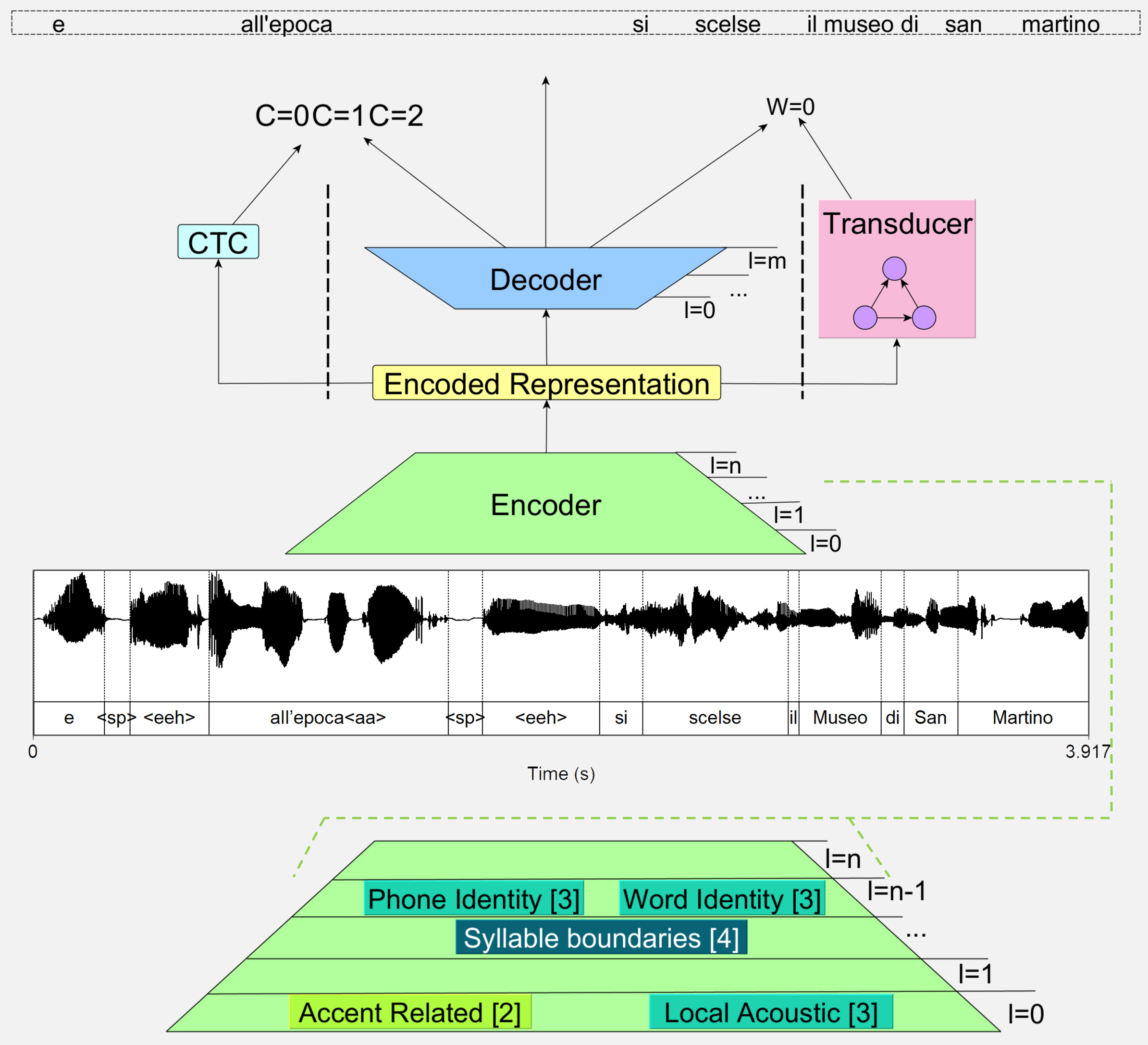
- reach high performances through increasingly complex Deep Neural Networks (DNNs)
  - require huge amounts of data
  - hardly interpretable
- obtain "clean" speech transcriptions
  - underrepresentation of phenomena characterising spoken communication (e.g. discourse markers, particles, pauses, disfluencies, ...)

- ➢ Studies on the interpretability of neural models investigated the systems' ability to model specific linguistic features
- ➢ Probing techniques are employed to investigate what is encoded in DNN layers at different "depths" [1-4]

AIM: investigating the ability of pre-trained E2E ASR systems to model distinguishing feature of hesitation phenomena

- fillers – FP, <eeh>, <ehm>
- segmental prolongations – PRL, the<ee>

RQ1: *Do E2E ASR encode information about disfluencies? To what extent?*
RQ2: *Is it possible to employ such information to identify and discriminate disfluencies?*



Disfluencies?

2. Materials and Method

Data

- ~ 80 minutes of informative speech CHROME corpus [5]
- ~ 90 minutes of descriptive speech Modokit-FROG corpus [6]
- ~ 40 minutes of dialogic speech CLIPS corpus [7]

Annotation

Manual annotation of disfluency phenomena [8]

- Prolongations (PRs), marked lengthening of segmental material
- Filled Pauses (FPs), non-verbal fillers realized as vocalization and/or nasalization

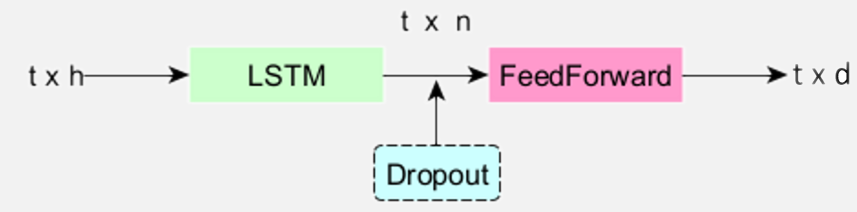
Cohen's κ, i.e., 0.92 for dialogic data and 0.82 for monologic data, 'high agreement'

Data Preparation

- ~ 1900 4-second segments containing PRL and/or FP including contextual information
- Two models [9] pre-trained on the NVIDIA ASRSet 2.0 for English:
  - Conformer Transducer
  - Conformer CTC
- For each segment, model, and encoding layer:
  - Collection of a sequence of intermediate layer emissions representing the input segment
  - Each emission = 40 ms of the input signal
  - A sequence of labels is associated with each emission belonging to a disfluency (FP and PR) or not (ND)

3. Probing Approach

- Corpus split in train (60%), validation (20%), and test (20%) sets
- Training of three differently-sized LSTM-based classifiers per encoding layer

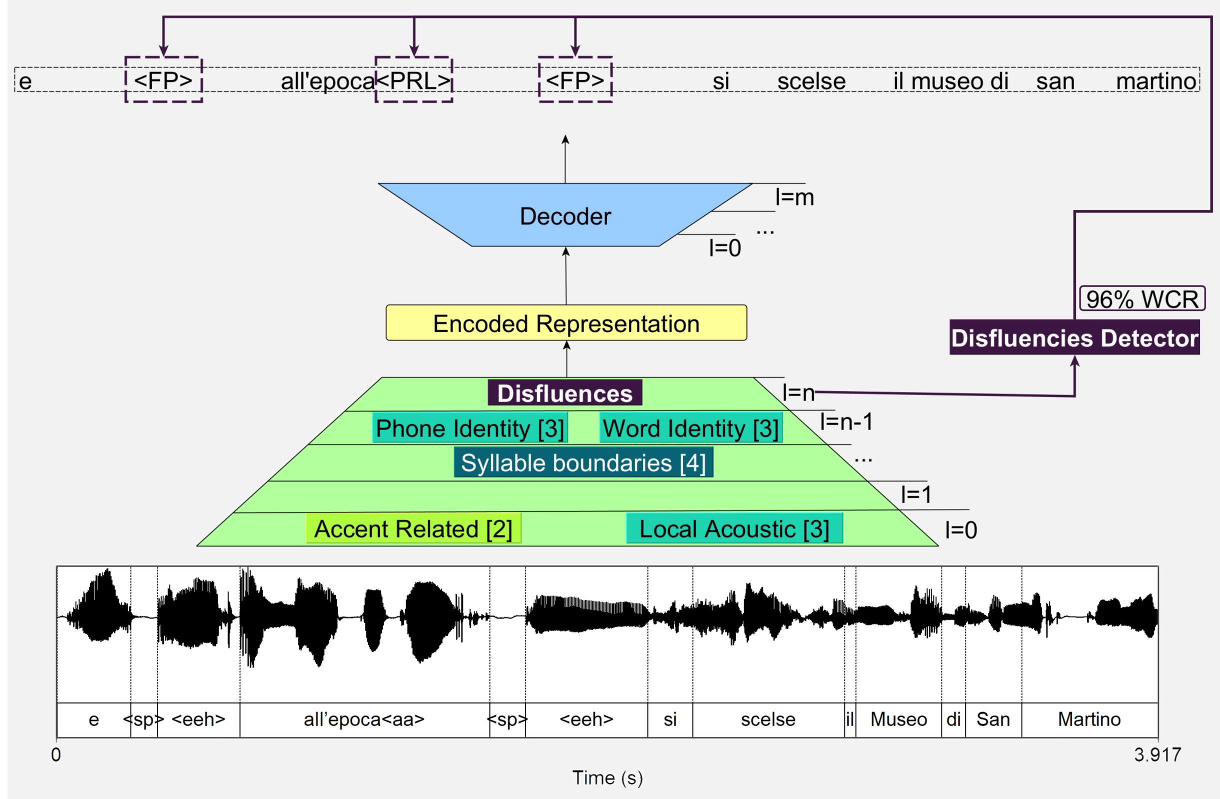


t = temporal index
h = emission vector size
n = hidden layer size
d = classification decision

- Possible hidden layer sizes:160/320/640
- Training setting:
  - Maximum training epochs=100
  - Adam optimizer with initial lr=1e^{-5}
  - Dropout neuron selection with a probability of 0.1.
  - Early stopping on validation-loss with threshold=0.001 and patience=20 epochs
- Classifiers' temporal sensitivity bound to pre-trained model

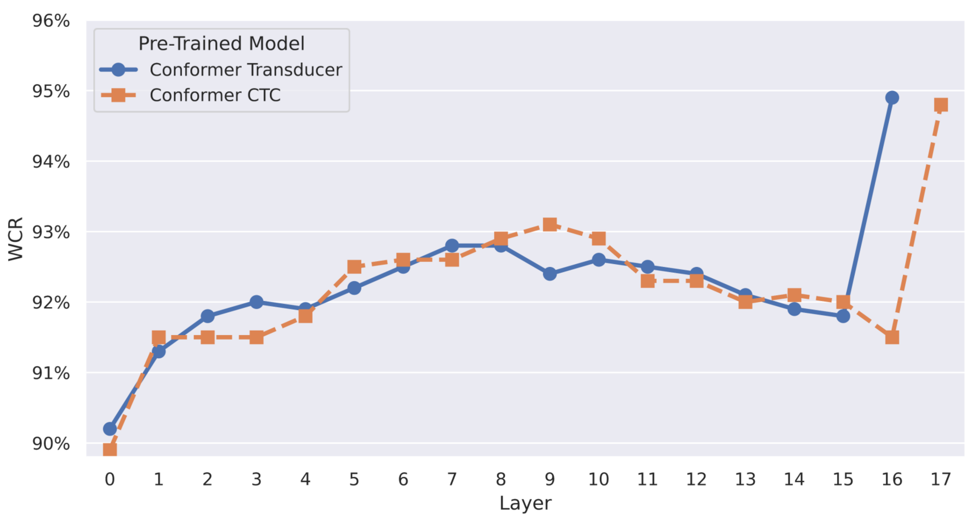
Metrics

- Word Correct Rate WCR (complementary to WER more fitting when considering time-mediated alignments
- NIST SCLITE toolkit to extract our metrics [10]

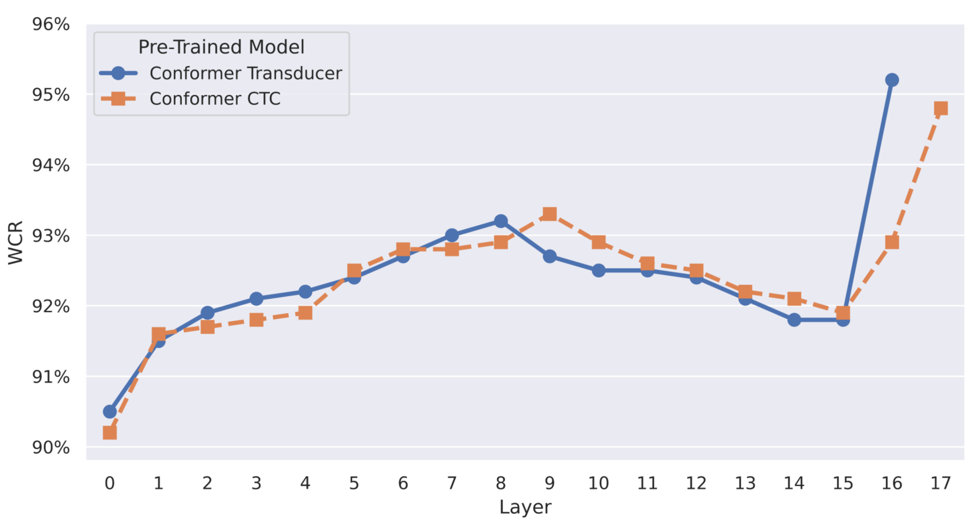


4. Results

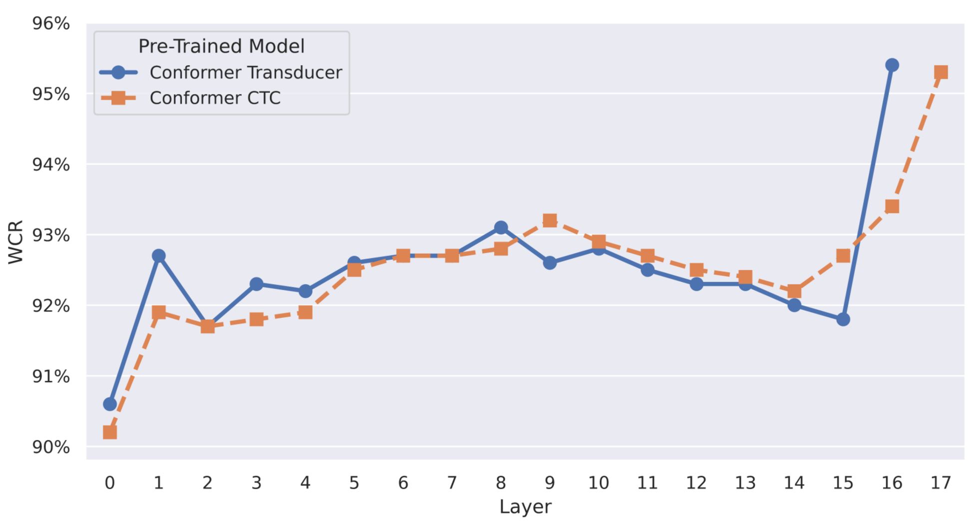
- Emerging general trend
  - the valuable information for identifying a disfluent segment increases while approaching the intermediate layers and then decreases until reaching the layer immediately preceding the decoding phase, where a significant peak is observed
- Last layer before the decoder (penultimate layer)
  - most informative
  - most stable



(a) WCR for the system with 160 hidden units



(b) WCR for the system 320 hidden units



(c) WCR for the system with 640 hidden units

Figure 1. WCR charts grouped by LSTM hidden-layer size (the x-axis indicates the probed layer depth expressed as layer index)

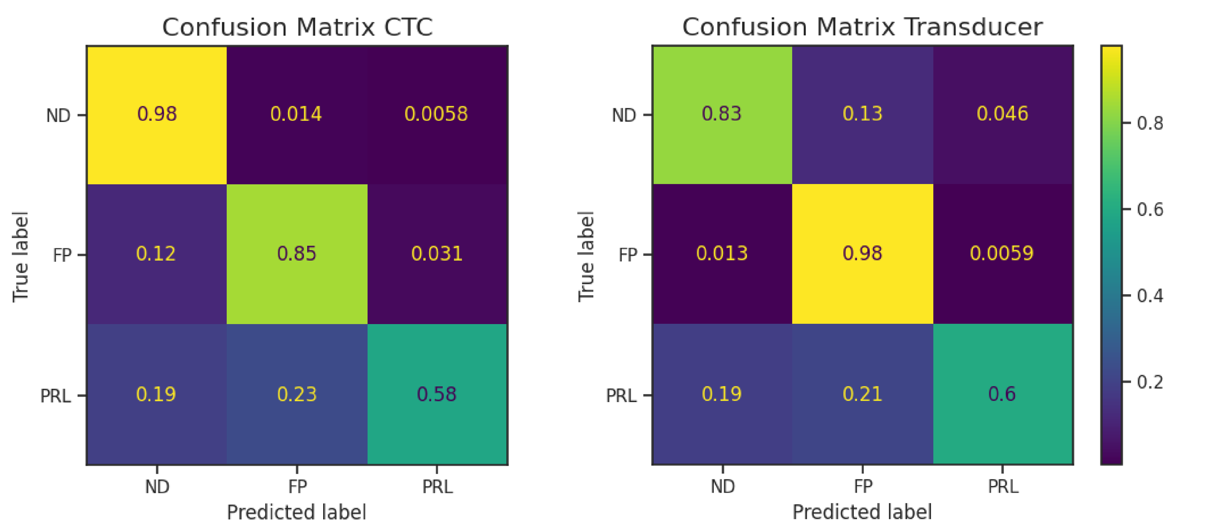
- Although the CTC-based model has an extra layer, its performance is comparable to the one of the Transducer-based model until the last layer
- for the last layer, the Transducer-based model almost always shows a better performance
- Confusion Matrices results corroborate the observation that the Transducer-based are better at identifying FP

5. Discussion and Conclusions

- pre-trained models with CTC and Transducer decoding strategies both capture features useful for the identification of disfluent features, although they are not originally trained for disfluency detection
- this seems to be especially favoured in layers closest to the objective function

Future application: 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 (by including hesitation phenomena)

! Note: this study focused on models trained in a supervised manner, particularly those based on the Conformer architecture (an extension of the Transformer architecture), which limits the observations compared to self-supervised models



CTC	ND	FP	PRL
ND	0.98	0.014	0.0058
FP	0.12	0.85	0.031
PRL	0.19	0.23	0.58

Transducer	ND	FP	PRL
ND	0.83	0.13	0.046
FP	0.013	0.98	0.0059
PRL	0.19	0.21	0.6

(a) CTC-based classifier with hidden size 640 trained on distilled features from layer 17

(b) Transducer-based classifier with hidden size 640 trained on distilled features from layer 16

Figure 2. Confusion matrix for the best classifiers obtained for each of the considered decoding approaches

References

[1] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in Interspeech, 2019, pp. 2140–2144. [2] A. Prasad and 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. [3] A. Pasad, J.-C. Chou, and 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. [4] V. N. Vitale, F. Cutugno, A. Origlia, and G. Coro, "Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique," Neural Computing and Applications, pp. 1–27, 2024. [5] A. Origlia, R. Savy, I. Poggi, F. Cutugno, I. Alfano, F. D'Errico, L. Vincze, and 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, vol. 2091, 2018, pp. 1–4. [6] G. Sarro, "The many ways to search for an italian frog. the manner encoding in an italian corpus collected with modokit." Master's thesis, Universit` a degli Studi dell'Aquila., 2023. [7] R. Savy and F. Cutugno, "Diatopic, diamesic and diaphasic variations in spoken Italian," in Proceedings of CL2009, The 5th Corpus Linguistics Conference, M. Mahlberg, V. González-Díaz, and C. Smith, Eds. 20–23 July 2009, Liverpool, UK, 2009, pp. 20–23. [8] L. Schettino, "The role of disfluencies in Italian discourse. Modelling and speech synthesis applications." Ph.D. dissertation, Università degli Studi di Salerno, 2022. [9] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models\stt_en_conformer_{transducer|ctc}_large. [10] https://www.nist.gov/itl/iad/mig/tools.

Poster PDF

Vitale

Schettino

Cutugno

E-mail:
{vincenzonorman.vitale,cutugno}@unina.it
lschettino@unibz.it
Web: *www.urbaneco.unina.it*

INTERSPEECH 2024