Interspeech 2024
1-5 September 2024, Kos, Greece

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

vincenzonorman.vitale@unina.it, lschettino@unibz.it, cutugno@unina.it

Abstract

Modern automatic speech recognition systems can achieve remarkable performances. However, they usually neglect speech characteristic phenomena such as fillers (<eeh> <ehm>) or segmental prolongations (the<ee>) which are still only considered as disrupting objects to be detected and removed, despite their acknowledged regularity and procedural value.

This study investigates the ability of state-of-the-art systems based on end-to-end models (E2E-ASRs) to model distinctive features of hesitation phenomena. Two types of pre-trained systems with the same Conformer-based encoding architecture but different decoders are evaluated: the Connectionist Temporal Classification (CTC) decoder and a Transducer decoder.

E2E-ASRs ability to model the acoustic information tied to such phenomena can be exploited rather than disregarded as a noise source, which would not only improve transcription and support linguistic annotation processes, but also deepen our understanding of the systems' working.[1]

Index Terms: disfluencies, speech recognition, probing, interpretability.

1. Introduction

Modern Automatic Speech Recognition (ASR) systems, among other Natural Language Processing (NLP) systems, achieve remarkable performances thanks to the computing potential enabled by Deep Neural Networks (DNN). In recent works, a great deal of effort went into incrementing the systems' performances by employing increasingly complex, hence hardly interpretable, DNN models that require huge amounts of data for the training. The resulting systems aim at 'clean'speech transcriptions rather than more accurate rich transcriptions including the phenomena that characterize spoken communication (e.g. discourse markers, particles, pauses, disfluencies, ...). However, a few studies on the interpretability of neural models found that state-of-the-art systems based on End-to-End models (E2E-ASRs) seem to model linguistic and acoustic features of spoken language. Hence, some studies have investigated the systems' modeling of specific linguistic features to explain their internal dynamics.

In particular, probing techniques have been employed to investigate what is encoded in DNN layers at different 'depths'. Some work investigated the influence of pronunciation diatopic variation in different English varieties on the performance of DeepSpeech2 [1, 2] and highlighted that the contextual phonetic information contained in intermediate representations played a role in the classification. Then, Pasad, Chou and Livescu [3] conducted a layerwise analysis of the encoding status of information related to acoustic features, phone identity, word identity, and word meaning based on the context of occurrence, which revealed that local acoustic features and word meaning information, rather than phone identity information, are retained on the last layer right before the decoding layers. A more recent study [4] probed the internal dynamics of three pre-trained E2E-ASRs by exploring the emergent syllables from an acoustic-syllable boundary detector.

Following this line of research, the present work aims to investigate the ability of pre-trained E2E ASR systems to model linguistic elements that are often overviewed although they characterize spoken language. Namely, hesitation phenomena such as fillers (<eeh> <ehm>) and segmental prolongations (the<ee>).

2. ASR systems and Disfluencies

In the last few decades, studies on ASR systems have acknowledged that speech consists not only of a sequence of words. The latter, in fact, co-occurs with a wide range of 'non-verbal'and non-lexical phenomena tied to the specific conditions of speech production, like breath sounds, tongue-clicks, laughter, fillers, prolongations, repetitions, and repairs, which should be taken into account as an integral part of the data that ASR systems commonly deal with.

Based on this insight, [5] highlighted that while, at the time, most ASRs were trained on large amounts of read speech, the latter considerably differs from spontaneous speech and advocated the need for studying speech-specific features. Among these, the so-called 'disfluency'phenomena emerge, a heterogeneous class of elements such as hesitation phenomena (pauses, fillers, prolongations) and repairs, that frequently occur in speech and seem to alter its ideal fluency, but were observed to regularly occur in spoken productions [6, 7]. In particular, [7] identified systematic variation in observable features of the data per type of disfluency, sentence position, distribution, e.g., the form of filled pauses (duration, f_0 properties). In the computational field, filled pauses were identified as disruptive elements that hindered the interpretation of the language-model component and the readability of a transcript, hence, they needed to be detected and removed to get the speaker's intended word sequences. The first studies on disfluencies and ASRs mostly concerned detection based on their specific acoustic features to enhance the performances of the recognizers. According to the acoustic analysis conducted by [5], filled pauses were characterized by a long, steady vowel produced with constant energy, little spectral change, low F0 to the speaker's baseline, flat pitch, and often bordered by a silence. Feature-based al-

[1]This work results from the collaboration among the authors. However, for academic purposes only, Vincenzo Norman Vitale is responsible for §§ 3.3, 3.4 and 4, Loredana Schettino for §§ 2, 3.1 and 3.2, all the authors are responsible for §§ 1 and 5.

10.21437/Interspeech.2024-2029

gorithms were used to detect filled particles before recognition considering fundamental frequency transitions and spectral envelope deformations [8], fundamental frequency, energy, duration, and spectral estimation [5], Mel-frequency cepstral coefficients (MFCCs) [9, 10], or formant-based analysis which was found to outperform standard cepstral stability-based or pitch-based detection techniques [11].

Later, studies delved into the analysis of the issues underlying prediction failures. Namely, elements with peculiar prosodic characteristics and non-lexical elements, e.g., discourse markers, filled pauses, vowel lengthening [12, 13, 14]. Also, Shriberg [15, 1781] mentioned disfluencies among the challenging spontaneous speech properties for spoken language application "because they violate assumptions often applied in automatic processing technology". She highlighted the positive role of disfluencies in speech as they reflect cognitive aspects of language production and argued that their modeling could improve speech recognition performances. Indeed, considering prosodic and textual information was found to positively contribute to the modeling performances compared to a purely word-based language model [16].

The advent of DNN has enhanced the potential of ASR systems of modeling acoustic features and contextual information. This, together with the acknowledgment of the positive role of hesitation phenomena, like filled pauses, for managing speech production and interactions, prompted studies concerning ASR modeling of filled pauses [17, 18].

However, with the introduction of the most recent E2E-ASR based on the Transformer architecture, a relevant part of the efforts was reverted to the increment of content-aimed transcription performances, disregarding non-lexical phenomena as merely disrupting elements to be detected and removed.

Another type of hesitation phenomena that realize a suspension in speech through vocalization and duration [19] is segmental lengthening, i.e., prolongations. The fewer studies on ASRs involving prolongation phenomena mostly concerned models developed to detect stuttering speech [20]. To help prolongation detection for corpus studies, [21] created a semiautomatic search tool that, however, requires phonemically annotated data, which is not always available and costly to obtain.

So, although phenomena such as fillers (<eeh> <ehm>) or segmental prolongations (the<ee>) have been observed to occur regularly and contribute to speech management by providing valuable information on discourse (planning, structure, the speaker disposition,...), they are still often disregarded as noise source in state-of-the-art ASR systems. This study aims to investigate the ability of E2E ASR systems to model features that distinguish such hesitation phenomena. In particular, the following research questions are addressed:

- RQ1 Whether and to what extent an E2E ASR encodes information about disfluencies.
- RQ2 Whether it is possible to employ such information to identify and discriminate disfluencies.

3. Materials and Method

3.1. Data

This study is based on approximately 80 minutes of informative speech [22], about 90 minutes of descriptive speech [23] and approximately 40 minutes of dialogic speech [24], that is dyads where two speakers recorded on different channels interact. These data were mostly produced by speakers of the Neapolitan variety of Italian.

The monologic section includes audio-visual recordings of guided tours at San Martino Charterhouse (in Naples) led by three female expert guides (CHROME corpus [22]). It consists of informative semi-monologic, semi-spontaneous speech characterized by a high degree of discourse planning and an asymmetrical relationship between the speakers. This section also includes audio-visual recordings of 10 speakers narrating 'Frog Stories'[23]. Each participant was given a picture book and asked to tell the story to a silent interlocutor, which elicited unplanned descriptive speech.

The dialogic section contains four task-oriented dialogues from the CLIPS corpus [24]. This setting provides mainly descriptive semi-spontaneous speech characterized by a low degree of discourse planning and a high degree of collaboration between the interlocutors.

3.2. Annotation

The phenomena considered as objects of analysis in this work were manually annotated along with other speech management phenomena according to the system described in [25]. Prolongations (PRs) were identified as marked lengthening of segmental material [26, 27]; filled pauses (FPs) as non-verbal fillers realized as vocalization and/or nasalization. Identifying these phenomena did not rely on absolute measures but on perceptual judgments given their specific contexts of occurrence. To test the reliability of the system the Inter-Annotator Agreement was evaluated by measuring the values of Cohen's κ, i.e., 0.92 for dialogic data and 0.82 for monologic data, both standing for 'high agreement'according to [28].

3.3. Data Preparation

Based on the annotation in TextGrid format, the dataset was divided into segments grouped first by speaker and second by type of disfluency (PR and FP), resulting in about 1900 segments. Each segment consists of a 4-second signal, including contextual information, both preceding and following each disfluency phenomenon. Then, for each of the considered pre-trained models (see Table1 for more characteristics), the following elements were extracted for each segment and each intermediate encoding layer:

- A *sequence of intermediate layer emissions* representing the input segment. Each emission belonging to the sequence represents a portion of 40 milliseconds of the input signal due to the considered models 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 PR) or not (label 0 if the segment does not belong to a disfluency).

The resulting dataset consists of pairs of sequences of emissions and corresponding labels, grouped by model and layer they belong to. 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.

3.4. The Probing Approach

Similar to previous studies investigating E2E-ASR models' internal behaviour [4, 3, 2], we employ a probing approach aimed

[2]Both models are publicly available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_{transducer|ctc}_large

Model	Version	Parameters
Conformer Transducer	1.6.0	$\sim$ 120M
Conformer CTC	1.6.0	$\sim$ 120M

Table 1: *Main characteristics of considered models[2], pre-trained on the NVIDIA ASRSet 2.0 for English.*

at investigating the pre-trained models' (m) ability to code disfluencies-related features during the encoding phase. The approach consists of training specific classifiers whose inputs are represented by intermediate emissions of the models' encoder layers (l), combined with the appropriate sequence of labels based on dataset annotation. The classifier architecture consists of a Long Short Term Memory (LSTM) module followed by a Feed Forward artificial neural network. The LSTMs are suited models for classifying a time series of observation vectors [29]; usually, an LSTM consists of one computational unit that iteratively processes all input time series vectors. This unit comprises three *gates* processing one vector at a time and combining it with information extracted from previous vectors.

Since the LSTM hidden-layer size is a crucial parameter, 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 if pre-trained E2E ASR models encode characteristics associated with disfluent speech segments, determining the extent to which these features facilitate their identification and subsequent differentiation. To this end, two publicly available Conformer-based models [30] were selected. Both are built with the NVIDIA Nemo toolkit[3], but differ in the decoding component (see Table 1 for more information). On the one hand, we considered a Conformer-based model with a Connectionist Temporal Classification (CTC) [31] decoder, one of the most popular decoding techniques. The CTC 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. As 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, we considered a Conformer-based model with the Recurrent Neural Network Transducer (RNN-T), commonly known as *Transducer*, which is an auto-regressive speech transcription technique that overcomes CTC's limitations, namely 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-decoding 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 with respect to CTC while requiring more training and computational resources [32].

Note that the temporal sensitivity of our disfluencies classifier/detector, namely the minimum difference between consecutive time steps, is 40 ms because the considered ASR models (see Table1) produce emissions at that rate.

4. Results

For each considered model and each encoding layer, three differently-sized LSTM-based[4] classifier were trained, resulting in $\sim$ 100 models. During the training phase, the considered corpus was split into train, validation, and test sets using 60%, 20%, and 20% percentages while ensuring that these sets did not share the same speakers. Each classifier has been trained for a maximum of 100 epochs using an Adam optimizer with an initial lr of 0.00001. To mitigate the risk of overfitting, we introduce a *dropout* neuron-selection strategy for the LSTM gates, which statistically excludes (with a 0.1 probability) one neuron and its weights during each training iteration [33]. Finally, an early stopping mechanism was used to avoid wasting computational resources. In particular, the training phase ends if the validation-loss does not decrease by a minimum of 0.001 during the last 20 epochs, which is the patience threshold.

As our goal is to investigate whether and to which extent pre-trained E2E models encode features letting the identification of disfluent speech segments, the chosen metric is the complementary to the Word Error Rate, the Word Correct Rate. The choice of using Word Correct Rate WCR (complementary to WER) is linked to the nature of the inputs consisting of temporal sequences (sec 3.3). Although preferred measures like F1 are useful for the evaluation of isolated items, WCR seems more fitting when considering time-mediated alignments and has been used in works concerning related tasks, e.g., boundary detection [4]. We rely on the NIST SCLITE toolkit[5] to extract our metrics.

In Figures 1 (a-c), we reported the WCR charts grouped by LSTM hidden-layer size. The x-axis indicates the probed layer depth expressed as layer index. The general trend that emerges from the graphs is that the valuable information for identifying a disfluent segment increases approaching the intermediate layers and then decreases until reaching the layer immediately preceding the decoding phase, where a significant peak is observed in all of the considered experimental conditions. Note that, in addition to being the most informative, the layer immediately preceding the decoder is also shown to be the most stable, as confirmed by the standard deviation values in Table 2.

Concerning the dimensions of the two models considered, although the CTC-based model has an extra layer, performance is comparable up to the last layer, for which the Transducer-based model almost always shows better performance.

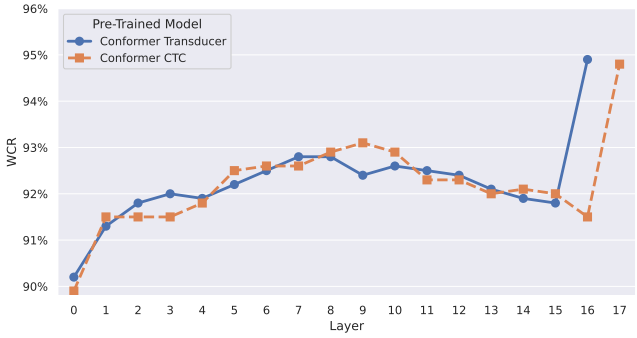
[3]Nemo version 1.21.0.

[4]Frameworks used to implement and train the classifiers: torch==2.2.1 and pytorch-lightning==2.0.7

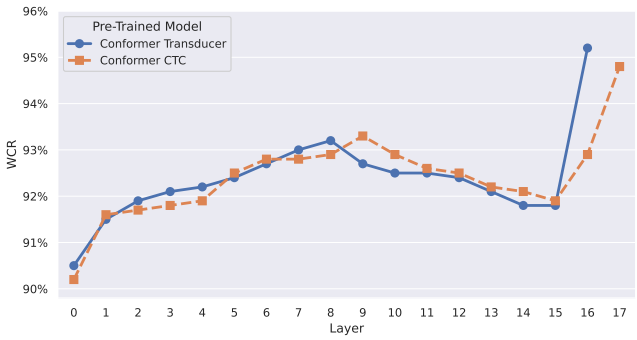
[5]NIST, "Speech recognition scoring toolkit (Sctk)," 2022, available at https://www.nist.gov/itl/iad/mig/tools.

Conformer Model	n	Probed Layer																	
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
CTC	160	7.8	6.6	6.8	7.6	7.2	6.8	6.8	6.9	6.8	6.8	6.6	6.4	6.3	6.4	6.2	6.4	7.1	***4.0***
CTC	320	7.6	6.9	7.0	7.1	7.1	7.2	6.8	7.0	6.5	6.8	6.6	6.3	6.4	6.4	6.2	6.6	6.2	***4.2***
CTC	640	7.7	6.7	6.9	7.5	7.3	7.2	6.9	7.1	6.8	6.7	6.5	6.1	6.5	6.2	6.3	6.5	5.6	***4.1***
Transducer	160	7.5	6.9	6.9	6.7	7.1	7.3	6.9	6.9	6.5	6.4	6.5	6.5	6.1	6.4	6.2	6.3	***4.6***	-
Transducer	320	7.1	6.9	6.7	7.2	7.3	7.1	6.6	6.7	6.6	6.4	6.3	6.5	6.3	6.3	6.4	6.4	***4.2***	-
Transducer	640	7.6	7.1	6.8	6.7	7.2	7.3	7.0	6.8	6.6	6.2	6.2	6.5	6.5	6.4	6.5	6.3	***4.6***	-

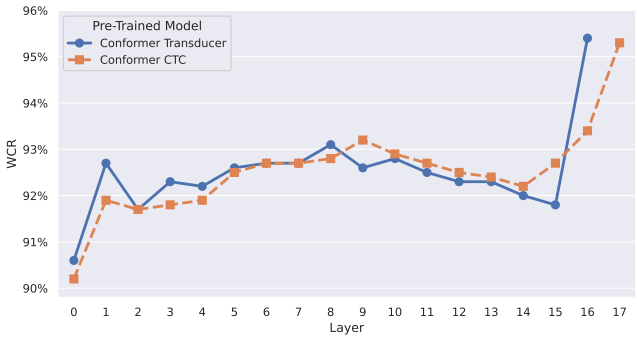
Table 2: *Standard deviations calculated through NIST SCLITE for each trained classifier, with size n, on the basis of intermediate layer emissions from the considered pre-trained Conformer models.*



(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: *Achieved Word Correct Rate (WCR) for each of the considered system configurations (n=160/320/640) and datasets. The x-axis indicates the origin of the intermediate emission of each dataset, which are used to train each system.*

5. Discussion and Conclusions

In this work, we presented a study aimed at understanding whether and to what extent modern ASR E2Es encode features related to disfluency phenomena, even if they are not directly trained to do so. We focused on pre-trained models with the same audio encoder but with two different state-of-the-art decoding strategies (CTC and Transducer), finding that in both cases, features useful for the identification of disfluent features are encoded, especially in the latest encoding layer.

The results suggest that, although the discussed models are not originally trained with the aim of detecting disfluencies, they tend to model some speech features that are functional for their identification. Although there seems to be a tendency to forget this information with subsequent layers, as the descending trend in figure 1 (a-c) would suggest, the last layers, which are those closest to the objective function, seem to favour the learning of characteristics useful for identifying disfluency phenomena. Interestingly, despite the differences between the two decoding modules, respectively CTC and RNN-T based, the performances for the chosen task are comparable.

These findings could find application in providing rich transcriptions (including hesitation phenomena) and/or filled pauses and prolongation annotation, which are still rather costly processes for studies concerning hesitation phenomena and (own) speech management. 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. Note that 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. Hence, further investigations would be devoted to further pursuing this aim.

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

[6] W. J. Levelt, *Speaking: From intention to articulation*. Cambridge and London: MIT press, 1993.

[7] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, Berkeley, 1994.

[8] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[9] F. Stouten and J. Martens, "A feature-based filled pause detection system for Dutch," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*. St Thomas, VI, USA: IEEE, 2003, pp. 309–314. [Online]. Available: http://ieeexplore.ieee.org/document/1318459/

[10] C.-H. Wu and G.-L. Yan, "Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 36, no. 2/3, pp. 91–104, Feb. 2004. [Online]. Available: http://link.springer.com/10.1023/B:VLSI.0000015089.17975.f4

[11] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan: IEEE, Apr. 2009, pp. 4857–4860. [Online]. Available: http://ieeexplore.ieee.org/document/4960719/

[12] G. Garg and N. Ward, "Detecting Filled Pauses in Tutorial Dialogs," 2006.

[13] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, Mar. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639309001599

[14] Palmerini, Maria and Savy, Renata, "Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare," in *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa*. pisa university press, 2014. [Online]. Available: http://clic2014.fileli.unipi.it/proceedings/Proceedings-CLICit-2014.pdf

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

[16] Yang Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1677974/

[17] D. Prylipko, O. Egorow, I. Siegert, and A. Wendemuth, "Application of image processing methods to filled pauses detection from spontaneous speech," in *Interspeech 2014*. ISCA, Sep. 2014, pp. 1816–1820. [Online]. Available: https://www.isca-archive.org/interspeech_2014/prylipko14_interspeech.html

[18] M. Bilac, M. Chamoux, and A. Lim, "Gaze and filled pause detection for smooth human-robot conversations," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. Birmingham: IEEE, Nov. 2017, pp. 297–304. [Online]. Available: http://ieeexplore.ieee.org/document/8246889/

[19] L. Schettino and 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.

[20] S. Gupta, R. S., R. K., and R. Verma, "Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020. [Online]. Available: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=9&Code=IJACSA&SerialNo=41

[21] S. Betz, J. Voße, S. Zarrieß, and P. Wagner, "Increasing recall of lengthening detection via semi-automatic classification," in *Proceedings of Interspeech*, 2017.

[22] 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.

[23] 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.

[24] 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.

[25] L. Schettino, "The role of disfluencies in Italian discourse. Modelling and speech synthesis applications." Ph.D. dissertation, Università degli Studi di Salerno, 2022.

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

[27] S. Betz, "Hesitations in spoken dialogue systems," Ph.D. dissertation, Universität Bielefeld, 2020.

[28] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] 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.

[31] A. Graves, S. Fernández, F. Gomez, and 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.

[32] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.