Toward Optimised Datasets to Fine-tune ASR Systems Leveraging Less but More Informative Speech.

Loredana Schettino[1,*], Vincenzo Norman Vitale[2] and Alessandro Vietti[1]

[1]Free University of Bozen-Bolzano, Piazza Università, 1, 39100 Bolzano, Italia
[2]University of Naples Federico II, C.so Umberto I, 40, 80138 Napoli, Italia

Abstract

Modern Automatic Speech Recognition (ASR) systems, based on Deep Neural Networks (DNN), have achieved remarkable performance modelling huge quantity of speech data. However, recent studies have shown that fine-tuning pre-trained models, despite providing a powerful solution in low-resource settings, lacks robustness across different speech styles, and this is not just related to the amount of training data, but to substantial differences in phonetic-prosodic characteristics. Therefore, this study aims to explore how modern E2E ASR systems' performance is affected by the amount of training data and the type of speech data and which acoustic-phonetic features most markedly exert an influence. To this aim, a k-fold cross-validation was performed by fine-tuning a pre-trained FastConformer model with datasets varying in type of speech data and size. Then we performed a correlation analysis between the values of the acoustic characteristics of the data and the recognition scores. The analyses allow the identification of an optimal combination of speech data type and amount of training data. Also, results show that using both more spontaneous speech or more controlled speech can be beneficial, provided that the speech rate is contained.

Keywords

Speech style, ASR, Sample Efficiency, Acoustic Features, K-fold Cross-Validation

1. Introduction

Spoken language is intrinsically variable. Speech produced to convey a message can vary widely depending on several internal and external factors, such as the communicative and contextual situation, the formality of the exchange, the speaker's disposition and individual choices of the forms and phonetic realisation deemed as most appropriate and functioning to convey the intended message given the specific condition of production and reception [1]. Thus, speech variability can be described as the synergetic contribution of linguistic, contextual, and social factors [2], which results in different types of speech, often referred to as *speech style*, characterized by varying levels of spontaneity, fluency, speaking rate, prosodic variation, degree of phonic specification [3, 1].

Modern ASR systems, based on Deep Neural Networks (DNNs), have achieved remarkable performance by modelling the linguistic and acoustic features of spoken language. However, these systems implicitly learn to model only a small proportion of the possible variation that characterises spoken language. As a result, error rates increase with the degree of linguistic and phonetic variation of the data considered. In fact, while most benchmarks consist of read or rather controlled speech productions, the interest in ASR applications in real contexts, such as human-machine-interactions or transcription of spontaneous conversation, led to the evaluation of ASR performance in different, less controlled and more spontaneous scenarios, which resulted in different performance values for other types of data, e.g, lower for more spontaneous datasets [4]. In particular, a recent study on the evaluation of ASR systems, based on state-of-the-art supervised, self-supervised, and weakly supervised End-to-End models, on Italian speech [5, 6], showed consistent performance differences across speech types: dialogic, monologic, and read speech. Namely, increasing performance from dialogic speech to monologic speech and from the latter to read speech.

Efforts devoted to overcoming this issue often consist of building complex and costly models that require large amounts of data and computational resources. However, this can be problematic, especially when working with so-called "low-resource languages". Different studies have provided evidence that a powerful solution is provided by fine-tuning pre-trained models (see [7]). However, [8] adopted this approach in a study on low-resource speech recognition and showed not only a lack of robustness in Word Error Rate (WER) distributions across different speakers and conversation contexts, but also that this was not related to the amount of training data, but to substantial differences in prosody, pronunciation and utterance length. This led to acknowledging that using more data and more complex techniques is not sufficient to address

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy.
*Corresponding author.
lschettino@unibz.it (L. Schettino);
vincenzonorman.vitale@unina.it (V. N. Vitale);
alessandro.vietti@unibz.it (A. Vietti)
0000-0002-3788-3754 (L. Schettino); 0000-0002-0365-8575 (V. N. Vitale); 0000-0002-4166-540X (A. Vietti)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

the problem of automatically recognising different types of data. Rather, we need to investigate how different types of data and their specific acoustic-prosodic features affect the performance of ASR systems to address this robustness issue [7].

Based on this body of research, this work aims to contribute to the study of how different types of speech data are modelled and how this affects the robustness of the model toward the definition of an optimal dataset to obtain robust recognition systems.

2. Related Work

Especially, but not exclusively, within the context of low-resource studies, the need to develop less resource-greedy ASR systems emerges. To this end, different data efficiency techniques, e.g., learning or data augmentation techniques, have been explored, such as multilingual transfer to provide robust acoustic word embeddings [9, 10], self-training, where an ASR system trained with the available human-transcribed data is used to generate transcriptions, which are then combined with the original data to train a new ASR system, or neural TTS synthetic data generation [11]. However, although it has been shown that the size of training data affects the performance of ASR systems, "[w]hether data augmentation is always beneficial is an open question." [11, 723].

Another way to help achieve high performance with minimal data may consist in relying on less but more informative data by investigating how different types of speech data are modelled and affect the robustness of the model, and which combination of different speech types and amount of data optimises the informativeness and efficiency of a sample to fine-tune pre-trained models.

To this end, a better understanding of the aspects of speech that challenge ASR architectures the most is required. In the last 20 years, various studies have investigated which phonetic features affect automatic recognition the most (see [7] for an overview). In particular, issues were observed to mostly concern features of conversational speech such as grammatical inconsistencies, self-interruptions, backchannels, lexical and non-lexical disfluencies, and the degree of pronunciation variation [12, 13]. ASR systems were also observed to struggle to recognise words with low intensity, high F0 value or shorter duration [14]. Then, a recent study aimed at gaining insight on which aspects of casual, conversational speech cause the largest challenges for different ASR HMM and transformer-based architectures showed that utterance length (in number of tokens), articulation rate and pronunciation variation exert a major influence, with higher recognition scores correlating with longer utterances, lower speech rates and lower phonetic variation [7].

The present study aims to contribute to this line of research by developing and validating a method to address the following research questions (RQs):

RQ1. If modern E2E ASR systems' performance is affected by the amount of training data and the type of speech data, can we identify the optimal combination of speech data and amount of training data?

RQ2. What acoustic-phonetic characteristics affect the most modern E2E ASR performance? To what extent?

3. Methodology

To investigate how data characterised by different features (data type) and varying amounts of training data (training data time) can affect the fine-tuning of modern ASR models, our method includes a K-fold cross-validation procedure [15]. This technique is used when there is a limited amount of data and provides insight into the model's performance across different data subsets. It consists of splitting the data into subsets (*folds*) and training different models, as many as the number of folds, each time considering a different combination of folds as training (potentially validation) and test sets. The approach follows these key steps:

- selection of data with different speech characteristics;
- fold splitting according to training-specific criteria, i.e., speech type and training fold size (minutes);
- selection of a pre-trained model for fine-tuning;
- evaluating model performance for the selected datasets;
- fine-tuning the pre-trained model by training it on the different folds;
- comparison of the performance of the fine-tuned models;
- Word Error Rate - acoustic features correlation analysis.

3.1. Data

Given the methodological focus of this study, we decided to work with a well-known, restricted dataset to gain clearer insights into the effectiveness of the method and the findings. Hence, we selected data from a corpus that was the object of previous phonetic studies [16, 17], namely the CHROME corpus [18]. The corpus comprises approximately 10 hours of speech produced by three female expert museum guides (G) leading visits at San Martino Charterhouse (in Naples). It consists of Neapolitan Italian, informative semi-monologic, semi-spontaneous speech characterised by a high degree of discourse planning and an asymmetrical relationship between the interlocutors. The three speakers show idiosyncratic speech

Table 1
Datasets duration, tokens, speech rate (SR) values.

dataset	duration	tokens	SR	m utterance duration (sd)	m utterance tokens (sd)
G01	192' 26"	27881	2,41	3,72 (2,76)	8,97 (7,23)
G02	216' 14"	39145	3,02	4,30 (2,50)	12,98 (8,08)
G03	181' 56"	29341	2,68	4,62 (3,31)	12,43 (9,04)

styles [19]. In particular, they use different speech rates and different "hesitation strategies". G01 produces approximately 159 words per minute and seems to privilege an "on the fly" production, using several non-lexical fillers (*eeh*, *ehm*) and prolongations to cover speech planning time; G02 shows a higher speech rate, producing about 174 words per minute, where utterances are juxtaposed to each other as she tends to avoid silent pauses altogether, avoid prolongations and non-lexical fillers, and prefer lexical fillers instead; G03 adopts a more controlled, "rhetorical" style, with a lower speech rate of about 146 words per minute and mainly using lexical fillers and silent pauses.

3.2. Data Preparation

Using the text annotation in TextGrid format [20], the dataset was split in Inter-Pausal Units based on pauses longer than 250 ms. This resulted in utterances with a mean duration of 4,81 seconds (standard deviation = 2,88, max length = 30 ms). The text was normalised by removing special characters, but leaving annotation of segmental phenomena such as fillers (eeh, ehm, mh) and prolongations (e.g., la**aa**). The final considered dataset consists of slightly more than 3h and 27881 tokens for G01, about 3h and a half and 39145 tokens for G02, and about 3 h and 29341 tokens for G03. G02 shows a higher speech rate than both G01 and G02. See Table 1 for total duration, tokens and speech rate (SR), and mean (m) and standard deviation (sd) of utterance duration and tokens.

3.3. Modelling

Selecting an appropriate pre-trained model is a critical decision that influences the success of subsequent downstream tasks. While many high-performing models are available, such as Whisper or Phi-4, our selection was guided by several practical requirements: language-specific support for Italian, computational efficiency, and public availability to ensure experimental reproducibility and democratic access. Accordingly, we chose the FastConformer model pre-trained on Italian by Nvidia [21]. The FastConformer is an efficient variant of the Conformer architecture, designed to significantly reduce the computational cost and latency of the standard Conformer model while maintaining high accuracy. This

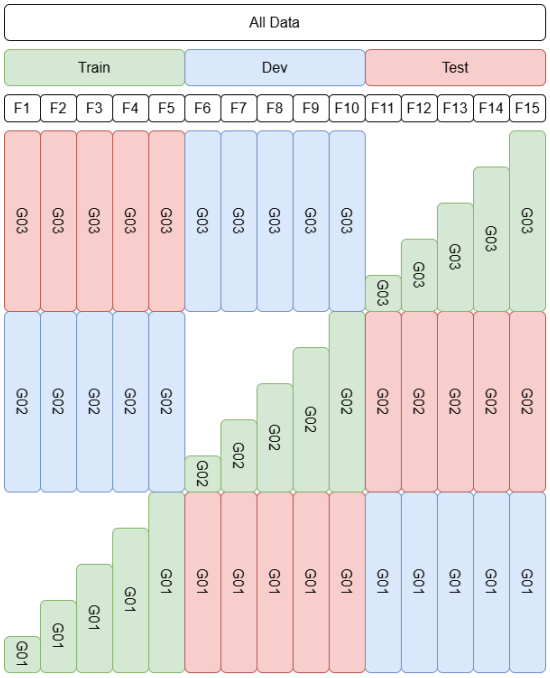

Figure 1: K-fold Cross-Validation Procedure.

makes it particularly suitable for real-time speech recognition tasks. Furthermore, the architecture is highly scalable, and indeed, FastConformer is at the core of top-performing Nvidia ASR systems like Canary and Parakeet.

The Group K-fold is a variation of k-fold cross-validation intended for scenarios where the data has a predefined group structure. The key constraint is to ensure that the same group is not represented within the same splits, namely training, validation and test sets. In our case, samples from the same speaker will be grouped in the same split. This method prevents data leakage by ensuring that the model generalises to new, unseen groups, not just new samples from existing groups. The corpus is split into three folds, one per speaker and idiosyncratic speech style (data set type), and these were further split into five sub-folds of different sizes (split size), resulting in 15 different fold combinations described in Figure 1.

3.4. Evaluation and correlation analysis

The model performance across the different folds was evaluated considering the Word Error Rate (WER) computed at the utterance level. Model comparison was conducted based on WER mean and distribution values per fold to observe which model performed better across the considered folds.

Then, correlation analysis between data character-

Table 2
Utterance count, mean, standard deviation, standard error of the mean, and confidence interval (default 95%) of WER per set and fold (train data sets*train time).

train set type	train set size	validation set	test set	N	wer	sd	se	ci
-	-	-	G01	3106	0.514	0.318	0.005	0.011
-	-	-	G02	3014	0.386	0.256	0.004	0.009
-	-	-	G03	2359	0.398	0.259	0.005	0.010
G01	15	G02	G03	2359	0.305	0.274	0.005	0.011
G01	30	G02	G03	2359	0.182	0.236	0.004	0.009
G01	60	G02	G03	2359	0.151	0.220	0.004	0.008
G01	120	G02	G03	2359	0.143	0.203	0.004	0.008
G01	all	G02	G03	2359	0.136	0.204	0.004	0.008
G02	15	G03	G01	3106	0.416	0.342	0.006	0.012
G02	30	G03	G01	3109	0.291	0.330	0.005	0.011
G02	60	G03	G01	3109	0.233	0.318	0.005	0.011
G02	120	G03	G01	3109	0.205	0.299	0.005	0.010
G02	all	G03	G01	3109	0.210	0.304	0.005	0.010
G03	15	G01	G02	3014	0.243	0.261	0.004	0.009
G03	30	G01	G02	3014	0.179	0.257	0.004	0.009
G03	60	G01	G02	3014	0.139	0.255	0.004	0.009
G03	120	G01	G02	3014	0.125	0.226	0.004	0.008
G03	all	G01	G02	3014	0.118	0.215	0.003	0.007

istics and WER was performed to examine the influence of acoustic features on the performance of different time folds. Feature values were automatically extracted for each utterance employing the OpenSmile toolkit [22]. The *Geneva Minimalistic Acoustic Parameter Set* (eGeMAPSv02) [23], i.e., a restricted set of features based on interdisciplinary evidence and theoretical significance, was selected as the feature set. The study focuses, in particular, on the features that could be considered as the most relevant, as reported in previous literature [7] and inspection of the data.

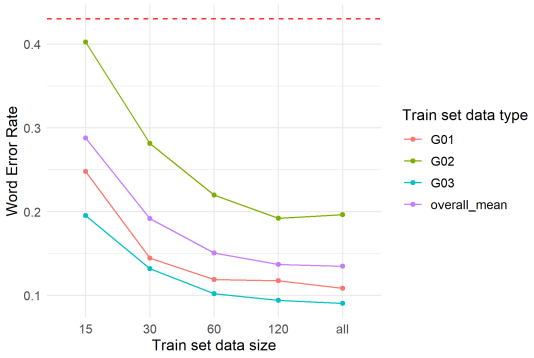

Figure 2: Word Error Rate (WER) per training time grouped by training data. The dashed red line indicates the mean baseline WER.

4. Results

4.1. Model performance and comparison

The analysis starts by evaluating the model's baseline performance on the defined datasets before applying k-fold cross-validation to establish a reference for comparison. The selected model performs less for the G01 dataset (mWER = 0.51, sd = 0.32) than for the G03 dataset (mWER = 0.40, sd = 0.26) and the G02 dataset (mWER = 0.39, sd = 0.26), see the first three rows of Table 2. The overall mean WER across different data type sets is 0,43 (sd = 0.26).

Then, we observe the model's performance on each fold. Figure 2 and Table 2 show the mean WERs per train set data type and size. The mean WERs across the data type sets (purple line) reach lower values than the baseline (red dashed line) already after fine-tuning with the smallest 15' sets (mWER_15 = 0.32, mWER_30 = 0.22, mWER_60 = 0.18, mWER_120 = 0.16, mWER_all = 0.16). The values decrease as the size of the training set increases. However, the magnitude of the WER difference between subsequent size groups progressively diminishes until it becomes trivial between the models trained on 60' speech and those trained on the entire datasets (about 3h). We then consider the mean WER values grouped by train set data type. Although models trained on G01, as well as G02 and G03 data, perform better than the baseline, we observe that the models trained on G02 data perform worse than the others, with WERs closer to the overall baseline. In particular, the models trained on G02 are tested on G03 and are closer to the G03 baseline (mWER =

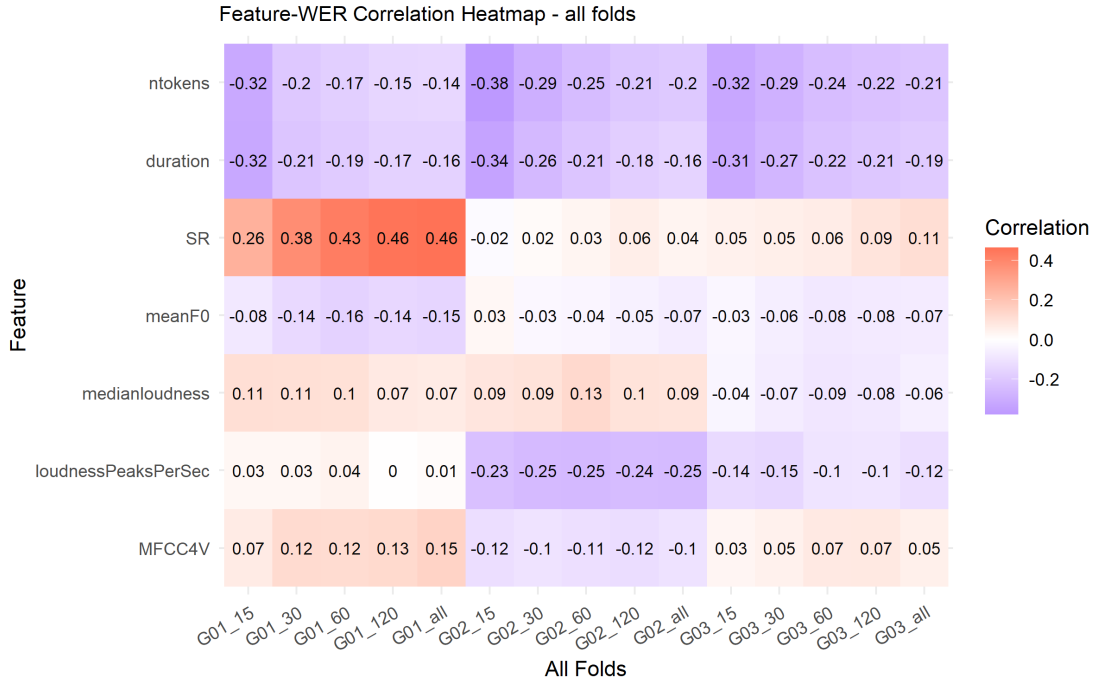

Figure 3: Correlation of feature values with WER per train set data type_size folds.

0.4). Instead, the models trained on G03 and tested on G01 show a larger difference with the G01 baseline (mWER = 0.51) than the difference between models trained on G01 and tested on G02 and the G02 baseline (mWER = 0.39).

Considering both the contribution of the train set data type and the size to the model performance improvement, the optimal fold is G03_120.

4.2. Features Correlation with WER

To explore how different datasets affect model performance, we observe which features correlate with the trained models. The heatmap in Figure 4.2 shows the Pearson coefficients resulting from the correlation between a selection of relevant acoustic features and the WER for each model. The colour of each tile represents the direction of the correlation, while its intensity indicates the strength of the correlation. Red denotes a positive correlation, meaning higher feature values correspond to higher WER, whereas blue indicates a negative correlation, where higher feature values align with lower WER. White represents a weak or no correlation.

We observe negative correlations between the WER values and both the utterance duration and tokens. The correlation becomes weaker, but still noticeable, with increasing train set data size, and the same trend is observed for each dataset. An opposite trend is observed for the speech rate values, the latter correlate with WERs positively and increasingly along the train set size. However, this trend is considerably stronger for the models trained on data from the G01 dataset (and tested on G02 dataset). Weaker correlations are observed for the mean values of F0, especially for the G02 and G03 models, with the strength slightly increasing with the size of the training set. Rather constantly weak correlations can be observed for median loudness, MFCC4 in voiced regions and WER values. Still rather constant but slightly stronger is the correlation between loudness peaks per second and WERs for the models trained on the G02 dataset.

5. Discussion and Conclusions

This study contributes to investigations on how the performance of modern E2E ASR models is affected by the type and amount of speech data used for training and aims to define a way to identify an optimal combination of type and amount of speech data. The investigation is supported by observation of how different speech acoustic features contribute to the model performance.

The Fast-Conformer WER on the selected semi-monologic, semi-sponetanous data presents overall lower values than the evaluation provided by a previous study on Italian monologic data, i.e., 12.8 WER [6]. More

specifically, lower recognition scores are reported for G01 speech, characterised by a more spontaneous speech style, including more features such as non-lexical fillers and prolongations than the other speakers, which is in line with the literature [12, 13].

The cross-fold evaluation shows that the models' performance improves with train set size; however, the magnitude of the improvement gradually decreases until becoming trivial between models trained on 120 minutes and about 3 hours of speech. This finding supports the claim that simply increasing the size of the training set is not always beneficial and not always enough to guarantee better performance. Although this trend stands across all datasets, variation can still be observed.

The models trained on speech produced by the second guide (G02) perform worse than the others, with recognition scores closer to the overall baseline. In particular, the models trained on G02 speech, that is characterised by higher speech rate and fewer pauses, are tested on G03 speech and achieve smaller improvement over the G03 baseline as compared to the models trained on G03 speech, showing a more controlled speech style, and G01 speech, defined by a more spontaneous speech style. It is particularly worth noticing that the models trained on G03 and tested on G01 show the best recognition scores over all size folds, thus overcoming the G01 baseline disadvantage. This seems to indicate that some speech data are more informative than others and may even overcome recognition issues related to more spontaneous and conversational speech styles; however, studies in this direction should be further developed.

Considering both the contribution of the train set data type and size to the model performance improvement, the dataset that optimises the combination of data type and amount is the one containing 120 minutes, i.e., two-thirds of the available dataset, of the more controlled, but still spontaneous, speech produced by G03 (RQ1).

In line with the literature [7], correlations between recognition scores and utterance durational features emerge. More specifically, higher length values (in terms of utterance tokens and duration) correlate with lower recognition errors, which indicates that providing a wider context enhances recognition. Conversely, higher speech rates hinder recognition. However, this effect is more or less mitigated according to the speech type in the training set (RQ2). This finding, as well as the constant and weak correlations observed for the other acoustic features, deserves further attention and needs to be explored in future works.

Overall, these findings show that using both more spontaneous speech and more controlled speech can be beneficial to fine-tune a pre-trained model, provided that the speech rate is not too high. More detailed analyses will be performed considering the values of the acoustic characteristics and their variation to gain deeper insight.

This study provides evidence corroborating the idea that less but more informative data can be used to fine-tune pre-trained models, which could be useful for fine-tuning in low-resource scenarios. Furthermore, the use of the Fastconformer highlights the value of architectures that offer a favorable trade-off between performance and computational resources. These models present a viable alternative for deployment on resource-constrained, privacy-oriented devices. At the same time, they can be quickly adapted to different low-resourced contexts, standing in practical contrast to larger-scale yet resource-demanding models.

In this study, we prioritised methodological soundness and understanding over immediate broad applicability. We selected a known dataset restricted in size and speaker diversity to enhance the interpretability of the results, verify the method's core effectiveness and establish a solid foundation for scaling to larger, more diverse corpora. Future work will be devoted to further exploring this direction by considering larger datasets that maximise differences in acoustic-phonetic features that were observed to be relevant for the modelling.

References

[1] B. V. Tucker, Y. Mukai, Spontaneous speech, Cambridge University Press, 2023.

[2] A. Vietti, Il ruolo della variabilità acustica nella costruzione del dato linguistico, in: Superare l'evanescenza del parlato: un vademecum per il trattamento digitale di dati linguistici, Bergamo University Press, 2021, pp. 45–70.

[3] P. Wagner, J. Trouvain, F. Zimmerer, In defense of stylistic diversity in speech research, Journal of Phonetics 48 (2015) 1–12.

[4] P. Gabler, B. C. Geiger, B. Schuppler, R. Kern, Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition, Information 14 (2023) 137.

[5] N. Vitale, E. Tanda, F. Cutugno, Towards a responsible usage of ai-based large acoustic models for automatic speech recognition: On the importance of data in the selfsupervised era, in: Atti quarto Convegno Nazionale CINI sull'Intelligenza Artificiale–Ital-IA 2024, 2024.

[6] T. Cimmino, E. Tanda, V. N. Vitale, F. Cutugno, Evaluating asr performance in italian speech, in: STUDI AISV, Milano: Officinaventuno, under review.

[7] J. Linke, B. C. Geiger, G. Kubin, B. Schuppler, What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures, Computer Speech & Language 90 (2025) 101738.

[8] J. Linke, P. N. Garner, G. Kubin, B. Schuppler, Conversational speech recognition needs data? experiments with austrian german, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 4684–4691.
[9] E. Hermann, H. Kamper, S. Goldwater, Multilingual and unsupervised subword modeling for zero-resource languages, Computer Speech & Language 65 (2021) 101098.
[10] H. Kamper, Y. Matusevych, S. Goldwater, Improved acoustic word embeddings for zero-resource languages using multilingual transfer, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1107–1118.
[11] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, M. Wieling, Making more of little data: Improving low-resource automatic speech recognition using data augmentation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, 2023, p. 715–729.
[12] B. Schuppler, M. Adda-Decker, J. A. Morales-Cordovilla, Pronunciation variation in read and conversational austrian german., in: INTERSPEECH, 2014, pp. 1453–1457.
[13] A. Lopez, A. Liesenfeld, M. Dingemanse, Evaluation of automatic speech recognition for conversational speech in dutch, english and german: What goes missing?, in: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), 2022, pp. 135–143.
[14] S. Goldwater, D. Jurafsky, C. D. Manning, Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase asr error rates, in: Proceedings of ACL-08: HLT, Association for Computational Linguistics, 2008, pp. 380–388.
[15] A. Burkov, The hundred-page machine learning book, volume 1, Andriy Burkov Quebec City, QC, Canada, 2019.
[16] L. Schettino, The role of disfluencies in Italian discourse. Modelling and speech synthesis applications, Ph.D. thesis, Ph. D. dissertation, Universita degli Studi di Salerno, 2022.
[17] N. Vitale, L. Schettino, F. Cutugno, Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers' ability to model hesitation phenomena, in: 25th Annual Conference of the International Speech Communication Association (INTERSPEECH 2024), ISCA, 2024, pp. 222–226.
[18] 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.
[19] L. Schettino, S. Betz, F. Cutugno, P. Wagner, Hesitations and individual variability in Italian tourist guides' speech, in: C. Bernardasci, D. Dipino, D. Garassino, S. Negrinelli, E. Pellegrino, S. Schmid (Eds.), Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications, STUDI AISV 8, Milano: Officinaventuno, 2021, pp. 243–262.
[20] P. Boersma, D. Weenink, Praat: doing phonetics by computer [computer program]. version 5.3. 51, Online: http://www. praat. org/retrieved, last viewed on 12 (1999-2022).
[21] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, Fast conformer with linearly scalable attention for efficient speech recognition, in: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2023, pp. 1–8.
[22] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.
[23] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.