Interspeech 2024
1-5 September 2024, Kos, Greece

On Disfluency and Non-lexical Sound Labeling for End-to-end Automatic Speech Recognition

Péter Mihajlik[1,2], *Yan Meng*[1], *Máté Kádár*[1,2], *Julian Linke*[3], *Barbara Schuppler*[3], *Katalin Mády*[2]

[1]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics; [2]HUN-REN Hungarian Research Centre for Linguistics, Hungary
[3]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{mihajlik, yan, kadar}@tmit.bme.hu, {linke, b.schuppler}@tugraz.at, mady@nytud.hu

Abstract

Spontaneous speech contains a significant amount of disfluencies and non-lexical sounds (e.g., backchannels, filled pauses), which are often difficult to transcribe. Disfluency labeling for automatic speech recognition (ASR) aims at editing these phenomena in the transcription to improve overall recognition accuracy. Such labeling techniques typically delete non-lexical/disfluent labels from the prediction, where classical ASR techniques either ignore or treat them as lexical items. Our results, obtained by systematic comparison and detailed evaluation of various disfluency labeling methods on two different language conversational corpora, suggest that neither of the previous approaches are optimal. We propose to distinguish between filled pauses and meaningful conversational grunts and show that keeping the non-lexical labels is not only possible but as low as 7% label error rates can be achieved for highly important categories (including 'mhm') while preserving a decent WER.

Index Terms: end-to-end speech recognition, disfluency, conversational speech, filled pauses, Hungarian, Austrian German.

1. Introduction

In automatic speech recognition (ASR) research, the modeling of various disfluencies and non-lexical sounds [1] has a long tradition [2, 3, 4]. As already pointed out by [5], short, less articulated spontaneous speech events can be a major source of ASR errors. With the advent of the deep neural network based end-to-end ASR approach [6], larger models based on larger speech datasets, the creation of transcriptions aware of non-lexical vocalizations and disfluencies have become more expensive. On the other hand, even in the case of basic CTC training [7], acoustic context is not limited to the neighbouring phones, but the entire utterance is taken into account. In other words, not transcribing explicitly certain audio events related to disfluencies or conversational grunts does not necessarily mean not modeling them in an end-to-end system: in an implicit way, e.g., associated to the 'blank' character, they can be learned – without generating dedicated output labels. In certain applications, however, there is a need for monitoring disfluent speech events, such as for various diagnostic applications [8, 9] and for the creation of corpora for linguistic analyses of conversational speech phenomena. Furthermore, some of the non-lexical conversational grunts may have an affirmative meaning [10], therefore skipping them in dialog may distort the semantic analysis of the conversation.

Based on annotated data like Switchboard [11] or CSJ [12], attempts have been made to label/edit the transcriptions before and/or after training end-to-end ASR models to improve recognition performance via disfluency (and non-lexical sound) labeling. [13] suggests that removing disfluencies from training (and reference) transcriptions is a viable alternative to model them directly for the English language data they used. [14] and [15] both introduced hesitation labeling for in-house English and public Japanese data sets, respectively, where mainly broken words were replaced by a uniform hesitation label for end-to-end RNN-T [16] training. These symbols were then deleted from the ASR output – and from the reference text – achieving significant word error reduction. [17] proposed a more general disfluency labeling approach where both hesitations and fillers (including interjections and responses) were replaced by symbols (@ and #) which were later removed from the ASR predictions. The paper claims their approach to be superior to the previous ones in terms of average WER (word error rate), however without reporting the error rates of the disfluent events. Furthermore, [17] considered backchannels and response-like non-lexical sounds as disfluencies to be removed, an approach we consider to be insufficient. In sum, all the reported techniques either delete disfluent and non-lexical events from the training transcriptions or replace them partially or fully by symbols that are also deleted after the inference, without addressing the potential information loss. None of the mentioned papers report a systematic comparison of various disfluency labeling techniques with respect to non-lexical sounds, nor do they evaluate their approach on more than one language. This paper fills these gaps and has the following main contributions:

- We report state of the art end-to-end ASR experiments with disfluency and non-lexical sound labeling for two conversational speech corpora from two languages, Hungarian and Austrian German.
- *Non-disfluent* non-lexical sound events are usually treated/labelled as disfluency in ASR. We, however, distinguish non-lexical fillers from backchannels, as they have different communicative functions.
- To systematically evaluate the effect of various labeling techniques for disfluent/non-lexical speech events, we introduce a 'reference token number aware' WER comparison approach.
- We report WERs separately for lexical and non-lexical tokens and show that the efficacy of disfluency labeling methods depends strongly on the error rate and distribution of various disfluent/non-lexical events.

This paper focuses on the acoustic modeling of disfluencies and non-lexical sounds in speech-to-text conversion. Disfluency types like (whole word) repetitions and restarts, that can be well transcribed automatically, are excluded from this study, as we believe that they can be treated better on the text level with NLP tools (such as LLM-s [18]). For simplicity, we may use the term 'disfluency labeling' in a general sense including implicit, removal-based approaches and labeling non-lexical events.

10.21437/Interspeech.2024-2157

2. Materials

2.1. BEA-Base V2 – Hungarian Conversational Speech

BEA (Spoken Hungarian Database) [19, 20] was recorded and transcribed for linguistic research purposes containing mostly conversational speech with an experimenter being present. For ASR benchmark, an 80 hours subset was selected from 140 speakers and train/dev/eval splits were defined [21]. In the experiments we used the 'train-114', 'dev-spont' and 'eval-spont' sets for training, development and evaluation, respectively. All experiments were performed on BEA-Base V2 where the splits and audio chunks were identical to V1 [21] but all the non-lexical acoustic tokens and broken/misbuilt (partial and restatement) word markers were reinserted that were previously removed in V1. A typical disfluent sentence: *"és akkor **ee elv-** elmennek **mhm**"* /and then they *uh lef-* leave *mhm*/. For lexical and non-lexical unit distributions see Table 1.

Table 1: *Major disfluency and non-lexical types and distributions in Hungarian BEA-Base V2.*

Category	Type	Train	Dev	Eval
Lexicals – disfluency ratio: 1.25%				
/non-disfluent/	words	$555k$	$28k$	$35k$
/disfluent/	partial words	7100	352	308
Non-lexicals – disfluency ratio: 94.6%				
Filled pause	*ee*	13916	824	1065
/disfluent/	*mm*	4293	303	185
Backchannel	*mhm*	815	29	121
/non-disfluent/	*aha*	126	16	21

2.2. GRASS – Austrian German Conversational Speech

Table 2: *Major disfluency and non-lexical types and distributions in Austrian German GRASS.*

Category	Type	Train	Dev	Eval
Lexicals – disfluency ratio: 0.94%				
/non-disfluent/	words	$167k$	$19k$	$9k$
/disfluent/	partial words	1573	180	89
Non-lexicals – disfluency ratio: 29.1%				
Filled pause	*ah*	705	80	22
	äh	351	42	37
/disfluent/	*ähm*	161	25	5
	ahm	206	27	9
Backchannel	*mhm*	2801	300	74
/non-disfluent/	*hm*	708	82	91

For our experiments, we used 19h of conversational speech from 38 Austrian speakers from the GRASS corpus [22, 23]. Closely related speaker pairs talked for one hour without interruption about topics of their choice while being recorded without the presence of an experiment leader, leading to a casual style, including many disfluent utterances (e.g., *"ein schöner **pu- äh** ein schöner **p-** platz oder so"* /a nice *pu- eh* a nice *p-* place or something). Overall, approx. 30% of all utterances in GRASS were single-word utterances and approx. 43% of them included non-lexicals tokens, of which approx. 80% were backchannels. For details on lexical and non-lexical unit distributions see Table 2. The data was split such that there was no overlap between eval and train speakers, the dev set is not speaker independent from train as it was obtained by a random 10% split from the non-eval recordings.

3. Methodology

3.1. Text Preprocessing

First transcriptions were normalized: we deleted all the speaker noise labels, and background noise, laugh or whisper annotations were also removed. Then we unified the labeling of partial (broken) or misbuilt words: a hyphen mark '-' was attached to them like *"**so-** solved"*. Finally all the special disfluent or non-lexical events – if written with special tags – were replaced by pure text tokens, as close to their pronunciation as possible. This transcription was applied both for training and reference texts served as the **Baseline (BL)** approach.

3.2. Disfluency and Non-lexical Labeling Methods

In the following we introduce all the applied disfluency (and non-lexical) labeling approaches with respect to their operations on the ASR training transcriptions. For operations on ASR output and reference text and for illustrative comparison and 'reference token number aware' evaluations, see Table 3.

Lexical Hesitation Labeling (LHL): proposed by [14] and [15] simply replaces each partial (and misbuilt) word by a @ symbol. We consider filled pauses also as hesitations but since they were not processed we added the 'Lexical' prefix.

Disfluency and Non-lexical Labeling (DNL): proposed by [17], beyond replacing each partial (and misbuilt) word by a @ symbol it also replaces all types of (transcribed) non-lexical sounds with a # symbol.

Disfluency and Non-lexical Labeling Plus (DNL+): while keeping @ symbols, non-lexical vocalizations are separated into two groups – filled pause like sounds with no relevant meaning, e.g. 'ee' or 'eh' are replaced by #, and backchannel or response like conversational grunts with some (affirmative) meaning, e.g., 'mhm', 'hm' are replaced by & symbols.

Non-lexical Labeling (NL): both disfluent (filled pauses) and normal (backchannel) like non-lexical sounds are replaced by # symbols.

Non-lexical Labeling Plus (NL+): disfluent non-lexicals – typically filled pauses – are replaced by # symbols whereas non-lexicals potentially having a role in the conversation are replaced by & symbols.

Disfluency and Non-lexical Removal (DNR): proposed originally by [13], this modified approach – according to [15] – eliminates all disfluencies and non-lexical units from the training text.

Non-lexical Removal (NR): in this technique we delete all the non-lexical units from the transcription so that only hesitations – broken or misbuilt words – are kept, marked with an attached '-' (as in the Baseline).

Phonetically Consistent (PC): applied by [21], this approach deletes all the non-lexical events and removes '-' marks from broken or misbuilt words. Thus, only the basic alphabet is used in the transcriptions, in a phonetically consistent manner.

Table 3: *Systematic comparison of disfluency and non-lexical labeling methods with Word Error Rates [%] on two speech corpora.*

Model	Deletion	Training text (expected output before cleaning)	Reference text (expected output after cleaning)	BEA-Base Dev WER	BEA-Base Eval WER	GRASS Dev WER	GRASS Eval WER
BL	–	uh ex- ex- excluding em the stuff mhm	uh ex- ex- excluding em the stuff mhm	18.36	19.26	22.26	24.95
LHL	–	uh @ @ excluding em the stuff mhm	uh @ @ excluding em the stuff mhm	18.09	19.15	22.20	24.47
DNL	–	# @ @ excluding # the stuff #	# @ @ excluding # the stuff #	18.19	18.85	*21.15*	*23.34*
DNL+	–	# @ @ excluding # the stuff &	# @ @ excluding # the stuff &	*17.76*	18.88	21.80	*23.25*
NL	–	# ex- ex- excluding # the stuff #	# ex- ex- excluding # the stuff #	18.16	19.03	21.49	*23.60*
NL+	–	# ex- ex- excluding # the stuff &	# ex- ex- excluding # the stuff &	18.27	19.18	21.64	23.90
LHL[14, 15]	@	uh @ @ excluding em the stuff mhm	uh excluding em the stuff mhm	17.86	18.96	22.01	24.32
DNL+	@	# @ @ excluding # the stuff &	# excluding # the stuff &	*17.55*	*18.67*	21.62	*23.16*
NL+	broken	# ex- ex- excluding # the stuff &	# excluding # the stuff &	17.77	18.79	*21.30*	*23.59*
NL	#	# ex- ex- excluding # the stuff #	ex- ex- excluding the stuff	*17.69*	*18.63*	21.91	23.99
NL+	#, &	# ex- ex- excluding # the stuff &	ex- ex- excluding the stuff	*17.75*	*18.64*	22.03	24.30
NR	–	ex- ex- excluding the stuff	ex- ex- excluding the stuff	*17.28*	*18.16*	22.22	24.14
NL	#, '-'	# ex- ex- excluding # the stuff #	ex ex excluding the stuff	*17.25*	*18.24*	21.74	23.85
NL+	#, &, '-'	# ex- ex- excluding # the stuff &	ex ex excluding the stuff	*17.27*	*18.25*	21.84	24.16
NR	'-'	ex- ex- excluding the stuff	ex ex excluding the stuff	***16.89***	***17.75***	22.05	24.04
PC[21]	–	ex ex excluding the stuff	ex ex excluding the stuff	***16.77***	***17.98***	22.05	24.14
DNL[17]	@, #	# @ @ excluding # the stuff #	excluding the stuff	*17.53*	*18.20*	*21.31*	*23.67*
DNL+	@, #, &	# @ @ excluding # the stuff &	excluding the stuff	***17.02***	***18.10***	22.02	*23.54*
NL	#, broken	# ex- ex- excluding # the stuff #	excluding the stuff	*17.15*	*18.19*	21.55	*23.63*
NL+	#, &, broken	# ex- ex- excluding # the stuff &	excluding the stuff	*17.22*	*18.21*	21.67	24.00
DNR[13]	–	excluding the stuff	excluding the stuff	***16.86***	***17.67***	21.67	23.70
NR	broken	ex- ex- excluding the stuff	excluding the stuff	***16.72***	***17.72***	21.91	23.80

Table 4: *Detailed evaluation of disfluency and non-lexical symbol emitting labeling methods by separated Error Rate (ER) calculation for text and symbol tokens. Symbols kept after cleaning the ASR predictions are displayed. N stands for token number.*

		BEA-Base V2 Dev				BEA-Base V2 Eval				GRASS Dev				GRASS Eval			
		Text word		Symbol		Text word		Symbol		Text word		Symbol		Text word		Symbol	
Model	*Symbols kept*	N	ER	N	ER	N	ER	N	ER	N	ER	N	ER	N	ER	N	ER
LHL	@	28 606	**17.43**	431	**61.95**	36 125	**18.65**	354	**69.77**	18 882	**21.78**	186	**65.05**	9 142	**23.99**	93	**72.04**
DNL	@, #	27 450	**16.91**	1 587	**40.34**	34 753	**17.91**	1 726	**37.66**	18 326	**20.97**	742	**25.61**	8 904	**23.29**	331	**24.47**
DNL	#	27 450	**17.43**	1 156	**31.09**	34 753	**18.16**	1 372	**29.18**	18 326	**21.20**	556	**11.69**	8 904	**23.63**	238	**7.14**
DNL	@	27 450	**16.98**	431	**65.43**	34 753	**17.94**	354	**74.86**	18 326	**21.10**	186	**69.89**	8 904	**23.36**	93	**69.89**
DNL+	@, #, &	27 450	**16.46**	1 587	**40.39**	34 753	**17.78**	1 726	**41.02**	18 326	**21.64**	742	**25.74**	8 904	**23.15**	331	**25.98**
DNL+	#, &	27 450	**16.94**	1 156	**31.92**	34 753	**18.05**	1 372	**34.57**	18 326	**21.90**	556	**12.41**	8 904	**23.48**	238	**11.34**
DNL+	@	27 450	**16.54**	431	**64.27**	34 753	**17.83**	354	**70.06**	18 326	**21.79**	186	**69.35**	8 904	**23.21**	93	**63.44**
NL	#	27 881	**17.59**	1 156	**31.87**	35 107	**18.51**	1 372	**32.39**	18 512	**21.85**	556	**9.71**	8 997	**23.93**	238	**10.92**
NL+	#, &	27 881	**17.69**	1 156	**32.35**	35 107	**18.55**	1 372	**35.15**	18 512	**21.91**	556	**12.59**	8 997	**24.23**	238	**11.34**

3.3. Model Training

Separate end-to-end ASR acoustic models were trained for each of the disfluency labeling methods, resulting in 9–9 models for Hungarian and Austrian German, respectively. All model variants were evaluated both on the dev and eval sets – note that dev sets were used for hyperparameter optimization only for the baseline models. For technical details, see Section 4.1.

3.4. Evaluation

In the evaluations in Table 3 – unlike in the previous studies – we applied a step-by-step disfluency/non-lexical symbol deletion (both from ASR output and reference texts) and evaluation approach. This way we could assess the contribution of certain disfluency or non-lexical sound types to the overall error rates. Also, this allowed theoretically more correct comparisons by grouping disfluency labeling approach emitting the same number of reference tokens.

Finally, in Table 4, for more refined error analysis, we report details on text and symbol error rates. Similarly to [13], we count deletions, substitutions and insertions separately for text words and disfluency symbols. In case of insertions, the identity of the predicted token determines whether it is a text word or symbol insertion.

4. Experimental results

4.1. Experimental setup

For the **Hungarian** ASR experiments on BEA-Base V2 we applied the Fast-Conformer [24] implementation of NVIDIA NeMo [25] v1.22.0 toolkit. The 121 Million parameter large model was fine-tuned on the 'train-114' training set initialized with the weights of an English language pretrained model [24]. CTC [7] training with 150 epochs on two RTX A6000 GPUs with a total batch size of 192 was performed using a maximum learning rate of 0.001, cosine annealing, 5% warmup ratio, AdamW optimizer [26], weight decay of 0.01, betas=[0.9, 0.98]. We kept the subword vocabulary size of 1024 of the pretrained model and trained the BPE-Unigram tokenizer [27] [28] for each disfluency labeling approach on the respective version of the training text. In all other aspects the default (hyper-) parameters of NeMo's fast-conformer CTC training were preserved. The speech recognition results are comparable to those achieved by [21] and [29] using similar size models.

Regarding the **Austrian German** ASR configuration, we employed the wav2vec2 large architecture with a light (two-layer feed-forward) decoder and used CTC loss, resulting in a model with 330 million parameters. The SSL-pretrained West Germanic model [30] served as the basis for weight initialization. We used the SpeechBrain v0.15 toolkit [31] with its default LibriSpeech CTC-only wav2vec2 recipe applying 150 epochs of training on a single RTX A6000 GPU with a batch size of 4. Similarly to the Hungarian scenario, a BPE-Unigram tokenizer with a vocabulary size of 1000 (inclusive of disfluency and non-lexical symbols) was trained for each experiment on the corresponding training set. To the best of our knowledge, these WER results stand as the lowest published thus far on the GRASS dataset, on the given split and without the integration of a language model.

4.2. Discussion

Based on 95% confidence intervals calculated for Table 3, weak improvements over the baseline (new results are outside of the baseline confidence interval) are denoted with italics, whereas clear improvements (no overlap between confidence intervals) are written in bold. Regarding **Hungarian**, a tendency of improvements can be observed: the more labels are deleted the lower error rates can be obtained. The four best results with clear improvements, however, are coming from removal-based models which lack the ability to predict any non-lexical sound. From the models with non-lexical/disfluency recognition capabilities only the DNL+ approach could achieve a clear improvement but only when all the symbols were deleted from the prediction. Applying the very same approaches to **Austrian German**, almost no tendencies could be observed and only weak improvements were measured. Interestingly, overall best results were yielded by methods not deleting (just replacing) disfluent or non-lexical tokens (DNL and DNL+). It is worth noting that when WERs with confidence intervals were compared within a 'same-reference-token-number' group there were no clear improvements at all (and only a few weak), independently of test set and language.

The detailed analysis of recognition errors in Table 4 reveals the cause of different behaviour of the two – Hungarian vs. Austrian German – data sets in the experiments: the recognition rates – and also the distribution (see Table 1 and 2) – of non-lexical sounds are remarkably different. The error rates for the @ symbols (partial or misbuilt words) are *only* surprisingly similar, around 70% across data sets and languages. The error rates for non-lexical sounds (see lines with # alone or with # and &), however, are in the range of 29–35% for Hungarian (almost the double of text WER), whereas they are in the range of 7–12% for Austrian German which is *about the half* of German text WER. The explanation looks straightforward: non-lexical tokens in GRASS are mostly 'mhm' grunts used for responses and backchanneling and are not disfluent at all, therefore their human and machine perception is relative easy. On the other hand, non-lexical sounds in the BEA-Base V2 do almost entirely belong to disfluencies, so their recognition is more challenging.

5. Conclusions

We compared several disfluency and non-lexical sound labeling approaches systematically on two different language corpora in parallel. Techniques published with significant improvements on other corpora [14, 15, 17] failed in achieving clear improvement on our data sets. Simpler methods trained with no explicit disfluency/non-lexical markings ([13], PC, NR) provided the best results for Hungarian with statistically significant improvements. However, all the methods failed to clearly improve WER on Austrian German raising an issue on portability of disfluency and non-lexical labeling techniques.

We have shown, though, by calculating error rates separately for disfluent/non-lexical tokens (and textual words) that keeping their labels can be useful: important conversational grunts could be recognized with high accuracy (down to 7% error rate) for Austrian German and even disfluent non-lexical sound recognition reached a reasonable (29%) error rate for Hungarian. The recognition of lexical disfluencies (partial or misbuilt words) is, however, difficult, as only the third of them could be identified correctly at best.

Even though the improvements due to disfluency and non-lexical labeling might be minor, it is worth mentioning that all the approaches numerically outperformed all the baselines. This suggests that using alphabetical letters to transcribe speech phenomena that are hard to transliterate, i. e., non-lexical sounds, is not optimal. Instead, distinct non-alphabetical symbols may be used for them in the training.

Our results suggest, researchers may decide themselves whether non-lexical/disfluent tokens are important for their ASR task and, accordingly, keep the respective tokens in the training or discard (or even do not transcribe) them, since the implicit modeling worked well even in the case of shallow decoders with CTC training. On the other hand, if labels for disfluencies and non-lexicals are already available it should not be a significant overhead to involve them in the training – after prediction they still can be removed. Keeping the labels, however, may facilitate novel applications such as monitoring and diagnostics for speech pathology, help in speech corpus creation and also can make question answering applications more efficient and convenient.

As a general conclusion we want to emphasize the importance of going beyond WER comparison in the evaluation of conversational speech recognition. Even if their contribution to WER can be minimal, the accurate recognition of some response type conversational grunts – like 'mhm' in our experiments – can be essential [10]. Therefore, separate error rate measurement for non-lexical tokens and/or the clarification on what type of tokens contribute to WER is highly relevant. As for future work we plan to drill down to analyze the impact of individual disfluency and non-lexical types on ASR performance.

6. Acknowledgements

This work was supported partially by the Hungarian NRDI Fund through projects NKFIH K143075 and K135038, NKFIH-828- 2/2021(MILAB) and by the NVIDIA Academic Hardware Grant. The work was further supported by grant P-32700-NB from FWF (Austrian Science Fund).

7. References

[1] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics & Cognition*, 2006.

[2] A. Stolcke, E. Shriberg, D. Z. Hakkani-Tur, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *EUROSPEECH*, 1999.

[3] R. Rose and G. Riccardi, "Modeling disfluency and background events in ASR for a natural language understanding task," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 1. IEEE, 1999, pp. 341–344.

[4] V. Rangarajan and S. Narayanan, "Analysis of disfluent repetitions in spontaneous speech recognition," in *2006 14th European Signal Processing Conference (EUSIPCO)*. IEEE, 2006, pp. 1–5.

[5] S. Goldwater, D. Jurafsky, and C. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates," *Computer Speech & Language*, vol. 22, no. 4, pp. 380–388, 2008.

[6] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[7] 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.

[8] K. López-de Ipiña, U. Martinez-de Lizarduy, P. M. Calvo, B. Beitia, J. García-Melero, E. Fernández, M. Ecay-Torres, M. Faundez-Zanuy, and P. Sanz, "On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment," *Neural Computing and Applications*, vol. 32, no. 20, pp. 15 761–15 769, 2020. [Online]. Available: https://doi.org/10.1007/s00521-018-3494-1

[9] A. Albanna, E. Edirisinghe, H. Fang, and W. Hadi, "Stuttering disfluency detection using machine learning approaches," *Journal of Information Knowledge Management*, vol. 21, 04 2022.

[10] B. Tran, K. Latif, T. Reynolds, J. Park, J. Elston Lafata, M. Tai-Seale, and K. Zheng, ""mm-hm," "uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology?" *J Am Med Inform Assoc*, vol. 30, no. 4, pp. 703–711, Mar 2023.

[11] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2 ldc97s62," Web Download, Philadelphia, 1993.

[12] T. N. I. for Japanese Language, "Construction of the corpus of spontaneous Japanese," *The National Language Research Institute Research Report No. 124*, 2006.

[13] P. J. Lou and M. Johnson, "End-to-end speech recognition and disfluency removal," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2051–2061.

[14] V. Mendelev, T. Raissi, G. Camporese, and M. Giollo, "Improved robustness to disfluencies in RNN-transducer based speech recognition," in *ICASSP 2021*. IEEE, 2021, pp. 6878–6882.

[15] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-end spontaneous speech recognition using hesitation labeling," in *2021 Asia-Pacific Signal and Information Processing Association ASC*. IEEE, 2021, pp. 1077–1081.

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

[17] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoga, "End-to-end spontaneous speech recognition using disfluency labeling," in *Proc. Interspeech 2022*, 2022, pp. 4108–4112.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[19] M. Gósy, "Bea–a multifunctional Hungarian spoken language database," *Phonetica*, vol. 105, pp. 50–61, 2013.

[20] T. Neuberger, D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *International Conference on Text, Speech, and Dialogue*. Springer, Cham, 2014, pp. 424–431.

[21] P. Mihajlik, A. Balog, T. E. Graczi, A. Kohari, B. Tarján, and K. Mady, "BEA-base: A benchmark for ASR of spontaneous Hungarian," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1970–1977. [Online]. Available: https://aclanthology.org/2022.lrec-1.211

[22] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of Read And Spontaneous Speech," in *LREC*, 2014, pp. 1465–1470.

[23] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[24] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, and B. Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," in *In arXiv: 2305.05084, eess.AS*, 2023.

[25] E. Harper, S. Majumdar, O. Kuchaiev, J. Li, Y. Zhang, E. Bakhturina, V. Noroozi, S. Subramanian, N. Koluguri, J. Huang, F. Jia, J. Balam, X. Yang, M. Livne, Y. Dong, S. Naren, and B. Ginsburg, "Nemo: a toolkit for conversational ai and large language models." [Online]. Available: https://nvidia.github.io/NeMo/

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019.

[27] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: ACL, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[28] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *CoRR*, vol. abs/1508.07909, 2015. [Online]. Available: http://arxiv.org/abs/1508.07909

[29] P. Mihajlik, M. S. Kádár, G. Dobsinszki, Y. Meng, M. Kedalai, J. Linke, T. Fegyó, and K. Mády, "What Kind of Multi- or Cross-lingual Pre-training is the most Effective for a Spontaneous, Less-resourced ASR Task?" in *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, 2023, pp. 58–62.

[30] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th AACL and the 11th IJCNLP (Volume 1: Long Papers)*. Online: ACL, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.