HAL open science

ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks

Valentin Pelloin, Franck Dary, Nicolas Hervé, Benoît Favre, Nathalie Camelin, Antoine Laurent, Laurent Besacier

To cite this version:

Valentin Pelloin, Franck Dary, Nicolas Hervé, Benoît Favre, Nathalie Camelin, et al.. ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks. Interspeech 2022, Sep 2022, Incheon, South Korea. hal-03770506

HAL Id: hal-03770506

https://hal.science/hal-03770506

Submitted on 6 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks

Valentin Pelloin[1], *Franck Dary*[2], *Nicolas Hervé*[3], *Benoit Favre*[2], *Nathalie Camelin*[1], *Antoine Laurent*[1], *Laurent Besacier*[4]

[1]Laboratoire d'Informatique de l'Université du Mans (LIUM), France
[2]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[3]Institut National de l'Audiovisuel (INA), France
[4]Naver Labs Europe (NLE), France
valentin.pelloin@univ-lemans.fr

arXiv:2207.01893v1 [cs.CL] 5 Jul 2022

Abstract

We aim at improving spoken language modeling (LM) using very large amount of automatically transcribed speech. We leverage the INA (French National Audiovisual Institute[1]) collection and obtain 19GB of text after applying ASR on 350,000 hours of diverse TV shows. From this, spoken language models are trained either by fine-tuning an existing LM (FlauBERT[2]) or through training a LM from scratch. New models (FlauBERT-Oral) are shared with the community and evaluated for 3 downstream tasks: spoken language understanding, classification of TV shows and speech syntactic parsing. Results show that FlauBERT-Oral can be beneficial compared to its initial FlauBERT version demonstrating that, despite its inherent noisy nature, ASR-generated text can be used to build spoken language models.

1. Introduction

Large pretrained language models (LMs) have been widely adopted as backbone for performing natural language processing tasks, whether on text [1] or on speech transcripts [2]. Those models are typically trained with massive quantities of text, and do not reflect well the specifics of spoken language. In this study, we investigate whether there is some benefit in better covering spoken-language phenomena in pretrained LMs when applying them to speech use cases.

Pretrained LMs, such as BERT [3], have been applied with success to a range of tasks that involve speech: ASR hypothesis rescoring [4], Spoken Language Understanding [5], Dialog State Tracking [6], Conversation Summarization [7], Emotion Recognition [8], Speaker and Language recognition [9]. Yet, LM training data is dominated by textual web content, which diverges from actual spoken language. Even though the web is filled with conversation-like interactions, such as forums and social media in which agrammaticality or colloquiallity are frequent, it does not include some fundamental properties that are inherent to spontaneous speech: floor holding and grabbing, conversational discourse markers, disfluencies such as repairs, adverse effects of noise and speaker overlap, among others. One may wonder if addressing such phenomena without having seen them in training is a capability of current techniques associated with LMs, and whether they matter in the context of everyday spoken language processing tasks.

Building LMs aware of speech phenomena is a difficult task. Recent work addresses the problem through direct learning of acoustic representations [10, 11], but have yet to show competitive performance compared to word-based models for high-level linguistic tasks. Instead, we aim at retraining LMs from speech using word transcripts, and are faced with the inherent challenge of gathering enough training material to be on par with text-only models. Therefore, we investigate the use of massive amounts of automatic speech recognition (ASR) generated text for spoken language modeling. This could bring diversity (oral/spontaneous style, different topics) to the LM training data and might also be useful for languages with fewer text resources but high availability of speech recordings. More precisely, we build and share oral LMs trained from a large amount (350,000 hours) of French TV shows, and evaluate them on three spoken language tasks: spoken language understanding, topic classification of TV shows, and spoken language syntactic parsing. Results suggest that ASR-generated text is a viable alternative to text not originating from speech when training spoken language models.

To our knowledge, this is the first time a LM is trained on such a massive amount of ASR transcripts. Indeed, [12] evaluated BERT based language models (BERT, RoBERTa) trained on spoken transcripts to investigate their ability to encode properties of spoken language, but most of their experiments are on small quantities of manual transcripts, except one which is trained on 2000h of ASR output (1k Librispeech + 1k proprietary dataset). We train LMs on larger quantities of ASR output, analyse their benefits on speech-specific tasks, and make them available for further research.

2. From FlauBERT to FlauBERT-Oral

2.1. Large speech corpus with ASR transcripts

The FlauBERT-Oral model is trained on 350,000 hours of speech collected from various French TV and radio channels collected at Institut National Audiovisuel (INA) from 2013 to 2020. TV news channels were continuously captured between 6am and midnight each day (BFMTV, LCI, CNews, France 24, France Info). For radio, the morning news were used (Europe1, RMC, RTL, France Inter) and for generalist TV channels we collected evening news (TF1, France 2, France 3, M6).

Collected speech is automatically transcribed with an ASR system built from Kaldi [13]. In that system, acoustic models are trained from French TV and RADIO corpora (ESTER 1&2 [14], REPERE [15] VERA [16]). A regular backoff n-gram model is estimated using manual transcripts augmented with several French newspapers [17], resulting in a vocabulary of 160k most frequent words. Automatic speech diarization of the

[1]https://www.ina.fr
[2]https://github.com/getalp/Flaubert

collection was performed using LIUMSpkDiarization [18]. The whole pipeline obtains the following ASR performance on different reference test corpora: REPERE (12.1%WER), ESTER1 (8.8%WER) and ESTER2 (10.7%WER).

As the ASR system outputs raw text without punctuation nor capitalization, speech turns from diarization serve as pseudo sentences for later processing. This results in 51M unique speech segments for a total of 3.5G words (19GB of data). It is important to note that this 'oral' data is highly biased towards radio and TV news type.

2.2. Fine-tuning or re-training FlauBERT

All the models generated in this study are built on the FlauBERT family of models, one of the first pre-trained LMs for French [19]. FlauBERT is trained with a masked language modeling (MLM) objective, and implements the RoBERTa architecture [20]. It is trained on 71GB of natural texts collected from various sources, among which a majority of crawled web pages. The baseline for our experiments is *flaubert-base-uncased* (**FBU**)[3] since we only have lowercase transcripts. We compare it to several learning configurations based on the same architecture in order to observe the impact of different parameters on the quality of the obtained models.

The first configuration, *FlauBERT-O-base_uncased* (**FT**), consists in fine-tuning *flaubert-base-uncased* for several epochs using the large corpus of ASR transcripts.

The second configuration *FlauBERT-O-mixed* (**MIX**) is a full model re-trained using a balanced mix of ASR transcripts and written text. Written texts come from two main sources: a French Wikipedia dump and press articles captured by the OTMedia research platform [21] (online press and AFP agency for the same time period). Overall, this learning dataset is also strongly news-biased. The written part of this training dataset is 94M randomly selected sentences representing 13G of data stripped from punctuation and capitalization to make it consistent with ASR data. For this mixed model, we also retrain the BPE tokenizer (50K sub-word units).

The third configuration, *FlauBERT-O-asr*, consists in re-training two LMs from scratch using ASR data only. For the first model (**ORAL**), we use the tokenizer provided with *flaubert-base-uncased* and for the second one (**ORAL_NB**) we retrain a BPE tokenizer (50K sub-word units). Both tokenizers share 52% of their tokens only.

The four models were trained on a single server with two 12-core Xeon CPUs, 256 GB of RAM and 8 Nvidia GeForce RTX 2080 Ti GPUs with 11 GB of memory. It lasted 15 days for 50 epochs of each model in the flaubert-base configuration (137M parameters) using the original FlauBERT code. The models are made available online.[4] The following sections describe experiments on various downstream tasks.

3. Spoken Language Understanding

Spoken Language Understanding (SLU) systems extract semantic information from speech utterances [22], which can be represented with slots or concepts tags on the token sequence. Pipeline SLU systems typically generate transcripts from spoken dialogues with an ASR system, and recognize slots with a Natural Language Understanding (NLU) module that only uses the word sequence [23, 24]. For example, in the recent approach by [25], a first ASR model finetuned on the domain obtains transcriptions, which is fed to a finetuned BERT-like model for classification. In order to reduce the impact of error propagation in the pipeline, alternative approaches have been proposed, such as end-to-end models that directly predict concepts from speech [26, 27, 28]. Our assumption with FlauBERT-Oral models is that the use of ASR-robust pretrained language models for NLU can decrease the impact of ASR errors without resorting to end-to-end training.

Table 1: *Sample instance from MEDIA. (a) corresponds to the transcribed sentence; (b) with associated concepts; (c) with concepts and normalized values; (d) concepts in BIO format; utterances were translated to English.*

(a)	i would like to book one double room in paris
(b)	<**cmd-task** i would like to book > <**nb-room** one > <**room-type** double room > in <**loc-city** paris >
(c)	<**cmd-task** booking > <**nb-room** 1 > <**room-type** double > in <**loc-city** paris >
(d)	**B-cmd-task I-cmd-task I-cmd-task I-cmd-task I-cmd-task B-nb-room B-room-type I-room-type O B-loc-city**

3.1. The MEDIA SLU benchmark

The French MEDIA dataset [29] contains $1,250$ telephone dialogues captured in Wizard-of-Oz settings for hotel booking and information requests in French. Transcribed segments are annotated with 76 concept tags that denote both domain-specific information (such as *hotel-services* or *room-type*) and generic concepts (like logical connectors as *connectprop*). Normalized concept values also have to be predicted. An example can be found in Table 1.

We use the official train/dev/test split of the MEDIA corpus, containing respectively 12.9k, 1.3k and 3.5k user utterances. SLU performance is evaluated with Concept Error Rate (CER), the rate of errors between force-aligned reference and hypothesis concepts, and Concept-Value Error Rate (CVER), the same for (concept, value) pairs. Values are extracted with the same rule-based system as in other studies on MEDIA [25, 27, 30].

3.2. Experiments

We use the different pre-trained models presented in section 2.2 as pre-trained Natural Language Understanding (NLU) components. We add a decision layer at the representations generated by the models, in order to tag each input token to its corresponding BIO concept. Each model is trained for 300 epochs, and the best checkpoint is selected according to CER on the development corpus. We release models and make them available online.[5]

Table 2 presents the *test* results for the different FlauBERT configurations. The baseline State-Of-The-Art (SOTA) results [25] are obtained with a cascade composed of wav2vec 2.0 [31] (finetuned on CommonVoice [32] and MEDIA), followed by a finetuned CamemBERT [33], similar to our approach. For transcription, we use a Kaldi system that reaches 9.1% WER on MEDIA (test)—ASR for the SOTA system has a WER of 8.5%.

We can see that all the FlauBERT-O uncased models (FT,

[3] https://huggingface.co/flaubert/flaubert_base_uncased

[4] https://huggingface.co/nherve

[5] https://huggingface.co/vpelloin/MEDIA_NLU_flaubert_uncased, https://huggingface.co/vpelloin/MEDIA_NLU_flaubert_finetuned

Table 2: *SLU Task: MEDIA test results according to Concept Error Rate and Concept-Value Error Rate (lower is better), using ASR transcriptions as input of NLU FlauBERT models. Confidence intervals are computed using the Student's t-distribution at* 95%.

Model	%CER	%CVER
w2v 2.0+CamemBERT [25]	11.2	17.2
FlauBERT cased	13.20 ± 0.71	18.25 ± 0.81
FlauBERT uncased (FBU)	12.40 ± 0.69	17.49 ± 0.79
FT	**11.98** ± 0.68	**17.00** ± 0.79
MIX	12.47 ± 0.69	17.66 ± 0.80
ORAL	12.43 ± 0.69	17.84 ± 0.80
ORAL_NB	12.24 ± 0.68	17.67 ± 0.80

MIX, ORAL and ORAL_NB) achieve comparable results with the baseline FlauBERT models (cased and uncased). It can be noticed that the FT model obtains an absolute improvement of 0.42% CER compared to FlauBERT uncased. This suggests that pretraining LMs on ASR transcripts can benefit downstream tasks even though the domain is quite different (TV shows vs telephone conversations). Although we use ASR transcripts with lower quality compared to [25] we obtain almost similar CER using the FT model.

One limitation of our experiments is that, even though SLU is applied to (noisy) ASR transcripts, during fine-tuning SLU models are updated using the (error free) reference transcripts of MEDIA. We believe that fine-tuning them on (noisy) ASR output could lead to even better results as it would match pre-training and test conditions. This improvement is left for future work but the next section will propose a classification task where ASR outputs are used during both fine-tuning and evaluation.

4. Automatic Classification of TV Shows

This section describes the evaluation of the different LMs on news classification. As part of the collection process, INA's documentalists finely segment newscasts of the main generalist TV/radio channels, and annotate them with content descriptions. This very rich metadata is used in particular to establish quantitative studies on the news in France. The InaStat barometer[6] has set up a stable methodology over time to classify news items into 14 categories (such as society, French politics, sport or environment). For classification experiments, we use the news items of 4 channels (TF1, France 2, France 3 and M6) for the years 2017, 2018 and 2019, which gives a total of 47,867 short TV shows, running on average 92 seconds.

The objective is to assess classification accuracy into the 14 categories solely on the basis of ASR transcripts. We compare results to a simple SVM classifier (with a non-parametric triangular kernel) on TF-IDF vectors baseline, with two vocabulary sizes of 5K and 20K words. A FlauBERT classifier is created for each model by appending a classification layer after mean pooling. Models are not selected and rather we report results for different number of training epochs. Since the categories are not well-balanced, we use weighted F1 to evaluate performance. The experiments are systematically performed on 10 different random splits of the dataset, taking into account the cardinality of the 14 categories, so as to have 38K examples for the training set and 5K for the test set. Mean performance results and standard deviation are shown in Table 3.

Table 3: *TV news classification task: train 38K, test 5K, F1 weighted metric. Higher is better.*

Model	Epoch 1	Epoch 3	Epoch 10
SVM-20K		0.777 ± 0.005	
FBU	0.780 ± 0.012	0.807 ± 0.010	0.809 ± 0.006
FT	0.784 ± 0.006	0.807 ± 0.006	0.817 ± 0.004
MIX	0.791 ± 0.005	0.812 ± 0.003	0.820 ± 0.004
ORAL	**0.811** ± 0.004	**0.829** ± 0.005	**0.824** ± 0.004
ORAL_NB	0.798 ± 0.004	0.814 ± 0.005	0.824 ± 0.004

If we look at the performance at the first epoch, the FBU (flaubert_base_uncased) model has almost equivalent performance to the SVM baseline (0.777). It is only after a few iterations of fine-tuning that the model fits the ASR data and reaches 0.809. On the other hand, the models that have already seen ASR data during training have better performance from the first epoch. The model trained only on ASR data is the best (ORAL). The ORAL_NB model is slightly worse: in this model, only the tokenizer is different, so it seems better to not adapt BPE to ASR transcripts, probably because it suffers from the closed vocabulary of the underlying system. After 10 epochs, all FlauBERT-Oral models converge to the same performance and are better than FBU for this task.

5. Syntactic Parsing of Spoken Dialogs

In this section, the downstream task consists in jointly predicting part of speech tags (POS) and building a labelled dependency tree from speech transcripts. We use our different LMs to generate contextual word representations and use them for predicting syntactic parses of speech transcripts. We contrast the obtained results to a baseline model trained using FastText non-contextual representations,[7] and a model learning its own representations without any pretraining.

5.1. The ORFEO dataset

Our study relies on the annotated subset of the speech corpus of the ORFEO project [34, 35], gathered with the goal of reflecting the contemporary usage of the French language. The audio recordings include work meetings, family dinners, narrations, political meetings, interviews, and goal-oriented telephone conversations. Their duration varies from 4mn to 1h. The corpus is annotated with part-of-speech (POS) tags, lemmas, labeled dependency trees and sentences boundaries. Annotations include 20 possible POS tags and 12 syntactic functions. We randomly split the corpus into train/dev/test sets of respective sizes 134,716/27,937/29,529 words; we sampled from each source so that the various genres of speech are equally represented in each split.

5.2. Parsing Model

The syntax-predicting model is a transition based parser using the arc-eager transition system, which has been extended for

[6] http://www.inatheque.fr/publications-evenements/ina-stat/ina-stat-sommaire.html

[7] https://fasttext.cc

the joint prediction of POS tags and parsing transitions [36]. A single classifier is trained to predict transitions from features extracted from the current state of the analysis (word embebdings, predicted tags, history of transitions), for stack elements and a sliding window of size [-3;2] centered on the current word. The POS sequence and dependency trees are greedily built from the sequence of transitions predicted by the classifier. A dropout of 50% is applied to the input vector, which passes through two hidden layers of size 3200 and 1600, both with 40% of dropout and ReLU activations. The two decision layers are linear with a softmax. Training runs for 40 epochs, using a dynamical oracle [37], with scoring against the development set to select the best checkpoint.

5.3. Experiments

The first set of experiments compares input representations from the FlauBERT variants (FBU, FT, MIX, ORAL) to non-contextual baselines. As with other experiments, LMs are applied on uncased transcripts without punctuation, which is not fair to the FBU model as it expects punctuation. Therefore we add a contrastive system with periods added at the end of each utterance prior to applying the LM (called FBU repunc), but added tokens are not passed to the parser, they just inform other embeddings. Note that except for random initialized embeddings, token representations are not fine-tuned during training.

As pre-processing, we deanonymize the transcripts by replacing masked proper name tokens with non-ambiguous names randomly chosen for each recording. In the fasttext setting, representations are computed for unknown words from their character n-gram factors. Token representations are computed at the whole recording level in chunks of 512 tokens without overlap. The parser is applied on the reference transcript and reference segmentation. We use mean pooling for words that are split in multiple tokens by BPE.

Parsing performance is evaluated with Labeled Attachment Score (LAS), the accuracy of predicting the governor of each word and its dependency label, Unlabeled Attachment Score (UAS), which ignores dependency labels, and Part-of-speech tagging accuracy (UPOS) (scoring script from CoNLL campaigns).

Results presented in Table 4 show that pre-training is valuable for syntactic parsing in that setting and that pretraining on ASR-generated text (**FT**, **MIX** and **ORAL**) leads to a substantial improvement in LAS over the text-only FlauBERT model (**FBU**) even though there is no domain overlap between the TV shows on which the LMs are trained and the data of the ORFEO corpus. As in the previous task, we see no benefit in retraining BPEs (**ORAL_NB**). Performance in the **FBU repunc** settings is on par with the best model, showing that absence of punctuation is an important factor affecting performance when applying text-only models, at least for predicting syntax.

Finally, in Table 5 we compare the drop in parsing performance between three models, when they are evaluated only against OOV words (according to the ASR system). Revealing that the **ORAL** model variants are more affected than the off-the-shelf FlauBERT baseline (**FBU**), despite their overall better performance (as seen in Table 4). This suggests future improvements of LMs trained on ASR-generated text if ASR output tokens were more consistent with LM tokens, and also suggests applying open vocabulary ASR.

Table 4: *Syntax prediction task: metrics are Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS) and Part-of-speech tagging accuracy (UPOS). Higher is better.*

Model	**LAS**	**UAS**	**UPOS**
No pretrain	84.92 ± 0.44	88.48 ± 0.37	94.51 ± 0.29
Fasttext	85.36 ± 0.10	88.76 ± 0.05	95.12 ± 0.03
FBU	85.55 ± 0.05	89.02 ± 0.16	93.36 ± 0.06
FBU repunc	87.48 ± 0.28	90.69 ± 0.19	95.03 ± 0.03
FT	86.81 ± 0.09	90.22 ± 0.06	94.99 ± 0.10
MIX	86.33 ± 0.15	89.79 ± 0.05	94.43 ± 0.29
ORAL	$\mathbf{87.65} \pm 0.11$	$\mathbf{90.92} \pm 0.10$	95.55 ± 0.04
ORAL_NB	87.54 ± 0.11	90.73 ± 0.08	$\mathbf{95.63} \pm 0.05$

Table 5: *Differences in morphosyntactic performances, when parsing models of Table 4 are evaluated only on OOV words (OOV according to the automatic transcription system used to generate the text our LMs were trained on).*

Model	**ΔLAS**	**ΔUAS**	**ΔUPOS**
FBU	-11.45	-6.82	-14.36
MIX	-11.93	-7.33	-14.07
ORAL	**-13.97**	**-8.11**	**-16.55**

6. Conclusion and future work

We investigated spoken language modeling using a massive amount of ASR-generated text (350,000 hours of diverse TV shows). Experiments suggest that, for the tasks we experimented with, systems relying on speech-informed LMs have similar and better performance than systems relying on text-only LMs. The models resulting from this work for processing French speech are made available to the community. Deeper analysis of these improvements, on when and why models are better for spoken language are items that are still to be addressed in the near future.

In this study, all texts are uncased as our ASR only generates lowercased, unpunctuated transcripts. We believe that applying massively re-capitalisation and restoring punctuation might be beneficial to train stronger LMs for spoken language. We also believe that leveraging open vocabulary ASR and more diverse speech sources might yield more versatile LMs. A more fine-grain analysis on speech-only phenomena such as disfluencies is also needed. Finally, some of the results obtained lead us to believe that it is important to further evaluate the impact of BPE units for spoken language modeling, as well as the consistency between ASR and LM tokens' vocabulary.

7. Acknowledgements

Some co-authors were funded by the AISSPER project supported by the French National Research Agency (ANR) under contract ANR-19-CE23-0004-01.

8. References

[1] M. Koroteev, "Bert: A review of applications in natural language processing and understanding," *arXiv*, p. 2103, 2021.

[2] L. Qin, T. Xie, W. Che, and T. Liu, "A survey on spoken language understanding: Recent advances and new frontiers," in *IJCAI*, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-

training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using bert for speech recognition," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 1081–1093.

[5] C. Liu, S. Zhu, Z. Zhao, R. Cao, L. Chen, and K. Yu, "Jointly encoding word confusion network and dialogue context with bert for spoken language understanding," *Proc. Interspeech 2020*, pp. 871–875, 2020.

[6] T. M. Lai, Q. H. Tran, T. Bui, and D. Kihara, "A simple but effective bert model for dialog state tracking on resource-limited systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8034–8038.

[7] B. Ma, H. Sun, J. Wang, Q. Qi, and J. Liao, "Extractive dialogue summarization without annotation based on distantly supervised machine reading comprehension in customer service," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 87–97, 2021.

[8] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[9] S. Ling, J. Salazar, Y. Liu, K. Kirchhoff, and A. Amazon, "Bertphone: Phonetically-aware encoder representations for utterance-level speaker and language recognition," in *Proc. Odyssey 2020 the speaker and language recognition workshop*, 2020, pp. 9–16.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] A. Kumar, M. Narayanan Sundararaman, and J. Vepa, "What BERT based language model learns in spoken transcripts: An empirical study," in *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. ACL, Nov. 2021, pp. 322–336.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition." in *LREC*, 2012, pp. 1102–1107.

[16] M. Goryainova, C. Grouin, S. Rosset, and I. Vasilescu, "Morpho-syntactic study of errors from speech recognition system." in *LREC*, vol. 14, 2014, pp. 3050–3056.

[17] P. Deléglise, Y. Esteve, S. Meignier, and T. Merlin, "Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate?" in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[18] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[19] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, "FlauBERT: Unsupervised language model pre-training for French," in *LREC*. Marseille, France: ELRA, May 2020, pp. 2479–2490.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[21] N. Hervé, "OTMedia, the TransMedia news observatory," in *FIAT/IFTA Media Management Seminar 2019*, 05 2019.

[22] G. Tur and R. D. Mori, "Chapter 1: Spoken language understanding for human/machine interactions," in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, 2011.

[23] F. Morbini, K. Audhkhasi, R. Artstein, M. Van Segbroeck, K. Sagae, P. Georgiou, D. R. Traum, and S. Narayanan, "A reranking approach for recognition and classification of speech input in conversational dialogue systems," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 49–54.

[24] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 530–539, 2015.

[25] S. Ghannay, A. Caubrière, S. Mdhaffar, G. Laperrière, B. Jabaian, and Y. Estève, "Where Are We in Semantic Concept Extraction for Spoken Language Understanding?" *Lecture Notes in Computer Science*, vol. 12997 LNAI, pp. 202–213, 2021.

[26] M. Dinarelli, N. Kapoor, B. Jabaian, and L. Besacier, "A data efficient end-to-end spoken language understanding architecture," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 8519–8523.

[27] V. Pelloin, N. Camelin, A. Laurent, R. De Mori, A. Caubrière, Y. Estève, and S. Meignier, "End2end acoustic to semantic transduction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[28] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards End-to-end Spoken Language Understanding," *ICASSP*, 2018.

[29] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *INTERSPEECH*, 2005.

[30] S. Hahn, M. Dinarelli, C. Raymond, F. Lefevre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing stochastic approaches to spoken language understanding in multiple languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1569–1583, 2010.

[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[32] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*. Marseille, France: ELRA, May 2020, pp. 4218–4222.

[33] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL, Jul. 2020, pp. 7203–7219.

[34] C. Benzitoun, J.-M. Debaisieux, and H.-J. Deulofeu, "Le projet orféo: un corpus d'étude pour le français contemporain," *Corpus*, no. 15, 2016.

[35] A. Nasr, F. Dary, F. Bechet, and B. Favre, "Annotation syntaxique automatique de la partie orale du CÉFC," *Langages*, Sep. 2020.

[36] F. Dary and A. Nasr, "The reading machine: A versatile framework for studying incremental parsing strategies," in *Proceedings of IWPT 2021*. Online: ACL, Aug. 2021, pp. 26–37.

[37] Y. Goldberg and J. Nivre, "A dynamic oracle for arc-eager dependency parsing," in *Proceedings of COLING 2012*, 2012, pp. 959–976.