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

Valentin Pelloin1, Franck Dary2, Nicolas Hervé3, Benoit Favre2,
Nathalie Camelin1, Antoine Laurent1, Laurent Besacier4

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

Abstract
We aim at improving spoken language modeling (LM) using very large amounts of automatically transcribed speech. We leverage the INA (French National Audiovisual Institute) 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) 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 grammaticality or colloquiality 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
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) \(^1\) 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 \(^2\) (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 retraining 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.\(^3\) 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],

\(^1\)https://huggingface.co/flaubert/flaubert-base-uncased

\(^2\)https://huggingface.co/nherve

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.

| Example | BIO Format |
|---------|------------|
| (a) i would like to book one double room in paris | B-room-type O-room-type B-city |
| (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` | |
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 transcriptions, 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 INA Stat barometer 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, 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

---

http://www.ina-theque.fr/publications-evenements/ina-stat/ina-stat-zommaire.html

7https://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 embeddings, 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 de-anonymize 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 | 88.48 | 94.51  |
| Fasttext    | 85.36 | 88.76 | 95.12  |
| FBU         | 85.55 | 89.02 | 93.36  |
| FBU repunc  | 87.48 | 90.69 | 95.03  |
| FT          | 86.81 | 90.22 | 94.99  |
| MIX         | 86.33 | 89.79 | 94.43  |
| ORAL        | 87.65 | 90.92 | 95.55  |
| ORAL_NB     | 87.54 | 90.73 | 95.63  |

### 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  | ∆LAS | ∆UAS | ∆UPOS |
|--------|------|------|-------|
| 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 lowercase, 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 AIISSPER 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-
