Detecting the terminality of speech-turn boundary for spoken interactions in French TV and Radio content

Rémi Uro, Marie Tahon, David Doukhan, Antoine Laurent, Albert Rilliard

▶ **To cite this version:**

Rémi Uro, Marie Tahon, David Doukhan, Antoine Laurent, Albert Rilliard. Detecting the terminality of speech-turn boundary for spoken interactions in French TV and Radio content. Interspeech 2024, Itshak Lapidot; Sharon Gannot, Sep 2024, Kos, Greece. pp.3560 - 3564, ⟨10.21437/interspeech.2024-1163⟩. ⟨hal-04694968⟩

HAL Id: hal-04694968

https://hal.science/hal-04694968v1

Submitted on 11 Sep 2024

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.

HAL Authorization

Interspeech 2024
1-5 September 2024, Kos, Greece

Detecting the terminality of speech-turn boundary for spoken interactions in French TV and Radio content

Rémi Uro[1,2], *Marie Tahon*[3], *David Doukhan*[1], *Antoine Laurent*[3], *Albert Rilliard*[2]

[1]French National Institute of Audiovisual (INA), Paris, France. [2]Université Paris Saclay, CNRS, LISN, Orsay, France. [3]LIUM, Le Mans Université, France

ruro@ina.fr, marie.tahon@univ-lemans.fr, ddoukhan@ina.fr,
antoine.laurent@univ-lemans.fr, albert.rilliard@lisn.fr

Abstract

Transition Relevance Places are defined as the end of an utterance where the interlocutor may take the floor without interrupting the current speaker –i.e., a place where the turn is terminal. Analyzing turn terminality is useful to study the dynamic of turn-taking in spontaneous conversations. This paper presents an automatic classification of spoken utterances as Terminal or Non-Terminal in multi-speaker settings. We compared audio, text, and fusions of both approaches on a French corpus of TV and Radio extracts annotated with turn-terminality information at each speaker change. Our models are based on pre-trained self-supervised representations. We report results for different fusion strategies and varying context sizes. This study also questions the problem of performance variability by analyzing the differences in results for multiple training runs with random initialization. The measured accuracy would allow the use of these models for large-scale analysis of turn-taking.

Index Terms: Spoken interaction, Media, TV, Radio, Transition-Relevance Places, Turn Taking, Interruption

1. Introduction

Turn-taking dynamics [1] and its relations to interruptions are important factors in describing broadcast interactions, as they hint at perceived power and dominance in a conversation [2, 3]. [4] introduced the notion of Transition Relevance Places (TRPs), which occur at the end of turn-construction units and allow for smooth transitions between speakers. TRPs are important because speakers can anticipate their occurrences in order to plan a potential turn-taking, optimizing the floor transfer offset [5]. Automatically detecting TRPs would allow for large-scale analysis of complex interruptions-related phenomena. As for other human science objects, allowing large datasets describing turn-tacking dynamics to be produced and analyzed is an important outcome for society. While [6] focuses on overlapping speech segments to study interruptions following a classification schema defined by [7] for French political interviews, interruptions may also occur without overlapping speech [8]. A better understanding of turn-taking also has implications for the development of human-machine interactions [9, 10]. [11] has annotated TRPs in 8 hours of Slovak TV discussions and reports a binary classification accuracy of 94.4% with an ensemble model using fundamental frequency (F0) and intensity curves on chunks of 1.5 s. [12, 13] propose an LSTM-based architecture using acoustic and linguistic features on English spontaneous dialogues to predict whether a speaker change would occur in the three following seconds. [14] presents a Transformer-based approach to the same task based on textual information only. [9] proposed a method for turn-taking prediction in spoken dialog systems. They report substantial inter-rater agreements for annotating TRPs in different acted dialogue scenarios. The proposed LSTM architecture using acoustic and linguistic features achieves binary accuracies between 79.3 and 89.5.

The work presented here focuses on media broadcasts, especially for shows proposing multispeaker interactions. This choice was made to investigate spontaneous spoken dialogs and to allow collaboration with sociologists analyzing roles and behavior in media. We present an automatic classification of spoken utterances as Terminal (i.e., ending with a TRP) or Non-Terminal (not ending with a TRP). The corpus was based on French TV and radio content, specifically multi-party conversation representing spontaneous interactions with various *levels of control* [15]. An existing corpus annotated with TRPs [16] was used to train multimodal models based on the Wav2Vec2 [17] and FlauBERT [18] pre-trained models. We report results comparing different fusion strategies and context sizes, evaluate the relative role of the different information modalities (lexical vs. acoustic) with respect to the length of the segments used for the inference, and discuss the relevance of these findings for spoken interaction modeling.

2. Method

2.1. Data

This study is based on an annotated corpus of multi-party interactions composed of French TV and Radio broadcasts from 1998 to 2015 [16] from ALLIES corpus [19]. Audio chunks corresponding to zones of turn changes were extracted from the complete show to encompass (1) the last speech *segment*[1] before the turn change and (2) the first segment after the change. The segment (1) comes before the turn change and does not include potential overlapped speech, while segment (2) may contain an overlap. These audio chunks were annotated regarding the terminality of the initial speaker's turn (terminal or not) and the second speaker's turn-taking category (interruption, backchannel, or smooth change). One annotator paid for this work (a student trained in linguistics) made the annotations. Two additional persons also annotated a subset of 338 samples to obtain an inter-rater agreement. A substantial agreement of 0.75 using Fleiss' Kappa [21] was observed for the Terminality annotation used in this paper. The work presented here focuses solely on the parts (1), preceding the speaker change, to avoid relying on information linked to the second speaker's intervention. This was decided because we aim to produce a model that classifies speech turn in relation to turn-taking management and the anticipation of a given speaker floor taking [5], thus without knowing what happened later during the dialogue. We se-

[1]defined by the Allies terminology as "sequences containing complete words which are syntactically and semantically coherent" [20]

10.21437/Interspeech.2024-1163

lected the segment (1) occurring before the speaker changes, which constitute the "samples" that are used in the remaining of this study. These samples amounted to a total of 1954 speaker changes annotated with TRP information with 128 different speakers. Table 1 shows the number of samples and their duration characteristics for each Terminality class.

	Nb	≤0.5	0.5<x≤1	1<x≤2	>2
Term.	839	13%	27%	35%	25%
Non Term.	1115	11%	19%	23%	47%

Table 1: *Number of samples and percentage of samples in different duration intervals (in seconds) for the Terminal and Non-Terminal classes*

Table 2 lists the names of the different shows used to select the samplesand the number and total duration of the annotated samples from each show.

Two settings for the chunks used as input of the model were tested: (1) the variable-size samples defined above, that comprises a complete "segment" as defined in the Allies corpus (thereafter *ref*); or (2) a fixed size chunk that correspond to the three seconds before the turn taking event occurring at the end of segment 1, and that could have been obtained using an automatic diarization (thereafter *3s*; these may include other speakers turns at the beginning, or incomplete turn). This was done to evaluate the possibility to apply a fully automatic approach.

For the fixed-sized chunks, other sizes have been tested in preliminary evaluations (2 and 5 seconds) with similar results. Three-second chunks were kept specifically to investigate if the model would work even if provided with parts of previous turns.

2.2. Data preprocessing

For each sample, we extracted the relevant audio waveform, either the annotated segment for *ref* or the last 3 seconds of the recordings before the speaker change for *3s*. All samples were then automatically transcribed using Whisper [22]. To evaluate the model in fully automatic processing pipeline conditions, we kept the Whisper transcriptions as is, even if easily detectable errors occurred (e.g., "Subtitles generated by…"). Manual transcriptions for the *ref* setting were also available from the Allies reference transcriptions and are used for comparison.

2.3. Model architectures

A total of 5 prediction models are proposed, operating on raw audio and/or textual transcriptions. The models are based on the pre-trained self-supervised models `wav2vec2-base` [17] for the audio and `flaubert_base` [18] for the text. They both output a classification token (CLS) of size 768 from variable-size audio or text inputs. Text-only (**TO**) and Audio-only (**AO**) consist of three linear layers after extracting the CLS token of the pre-trained model (see Figure 1). Three additional fusion processes inspired by [23] are defined: Early Fusion (**EF**), with one linear layer after each pre-trained model before concatenation and another three linear layers (see Figure 2); Late Fusion (**LF**) with one linear layer on the concatenation of the TO and AO outputs; and Average Fusion (**AF**) which takes the average of the logits of TO and AO. A dropout of 0.30 was applied between each linear layer during training.

Show	Samples	Dur. (s)
BFMStory (BS)	200	853.34
CaVousRegarde (CR)	269	969.84
CultureEtVous (CV)	7	8.69
DEBATE (D)	128	266.22
EntreLesLignes(EL)	307	1179.53
LaPlaceDuVillage (PV)	770	1316.76
PileEtFace (PF)	150	573.33
PlaneteShowbiz (PS)	10	13.94
TopQuestions (TQ)	5	21.88
fm (FM)	108	254.25

Table 2: *Number samples from each show, and their total duration*

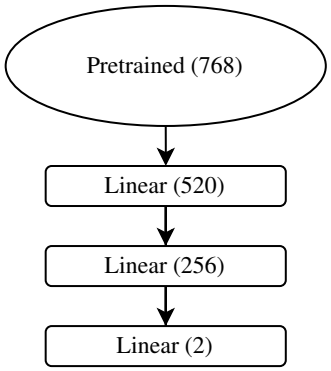

Figure 1: *Single modality model architecture*

3. Results

For better convenience, the model's input used during training or at inference time will be referred to using the following symbols: **ref_auto**: reference speaker segmentation and their automatic speech transcription; **ref_man**: reference speaker segmentation and their manual speech transcription; **3s_auto** fixed 3-second duration excerpts and their automatic speech transcription.

3.1. Evaluation protocol

Models were trained and cross-tested for each train and inference input combination (*ref_auto*, *ref_man* and *3s_auto*), resulting in nine evaluation configurations for each architecture. Each model was evaluated using K-Fold grouped by show from which the examples were extracted. This was done to prevent having samples of the same show in the training and testing sets and to allow for a comparison of the performance degradation associated with materials of different natures (e.g., different types of shows). A random split of 30% of the training set was used as a validation set. A patience of five epochs was used on validation accuracy. This process was run 10 times to estimate the variability due to random initialization, allowing for more reliable accuracy measures (mean and confidence intervals over the 10 runs are given). The training was done on an RTX 4090 GPU, and it took about 80 hours to train the 1500 models for the whole experiment. In order to investigate the significance of accuracy variation and the impact of the different settings, a linear mixed model was fit on the accuracy of these model architectures across training and test sets, with the samples nested

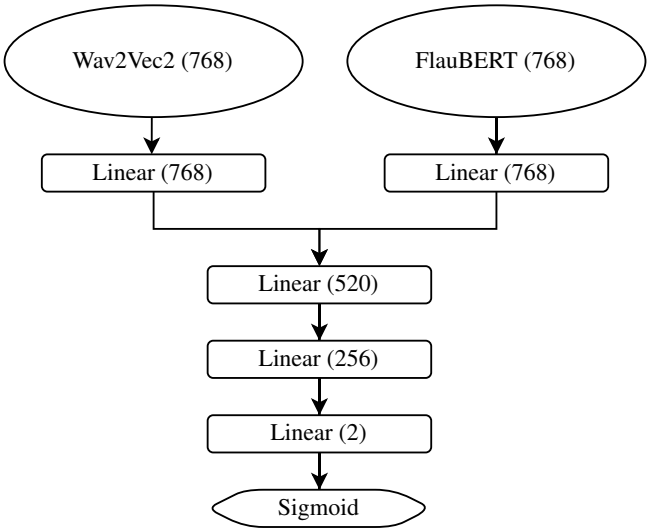

Figure 2: *Early fusion model architecture*

in the show as random factors, as in equation 1.

$$accuracy \sim model*(test+train)+(1|show/sample) \quad (1)$$

where the $accuracy$ is explained by the $model$ architecture and its interactions with the $test$ and $train$ settings as fixed factors, and the $sample$ nested in $show$as random factors. The model was fitted using R's *lme4* library [24, 25]. The mean and confidence interval fitted by the models are plotted in Figure 3. A post hoc comparison of the models' accuracies (using Bonferroni correction) for each level of test and train settings was done.

3.2. Variation across models

All models achieved an average accuracy score above 0.85 as can be seen on Figure 3. This indicates that the proposed approach reached coherent and qualitative results. Figure 3 shows the mean accuracies with their confidence intervals for each combination of model architecture, training set, and testing set, as estimated by the linear mixed model across all random initialization runs. We observed that the Text Only approach performs significantly worse than the other models. Then comes the Late and Average Fusion models, which significantly outperform the TO approach and achieved comparable performances (in terms of accuracy) but are significantly outperformed by the Early Fusion and Audio Only approaches. AO and EF achieve comparable performances, hinting that the audio signal carries an important share of information for this task, including non-linguistic cues. These trends are similar for each training and testing setting. The training settings do not seem to have much impact, all models achieved similar accuracies across training configurations, with the ones trained on the *ref_man* performing slightly better.

Tables 3 and 4 report the mean accuracy for each show, respectively, with the *ref_man* and *3s* models for the different test settings. The EF model performs slightly better than the AO for most shows. Meanwhile, considering the confidence intervals obtained from the ten random runs, the difference does not reach significance. The worst accuracies for both training settings are observed on the *PlanetShowbiz* show, which contains fewer shows (10) and displays an interaction style that differs from the others, with presenters often reading prepared scripts.

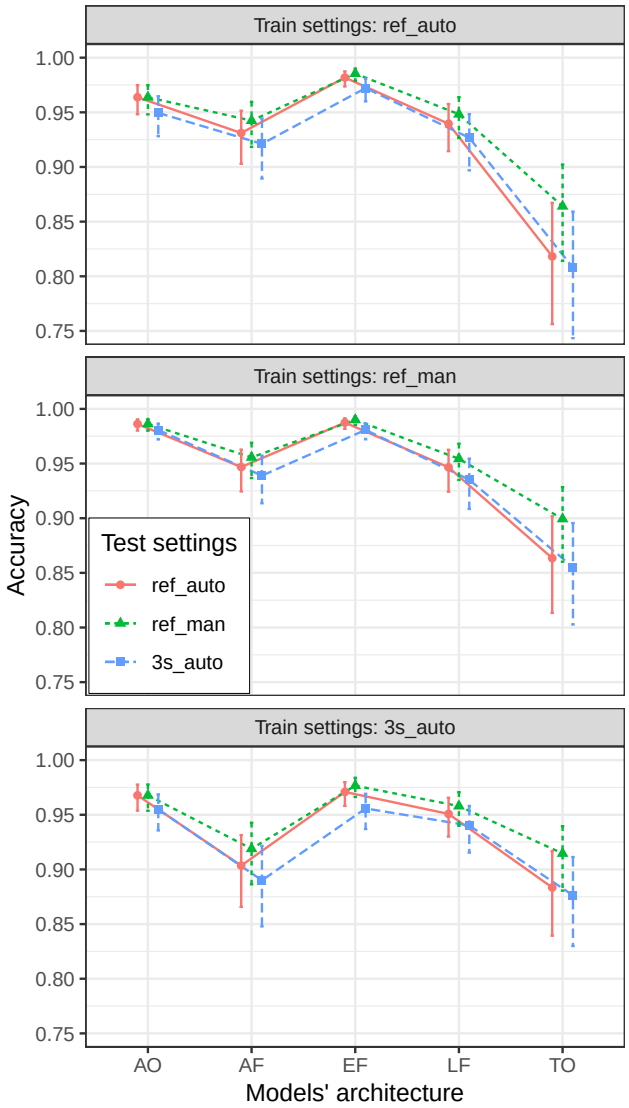

Figure 3: *Mean accuracies for each model's Test settings (lines) with their confidence intervals, for each Train setting (subplots), for the different model architectures (x-axis).*

Using the reference segmentation leads to better results than using a fixed size of three seconds, but a fully automatic approach still reaches accuracies over 90%.

Figure 4 represents the mean accuracy for each model architecture depending on the training setting and the duration of the speech segment in the training sample. We observe that textual information on shorter segments is less reliable than the audio signal, but tends to allow similar performances on longer segments. This is especially true with the automatic settings, as very short segments may not include enough data for a reliable transcription.

4. Discussion & conclusions

We compared different models and data settings to classify the terminality status of speech samples extracted from spontaneous interactions in media archives. While the proposed

Show	AO		EF		
	ref	3s	ref	ref_man	3s
(BS)	97.86	96.21	98.14	**98.21**	95.79
(CR)	**96.49**	94.74	95.27	96.02	93.73
(CV)	**97.96**	87.76	95.92	95.92	89.80
(D)	**96.54**	94.42	96.09	**96.54**	94.20
(EL)	**97.16**	94.23	93.62	**97.16**	88.83
(PV)	91.41	87.25	93.75	**95.84**	90.95
(PF)	98.76	97.43	95.90	**99.33**	93.81
(PS)	**84.29**	77.14	82.86	**84.29**	74.29
(TQ)	91.43	91.43	**100.00**	**100.00**	**100.00**
(FM)	**96.16**	94.18	**96.16**	95.77	95.24
Mean	94.80	91.48	94.77	**95.90**	91.66

Table 3: *Mean accuracies for each show with the AO and EF model trained in the ref_man setting*

Show	AO		EF		
	ref	3s	ref	ref_man	3s
(BS)	97.86	97.64	**98.29**	**98.29**	97.50
(CR)	96.49	96.23	**96.81**	**96.81**	95.75
(CV)	**93.88**	**93.88**	**93.88**	**93.88**	91.84
(D)	96.76	**97.66**	95.87	95.76	96.76
(EL)	97.86	95.86	**97.95**	97.91	96.04
(PV)	88.09	85.77	96.07	**96.38**	93.65
(PF)	98.86	98.86	98.48	**98.95**	97.81
(PS)	**87.14**	81.43	**87.14**	**87.14**	81.43
(TQ)	**94.29**	88.57	**94.29**	**94.29**	91.43
(FM)	96.16	94.58	**96.30**	**96.30**	95.24
Mean	94.74	93.05	95.50	**95.56**	93.75

Table 4: *Mean accuracies for each show with the AO and EF model trained in the 3s setting*

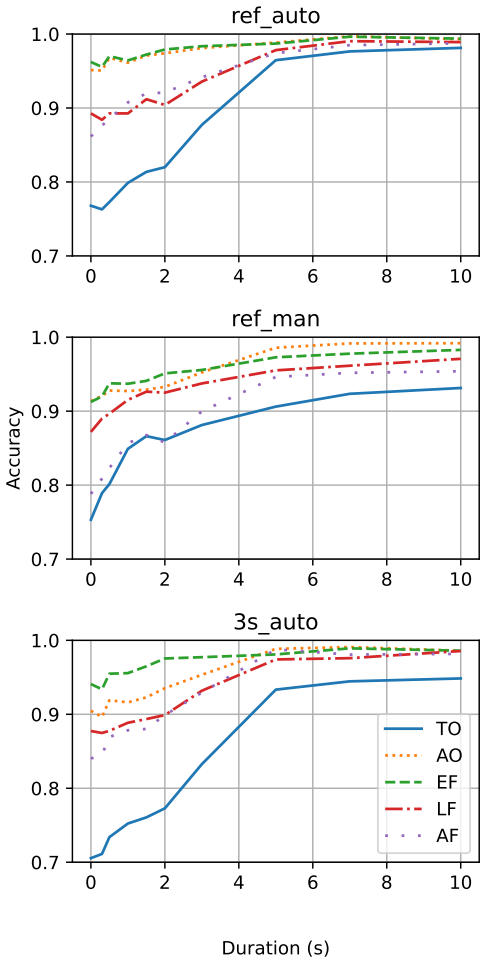

Figure 4: *Mean accuracies as a function of the duration of the testing sample for the different three training settings.*

approaches achieved highly encouraging performances, these models are based on a corpus of about 2000 audio samples from 10 different shows with limited types of interactions and have not yet been tested on other corpora. One important takeaway of this study is that the *Wav2Vec2* representation alone seems sufficient for classifying Transition Relevance Places, as Early Fusion and Audio Only models achieve comparable performances. However, the AO model is smaller and does not require transcription as a preprocessing step, making it less resource-demanding and thus more easily usable in real-life settings. It also potentially has fewer language dependencies than a model (partially) based on text, albeit we didn't test this. The Text Only models also achieve satisfying performances, especially on longer speech turns, showing that lexical information is relevant for this task but, as pointed out by [14], more work on language modeling, specifically focused on the varying levels of boundaries (e.g., [26]), would be important. Comparing different settings for data preprocessing (manual segmentation vs. fixed-size window; manual vs. automatic transcription) shows that the proposed models are fairly invariant to the preprocessing setting, hinting that they could be used in a fully automatic setting. Applying automatic diarization may also help to reduce the gap between manual and automatic settings at inference. Testing the proposed models on other languages and interaction types could prove useful in better understanding the differences in turn-taking dynamics depending on context. The performances that the proposed models achieved are coherent with the most similar works in the literature [11] which is based on prosodic features alone. We believe rendering our dataset available will improve evaluation homogeneity among similar work. We think the notion of terminal speech turn is important for better-describing interactions, and the results we obtained here encourage us to assert that it is a notion that can be automatically predicted from spontaneous speech. This work can have applications for large-scale semi-automatic media analysis. The annotation of interruptions in the media is a subjective task, with multiple studies not necessarily obtaining the same conclusions [27, 28]. The proposed system would allow for a reproducible method for automatically highlighting places of interest in a long document and thus enable researchers to focus on the speaker, specifically changes that happen outside of a detected TRP on French TV and Radio content. This corpus is also annotated with backchannels and interruption information, future works will focus on the classification of turn-taking with regard to these phenomena.

5. Availability

Code and datasets used in this paper are available at `https://github.com/ina-foss/termClassif`.

6. Acknowledgements

This work has been partially funded by the French National Research Agency and the German DFG (GEM project - ANR-19-CE38-0012 and CLD 2025 - ANR-19-CE38-0015).

7. References

[1] E. Couper-Kuhlen, *English Speech Rhythm*, ser. Pragmatics Beyond New Series. Amsterdam, Netherlands: John Benjamins Publishing, Apr. 1993.

[2] D. H. Zimmerman and C. West, "Sex roles, interruptions and silences in conversation," 1996.

[3] G. W. Beattie, "Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted," *Semiotica*, vol. 39, no. 1–2, 1982. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/semi.1982.39.1-2.93/html

[4] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.

[5] S. C. Levinson, "Turn-taking in human communication – origins and implications for language processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, p. 6–14, Jan. 2016.

[6] M. Lebourdais, M. Tahon, A. Laurent, S. Meignier, and A. Larcher, "Overlaps and gender analysis in the context of broadcast media." LREC 2022, 2022, p. 7.

[7] M. Adda-Decker, C. Barras, G. Adda, P. Paroubek, P. Boula de Mareüil, and B. Habert, "Annotation and analysis of overlapping speech in political interviews," in *LREC 2008*, Marrakech, Morocco, 2008, pp. 1–7.

[8] S. C. Levinson, *Pragmatics*. Cambridge University Press, 1983.

[9] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Turn-taking prediction based on detection of transition relevance place," in *Interspeech 2019*. ISCA, Sep. 2019, p. 4170–4174. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/1537.html

[10] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech Language*, vol. 67, p. 101178, May 2021.

[11] S. Ondas, M. Pleva, and S. Bacikova, "Transition-relevance places machine learning-based detection in dialogue interactions," *Elektronika ir Elektrotechnika*, vol. 29, no. 3, p. 48–54, 2023.

[12] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics, 2017, p. 220–230. [Online]. Available: http://aclweb.org/anthology/W17-5527

[13] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Oct. 2018, p. 186–190, arXiv:1808.10785 [cs]. [Online]. Available: http://arxiv.org/abs/1808.10785

[14] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, p. 2981–2990, arXiv:2010.10874 [cs]. [Online]. Available: http://arxiv.org/abs/2010.10874

[15] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, p. 1–12, 2015.

[16] R. Uro, M. Tahon, J. Wottawa, D. Doukhan, A. Rilliard, and A. Laurent, "Annotation of transition-relevance places and interruptions for the description of turn-taking in conversations in French media content," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 1225–1232. [Online]. Available: https://aclanthology.org/2024.lrec-main.110

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," no. arXiv:2006.11477, Oct. 2020, arXiv:2006.11477 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2006.11477

[18] 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 *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 2479–2490. [Online]. Available: https://aclanthology.org/2020.lrec-1.302

[19] M. Tahon, A. Larcher, M. Lebourdais, F. Bougares, A. Silnova, and P. Gimeno, "Allies: a speech corpus for segmentation, speaker diarization speech recognition and speaker change detection," in *LREC/COLING*, 2024.

[20] A. Larcher, A. Mehrish, M. Tahon, S. Meignier, J. Carrive, D. Doukhan, O. Galibert, and N. Evans, "Speaker Embedding For Diarization Of Broadcast Data In The ALLIES Challenge." ICASSP, 2021, pp. 5799–5803.

[21] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, p. 378–382, 1971.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," no. arXiv:2212.04356, Dec. 2022, arXiv:2212.04356 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2212.04356

[23] M. Macary, M. Tahon, Y. Estève, and D. Luzzati, "Acoustic and linguistic representations for speech continuous emotion recognition in call center conversations," no. arXiv:2310.04481, 2023. [Online]. Available: https://arxiv.org/abs/2310.04481v1

[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, 2015.

[26] P. A. Barbosa and T. Raso, "Spontaneous speech segmentation: Functional and prosodic aspects with applications for automatic segmentation/a segmentação da fala espontânea: aspectos prosódicos, funcionais e aplicações para a tecnologia," *Revista de Estudos da Linguagem*, vol. 26, no. 4, pp. 1361–1396, 2018.

[27] M. Sandré, "Analyse d'un dysfonctionnement interactionnel – l'interruption – dans le débat de l'entre-deux-tours de l'élection présidentielle de 2007," *Mots. Les langages du politique*, no. 8989, p. 69–81, Mar. 2009.

[28] H. Constantin de Chanay and C. Kerbrat-Orecchioni, "Les interruptions dans les débats médiatiques: une stratégie interactionnelle," *Pratiques. Linguistique, littérature, didactique*, no. 147–148147–148, p. 83–104, Dec. 2010.