Unraveling Spontaneous Speech Dimensions for Cross-Corpus ASR System Evaluation for French

Solène Evain, Solange Rossato, François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{firstname.surname}@univ-grenoble-alpes.fr

Abstract

Many papers on speech processing use the term 'spontaneous speech' as a catch-all term for situations like speaking with a friend, being interviewed on radio/TV or giving a lecture. However, Automatic Speech Recognition (ASR) systems performance seems to exhibit variation on this type of speech: the more spontaneous the speech, the higher the WER (Word Error Rate). Our study focuses on better understanding the elements influencing the levels of spontaneity in order to evaluate the relation between categories of spontaneity and ASR systems performance and improve the recognition on those categories. We first analyzed the literature, listed and unraveled those elements, and finally identified four axes: the situation of communication, the level of intimacy between speakers, the channel and the type of communication. Then, we trained ASR systems and measured the impact of instances of face-to-face interaction labeled with the previous dimensions (different levels of spontaneity) on WER. We made two axes vary and found that both dimensions have an impact on the WER. The situation of communication seems to have the biggest impact on spontaneity: ASR systems give better results for situations like an interview than for friends having a conversation at home.

Keywords: Spontaneous Speech, Automatic Speech Recognition, Evaluation

In the field of automatic speech recognition, spontaneous speech is very often described as opposed to read or prepared speech to explain the difficulties in recognizing it. At the moment, the literature appears to show a significant disparity in performance on spontaneous speech: results on ASR (Automatic Speech Recognition) Benchmarks have been reported by Gabler et al. (2023) and shows a ≈3% WER (Word Error Rate) on Switchboard, a ≈10% WER on CallHome and a ≈65% WER on two meeting corpora (Meeting SDM or MDM - single or multiple speakers).

The literature shows a correlation between speech spontaneity and WER (Gabler et al., 2023; Dufour et al., 2010; Deléglise and Lailler, 2020; Szaszák et al., 2016) when spontaneity levels are labeled by humans. But no to few details are given about what makes speech more or less spontaneous. Some of its characteristics, such as hesitations, repetitions, restarts, and word fillers are more and more observed when spontaneity arises (Hoesen et al., 2016; Szaszák et al., 2016; Candido Junior et al., 2023; Johnson, 2004; Bigi and Meunier, 2018a). However, one problem is that the dichotomy between prepared speech and spontaneous speech is not so simple when considering it on the scale of corpora: spontaneous speech can appear in prepared contexts, such as debates and TV or radio shows for example (Garnerin, 2022) (Dufour et al., 2010), but also in lectures (Glass et al., 2004). Fügen et al. (2007) also mentions a corpus of "speeches which are usually prepared in advance and therefore less spontaneous, i.e., planned speech".

In this paper, we intend to train an ASR system for spontaneous speech that would perform well for speech in interaction. For this purpose, we categorize levels of spontaneity in order to (1) evaluate the relation between categories of spontaneity and ASR systems performance and (2) improve the recognition on those categories. We first present our review of the literature on spontaneous speech (sec. 1). This led us to explore four dimensions with the aim of assessing the degree of spontaneity in a corpus or recording. In section 2, we present the gathered French spontaneous speech corpora, and the selection process we implemented to extract four instances of face-to-face interaction, thus representing various levels of spontaneity. This section also introduces the ASR systems implemented and the experiments conducted to measure the impact of our spontaneity levels on WER. Results are given in section 3 and discussed in section 4.

1. Spontaneous Speech

If spontaneous speech is often described through its characteristic, several studies (Swerts and Collier, 1992; Shriberg, 2005; Luzzati, 2007; Bigi and Meunier, 2018b) show that spontaneous speech is elaborated during its production, showing traces of its process with hesitations, repairs, pauses that reveal the cognitive processes in progress.

LREC-COLING 2024, pages 17165–17175

1.1. Spontaneity and the WER

Gabler et al. (2023) show the historical progress of English ASR benchmarks over the time and assess that "performance gains become flatter [when] the more spontaneous the speaking style becomes".

Deléglise and Lailler (2020); Szaszák et al. (2016); Dufour et al. (2010) also demonstrate bad WER results with a high level of spontaneity.

The levels of spontaneity used in ASR have mainly been determined by human judgment. This labeling has been done on different levels: the corpus (Gabler et al., 2023; Szaszák et al., 2016), the broadcast in an audiovisual broadcasting corpus (Deléglise and Lailler, 2020) and the speech segment (with excellent inter-annotator agreement) (Dufour et al., 2010).

This time-consuming labeling task have been automated by Dufour et al. (2010) however they report that this task remains complex. With the aim of rapidly labeling corpora or recordings as more or less spontaneous (the full recording, the corpus), we analyzed the literature in search of factors influencing spontaneity.

1.2. Predominant factors

Among the factors that may influence the production of more or less spontaneous speech, context appears to be important. Labov (1973) distinguished as different contexts: reading, interview (careful speech) and conversational speech (casual speech). To Beckman (1997), spontaneous speech includes several types of speech that depend on social and rhetorical contexts of the recording. We analyzed six papers introducing french spontaneous corpora to determine the elements related to the context that could help categorizing the recordings as more or less spontaneous.

Cresti et al. (2004) collected in the C-ORAL-ROM corpus recordings representing a variety of speech acts by varying the formality level, the public/private dimension (that they call the sociological context), and speech genres (political speech, teaching, conference, talk show, news, private conversation...) that they gathered under "natural context", "media" or "telephone" categories.

In the ESLO corpus (Baude and Dugua, 2011; Eshkol-Taravella et al., 2011), the authors collected recordings with different "degree of speech planning" (spontaneous vs written discourse). The dataset gathers different recording situations such as face-to-face interviews, work meetings, spontaneous conversations, free recordings, private or professional situations, in places like a medico-psychopedagogical center or public spaces (stores, market, street...), with different formality levels (based on a social framework involving status, roles and language behavior). They also give details about the social distance between speakers (level of education, profession), if they know each others and the role of the interviewed speaker in the society.

In the PFC corpus (Laks et al., 2009), the speakers are selected for their proximity to one of the interviewers, in order to bring out informal and formal speech, depending on the interviewer the speaker is speaking to. They recorded face-to-face interactions and peer group meetings, at home or in places they call "neutral" like university. Speakers are invited to talk about their activities, childhood, the news...

André and Canut (2010) in TCOF had the objective to record speakers in situations "as natural as possible" including interviews with at least two speakers speaking about their life, events, experiences or explaining a skill they have, but also free or theme-based conversations and public meetings or professional activities. Their metadata also include pieces of information such as the relation between the speakers, their role in the interaction, their study level and profession, the channel of communication, discourse genre and the place of recording. They also specify four degrees of speech planning: planned, semi-planned, unplanned, unknown.

As for the CRFP corpus (Equipe Delic et al., 2004), it includes private, public and professional speech. People may be talking about their life or introducing a skill they have, but the corpus also includes political or association meetings, lectures, conferences or broadcast speech. Some of them have been recorded at work, bringing forth what they call "institutional speech". The metadata includes information about the level of education, profession and roles (interviewer, interviewee...).

Finally, the CLAPI corpus (Baldauf-Quilliatre et al., 2016) gathers social situations like work meetings, commercial interactions, dinner with friends or family, medical consultations, private and professional phone calls and online conversations, that can happen in different institutions, public services, private companies, home or at the doctor.

The literature abounds with elements that may influence spontaneous speech levels. Some elements are intertwined: indeed, the formality level depends on the relationship between the speakers, but also on the situation (official discourse, interview, conversation). Likewise, there are different types of interviews depending on whether you are speaking with a friend or a stranger.

1.3. Unraveling spontaneous speech elements

In order to improve ASR performances on spontaneous speech, we consider four dimensions unraveling the yarn ball of factors influencing spontaneity.

Spontaneity	+ + +			- - -
Situation of communication	Usual	Strong place or role	Strong place and role	
Intimacy level	Close friends or family members	Colleagues	Acquaintances	None
Channel of communication	Face-to-face	Distant and video	Distant, no video	
Type of communication	Interpersonal	Group	Mass or public	

Table 1: Four dimensions to unravel spontaneous speech

This simplifying attempt focuses on the following prominent factors:

Situation of communication: The aim is to capture the level of constraint or control in the interaction by catching (i) the existence of roles (social role like *politician or professor* or in discourse like *interviewer/interviewee*) and (ii) the significance of a place. Public spaces (parks, street...) and home are considered has places not involving any constraint or control on speech, contrary to social institutions[1], public services, private companies or workplace as mentioned by Equipe Delic et al. (2004).

Intimacy level between speakers: The second axis is based on the fact that the more two people know each others, the more “they share experiences that create a cultural code between them” (Romera Ciria, 2019).

Channel of communication: This axis include face-to-face, distant with video (visioconference) or distant without video (phone) modalities.

Type of communication: This axis sets apart interpersonal, group and mass or public speaking. When the communication is interpersonal (two people speaking), there is underlying stakes of not breaking the relationship between speakers (Romera Ciria, 2019) (Agha, 2006). Whereas when the speech is public, it is deeply linked to performance and power [2], with well-defined aims (like entertain, appeal, convince) and heterogeneous audience. Public speaking is like a one-shot with less spontaneity than a dialog, caused by the fact that an error is less easy to correct.

2. ASR on French spontaneous speech

The goal of the experiments is to determine whether ASR systems benefit from adaptation to spontaneous speech.

[1]organizations, structures, or systems within society that fulfill various functions, such as education, government, family, and healthcare, to help maintain social order and meet the needs of individuals and communities

[2]Speech is very often compared to a weapon. See (Périer, 2017) and (Viktorovitch, 2021) books for instance.

2.1. Experimental Design and Objectives

First, we train a baseline system (i) on the official CommonVoice (Ardila et al., 2020) (Ardila et al., 2020) 10.0 datasets (train: 660h, dev: 25h, test: 26h). It serves as a check to ensure that the ASR system we will use as a base is state-of-the-art.

Next, we (ii) train a domain-adapted system for spontaneous speech with a large dataset of spontaneous speech ("All_spont") specifically elaborated for this task and described in section 2.3.

Finally, the last experiment involves (iii) fine-tuning the domain-adapted system on sub-datasets one-by-one characterized based on our dimensions: "Usual_close", "unusual_close", "Usual_distant", "Unusual_distant", and "All_cases" as a sum of the four (section 2.4).

We chose to stabilize the canal of communication and the type of communication to face-to-face and interpersonal communications.

2.2. ASR systems architecture

The ASR systems are trained using Speechbrain v0.11 (Ravanelli et al., 2021), CommonVoice ASR CTC (Connectionist Temporal Classification (Graves et al., 2006)) recipe. The architecture was: a pre-trained model (always fine-tuned on train data), followed by 3 DNN layers and CTC loss. We used LeBenchmark’s Wav2Vec2 7k-large pre-trained model for French[3] (Evain et al., 2021) which we will refer to as LB7K from now on. This model was trained on 7,000h of speech including 1,626 h of radio broadcast, 1,115 h of read speech, 127 h of spontaneous speech, 38 h of acted telephone dialogue and 29 h of acted emotional speech. The learning rates were 0.0001 for LB7K (Adam optim. (Kingma and Ba, 2014)) and 1.0 for the rest of the model (Adadelta optim. (Zeiler, 2012)) with annealing factors of 0.9 and 0.8 respectively. Batch sizes are 2 for train and dev, and 4 for the test. Utterances of more than 30 seconds were not taken into account in training, validation and testing. Greedy decoding is used.

[3]https://huggingface.co/LeBenchmark/wav2vec2-FR-7K-large

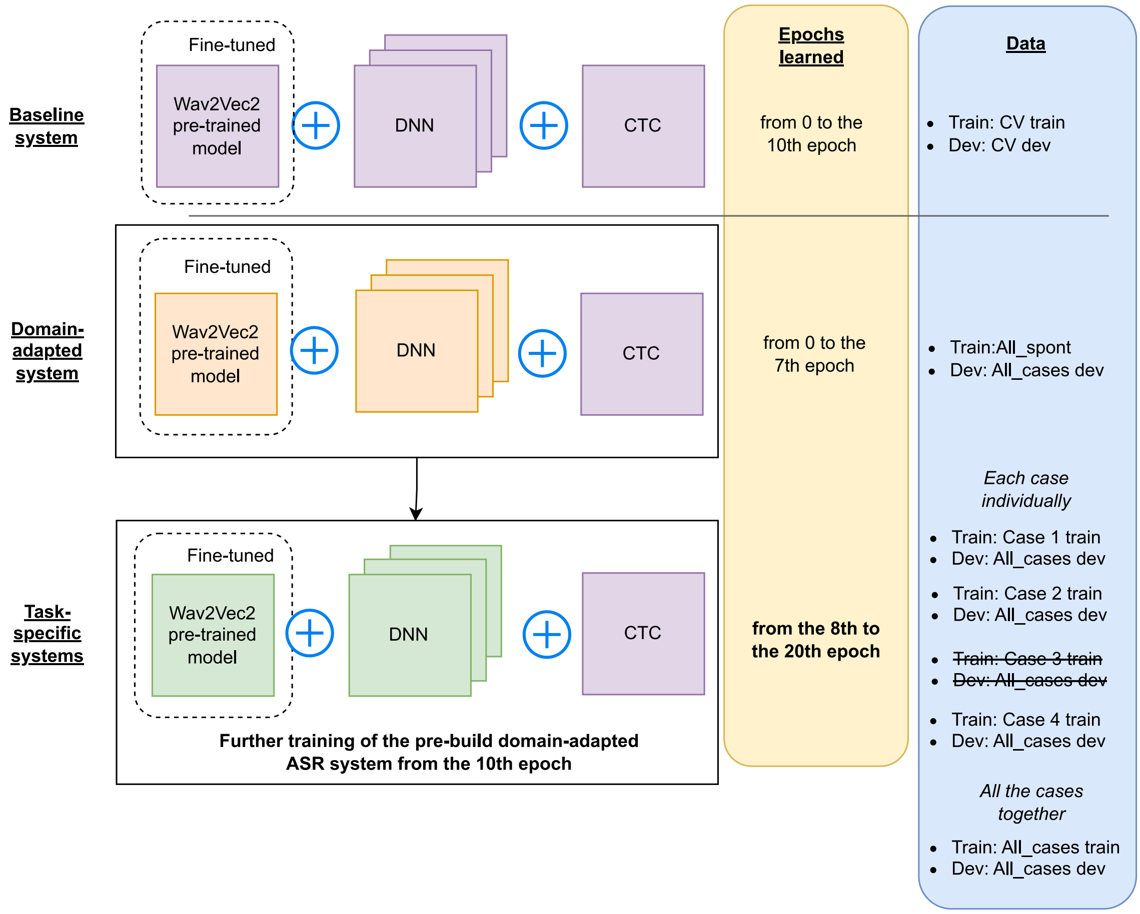

Figure 1: ASR architectures

It took around 7 h per epoch for the ASR system trained with All_spont dataset. The system was trained for 7 epochs on 4 Nvidia A100 40GB GPUs, using Distributed Data Parallel. The specific fine-tuning for the task-specific system was done by further training the pre-built domain-adapted system for an additional 13 epochs, so it was not learned from scratch. It took around 46 min per epoch for ASR system trained with AllCases dataset.

Domain-adaptation and task-specific tasks scripts can be found in https://gitlab.com/solene-evain/lrec.

2.3. Spontaneous Data gathering

We conducted extensive work in collecting and preparing corpora of spontaneous speech, along with their transcriptions and metadata. The selected corpora had to have a license permitting data reuse for research, as well as be freely and easily accessible (via the completion of a simple form). We excluded data typically used for automatic speech recognition in French spontaneous speech as these datasets contain an unknown proportion of prepared speech.

The corpora that have been used are as follows: **CFPB***[4] (Dister and Labeau, 2017), (Dister and Labeau, 2017), **CFPP****[5] (Branca-Rosoff et al., 2012), (Branca-Rosoff et al., 2012), **CID** (Bertrand et al., 2008), (Bertrand et al., 2008), **CLAPI**** (Baldauf-Quilliatre et al., 2016), (Baldauf-Quilliatre et al., 2016), **C-ORAL-ROM*** (Cresti et al., 2004), (Cresti et al., 2004), **CRFP*** (Equipe Delic et al., 2004), **ESLO2**[6](Baude and Dugua, 2011), (Baude and Dugua, 2008), **FLEURON** (André, 2017),**MPF** (Gadet and Guerin, 2016), (Gadet and Guerin, 2016), **OFROM*** (Avanzi et al., 2016), (Avanzi et al., 2016), **PFC** (Laks et al., 2009), (Laks et al., 2009), **Réunions_de_travail***, **TCOF**** (André, 2017), (André and Canut, 2010) and **TUFS*** (Akihiro and Kawaguchi, 2014). See section 6 for more information about each corpus.

The number of collected corpora amounts to 14 (see Figure 2), totaling ≈ 370 hours of speech,

[4]Corpora marked with * are entirely included in the CEFC corpus (Benzitoun et al., 2016), (Benzitoun et al., 2016) and have been used as such.

[5]Corpora marked with ** have been completed.

[6]Files were downloaded manually one-by-one under the following conditions: audio quality label equivalent to 'excellent' (*excellente*), 'good' (*bonne*), or 'fair' (*passable*), and transcription level 'C' (validated).

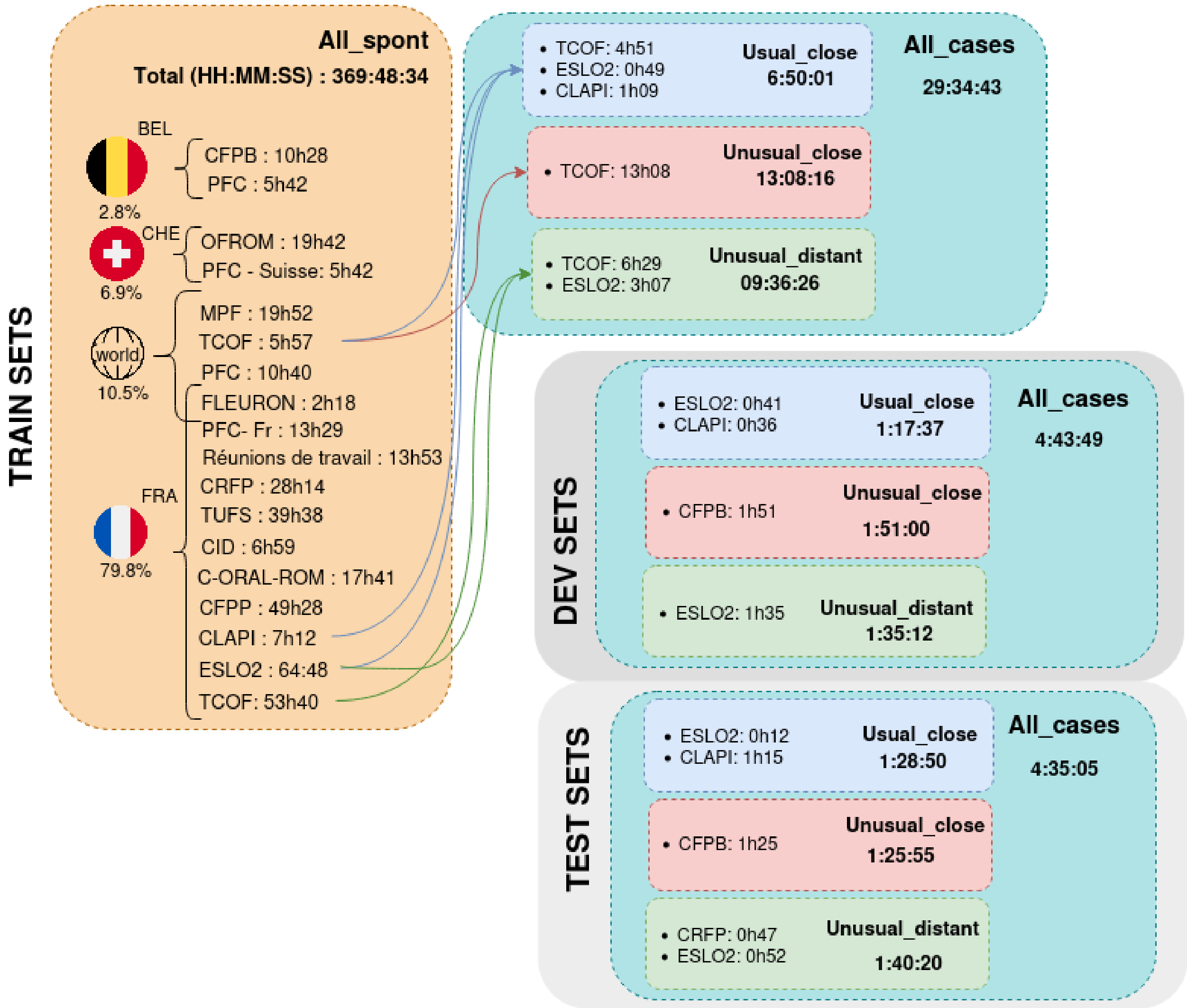

Figure 2: Train, dev and test sets for the domain-adapted system and the task-specific systems

including 2.8% from Belgian French, 6.9% from Swiss French, 10.5% from French spoken in the francophone community, including France[7], and 79.8% from French spoken in France.

Every audio file was converted to 16KhZ, mono channel, 16 bits and wave format. Every transcription file (*.textgrid*, *.csv*, *.orfeo*, *.trs*) was normalized and converted to *.json* format. We managed to remove overlapping speech for every corpus except MPF and PFC because of the TextGrid transcription file format. Data preparation scripts as well as every train, dev and test *.json* files can be found on https://gitlab.com/solene-evain/lrec.

[7]The MPF corpus is somewhat unique as it was created to study French spoken by people in Paris of diverse ethnic backgrounds.

2.4. Sub-datasets: exploring two dimensions

We want to study the influence of two of our dimensions: the **situation of communication** and the **level of intimacy between speakers** on spontaneity on the performance of an ASR system. We supposed that ASR performance will increase when the intimacy between speakers increases and when the situation of communication becomes more casual.

We researched, in the spontaneous data, recordings that would form homogeneous case studies that correspond to the following situations:
Usual_close Close relationship between speakers and usual situation of communication.
e.g.: *Friends chatting at one's place.*
Unusual_close : close relationship between speakers with strong roles.

e.g.: *Someone interviewing a friend.*
Usual_distant: a usual situation of communication, between speakers that do not know each others.
e.g.: *Two people that do not know each others chatting in the street.*
Unusual_distant strong role and people do not know each others.
e.g.: *An interview between people that do not know each others*

We tried gathering as many recordings as we could for each case as we needed at least 10 h of data for fine-tuning (Baevski et al., 2020).

We analyzed each corpus. If a corpus was created with the same protocol, the labeling of data was done accordingly to the information available. When there were multiple protocols within the same corpus, resulting in a wide variety of recordings, the metadata files were individually examined. Unfortunately, we couldn't manage to gather enough data for the Usual_distant case.

2.5. Sub-datasets partitioning

Train, dev and test datasets are:

All_spont train set: This one is the collection of spontaneous speech we gathered in section 2.3, including 369h48 of various spontaneous speech. This is only used for training. [8]

Study_cases train, dev and test sets: This is the collection of specific speech recordings we gathered in section 2.4. The recordings from each of the case studies were divided into training, development, and test sets. We excluded from dev and test sets the audio files already used to train the LeBenchmark model. The ESLO2 data we put on dev and test sets is supplementary data found on the website that was not used for LeBenchmark models training. The train, dev and test sets do not contain overlap speech since they do not include PFC or MPF files. In the end, there is 6h50 of Usual_close speech, 13h08 of Unusual_close speech and 9h36 of Unusual_distant speech in train, 1h17 of Usual_close speech, 1h51 of Unusual_close speech and 1h35 of Unusual_distant speech in dev, and 1h28 of Usual_close speech, 1h25 of Unusual_close speech and 1h40 of Unusual_distant speech in test. It is worth noticing that the development and test datasets were removed from All_spont before training of any ASR system.

All_cases train, dev and test sets: These datasets are formed by combining the training, development, or test sets of each of the previous case studies. This gives 29h34 of speech for train, 4h43 of speech for dev and 4h35 of speech for test.

All these sets are summarized in Figure 2.

3. Results

3.1. Baseline system

The baseline system achieves results very close to state-of-the-art[9]. It is observed that while the system performs well on Common Voice (11.92%), the WER deteriorates on spontaneous speech (61.34%), which varies depending on spontaneous speech characteristics (Usual vs Unusual, Close vs Distant).

As Dufour et al. (2010); Gabler et al. (2023); Deléglise and Lailler (2020); Szaszák et al. (2016), we observe an increase in WER as spontaneity increases: a difference of 22.18 WER points between the least spontaneous case (Unusual_distant) and the moderately spontaneous case (Unusual_close), and 25.56 WER points between the moderately spontaneous case (Unusual_close) and the most spontaneous one (Usual_close). This results in a total difference of 47.74 WER points between situations like having a drink with a friend at home (Usual_close) and interviews between two people that don't know each others (Unusual_distant).

3.2. Domain-adapted system

The system adapted to the domain of spontaneous speech degrades results on Common Voice, with 24.90% WER. Note that the system has no reading recordings in train. However, the results on spontaneous speech datasets are noticeably better (33.59%, an 27.75 points improvement). The system shows a 15.46% WER on Unusual_distant, the least spontaneous case, a 28.09% WER on Unusual_close, the moderately spontaneous one and 58.25% on Usual_close, the most spontaneous one. We still observe a degradation of the WER when speakers know each others (+ 12.63 points) and when the situation is casual (+30.16 points).

3.3. Task-specific systems

By continuing the learning process with specific data for each case, we observe a very slight improvement in performance. This results in a 14.44% WER for the "Unusual_distant" case after fine-tuning on this type of speech, compared to 15.46% with the domain-adapted system. In the "Unusual_close" case, we achieve a 25.21%

[8] The model always stopped training when NCCFr (Torreira and Ernestus, 2010), (Torreira et al., 2010) and ESLO2-cinema files were encountered. We do not reconsider the quality of the audio files included in those corpora but still had to remove them from our experiments after no explanation was found, not to loose too much time.

[9] see https://huggingface.co/speechbrain/asr-whisper-large-v2-commonvoice-fr

WER after fine-tuning on similar data (compared to 28.09% previously). Finally, for the "Usual_close" case, we obtain a 53.02% WER (compared to 58.25% previously).

A system adapted to a specific case of spontaneous speech does not degrade the performance achieved on speech with different levels of spontaneity.

When combining the three levels of spontaneity, the following performances are obtained: 14.28% WER on Unusual_distant, 25.21% WER on Unusual_close, and 53% on Usual_close, which corresponds to the best performance for each case.

4. Discussion

Domain-adaptation clearly improves ASR performances for spontaneous speech. LeBenchmark pre-trained model already includes spontaneous speech (165 h[11]), but this is not representative in respect with the 7 000 hours in total.

Task-specific systems slightly improve the WER, and the best system is obtained with fine-tuning on the All_cases dataset.

First, there is a tendency for the WER to vary according to the levels of spontaneity. The best WER is obtained on Unusual_distant. When the Intimacy level changes from acquaintances or none (grouped together in Unusual_distant) to close friends/family members (Unusual_close), the WER increases by 10.9 points, showing the impact of this dimension on spontaneity. Then, when the Situation of communication changes from a strong place or/and role (Unusual_close) to usual (Usual_close), the WER increases by 27.81 points. This suggests that the Situation of communication dimension has more impact on the spontaneity than the Intimacy level between speakers. It seems therefore that two friends change their speech depending on the situation they're on, even if they know each other very well.

Those results should be treated cautiously as they are obtained on test sets with only a few recordings and the performances may depend on the recordings quality in themselves. We plan to use k-fold cross-validation in order to test our systems on more data and then verify the homogeneity of the recordings labeled as Unusual_distant, Unusual_close and Usual_close.

The task-specific systems are fine-tuned on very small datasets (from 6h50 to 13h08). This was limited by what we could label in the 369 hours of spontaneous speech. This is surely a limit and may explain the very little improvement we achieved.

5. Conclusion

We introduced in our paper four dimensions (situation of communication, relationship between speakers, canal and type of communication) to describe spontaneous speech variation.

We focused in this study on 2 over 4 dimensions, the situation of communication and the intimacy level between speakers, focusing on face-to-face interactions. We aimed at proposing a domain-adapted ASR system and task-specific ASR systems and evaluated them on sub-datasets representative of some types of spontaneous speech. Thus, we gathered 14 spontaneous speech corpora (nearly 400 hours of speech) and identified three study cases: Usual_close being the most spontaneous one, Unusual_close the moderately spontaneous one, and Unusual_distant the least spontaneous one. We found that the lower the spontaneity, the lower the WER, just as Dufour et al. (2010); Gabler et al. (2023); Deléglise and Lailler (2020); Szaszák et al. (2016). With no surprise, a domain-adaptation of the ASR system to spontaneous speech was highly beneficial. There also seems to be a benefit from using task-specific fine-tuning on cases.

Our results show that the situation of communication has a high impact on ASR performance: we never obtained less than a 53% WER for the Usual_close case. Moreover, the intimacy level between speakers also has an impact, even if less than the situation of communication. However, those results should be treated cautiously as we tested on a small amount of data.

To go further, we would need much more well detailed spontaneous speech data, which we do not have for now. Also, the study can also be continued with different study cases, including different dimensions (canal of communication, type of speech) or different levels on each axis. Finally, it would be interesting to complete this study with a comparison on different languages.

6. Datasets

CFPB - Corpus du Français Parlé à Bruxelles*[12]: [CC-BY-NC-SA 3.0] Corpus of French as spoken in the 19 Brussels communes. Same method as the CFPP2000.

CFPP - Corpus du Français Parlé à Paris*: [CC-BY-NC-SA 4.0] The Corpus of Parisian Spoken

[10] Error margins corresponding to 95% confidence intervals were computed using bootstrap re-sampling as proposed in (Bisani and Ney, 2004).

[11] 1,791 h if we include broadcast data, but remember that those datasets are mostly prepared speech

[12] The corpora marked with an asterisk (*) are included in the CEFC. When the data comes from the CEFC, the CC-BY-NC-SA 4.0 license prevails.

Train. data	CV	Usual_close	Unusual_close	Unusual_distant	All Cases
Baseline system					
CV train	**11.92**	86.29	60.73	38.55	61.34
Domain-adapted system					
All_spont	24.90	58.25	28.09	15.46	33.59
Task-specific systems					
Usual_close	25.95	53.65 ± 1.29	26.19 ± 0.86	14.84 ± 0.56	31.51
Unusual_close	25.27	55.03 ± 1.31	25.67 ± 0.86	14.77 ± 0.57	31.92
Unusual_distant	25.09	56.76 ± 1.26	25.95 ± 0.85	14.44 ± 0.54	32.23
AllCases	24.94	**53.02** ± 1.29	**25.21** ± 0.87	**14.28** ± 0.54	**30.66**

Table 2: WER results on test sets for each ASR system (in %).
Cells in light gray show the correspondence between ASR systems trained on each case and the result on the same case. In bold and green frame is the best model for each test set. Gray numbers indicate 95% confidence intervals.[10]

French (CFPP2000) consists of a collection of non-directive interviews about the neighborhoods of Paris and its close suburbs.

CLAPI - Corpus de Langue PArlée en Interaction*: [CC-BY-NC-SA 4.0] Multimedia database of recorded corpora in real-life situations, in various contexts: professional, institutional, private, commercial, educational, medical...

C-ORAL-ROM*: [CC-BY-NC-SA 4.0 (licence CEFC)] A set of comparable oral corpora for 4 Romance languages, including French. In the context of this project, oral corpora of spontaneous speech for Romance languages have been developed.

CRFP - Corpus de Référence du Français Parlé*: [CC-BY-NC-SA 4.0 (licence CEFC)] : A testament to the French language spoken in France, the CRFP consists of 134 recordings sampled based on various speech situations and the educational levels of the speakers, collected in around forty different cities.

FLEURON* : [CC-BY-NC-SA 4.0 (licence CEFC)] Actions and interactions in various university situations (in classes, at the university library, at CROUS...) as well as in everyday life situations (in shops, at the museum, in private...). Other resources provide testimonies from French and foreign students who share anecdotes and explanations (the functioning of associations, the social security system, the university system...).

OFROM - Corpus Oral de Français de Suisse Romande*: [CC-BY-NC-SA 4.0 (licence CEFC)] The OFROM corpus contains hundreds of recordings of the Swiss Romandy dialect.

Réunions de travail*: [CC-BY-NC-SA 4.0 (licence CEFC)] This corpus, overseen by Magali Husianycia (ATILF), was recorded in 2007-2008 as part of a doctoral thesis. It features workplace interactions, including meetings, work sessions in the nonprofit sector, and conversations among colleagues before meetings.

TCOF - Traitement des Corpus Oraux en Français*: [CC-BY-NC-SA] The 'Treatment of Oral Corpora in French' (TCOF) project emerged from the desire to preserve oral corpora collected in the 1980s and 1990s for personal research purposes. The provided corpus comprises two main categories: recordings of adult-child interactions (children up to 7 years old) and recordings of interactions between adults.

TUFS - Tokyo University of Foreign Studies*: [CC-BY-NC-SA 4.0 (licence CEFC)] The Tokyo University of Foreign Studies (TUFS) corpus, overseen by Y. Kawaguchi (Tokyo University of Foreign Studies), was compiled in several waves between 2005 and 2011, mostly in French universities (Aix-Marseille and Paris XIII) with students. The recordings are lengthy (average of 50 minutes), which generally allows for a gradual ease of speakers and increasingly spontaneous production.

CID - Corpus of Interactional Data: [CC-BY-NC-SA 4.0] It is a corpus of dyadic conversational interactions in French (8 hours, including 3 audio-visual recordings).

ESLO2 - Enquêtes SocioLinguistiques à Orléans: [CC-BY-NC-SA] A linguistic corpus consisting of audio recordings and their transcriptions conducted in Orléans between 1968 and 1974 (ESLO1) and from 2008 onwards (ESLO2).

MPF - Multicultural Paris French: [CC-BY-NC-SA] The MPF project aligns with discussions on linguistic processes at play in the ways of speaking the dominant language in Western metropolises, due to the presence of a significant immigrant population, and comparing them (in this case, in relation to MLE, the London corpus). The corpus presented here contributes to this reflection for the Paris region, featuring recordings of young individuals of 'ethnic' origins.

PFC - Phonologie du Français Contemporain: [CC-BY-NC] PFC (Phonology of Contemporary French) is a research program providing a database of contemporary spoken French in the French-speaking world.

7. Acknowledgements

This work was supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

8. Bibliographical References

Asif Agha. 2006. *Language and Social Relations*. Studies in the Social and Cultural Foundations of Language. Cambridge University Press.

Hisae Akihiro and Yuji Kawaguchi. 2014. Présentation du corpus oral en français de TUFS et son application pour l'analyse linguistique. http://www.tufs.ac.jp/ts/personal/ykawa/art/2014_Waseda_Corpus_TUFS.pdf.

Virginie André. 2017. Un corpus multimédia pour apprendre à interagir en situations universitaires en France. In *proceedings of ATPF conference « Enseigner le français : s'engager et innover »*, Bangkok, Thaïlande.

Virginie André and Emmanuelle Canut. 2010. Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques [Online]*, (147-148).

Rosana Ardila et al. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *proceedings of the 12th LREC conference*, Marseille, France.

Mathieu Avanzi, Marie-José Béguelin, and Federica Diémoz. 2016. De l'archive de parole au corpus de référence : la base de données orales du français de Suisse romande (OFROM). *Corpus*, (15).

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.

H. Baldauf-Quilliatre et al. 2016. CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, (15).

Olivier Baude and Céline Dugua. 2011. (Re)faire le corpus d'Orléans quarante ans après :quoi de neuf, linguiste ? *Corpus*, (10).

Mary E. Beckman. 1997. A typology of spontaneous speech. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 7–26. Springer US.

Christophe Benzitoun et al. 2016. Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, (15).

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Revue TAL : traitement automatique des langues*, 49(3):pp.105–134.

Brigitte Bigi and Christine Meunier. 2018a. Automatic Segmentation of Spontaneous Speech. *Revista De Estudos Da Linguagem*, 26(4):1489–1530.

Brigitte Bigi and Christine Meunier. 2018b. euh, rire et bruits en parole spontanée : application à l'alignement forcé. In *Proceedings of XXXIIe Journées d'Études sur la Parole, JEP*.

Maximilian Bisani and Hermann Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409. IEEE.

S. Branca-Rosoff, S. Fleury, F. Lefeuvre, and M. Pires. 2012. Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000). Http://cfpp2000.univ-paris3.fr/CFPP2000.pdf.

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2023. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*, 57(3):1139–1171.

Emanuela Cresti, Fernanda Bacelar do Nascimento, Antonio Moreno Sandoval, Jean Veronis, Philippe Martin, and Khalid Choukri. 2004.

The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages. In *proceedings of the 4th LREC conference*, Lisbon, Portugal.

Paul Deléglise and Carole Lailler. 2020. Quel type de systèmes utiliser pour la transcription automatique du français ? Les HMM font de la résistance (What system for the automatic transcription of French in audiovisual broadcasts ?). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, Nancy, France.

Anne Dister and Emmanuelle Labeau. 2017. Le corpus de français parlé à Bruxelles: origines, hypothèses, développements et prédictions. *Cahiers de l'AFLS*, 21(1).

Richard Dufour, Yannick Estève, and Paul Deléglise. 2010. Automatic indexing of speech segments with spontaneity levels on large audio database. In *Proceedings of the 2010 international workshop on Searching spontaneous conversational speech - SSCS'10*, Firenze, Italy.

Equipe Delic, Sandra Teston-Bonnard, and Jean Véronis. 2004. Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, 18:11–42.

Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Celine Dugua, and Isabelle Tellier. 2011. Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Traitement Automatique des Langues*, 53(2):17–46.

Solène Evain et al. 2021. Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *NeurIPS 2021 Datasets and Benchmarks Track*, online.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine Translation*, 21(4):209–252.

Philipp Gabler, Bernhard C. Geiger, Barbara Schuppler, and Roman Kern. 2023. Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information*, 14(2):137.

Françoise Gadet and Emmanuelle Guerin. 2016. Construire un corpus pour des façons de parler non standard : « Multicultural Paris French ». *Corpus*, (15).

Mahault Garnerin. 2022. *Des données aux systèmes : étude des liens entre données d'apprentissage et biais de performance genrés dans les systèmes de reconnaissance automatique de la parole*. phdthesis, Université Grenoble Alpes.

James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang. 2004. Analysis and Processing of Lecture Audio Data: Preliminary Investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, Boston, Massachusetts, USA.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*.

Devin Hoesen, Cil Hardianto Satriawan, Dessi Puji Lestari, and Masayu Leylia Khodra. 2016. Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models. *Procedia Computer Science*, 81:167–173.

K. Johnson. 2004. Massive reduction in conversational american english. In *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th InternationalSymposium.*, Tokyo, Japan.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*.

William Labov. 1973. The isolation of contextual styles. In *Sociolinguistic patterns*, 2nd edition edition, pages 70–110. Philadelphia, University of Pennsylvania Press.

Bernard Laks et al. 2009. Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. In *Phonologie, variation et accents du français*. Hermès.

Daniel Luzzati. 2007. Le dialogue oral spontané: quels objets pour quels corpora. *Revue d'Interaction Homme-Machine*, 8(2).

Bertrand Périer. 2017. *La parole est un sport de combat*. Broché.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and

Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624.

Magdalena Romera Ciria. 2019. Relationships as regulators of discourse interaction in Spanish. *Círculo de Linguística Aplicada a la Comunicación*, numéro 79:pages 297–322.

Elizabeth Shriberg. 2005. Spontaneous speech: How people really talk and why engineers should care. In *proceedings of INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, pages 1781–1784, Lisbon, Portugal.

Marc Swerts and René Collier. 1992. On the controlled elicitation of spontaneous speech. *Speech Communication*, 11(4):463–468.

György Szaszák, Máté Ákos Tündik, and András Beke. 2016. Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Setubal, PRT.

Francisco Torreira and Mirjam Ernestus. 2010. The Nijmegen Corpus of Casual French. In *proceedings of LREC 2010*, Valletta, Malta.

Clément Viktorovitch. 2021. *Le pouvoir rhétorique*. Seuil.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. ArXiv:1212.5701.

9. Language Resource References

André, Virginie and Canut, Emmanuelle. 2010. *TCOF : Traitement de Corpus Oraux en Français*. distributed via Ortolang: https://hdl.handle.net/11403/tcof.

Ardila, Rosana and others. 2020. *Common Voice corpus*. PID https://commonvoice.mozilla.org/fr/datasets.

Avanzi, Mathieu and others. 2016. *Corpus Oral de Français de Suisse Romande*. distributed via Cocoon: https://doi.org/10.34847/COCOON.114A10E8-B61C-42FB-8A10-E8B61C72FBB1.

Baldauf-Quilliatre, H. and others. 2016. *CLAPI*. also distributed via Ortolang: https://hdl.handle.net/11403/clapi. PID http://clapi.icar.cnrs.fr/.

Baude, Olivier and Dugua, Céline. 2008. *ESLO*. also distributed via ORTOLANG: https://hdl.handle.net/11403/eslo. PID http://eslo.huma-num.fr/.

Benzitoun, Christophe and others. 2016. *CEFC*. distributed via ORTOLANG: https://hdl.handle.net/11403/cefc-orfeo.

Bertrand, Roxane and Blache, Philippe and Espesser, Robert and Ferré, Gaëlle and Meunier, Christine and Priego-Valverde, Béatrice and Rauzy, Stéphane. 2008. *Transcriptions of CID*. distributed via Ortolang: https://hdl.handle.net/11403/sldr000720/.

Branca-Rosoff, S. and others. 2012. *Corpus de Français Parlé Parisien des années 2000 (CFPP)*. also distributed via Cocoon: https://doi.org/10.34847/COCOON.8BC96A4E-9899-30E4-99BE-C72D216EB38B. PID http://cfpp2000.univ-paris3.fr/.

Cresti, Emanuela and others. 2004. *C-ORAL-ROM - Integrated reference corpora for spoken romance languages. Multi-media edition; tools of analysis; standard linguistic measurements for validation in HLT*. distributed via ELRA: ELRA-S0172, ISLRN 318-977-046-077-4.

Dister, Anne and Labeau, Emmanuelle. 2017. *Corpus de Français Parlé à Bruxelles (CFPB)*. PID http://cfpp2000.univ-paris3.fr/cfpb.html.

Gadet, Françoise and Guerin, Emmanuelle. 2016. *MPF*. distributed via ORTOLANG: https://hdl.handle.net/11403/mpf.

Laks, Bernard and others. 2009. *PFC*. distributed via ORTOLANG: https://hdl.handle.net/11403/pfc.

Torreira, Francisco and others. 2010. *Nijmegen Corpus of Casual French*. PID https://mirjamernestus.nl/Ernestus/NCCFr/index.php.