Forced Alignment of the *Phonologie du Français Contemporain* Corpus

George Christodoulides(✉)

Language Sciences and Metrology Unit, Université de Mons,
Place du Parc 18, 7000 Mons, Belgium
george@mycontent.gr

Abstract. The *Phonologie du Français Contemporain* project is an international, collaborative research effort to create resources for the study of contemporary French phonology. It has produced a large, partially transcribed and annotated corpus of spoken French, consisting of approximately 300 h of recordings, and covering 48 geographical regions (including Metropolitan France, Belgium, Switzerland, Canada, and French-speaking countries of Africa). Following a detailed protocol, speakers read aloud a word list and a short text and engage in guided and spontaneous conversation with an interviewer. The corpus presents several challenges: significant regional accent variation; variable recording quality and different types of environment noise; variation in speaker characteristics (age, sex); and interspersed segments of overlapping speech. In this article, we describe the procedure followed to address these challenges and produce an automatic forced alignment of the corpus at the phone, syllable and token level, starting from the initial transcriptions.

Keywords: Forced alignment · Speech recognition · Sociophonetics
Regional variation · French · Corpus linguistics · Language resources

1 Introduction

The *Phonologie du Français Contemporain* project is an international, collaborative research effort to create resources for the study of contemporary French phonology [7]. It has brought together almost a hundred researchers, who performed recordings of over 600 speakers of French, covering 48 geographical regions, in Metropolitan France, Belgium, Switzerland, Canada and several countries in Africa. The PFC Project recordings follow a strict protocol, inspired by Labov's work on sociolinguistics, including two reading tasks and two conversations. The project has produced hundreds of hours of recordings, and approximately 300 h have been transcribed to date. As part of the transcription conventions of the project, expert annotators code two important phonological phenomena of French, namely the realisation of liaisons and the presence or absence

T. Dutoit et al. (Eds.): SLSP 2018, LNAI 11171, pp. 47–55, 2018.
https://doi.org/10.1007/978-3-030-00810-9_5

of schwa, in positions where these phenomena are optional and reflect sociolinguistic factors (regional variation, speaking style variation and sociolinguistic variation).

The corpus has been transcribed in short (5 to 20 s) segments, which contain an orthographic transcription of one or more speakers (i.e. speaker overlaps are only approximately indicated). Two additional coding tiers contain the information on liaison and schwa. The PFC corpus presents several challenges to automatic speech processing: there is significant regional accent variation, since this is one of the objectives of the project; there is significant variation in recording quality, as is often the case with sociolinguistic fieldwork; there is also variation in speaker characteristics, as the project strives to keep a gender balance and cover four age groups from 20 to 80 years old; and finally, the speaking style of free conversation is always challenging to automatic speech recognition (ASR) systems.

In this article, we present our efforts to produce a reliable automatic forced alignment of the entire corpus, at the phoneme, syllable and token level, starting from the source recordings and transcription data. This project is complementary to previous work to provide an automatic part-of-speech and disfluency annotation for the PFC corpus, outlined in [6]. In the following section, we will review relevant work on forced alignment of French speech, and the PFC corpus. In Sect. 3, we present the main characteristics of the PFC corpus which are relevant to our endeavour. In Sect. 4 we present the method used, followed by preliminary evaluation results in Sect. 5, and finally we outline the perspectives of this work, both with respect to improving the forced alignment tools for French and with respect to new uses for an aligned PFC corpus.

2 Related Work

Several automatic forced alignment tools have been developed for French, over the past two decades. Among these, we can cite *EasyAlign* [9], *SPPAS* [2], *Train&Align* [4], the *Montreal Forced Aligner* [13] and *SailAlign* [12].

EasyAlign is based on the HTK toolkit [16] and a monophone model trained on a relatively small corpus; it operates as a plug-in under Praat [3] but only under Windows (due to a dependency on an external DOS-based phonetiser). *Train&Align* is also based on the HTK toolkit, and can be used to produce monophone and triphone models, but is available only for use on the web: due to restrictions in the HTK license, it is not possible to redistribute the files necessary for recognition, training new acoustic models or for performing speaker adaptation as part of an open-source project. For this reason, *SPPAS* uses the Julius open-source toolkit for the aligner, while its models are trained in house using HTK; the triphone model for French is based on a corpus of approximately 10 h. *MFA* is the newest of the tools and is a collection of Python scripts around the Kaldi ASR system [14]. It can generate monophone and triphone models and perform speaker adaptation; an acoustic model is provided for French, albeit without a pronunciation lexicon. Finally, *SailAlign*, which is based on Sphinx [15],

focuses on the problem of long sound alignment (finding initial anchor points for a transcription of a long recording).

From the short description of available tools above, it is understood that none of them could cover the needs of our project "out of the box". We have therefore opted to develop a new system in C++ using the Kaldi ASR system; our system is modular and uses the *Praaline* Core Library for corpus and annotation management operations, and a Qt user interface. A phonetisation module is also provided, and has been adapted to the particular needs of this project, as will be explained in Sect. 4.3.

It should also be noted that the C-PROM-PFC corpus [1] comprises of 3-min samples from the PFC corpus. The C-PROM-PFC corpus is approximately 10 h long and its alignment to the phone, syllable and token label has been manually verified by an expert annotator.

3 Corpus Description

3.1 Corpus Composition

The PFC corpus consists of four speaking tasks which are recorded for each participant: reading a list of 94 words, that have been carefully chosen to study phonetic variation; reading a short 300-word text, a fictitious newspaper article that contains multiple points of interest where phonological variation may appear; engage in a guided interview with the researcher; and having a more spontaneous, open-ended conversation with the researcher. Roughly 10 min per conversation are transcribed per speaker.

Table 1 shows the corpus composition, at its current state of transcription. The number of samples per region and task is given, along with their duration in minutes. For the two conversation tasks (guided and free conversation), the percentage of single-speaker utterances in the corpus is indicated: this percentage is calculated as the ratio of the duration of single-speaker transcription segments over the total duration of all transcription segments (after performing the preprocessing steps outlined in Sect. 4.2).

3.2 Available Annotations and Coding Schemes

Information on schwa and liaison realisation is coded based on a common systematic methodology. Schwa coding consists of four fields for each potential schwa realisation in a token: field 1 indicates the presence or absence of the schwa, field 2 the position of the schwa within the word, field 3 its left context and field 4 its right context. Liaison coding consists of four fields: field 1 indicates whether the word is mono-syllabic or poly-syllabic, field 2 indicates the presence or absence (and the type) of liaison, and field 3 indicates the liaison consonant and field 4 gives information about the context. For more information on the coding schemes, refer to [8]. The schwa and liaison coding is valuable for the phonetisation procedure outlined in Sect. 4.3.

Table 1. PFC corpus contents. For each of the four tasks (guided conversation, free conversation, text reading and word reading) the number of speakers is given, along with the duration of the transcribed part of the corpus in minutes. For the two conversational tasks, the percentage of non-overlap utterances (calculated as the ratio of their duration over the total transcription duration) is indicated.

Code	Region	Guided Conv			Free Conv			Text		Words	
		Spk	Dur	Mono	Spk	Dur	Mono	Spk	Dur	Spk	Dur
11a	Douzens	10	255.2	91.2%	5	113.9	83.0%	10	29.3	9	25.0
12a	Rodez	8	236.8	97.2%	8	161.0	87.5%	9	23.0		
13a	Marseille Centre Ville	9	193.0	94.0%	9	175.0	88.5%	9	21.4	9	29.1
13b	Aix-Marseille	7	178.9	97.6%	8	288.7	96.0%	8	21.1	8	34.3
21a	Dijon	7	73.2	84.4%	8	84.8	88.3%	8	19.2	8	25.6
31a	Toulouse	14	296.3	91.6%	9	408.6	94.4%	14	53.4	14	44.5
38a	Grenoble	8	116.9	96.5%	8	109.8	96.1%	7	18.8	9	26.5
42a	Roanne	8	107.9	93.4%	8	148.0	91.9%	8	20.7	8	26.3
44a	Nantes	11	207.1	93.8%	9	289.2	94.7%	10	25.6	11	40.3
50a	Brécey	11	122.6	44.6%	6	61.9	49.6%	9	24.6	11	45.6
54b	Ogéviller	11	269.3	96.2%	11	250.3	96.5%	9	22.1	10	28.0
61a	Domfrontais	12	175.4	74.7%	12	150.3	70.0%	12	34.8	12	40.7
64a	Biarritz	12	204.2	86.7%	4	66.8	67.5%	11	27.4	12	36.4
69a	Lyon	10	232.0	96.8%	11	209.0	96.8%	10	20.6	8	17.6
75c	Paris Centre Ville	12	121.2	49.2%	11	114.9	50.6%	12	27.9		
75x	Aveyronnais à Paris	12	308.3	92.9%	10	286.9	87.8%	8	21.7	12	36.4
80a	Amiens	5	50.0	90.5%	6	60.0	86.7%				
81a	Lacaune	13	172.3	90.9%	11	85.4	68.2%	11	33.9	13	54.8
85a	Vendée	7	71.6	84.1%	8	93.1	86.7%	8	19.0	8	25.2
91a	Brunoy	1	5.2	71.6%	9	23.5	100.0%				
92a	Puteaux-Courbevoie	6	133.9	97.9%	5	155.9	98.0%	5	11.9	4	12.0
974	Ile de la Réunion	7	162.4	97.6%	7	170.4	97.6%	9	28.0	8	32.4
aba	Béjaia	11	250.0	90.4%	10	221.1	90.3%	11	29.6	11	37.5
aca	Chlef	12	213.0	97.3%	12	194.4	96.4%	11	31.7	12	37.9
bfa	Burkina Faso	12	283.7	90.0%	11	282.2	88.6%	9	37.0	11	43.2
bga	Gembloux	12	296.5	92.5%	12	237.3	89.8%	9	27.2	11	27.0
bla	Liège	11	244.0	93.7%	11	281.4	95.7%	12	35.5	11	26.2
bta	Tournai	11	264.0	95.6%	11	253.0	94.7%	12	35.3	12	29.6
caa	Peace River	10	109.0	57.7%	7	22.9	100.0%	9	29.7		
cia	Abidjan	14	267.6	90.6%	12	321.9	92.5%	12	45.0	13	68.1
cqa	Québec ville (université)	9	148.0	93.0%	7	64.2	87.2%	7	18.4	8	28.0
cqb	Saguenay	11	167.3	94.4%	10	321.4	79.9%	11	32.7	10	35.6
cya	Cameroun	6	52.5	91.3%	6	89.9	98.1%	6	29.1	5	20.9
maa	Bamako	10	211.5	66.8%	12	235.5	65.9%	12	61.0	12	52.1
rca	Bangui	11	341.6	99.7%	12	262.7	95.3%	12	51.9	12	43.8
sca	Neuchâtel	12	433.8	89.3%	13	490.6	80.4%	12	33.4	12	47.1
sga	Genève	8	167.8	90.4%	9	206.0	88.5%	9	23.8	9	29.5
sna	Sénégal Dakar	12	235.2	94.5%	11	187.6	88.0%	11	34.0	11	44.9
sva	Nyon	12	147.0	96.9%	9	117.8	76.5%	11	28.5	11	45.2

4 Method

An outline of the method employed in order to align the corpus is shown in Fig. 1. We used audio processing software to enhance and restore the original audio recordings, and decrease the variation in the audio properties (e.g. levels). The original transcriptions were checked for consistency with the annotation protocol, and were manually corrected where necessary. The transcriptions were then separated in sequences of segments corresponding to different speakers; overlapping segments were identified at this step. The segments were tokenised and a phonetic transcription was added: the phonetisation includes pronunciation variants, but these are limited based on the PFC schwa and liaison coding. Forced alignment of all segments was performed, and acoustic models were trained based on the data. Special processing was performed on overlapping segments. Finally, the results of the automatic alignments were combined into the end result.

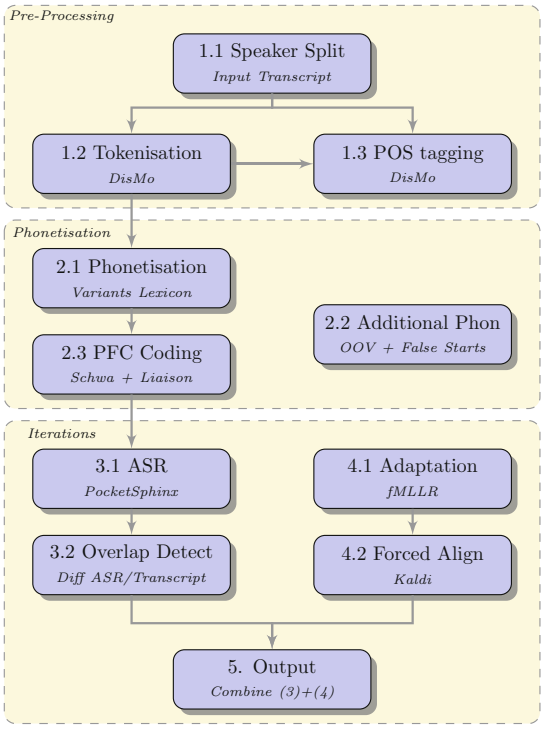

Fig. 1. Methodology used to align the corpus

4.1 Audio Processing and Restoration

We performed an audio enhancing and restoration procedure on all recordings of the corpus, using the commercial iZotope RX 6 Audio Editor. The following

filters were applied in sequence: De-clip (restore clipped samples at high quality, new maximum level −1 dB), De-click (remove random clicks), De-hum (remove 50 Hz noise and its harmonics), De-reverb (remove reverberation, in light vocal processing mode), Voice De-noise (adaptive noise reduction), De-plosive (light removal of hard microphone puffs), Phase correction (if needed), Equaliser Match (using the "full dialogue" preset) and Leveler (normalisation of audio levels, respecting dialogue dynamics). All subsequent application of ASR models was performed on the processed audio files.

4.2 Pre-processing and Transcription Protocol Validation

According to the PFC transcription protocol, the orthographic transcription shall always indicate the speaker. In cases of overlaps, the overlapping speaker's utterance is transcribed inside angle brackets. Each recording is transcribed on one Praat TextGrid: the first tier contains the orthographic transcription, a second tier the schwa coding and a third tier the liaison coding (the order of tiers was found not to be consistent). For each transcription segment, the same number of words (separated by whitespace) shall exist on each tier. Parentheses are used for events and comments. A peculiar pronunciation can be given in brackets using the SAMPA phonetic alphabet.

As can be easily understood, human transcribers can easily violate this set of rules. We have therefore written scripts for data quality assurance. The scripts verify the number of speakers, the correspondences of tokens across the three tiers, the use of punctuation etc. Several corrections were performed automatically; based on the problems detected by the scripts, we performed approximately 2700 corrections manually. For this process, all TextGrids were imported into a *Praaline* [5] database, and the scripts operated on the database. We developed an interactive editor to accelerate the manual corrections.

Subsequently, segments were split into different timelines for each speaker. As part of this process, single-speaker segments (utterances) and overlaps (multiple speakers transcribed within the same segment) were identified. According to the PFC transcription protocol only very long pauses (over 5 s in length) are transcribed; therefore, normal reading and conversational speech pauses are not transcribed and will have to be detected as part of the forced alignment process.

The final pre-processing step was to tokenise the entire corpus and to annotate it using the *DisMo* part-of-speech tagger [6]. These tokens are the basis for the next step.

4.3 Phonetisation

A dictionary of pronunciation variants for French has been constructed, based on the lexicon distributed with Sphinx ASR (converted to the SAMPA alphabet) and the GLÀFF [10] lexicon. The part-of-speech tags produced by *DisMo* were used to limit the possible pronunciation variants. However, the PFC coding schemes were the most important aid in improving the phonetisation.

For each token, the corresponding tokens from the schwa and liaison tiers were examined, and the pronunciation variants were adjusted accordingly. This reduces the size of the graph of possible pronunciation variants that the forced aligner will have to consider, and it also ensures that the resulting alignment will be coherent with the PFC corpus coding.

4.4 Forced Alignment

The Kaldi automatic speech recogniser [14] was used to perform the main forced alignment of the corpus. In each batch, we first train a monophone model on the data to align, followed by a triphone model, and finally a speaker-adapted triphone model. The acoustic model features consist of Mel-frequency cepstral coefficients (MFCCs) and their deltas. Cepstral mean and variance normalization (CMVN) is applied to all models. The speaker adaptation is performed using Feature space Maximum Likelihood Linear Regression (fMLLR).

First, a separate model is trained for each combination of region and speaking style (reading vs conversation). These models are used to align all data, and perform cross-validation. Aggregate models per speaking style are then trained and used to align the data.

A special procedure is followed for the transcription segments of overlapping speech. A quick constrained speech recognition is performed on the overlapping segment, using PocketSphinx [11], in an effort to detect the overlap in the recording, with a better temporal precision than the one given by the transcription. In cases of success, the utterance boundaries are adjusted (speech correctly recognised as non-overlapping is concatenated with the previous or next utterance as appropriate).

The entire process is automated in a C++ plug-in for *Praaline*, which calls the appropriate external programmes.

5 Evaluation

In the absence of a gold-standard alignment, against which we could compare the outputs of the automatic forced alignment system, we had to devise indirect methods of evaluation. These methods essentially indicate how to improve the process, and can help isolate these utterances in the corpus that may have been incorrectly aligned.

Table 2 shows the preliminary results of cross-validation. The data of each region and speaking style (e.g. 11a-reading) is aligned using the acoustic models trained on each of the other regions, for the same speaking style (e.g. 12a-reading). The alignments are compared by checking the temporal difference between the center of each phoneme; the table indicates the percentage of phonemes where this difference is less than 40 ms.

As expected, read speech is less variable (in phone duration, other prosodic characteristics) than spontaneous speech, and therefore the models achieve better results. However, the results are overall encouraging. This procedure also identifies utterances with important differences between the boundaries of phones,

Table 2. All samples of each region are aligned with each of the acoustic models trained on other regions. The table shows the percentage of phonemes whose center is within 40 ms of the center of the original alignment results.

Region to align	Text			Region to align	Conversation		
	11a	12a	13a		11a	12a	13a
11a		84.8%	84.8%	11a		81.7%	81.8%
12a	87.3%		89.7%	12a	83.5%		84.2%
13a	89.9%	90.1%		13a	82.8%	83.3%	

suggesting a potential problem in the transcription or a particularly difficult to align utterance. We intend to explore whether excluding these "problematic" utterances from the training of the final aggregate models improves the overall performance; however a small dataset of manually checked alignments will be needed for this evaluation.

6 Conclusion and Perspectives

We have presented a procedure for producing an automatic forced alignment of the *Phonologie du Français Contemporain* Corpus at the phone, syllable and token level, starting from the initial transcriptions. As part of this effort, the audio recordings were enhanced and restored, the transcriptions were checked for consistency, the data already coded in the corpus were used to improve the input to the ASR system, and multiple iterations of forced alignment using the Kaldi recogniser were performed.

The PFC corpus has been a valuable resource for studies in French phonology. We hope that this work will allow researchers to use the corpus in new ways and in investigating new research questions. For example, as part of the alignment process, speech pauses were detected with an improved precision: the corpus could be used for studying the dialogue dynamics in socio-linguistic interviews, or in similar studies in prosody. Concordances (text along with the corresponding sound segment) can now be extracted for downstream processing.

We plan to distribute the aligned version of the corpus to the community. To this end we plan to use institutional repositories (such as Ortolang) and also create a custom website using *PraalineWeb* (a tool generating *Django* websites for presenting speech corpora).

Finally, this project resulted in the development of a new tool for speech-to-text alignment of French spoken corpora, that we plan to release in the near future, along with the acoustic models trained on the PFC corpus.

References

1. Avanzi, M.: A corpus-based approach to French regional prosodic variation. Nouveaux Cahiers de Linguistique Française **31**, 309–332 (2014). (Proceedings of the 3rd SWIP)
2. Bigi, B., Hirst, D.: Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In: Proceedings of the 6th Speech Prosody Conference, 22–25 May, Shanghai, China (2012)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer, ver. 6.0.37 (2018). http://www.praat.org
4. Brognaux, S., Roekhaut, S., Drugman, T., Beaufort, R.: Train & Align: a new online tool for automatic phonetic alignment. In: 2012 IEEE Spoken Language Technology Workshop (SLT), pp. 416–421, December 2012
5. Christodoulides, G.: Praaline: integrating tools for speech corpus research. In: LREC 2014—Proceedings of the 9th International Conference on Language Resources and Evaluation, 26–31 May, Reykjavik, Iceland, pp. 31–34 (2014). http://www.praaline.org
6. Christodoulides, G., Barreca, G.: Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo. Corela: Cognition, Représentation, Langage HS-21 (2017). https://journals.openedition.org/corela/4867
7. Durand, J., Laks, B., Lyche, C.: Phonologie, variation et accents du français. Hermes, Paris (2009)
8. Durand, J., Lyche, C.: French liaison in the light of corpus data. J. Fr. Lang. Stud. **18**(1), 33–66 (2008)
9. Goldman, J.P.: EasyAlign: an automatic phonetic alignment tool under Praat. In: INTERSPEECH 2011—Proceedings of the 12th Annual Conference of the International Speech Communication Association, 27–31 August , Florence, Italy, pp. 3233–3236 (2011)
10. Hathout, N., Sajous, F., Calderone, B.: GLÀFF, a large versatile French lexicon. In: LREC 2014—Proceedings of the 9th International Conference on Language Resources and Evaluation, 26–31 May, Reykjavik, Iceland (2014)
11. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: PocketSphinx: a free, real-time continuous speech recognition system for hand-held devices. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, pp. I-I, May 2006
12. Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., Narayanan, S.S.: SailAlign: robust long speech-text alignment. In: Proceedings of the Workshop on New Tools and Methods for Very-Large Scale Phonetics Research (2011)
13. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Proceedings of the 18th Conference of the International Speech Communication Association (2017)
14. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, December 2011. IEEE Catalog No. CFP11SRW-USB
15. Walker, W., et al.: Sphinx-4: a flexible open source framework for speech recognition. Technical report, Sun Microsystems Inc., Mountain View, CA, USA (2004)
16. Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press, Cambridge (2006)