Is automatic phoneme recognition suitable for speech analysis? Temporal and performance evaluation of an Automatic Speech Recognition model in spontaneous French

Vincent Martin, Colleen Beaumard, Jean-Luc Rouas, Yaru Wu

Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Is automatic phoneme recognition suitable for speech analysis? Temporal and performance evaluation of an Automatic Speech Recognition model in spontaneous French

Vincent P. Martin[1], *Colleen Beaumard*[2,3], *Jean-Luc Rouas*[3], *Yaru Wu*[4]

[1]DDP Research Unit, Department of Precision Health, LIH, 1445 Strassen, Luxembourg
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[3]Univ. Bordeaux, CNRS, SANPSY, UMR 6033, F-33000 Bordeaux, France
[4]CRISCO/UR4255, 14032 Caen, France

vincentp.martin@lih.lu, {colleen.beaumard, jean-luc.rouas}@labri.fr, yaru.wu@unicaen.fr

Abstract

The correct automatic identification and segmentation of phonemes is crucial for a more in-depth exploration of prosodic parameters on a syllabic level. As such, automatic phonemic transcription from spontaneous speech recordings has numerous applications, such as teaching or health monitoring. Such transcriptions are usually evaluated either in terms of correct phoneme estimation or temporal segmentation, each task being addressed by a dedicated system. However, no system to our knowledge has ever been evaluated on doing correctly the two tasks at the same time. This article evaluates a state-of-the-art Kaldi-based phonetic transcription system for spontaneous French. We use the Rhapsodie database, composed of spontaneous speech recordings with diverse levels of planning. Our phoneme recognition system obtains good results on phoneme and phoneme category identification (respective error rates of 19.2% and 13.4%), performed poorly on phonemes and category segmentation: an average of 40% of phoneme duration and 34% of phonetic categories duration have not been detected by it. On both metrics, the performances of the system increase with the degree of planning of the spontaneous speech. These results suggest that improvements are necessary for designing truly reliable automatic phonetic transcription systems to be useful for further analysis.

Index Terms: speech recognition, phoneme recognition, phoneme segmentation

1. Introduction

1.1. Context

Segment boundaries provide valuable information for prosodic analyses and for speech analysis in general. Depending on the studied language, phrases are delimited using silence, pitch, intensity, and duration, with different contributions [1]. However, pauses have been observed to always be the most prominent feature, independently from the languages [1]. As such, pauses play a critical role in prosodic phrases automatic estimation, as shown for example by the work of Biron et al. [2] who automatically estimated them using speech rate and silent pauses. Incorporating phone boundaries and identification is crucial for a more in-depth exploration of prosodic parameters on a syllabic level, such as duration or pitch. Such prosodic parameters on the phoneme scale, extracted from spontaneous speech, open up a wide range of applications, from health monitoring (e.g. identifying depression by analyzing verbal fluency deficits [3]) or optimizing the learning of foreign languages [4]). However, all these tasks are language-dependent: our focus in this article is on French.

Since the foundational models in the 1970's based on statistical models [5, 6], the field of speech transcription has gone a long way towards end-to-end deep-learning models [7]. While the transcription systems were initially based on phoneme modeling, the latest and most performant models directly provide word transcription, either directly based on words or characters [7]. However, a small subset of applications is still interested in phonemic transcription, in order to assess the correctness of the pronunciation when learning a language, to detect words that are outside dictionaries, (e.g. children's speech [8]) or to evaluate the impact of pathologies on articulation [9, 10]. The field of automatic speech-to-phoneme transcription is divided between two different sub-applications: the correct estimation of the phoneme sequence, as measured by the Phoneme Error Rate (PER) on one side; and the correct segmentation of the audio file, delineating the location of the different phonemes (usually measured in terms of sensitivity, specificity and F1-score) on the other side.

On the automatic estimation of phonemes, the latest model for French to our knowledge has been released on Huggingface by the CNAM-LSSM[1], and is a finetuned version of a Facebook's Wav2Vec2 model. It has been trained on Common Voice v13 [11] and reached PERs of 5.5% and 4.4% on Common Voice v13 and the French subset of the Multilingual LibriSpeech (MLS) [12] respectively, which are both read speech corpora. We did not find any recent report of phoneme transcription performance for spontaneous speech in French. Regarding the segmentation of speech signals into phonemes, the latest approaches encompass self-supervised learning [13] and autoregressive models[14], achieving F1-scores around 90% on the TIMIT [15] and Buckeye [16] corpora. However, to our knowledge, these systems have only been evaluated on one of these tasks, and no system has been evaluated both in terms of phoneme recognition and phoneme segmentation.

1.2. Objective

Our objective is to evaluate a standard speech-to-phone transcription system for different styles of spontaneous speech in French, both in terms of phoneme recognition (phoneme error rate) and temporal accuracy (recall, precision and F1-score) This dual evaluation will be useful for any kind of analysis of

[1]https://huggingface.co/Cnam-LMSSC/wav2vec2-french-phonemizer

10.21437/SpeechProsody.2024-226

Source	Description	#rec.	#speakers	Duration	Style (#files)
CFPP2000	*Corpus de Français Parlé Parisien*, interviews about Paris district [17]	3	2 M / 5 F	15min.	Semi-spt (3)
Avanzi	Collected by M. Avanzi for the intonosyntactic study of macrosyntactic phenomena [18]	17	7 M / 15 F	14min	Spontaneous (17)
Lacheret	Collected for the continuous and functional modeling of French modeling [19]	2	3 M / 1 F	9 min.	Planned (1), Spontaneous (1)
Mertens	Collected for the intonosyncratic modeling of French [20]	2	4 M / 0 F	10 min	Planned (1), Semi-spt (1)
C-Prom	Collected to study the syllable prominences in French [21]	1	1 M / 0 F	3 min.	Planned (1)
ESLO	*L'Enquête Sociolinguistique à Orléans*, gathered in Orleans, France in 1968-74 with a sociolinguistic aim [22]	1	2 M / 0 F	7 min.	Planned (1)
PFC	*Phonologie du français contemporain*, directed conversations between a subject and an interviewer and informal conversations between two persons belonging to a dense social network, [23]	3	2 M / 4 F	14 min.	Spontaneous (3)
Movie	Monologues in which 7 different speakers are invited, in an in- formal setting, to describe a short scene from a Charlie Chaplin movie collected for the Rhapsodie project	7	4 M / 3 F	9 min.	Spontaneous (7)
Professional	Monologues and dialogues in a professional context collected for the Rhapsodie project	3	2 M / 2 F	8 min.	Spontaneous (3)
Broadcast	13 broadcasted monologues, dialogues and conversations downloaded from the Internet for the Rhapsodie project	13	22 M / 6 F	67 min.	Planned (7), Spontaneous (6)
All		54	49 M / 36 F	2h 41m	

Table 1: *Description of the Rhapsodie corpus: number of samples, number of speakers, duration of the corpus.*

automatically extracted phonemes, i.e. pronunciation assessment for language learners or pathological speech analysis.

This paper is organized as follows. We introduce the Rhapsodie corpus, our model, and the performance metrics in Section 2. We report and discuss the results of the designed system in Section 3 and draw conclusions in Section 4.

2. Methods

2.1. Phoneme recognition system

In this study, we use a standard Kaldi-based Automatic speech recognition system. It is a chain-based TDNN-HMM model trained with the LF-MMI objective function. We chose this approach to maintain the time stamp for each phoneme which can be a complex task when considering more recent end-to-end systems [24]. The neural network is based on a subsampled time-delay neural network with 7 TDNN layers, each one having 1024 units. The time stride value is set to 1 for the first three layers, 0 for the fourth, and 3 in the following ones. The acoustic model is based on a 40-dimensional high-resolution MFCC vector concatenated with a 100-dimensional i-vector [25]. It was trained using the Kaldi toolkit [26] on a sub-corpora of ESTER 1 and 2 (French) [27]. This system achieves a Word Error Rate of 13.7% on the test set of the ESTER corpus [28], which is close to state-of-the-art performances on the same corpus (slightly below 12% WER [29]). The phonetic symbols and their alignment are obtained using the `lattice-align-phones` command, resulting in the segmentation and annotation of 35 phonemes.

2.2. Rhapsodie corpus

Our analyses were carried out on the Rhapsodie corpus [30], a multigenre corpus of spoken French. The corpus contains three hours of speech in total ($\sim$33000 words), made up of 55 short samples (5 minutes on average). Face-to-face interviews, radio and TV broadcasts were covered and 89 speakers were included. The phonetic transcriptions are obtained using an automatic grapheme-to-phoneme (g2p) conversion tool (Easyalign [31] in Praat [32]), followed by manual verification [33]. Pauses were detected automatically. Two recordings, D0001 and D1003 (respectively in the CFPP2000 and Rhapsodie Professional subcorpora) have been excluded due to their poor acoustic quality. Another file, M2006 (Broadcast subcorpus), has been excluded due to mistakes in the timestamps of the ground-truth phonetic annotation. All the further results and statistics do not include these files. The different data sources that are used in the Rhapsodie corpus are described in Table 1.

While the corpus contains several variables to represent the discourse features of each sample, we focus in this study on analyzing the results of the automatic phoneme transcription system in light of the degree of speech planning that includes three categories: planned, semi-spontaneous and spontaneous.

2.3. Ground-truth: phonemes and categorization

The 54 files of the corpus represent a total of 96756 phonemes, which have an average duration of 81.2ms. To make it easier to gain insight from the results, the 35 different phonemes have been grouped into 10 standard categories: 5 categories for consonants (stops, fricatives, nasals, liquids, glides) and 5 categories for vowels (i.e. front rounded vowels, front unrounded vowels, central vowels, back rounded vowels, nasal vowel).

2.4. Performance metrics

In order to evaluate our system as comprehensively as possible, we measured recognition and correct segmentation performance according to the following metrics.

Phone Error Rate (PER) is the metric used in the field of automatic speech recognition (ASR) to measure the accuracy of the phonetic transcription of spoken language. The formula

for Phone Error Rate involves adding the number of substitutions (S), insertions (I), and deletions (D) of phonemes in the recognized output compared to the total number of phonemes in the reference transcription (N), and to compute the ratio $PER = 100 \times (S + I + D)/N$. A low value for Phone Error Rate indicates high accuracy in phoneme automatic recognition. The PER does not however take into account errors due to misplaced boundaries.

To complete the PER measurements, we also wish to measure the phoneme labeling in terms of duration. To this end, we use the *trackeval* tool (margin error = 0) that was used during the ESTER evaluation campaign for estimating the performances of audio event detection [27]. Here, the events to be correctly identified are phonemes. Doing so, we measured three metrics: *Recall*, which is the ratio of the duration of correct detection of a phoneme over the total duration of said event in the reference file: $R = \hat{d}_{corr}(phon)/d_{ref}(phon)$; *Precision*, which is the ratio between the duration of correct detections of a phoneme over the total duration of detection of this phoneme (including insertions): $P = \hat{d}_{corr}(phon)/\hat{d}_{corr+ins}(phon)$; and *F-measure*, a metric combining both precision and recall into a single value: $F = 2 \times (P \times R)/(P + R)$.

A high value for *Recall* indicates that the considered phoneme is well detected, while a high value for *Precision* shows that the system detects the phoneme mostly when it is really present (few insertions). Ideally, a good phoneme segmentation system should have high values for both *Precision* and *Recall*, hence a high *F-measure*.

Both the PER and segmentation metrics are then recomputed on phonetic categories.

3. Results and discussion

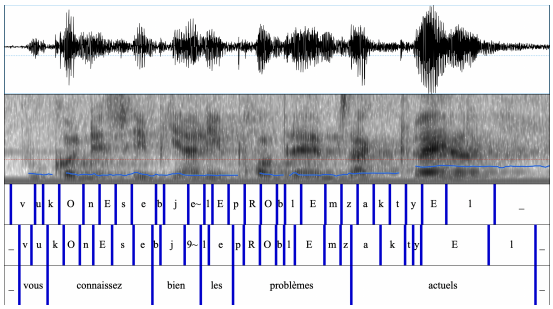

Figure 1: *Example output from our phonetic recognition system. Upper Tier: results of the automatic phoneme detection, Middle Tier: reference phonetic annotation, Lower Tier: reference word transcription. The example reads "vous connaissez bien les problèmes actuels" (you know well the current problems)*

An example of the output of our automatic phonetic recognition system is given on Figure 1. This example is an excerpt from file D1001 of the Rhapsodie corpus, originating from the ESLO database [22]. This recording is from 1968 and is a monologue from a male speaker. Most phonemes are detected correctly with the exception of two substitutions: /9~/ with /e~/ and /e/ with /E/, two couples of phones that are close to each other. Given the low number of phonetic token detection errors, if the phonetic recognition system behaves in the same way for other files, we hope to obtain low values for PER, demonstrating the efficiency of the automatic system. This is discussed in section 3.1. However, while the sequence of phones is well detected, we observe quite a lot of misplacements of boundaries, particularly at the end of the excerpt ("actuels"). These misplacements can be problematic for using this automatic segmentation for further analysis, such as prosodic or voice quality analysis on specific phonemes. The quality of the segmentation is discussed in section 3.3.

3.1. Phoneme recognition performances

The performances of our system on each phoneme are reported in Figure 2. The most common error type is substitutions, except for /2/ and /9/ which are mostly deleted. In particular, the vowel /e/ has a quite high substitution rate. Indeed, in continuous speech, /e/ is interchangeable with /E/ by native speakers of French. Since the system does not allow free choices between the two phonemes in the dictionary, we hypothesize that the system tends to label /e/s when it comes across a sound similar to /e, E/. On another note, the high insertion rate for schwa /@/ is highly related to the fact that the system is not informed about the vowel being optional in French.

When considering phonetic categories (Figure 3), we observe a drastic reduction of the number of substitutions, showing that most of the substitution errors are made on phonemes belonging to the same category. Moreover, interesting patterns are found for the error rates of the consonants: the more sonorous the consonant, the more difficult it is for the system to identify it.

3.2. Phoneme recognition performances depending on style

The performances of phoneme recognition in terms of phoneme and phoneme category depending on each subcorpus are reported in Tables 2 and 3. For both units of measurement, the error rate decreases with the degree of preparation of spontaneous speech.

However, since most of the files of the planned sub-corpus come from the "broadcast" source of the Rhapsodie database, we expected the automatic transcription system to better perform on this data than the other subcorpora since it is trained on similar broadcast samples (although not from the same period). It is not the case, partly due to one sample from the broadcast sub-corpus which is more of spontaneous nature than planned speaking (D2002, discussion on a book) resulting in poor performances (PER=24.9%). Still, the system performs quite well on the other semi-spontaneous subcorpora. While the error rate increases with more spontaneity in speech, insertion rates are relatively constant, while substitution rates slightly increase. The error type that increases the most is the deletion rate (from 3% on planned speech to 8.5% on spontaneous speech), reaching a maximum of 24.9% on the D2004 file from the Lacheret source (speaker with a strong regional accent).

Regarding the fact that even for spontaneous speech, more than 80% of the phonemes are detected correctly (87% for phonetic categories) and that errors are mainly due to deletions, we may assume that the automatic phoneme detection can be suitable for phonetic analysis on spontaneous speech. Nevertheless, if we wish to study for example the acoustic properties of phonemes, we want to make sure that the phonemes are detected correctly not only in terms of tokens, but also in terms of duration (or boundaries). This is the topic of the next section.

3.3. Phoneme segmentation performances

Tables 4 reports the performances of phoneme segmentation according to metrics detailed in section 2.4.

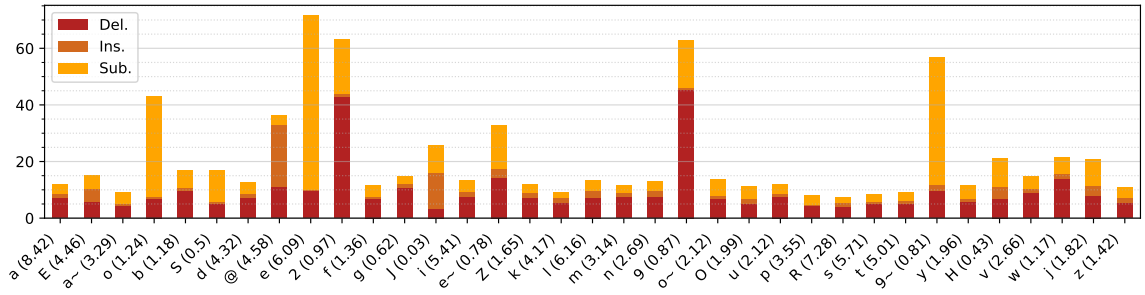

Figure 2: *Performances of the automatic phonetic recognition system for each phoneme. Values between brackets denote the ratio of the number of occurrences of the phoneme to the total number of phonemes*

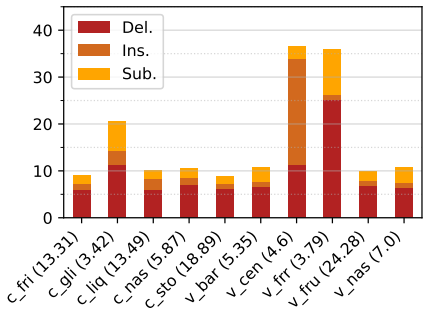

Figure 3: *Performances of the system on phoneme categories. Values between brackets denote the ratio of the number of occurrences of the phoneme belonging to the category to the total number of phonemes*

style	# files	# phones	Corr	Sub	Del	Ins	Err
All	54	96756	83.9	9.4	6.7	3.4	19.5
planned	11	30549	89.2	7.6	3.2	3.0	13.8
semi-spt	4	15302	82.8	9.3	7.9	3.3	20.5
spontaneo	39	50905	81.0	10.5	8.5	3.6	22.6

Table 2: *Error details for phonetic tokens detection for different speaking styles*

style	# files	# phones	Corr	Sub	Del	Ins	Err
All	54	96756	89.7	3.5	6.7	3.4	13.6
planned	11	30549	94.9	1.9	3.2	3.0	8.1
semi-spt	4	15302	88.7	3.5	7.8	3.3	14.6
spontaneo	39	50905	87.0	4.5	8.6	3.6	16.7

Table 3: *Error details for phonetic categories detection for different speaking styles*

Of the 7912 seconds to be detected, only 4746 s. (R=60.0%) are correctly detected at the phoneme level, with a precision of P=68.2%, leading to a F-score of 0.62. This value reaches 5232 s. when considering phonetic categories (66.1%), with a greater precision P = 71.0%, leading to a corresponding F-score of 0.68, which is still under 0.7.

Regarding the effect of the degree of planning, similarly to what we observed in section 3.1, adding more spontaneity degrades the results, both when considering phonemes (from an F1-score of 0.67 for planned discourse to 0.58 for spontaneous discourse) and phonetic categories (from F=0.73 to F=0.68).

Moreover, our system performs unevenly depending on the phonetic classes: while it segments with good performances nasal vowels (target duration=823 s., R=71%, P=81%, F=0.76) and fricative consonants (target duration=1117 s., R=71%, P=78%, F=0.74), it struggles to estimate the borders of the central (target duration=330 s., R=60%, P=47%, F=0.53), front rounded vowels (target duration=595 s., R=30%, P=72%, F=0.42) and glide consonants (target duration=170 s., R=57%, P=48%, F=0.52). Other phonetic categories are segmented with F-scores between 0.60 and 0.68.

		Phonemes			Phoneme categories		
	target	%R	%P	F	%R	%P	F
planned	2596s	66.5	68.2	0.67	72.3	74.1	0.73
semi-spt	1198s	59.0	64.6	0.62	65.2	71.3	0.68
spontan.	4118s	55.5	61.5	0.58	62.5	68.8	0.65
All	7912s	60.0	64.4	0.62	66.1	71.0	0.68

Table 4: *Segmentation evaluation details for phonetic tokens detection for different speaking styles*

4. Conclusion

This paper evaluated the performance of a state-of-the-art phoneme recognition system for spontaneous French, not only in terms of token detection but also on correct segmentation. Based on the Rhapsodie corpus, which contains spontaneous speech from several sources with three degrees of spontaneity, we have computed identification and segmentation metrics both on the phonemic level and depending on ten standard phonetic categories (5 types of vowels and 5 types of consonants). We have shown that a given ASR system could at the same time obtain satisfactory phoneme identification performances (global PER of 19.5%, error rate of 13.6% on categories) and unsatisfactory segmentation performances (F-scores of 0.62 and 0.68 for phonemes and categories respectively). However, on both evaluation, a lot of disparities depending on the type of speech and the considered phonemes were observed. Furthermore, considering phoneme categories enhances the performances on both evaluations, suggesting that the substitutions are done in the same phonetic group. Moreover, all performance metrics increase with the degree of planning of the spontaneous speech under consideration. As most phonetic recognition systems are only evaluated on read speech, our results call for caution when using these systems for linguistic or prosodic analysis of spontaneous speech.

5. Acknowledgements

This research is supported by the CNRS through the MITI PRIME 80 DSM-HEALTH and the French Research Agency ANR through the axis "Autonom-Health" of the PEPR "Santé Numérique", Grant agreement n°ANR-22-PESN-000X and by the European Union's Horizon Europe research and innovation program through the Marie Skłodowska-Curie grant agreement No. 101106577.

6. References

[1] V. Soto, E. Cooper, A. Rosenberg, and J. Hirschberg, "Cross-language phrase boundary detection," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 8460–8464.

[2] T. Biron, D. Baum, D. Freche, N. Matalon, N. Ehrmann, E. Weinreb, D. Biron, and E. Moses, "Automatic detection of prosodic boundaries in spontaneous speech," PloS one, vol. 16, no. 5, p. e0250969, 2021.

[3] J. D. Henry and J. R. Crawford, "A meta-analytic review of verbal fluency deficits in depression," Journal of Clinical and Experimental Neuropsychology, vol. 27, no. 1, p. 78–101, Feb. 2005. [Online]. Available: http://dx.doi.org/10.1080/138033990513654

[4] G. Bailly, E. Godde, A.-L. Piat-Marchand, and M.-L. Bosse, "Automatic assessment of oral readings of young pupils," Speech Communication, vol. 138, 02 2022.

[5] J. K. Baker, Stochastic modeling as a means of automatic speech recognition. Carnegie Mellon University, 1975.

[6] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, vol. 64, no. 4, pp. 532–556, 1976.

[7] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil, "Automatic speech recognition: Systematic literature review," IEEE Access, vol. 9, pp. 131 858–131 876, 2021.

[8] L. Gelin, T. Pellegrini, J. Pinquier, and M. Daniel, "Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition," in Proc. Interspeech 2021, 2021, pp. 3860–3864.

[9] M. Huckvale, Z. Liu, and C. Buciuleac, "Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech," Biomedical Signal Processing and Control, vol. 86, p. 105201, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423006341

[10] C. Beaumard, V. P. Martin, Y. Wu, J.-L. Rouas, and P. Philip, "Automatic detection of schwa in French hypersomniac patients," in Journée Santé et Intelligence Artificielle (Evènement affilié à PFIA 2023), 2023.

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[12] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in Interspeech 2020. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2826

[13] L. Strgar and D. Harwath, "Phoneme segmentation using self-supervised speech models," in IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 1067–1073.

[14] H. Kim and H.-S. Choi, "Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[15] Garofolo, John S., Lamel, Lori F., Fisher, William M., Pallett, David S., Dahlgren, Nancy L., Zue, Victor, and Fiscus, Jonathan G., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," p. 715776 KB, 1993.

[16] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Communication, vol. 45, no. 1, pp. 89–95, Jan. 2005.

[17] S. Branca-Rosoff and F. Lefeuvre, "Le CFPP2000 : constitution, outils et analyses. Le cas des interrogatives indirectes," Corpus, no. 15, Oct. 2016.

[18] M. Avanzi, L'interface prosodie/syntaxe en français, Jan. 2013.

[19] A. Lacheret-Dujour, La prosodie des circonstants en français parlé, ser. Collection linguistique (Paris). Paris: Peeters, 2003.

[20] P. Mertens, "L'intonation du français : De la description linguistique à la reconnaissance automatique," Ph.D. dissertation, Jan. 1987.

[21] M. Avanzi and A. C. Simon, "C-PROM: An Annotated Corpus for French Prominence Study," Speech Prosody, 2010.

[22] I. Eshkol-taravella, O. Baude, D. Maurel, L. Hriba, C. Dugua, and I. Tellier, "Un grand corpus oral disponible : le Corpus d'Orléans 1968-2012 [A Large available oral corpus: Orleans corpus 1968-2012]," Traitement Automatique des Langues, vol. 52, no. 3, pp. 17–46, 2011.

[23] J. Durand, B. Laks, and C. Lyche, "Phonologie, variation et accents du français," J. Durand, B. Laks, and C. Lyche, Eds. Hermès, 2009, ch. Le projet PFC: une source de données primaires structurées, pp. 19–61.

[24] F. Boyer and J.-L. Rouas, "End-to-End Speech Recognition: A review for the French Language," arXiv, 2019.

[25] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in ICASSP, 2014, pp. 6334–6338.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii.

[27] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," in Interspeech 2009, 2009, pp. 2583–2586.

[28] F. Boyer, "Reconnaissance de parole pour le français et intégration dans un système de compréhension du langage parlé," Ph.D. dissertation, Université de Bordeaux, 2021.

[29] A. Heba, "Reconnaissance automatique de la parole à large vocabulaire : Des approches hybrides aux approches End-to-End," Theses, Université toulouse 3 Paul Sabatier, Mar. 2021.

[30] A. Lacheret-Dujour, S. Kahane, and P. Pietrandrea, Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French. John Benjamins, 2019.

[31] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under praat," in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[32] P. Boersma, "Praat, a system for doing phonetics by computer," Glot. Int., vol. 5, no. 9, pp. 341–345, 2001.

[33] A. Lacheret, S. Kahane, J. Beliao, A. Dister, K. Gerdes, J.-P. Goldman, N. Obin, P. Pietrandrea, and A. Tchobanov, "Rhapsodie: a prosodic-syntactic treebank for spoken French," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 295–301. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/381_Paper.pdf