Using a Knowledge Base to Automatically Annotate Speech Corpora and to Identify Sociolinguistic Variation

Yaru Wu, Fabian Suchanek, Ioana Vasilescu, Lori Lamel, Martine Adda-Decker

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1054–1060
Marseille, 20-25 June 2022

Using a Knowledge Base to Automatically Annotate Speech Corpora and to Identify Sociolinguistic Variation

Yaru Wu[1,3,4], Fabian Suchanek[2], Ioana Vasilescu[3], Lori Lamel[3], Martine Adda-Decker[4]
[1]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France,
[2]Télécom Paris, Institut Polytechnique de Paris, France,
[3]LISN, Univ.Paris-Saclay, 91405 Orsay cedex, France,
[4]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
yaru.wu@unicaen.fr, suchanek@telecom-paris.fr,
{ioana.vasilescu, lori.lamel}@lisn.upsaclay.fr, martine.adda-decker@sorbonne-nouvelle.fr

Abstract

Speech characteristics vary from speaker to speaker. While some variation phenomena are due to the overall communication setting, others are due to diastratic factors such as gender, provenance, age, and social background. The analysis of these factors, although relevant for both linguistic and speech technology communities, is hampered by the need to annotate existing corpora or to recruit, categorise, and record volunteers as a function of targeted profiles. This paper presents a methodology that uses a knowledge base to provide speaker-specific information. This can facilitate the enrichment of existing corpora with new annotations extracted from the knowledge base. The method also helps the large scale analysis by automatically extracting instances of speech variation to correlate with diastratic features. We apply our method to an over 120-hour corpus of broadcast speech in French and investigate variation patterns linked to reduction phenomena and/or specific to connected speech such as disfluencies. We find significant differences in speech rate, the use of filler words, and the rate of non-canonical realisations of frequent segments as a function of different professional categories and age groups.

Keywords: diastratic variation, knowledge base, socio-linguistic factors, journalistic French, large corpora, non-canonical realisations, automatic alignment

1. Introduction

Speech characteristics depend on many factors. While some variations are due to the communicative context (e.g., formal *vs* familiar), others are due to individual traits of the speaker such as gender, origin, age, and socio-economic status, or to the emotional state of a speaker in a particular situation.

How speech varies according to the speakers' characteristics has been a key focus in variationist studies over the last fifty years, starting with Labov's seminal work (Labov et al., 2013; Labov and Rosenfelder, 2011). More recently, speech technology has also taken on the subject, as speech-based applications can be highly impacted by uncontrolled variation (Benzeghiba et al., 2007; Goldwater et al., 2010; Adda-Decker and Lamel, 2017). Speaker-related patterns of variation are thus of interest for linguists and speech technologists alike, for example to improve the performance of speech-to-text transcription systems, of speaker diarization and/or identification algorithms, of emotion recognition systems, or to customise the output of speech generation systems. However, a principled analysis of such patterns relies on the availability of speech corpora with annotations for speaker identities and profiles, or on the recruitment of volunteers, obtaining their profiles, and acquiring speech corpora from them in various communicative settings. Both are time and money consuming endeavours, and in terms of methodology tributary to the Observer's paradox that may impact the factors responsible for the linguistic variation (Labov et al., 2013).

This paper has two principal aims. First, we propose a novel method that can be used to automatically enrich existing corpora with annotations that cover speakers' profiles in a broad sense. Second, we assess the utility of the method on existing corpora by correlating the annotations with an array of linguistically motivated features. We estimate the statistical relevance of the features for the characterisation of the speakers in the database.

The input to our method is a corpus of recorded speech that has been orthographically transcribed, and with segments (roughly speaker turns) annotated with the speakers' names. A knowledge base (KB) is used to retrieve the characteristics of the speakers, such as their age at the date of the recording, their profession, their gender, their place of birth. We then extract the characteristics of their speech using off-the-shelf tools. The proposed method thus enables highlighting links between linguistic knowledge and various extra-linguistic meta-data *via* the correlation between speech features such as non-canonical pronunciations, segment duration, and disfluencies, and individual speaker characteristics (eg., age, profession, gender). It allows the discovery of patterns between socio-cultural speaker traits and instances of speech variation. In this study our method is applied to a large set of French broadcast news extracts that contain speech from 295 known

adult speakers covering a variety of professions. We are able to discern several interesting correlations, such as the relation between age and non-canonical pronunciations of voiced and voiceless consonants, or the profession and the speech rate.
The remainder of this paper is structured as follows: Section 2 discusses related work, Sections 3 and 4 describe our method and the corpus. Section 5 is dedicated to the results, followed by some conclusions in Section 6.

2. Related Work

Identifying speakers based on individual speech patterns is a key topic for many speech technology applications, aiming not only to transcribe words but also to say "who spoke when" and possibly under the influence of which emotional state. Speech recognition and speaker diarization are thus two domains concerned with speaker's characteristics (Lamel et al., 2004; Tranter and Reynolds, 2006; Park et al., 2021). In order to train algorithms dedicated to such tasks, data are needed with not only reference transcriptions of what is being said, but also segmentation into speaker turns and annotation of the speaker identities. (Broux et al., 2018). As for emotion recognition from speech, both external emotional triggers and speakers' *a priori* communicative patterns appear to be of interest to accurately link speech chunks and specific emotion labels (Koolagudi and Rao, 2012). The annotation with corpus dependent emotion labels, attitudes and/or cognitive states is a laborious task challenged by the intrinsic subjectivity of such labels. The consequence is that even today few "real-life" corpora portraying actual emotions are available.

As for the linguistic analysis *per se*, speech variation has been a topic of study in phonetics and laboratory phonology. Prominent variations include differences between the expected canonical pronunciations and the effective realisation of words by speakers in various communicative contexts. Among the observed phenomena, reductions are frequent patterns of variation, resulting, for instance, in consonant weakening (Jatteau et al., 2019a; Jatteau et al., 2019b) or in vocalic centralisation (Audibert et al., 2015) as a consequence of reduced durations. Reductions occur as a consequence of various linguistic and extra-linguistic factors, including speech rate, speaking style, lexical frequency, morphological properties of the words, etc. (Ernestus, 2011). Reductions are of interest for linguistics, for instance for laboratory phonology studies dedicated to the relation between synchronic variation and sound change (Vasilescu et al., 2020). They are also of interest for speech technology as such processes in connected speech may impact the performance of speech technology, eg. automatic speech recognition (Adda-Decker and Lamel, 2017; Vasilescu et al., 2018). In addition to reduction phenomena, connected speech, in particular in spontaneous settings, is also characterised by numerous disfluent events (Shriberg, 1994). Disfluencies occur as empty or filled pauses (such as *euh* in French and *uh, um* in English), and as a range of reformulation strategies that help build the verbal message and keep the communicative connection. The proportion and type of disfluencies can be correlated with the corpus, speaker's specificities (Clark and Fox Tree, 2002) and emotion labels (Tian et al., 2015).

Sorting out the triggering factors of variation phenomena has thus been a long-term activity in the linguistic community for which correlating speaker features with variation patterns remains a challenge. The reason is that building suitable corpora requires the recruitment of speakers meeting the diastratic criteria under investigation, and a large amount of time and money to transcribe, annotate and analyse the corpora. Such analyses are thus constrained both by the type and quantity of annotations and in terms of available socio-linguistic metadata.

In this paper, we address the limitations observed in both speech technology and linguistic approaches, by resorting to the use of a knowledge base instead of manual annotations to obtain the socio-linguistic annotations. Combined with speech processing technologies, this allows us to automatically extract linguistic speech variation knowledge and to analyse socio-linguistic variations on a large scale.

3. Approach

Our method takes as input a corpus of recorded speech that has been orthographically transcribed, as well as annotated with the speakers' names. It also requires a knowledge base (KB), that is a computer-processable collection of knowledge about the real world, with information about these speakers. There exist numerous general-purpose KBs (Weikum et al., 2021), and they usually store information about entities such as people, organizations, or locations. This information takes the form of triples, each of which specifies a relation between an entity and another entity, or between an entity and a literal – as in ⟨*EmmanuelMacron, presidentOf, France*⟩ or ⟨*EmmanuelMacron, birthDate, "1977-12-21"*⟩.
The problem of determining which name in the corpus refers to which name in the KB is known as *entity linking*, and several approaches have been developed for this (Weikum et al., 2021). We use a particularly simple approach, which just links every speaker name in the corpus to the person in the KB of the same name. In case of homonymy, or in the absence of a suitable person in the KB, we discard the utterances of that speaker from our corpus. This approach can obviously be refined, but works surprisingly well, as our analysis shows.

We then extract an array of instances of speech variation from our corpus, based on speech-to-text forced alignments permitting pronunciation variants. For the current study we focus on some frequent instances of variation concerning speech rate (Adda-Decker and Lamel, 2017), voicing alternation processes (Jatteau et al., 2019b; Vasilescu et al., 2020), and hesitation rate (Vasilescu and Adda-Decker, 2006). Finally, we correlate the speaker characteristics extracted from the knowledge base with the measured patterns of speech variation, so as to link the specific speech variations to categories of age, region of birth, and profession.

4. Data and Methodology

This section provides more detail about the data and methodology used to demonstrate the validity of our proposed approach.

4.1. Speech Corpus

Our study makes use of two corpora of broadcast speech in French, totalling over 120 hours. The ESTER (Galliano et al., 2005) corpus contains 80 hours of semi-prepared or prepared formal speech (radio broadcast news). The two-part ETAPE corpus, ETAPE-1 and 2 (Gravier et al., 2012) contains 13.5 hours of radio data and 29 hours of TV data in French including more interactive data than ESTER such as debates and interviews. Both corpora come with manual word-level transcriptions, and automatic alignments with the speech were carried out using an available speech recognition system for French (Gauvain et al., 2005). The method of (Vasilescu et al., 2020) was adopted in order to allow the alignment of non-canonical pronunciation variants for French stops /ptkbdg/ (e.g. permitting voiced variants [bdg] for /ptk/, and voiceless [ptk] for /bdg/). Duration patterns and filler word rates (here, filler words are the French hesitations "euh" and "hum") are computed from the automatic segmentations provided by the system.

4.2. Knowledge Base

We use the Yago 4 knowledge base (Pellissier-Tanon et al., 2020), because it combines the wealth of Wikidata, the largest public knowlegde base, with a particularly simple schema. Yago 4 contains more than 50 million entities and 2 billion facts about them. It includes in particular the gender, birth date, place of birth, nationality, and profession of many people, among other things. For this study, the birth places were grouped into 4 broad categories *Northern France*, *Parisian Region*, *Southern France*, and *Maghreb*. Although a priori Maghrebian speakers are not native speakers of French, they often exhibit a near-native level of French, and are frequent in our data. We thus decided to keep this as a category of speaker origin. Yago lists 68 professions for the speakers of our corpus, which we manually grouped into the 7 broad categories: *Journalist, Movie personality, Politician, Scientist, Lawyer, Artist/Writer* and *Business person*. Table 1 shows the number of speakers and word tokens for each profession in our corpus.

Profession	Speakers	Word Tokens
Politician	102	81541
Journalist	61	121083
Artist/writer	33	47889
Movie personality	29	36555
Scientist	27	34326
Lawyer	27	19744
Business person	16	8337

Table 1: Number of speakers and word-tokens by profession.

4.3. Entity Linking

Our corpus contains 342 distinct named speakers, all of whom are in YAGO 4. While none have an ambiguous name, a mapping can still be erroneous if a (lesser-known) speaker happens to have the same name as a (better-known) person in the KB. We verified the mapping on a sample of 45 randomly selected speakers, and 3 were mapped erroneously, giving us an estimated error rate of 7%.

4.4. Preprocessing

The age of a speaker is calculated by subtracting the year of birth from the year of the recording, and speakers were grouped into three age categories: 20-29 years, 30-59 years, and 60+ years. None of the speakers were younger than 20 years old and there were too few speakers over the age of 60 to be included in our analyses. While we are aware that this grouping can be refined, it provides a preliminary basis to highlight age-specific variation patterns. The speech rate is estimated for speech chunks between pauses as the number of words per second. Speech rate was calculated for speech utterances between pauses (including pauses/hesitations/breaths) that are equal to or greater than 200ms. Voicing alternation, determined as described above, is taken as an illustrative example to study non-canonical pronunciation variation as a function of speaker traits.

4.5. Statistical Significance

Linear mixed models (LMM) with the *lme4* packages (Bates et al., 2011) in R (R Core Team, 2013) were used to assess the statistical significance of our results. Two LMMs were used for our analyses on speech rate, given that slightly different datasets were used due to data selection (e.g. analyses related to birthplace only concern speakers born in the North/South of France, in Paris or in the Maghreb). The first LMM was used to test the effect of profession and age group. The fixed effects considered were: profession (Politician, Journalist, Artist/writer, Movies, Lawyer, Scientist and

Businessperson; reference: Politician) and age group (20-29 years, 30-60 years and 60+ years; reference: 20-29 years). As random effects, intercepts were included for the subject. For the second LMM, birth place was included as fixed effect (North, Paris, South, Maghreb; reference: Paris). Profession and age were included as control variables and intercepts were included for the subject in this model. Post-hoc tests were performed for each fixed effect based on each model.

Generalized linear models (GLM) were used to assess the significance of the analyses of filler words and voicing alternation. Two GLMs for each were conducted since the analyses of the post-lexical contexts each concern two different subsets of the data. The first GLM was used to test the effect of profession and age group on filler words. Similar to LMMs, the fixed effects were profession and age group. As for the second GLM on filler words, birth place was included as a fixed effect, with profession and age as control variables. Post-hoc tests were performed for each fixed effect based on each model. The third and fourth GLMs carried out for voicing alternation used the same fixed effects and control variables as were used in the first and second GLMs.

5. Results

In the following we present the results concerning the correlation between speaker characteristics and speech and acoustic measurements. We focus on the effect of speaker *profession*, *provenance* and *age*, on *speech and filler words (French "euh" and "hum") rates*, and on *+/-canonical pronunciation of some frequent segments* rates. The results are grouped as a function of the socio-linguistic factors of profession, age, and region of birth. We do not report on gender differences, as these have been studied before (Hutin et al., 2020), and do not require a knowledge base.

5.1. Speech rate

Figure 1 shows the speech rate for the different professions (top) and age groups (bottom). We also studied the speech rate as a function of the speaker's birth region, but saw no clear evidence of the influence of birth place on speech rate. While we did observe that politicians born in the South tend to have a slower speech rate than other regions, the LMM results suggest that the impact of the birthplace on the speech rate is not statistically significant.

5.1.1. Profession *vs* speech rate

The LMM results indicate that Movie personalities ($p < 0.001$), Scientists ($p < 0.05$), and Businesspersons ($p < 0.01$) have a significantly higher speech rate than Politicians. Post-hoc tests show significant differences for Journalist *vs* Movie personality ($p < 0.001$), Movie personality *vs* Artist/writer ($p < 0.001$), Movie personality *vs* Lawyer ($p < 0.001$) and Movie personality *vs* Scientist ($p < 0.05$). This may be because politicians

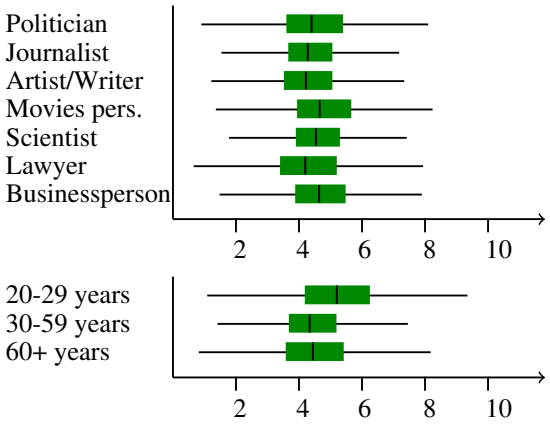

Figure 1: Speech rate (words per second) as function of profession (top) and age (bottom).

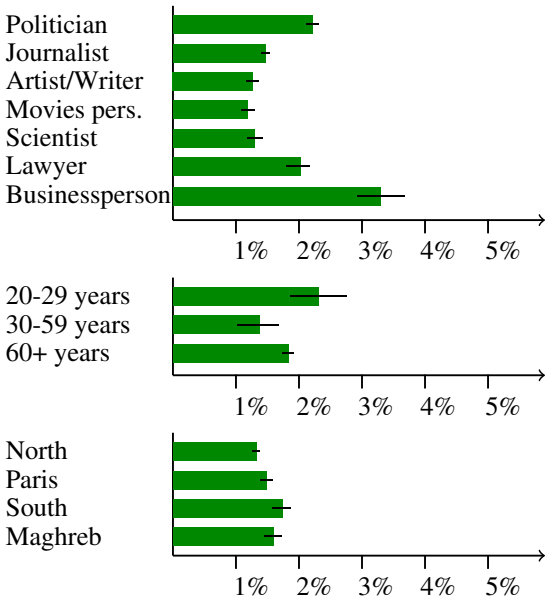

Figure 2: Percentage of filler words (Wilson score interval at $\alpha = 5\%$) according to profession, age and birth region.

and journalists tend to be professional public speakers. Concerning journalists, we observed that the data consist mainly of (semi-) prepared speech. Thus, the slower speech rate may be chosen consciously to address a large and socially diverse audience.

5.1.2. Age *vs* speech rate

As seen in the bottom plot in Figure 1, the younger group exhibits a higher speech rate than the two other groups, and this difference is statistically significant. A post-hoc analyses based on the model confirmed that the difference in speech rates between the middle and older group are not significant.

5.2. Filler words

Figure 2 plots the percentage of filler words as a function of profession (top), age group (middle) and birth

region (bottom).

5.2.1. Profession *vs* filler words

Filler words are classically linked to a spontaneous speech style (Clark, 2002). As can be seen in the top plot, business people tend to use more filler words than other professions, followed by politicians and lawyers. Our GLM indeed shows that the categories Businessperson, Scientist, Movie personality, Artist / writer and Journalist produce significantly fewer filler words than Politician ($p < 0.001$ for all comparison). Post-hoc tests also show significant differences between Businessperson vs Laywer / Scientist/Movie personality, Artist/writer and Journalist ($p < 0.001$ for all comparisons). A significant difference in filler word production is also found for journalists *vs* scientists. Journalists, artists / writers and movie personalities rarely use filler words in our data (<2%). We computed the Wilson score interval at $\alpha = 5\%$, and these also confirm that the above results are statistically significant.

5.2.2. Age *vs* filler words

The rate of filler words by age group is shown in the middle plot. Younger speakers tend to produce more filler words, and our GLM model shows that this difference is statistically significant.

5.2.3. Region of birth *vs* filler words

The bottom plot in Figure 2 shows the rate of filler words by region of birth. The results show an interesting pattern, where Southern French-born people tend to produce more filler words than Northern-born people. These differences are statistically significant. Speakers from the Maghreb region tend to produce more filler words than Parisian speakers and fewer than speakers born in the South of France.

5.3. Non-canonical surface forms

In our study, non-canonical surface forms correspond to stops that changed their voicing category, that is canonical /ptk/ pronounced [bdg] and canonical /bdg/ pronounced [ptk]. Both phenomena correspond to acoustic target undershoots that can be associated with different factors such as the influence of the context or the position in the word (Vasilescu et al., 2020).

5.3.1. Profession *vs* Non-canonical surface forms

The two top plots of Figure 3 show the number of canonical voiceless consonants /ptk/ (top plot) and voiced /bdg/ (second plot) that were pronounced non-canonically. Linguistic literature associates /ptk/>[bdg] with weakening and /bdg/>[ptk] with strengthening, however both are reduction phenomena in the sense of the production of segments that are not expected and may be due to faster and less carefully articulated speech. On average, in our data voicing is more frequent than devoicing: /ptk/ are more often pronounced [bdg] (change rate at 9.5%) than /bdg/ pronounced [ptk] (change rate at 7.8%). Voicing is a reduction phenomenon associated to decrease in duration (Priva and Gleason, 2020) and professions such as politician, journalist, lawyer and business person show a stronger correlation with voicing. In terms of statistical significance, results of the GLM model underline that the categories Businessperson, Movie personal-

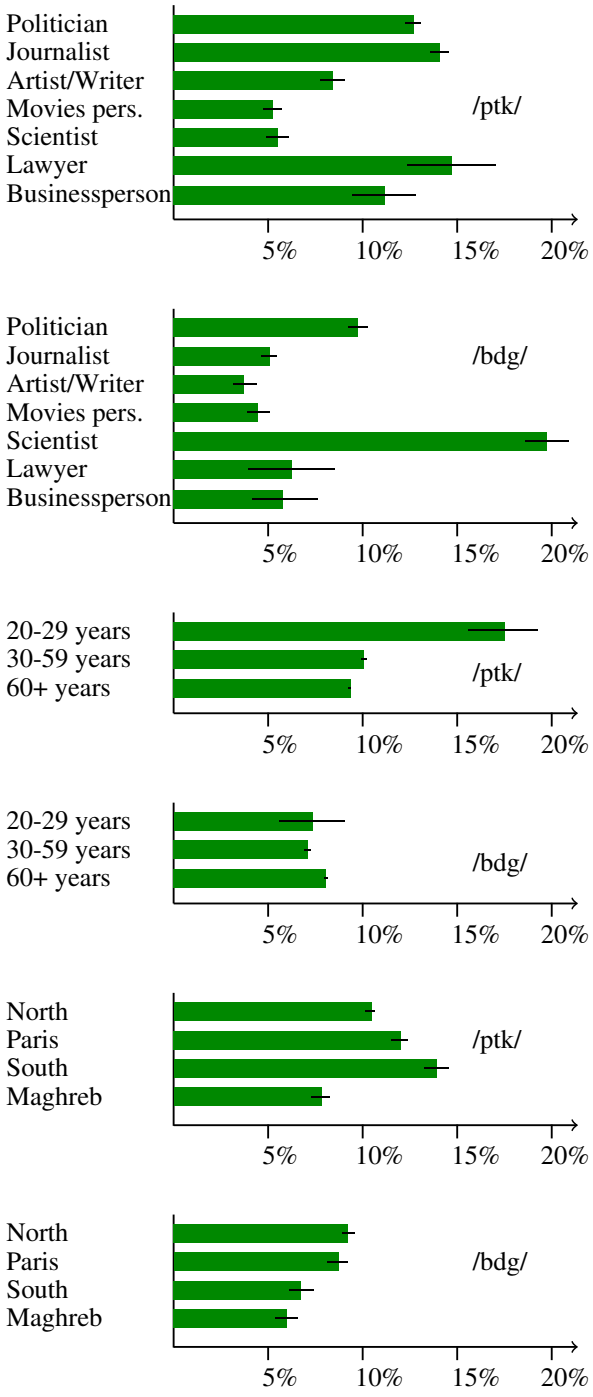

Figure 3: Voicing alternation rate of French stops /ptk/ and /bdg/ (Wilson score interval at $\alpha = 5\%$) according to profession, age and birth region.

ity, Artist/writer and Journalist show significantly less voicing alternation than the category Politician ($p < 0.001$ for all comparisons, except for politician *vs* journalists, which has $p < 0.01$). Post-hoc tests also show significant differences for Businessperson *vs* Scientist ($p < 0.001$), Businessperson *vs* Movie personality ($p < 0.001$), Businessperson *vs* Artist/writer ($p < 0.01$), Scientist *vs* Movie personality ($p < 0.001$), Scientist *vs* Artist/writer ($p < 0.001$), Scientist *vs* Journalist ($p < 0.05$), Movie personality *vs* Artist/writer, Movie personality *vs* Journalist and Artist/writer *vs* Journalist (all $p < 0.001$).

5.3.2. Age *vs* Non-canonical surface forms

The middle two plots of Figure 3 illustrate the rate of non-canonical realization of /ptk/ and /bdg/ as a function of the three age groups. As before, it can be seen that voicing is more frequent than devoicing. A particularly high voicing alternation rate is observed for young speakers. Our GLM results suggest that this difference is statistically significant ($p < 0.001$), a finding in line with general socio-phonetic predictions about young speakers initiating sound change (Labov, 1963).

5.3.3. Region of birth *vs* Non-canonical surface forms

The bottom two plots of Figure 3 show the rate of non-canonical stop realisations as a function of the region of birth. Speakers born in the South of France tend to have a higher non-canonical realisation rate of /ptk/, whereas those from Maghreb have the lowest. The opposite pattern is found for the non-canonical realisation rate of /bdg/ where the Southern speakers and those from Maghreb show the lowest proportion of non-canonical realisations. Significant differences are found for all pairwise comparisons of the region of birth ($p < 0.001$), based on the GLM outputs.

6. Conclusions

In this paper we have proposed a novel method to automatically enrich existing corpora with speaker characteristics extracted from a knowledge base. In order to estimate the benefits of the method for linguistic analysis and for in-depth characterisation of the speakers to potentially support speech applications, we extracted several speech features such as speech rate, filler words, and non-canonical pronunciation of frequent segments that we correlated with diastratic information about the speakers (such as profession, age, and region of birth). The proposed procedure allowed us to go beyond "classical" speaker features that are usually studied in naturalistic speech (e.g. gender), and to associate speech features and diastratic speaker information on a large scale.

We applied our method on a corpus of French broadcast news, and studied the relation between selected speech patterns and several socio-linguistic features. We observe interesting patterns for all investigated variation features that can be of interest for both socio-linguistic studies and speech technology. For the former, they can help to refine the link between fine-grained phonetic patterns and various speaker characteristics. For the latter, a direction to follow may concern the automatic modelling of speakers' profiles given extracted linguistic knowledge.

For instance, the variation is linked to younger speakers who are generally initiators of sound change. This variation also results in increasing non-canonical realisations that are responsible for some automatic processing errors (Vasilescu et al., 2018). In further work we will extend our analysis to other corpora, languages and more socio-linguistic criteria. We will also explore to what extent other documented features that are not *a priori* intended to be used as meta-data can be of help for to extract new types of linguistic knowledge.

We believe that our methodology can be ported to such new settings, and be of help for both speech technology and the humanities research.

7. Acknowledgements

This research was supported by the DATAIA / MSH Paris-Saclay grant "OTELO" and the ANR-20-CHIA-0012-01 ("NoRDF").

8. Bibliographical References

Adda-Decker, M. and Lamel, L. (2017). Discovering speech reductions across speaking styles and languages. In *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. De Mouton Gruyter.

Audibert, N., Fougeron, C., Gendrot, C., and Adda-Decker, M. (2015). Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation. In *International Congress of Phonetic Sciences*, page 5.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., and Grothendieck, G. (2011). Package 'lme4'. *Linear mixed-effects models using S4 classes. R package version*, 1(6).

Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: a review. *Speech Communication*, 49(10–11):763–786.

Broux, P.-A., Doukhan, D., Petitrenaud, S., Meignier, S., and Carrive, J. (2018). Computer-assisted Speaker Diarization: How to Evaluate Human Corrections. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, (84 (1)):73–111.

Clark, H. (2002). Speaking in time. *Speech Communication*, (36):5–13.

Ernestus, M. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, (39):253–260.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Interspeech*, pages 1149–1152.

Gauvain, J.-L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., and Schwenk, H. (2005). Where are we in transcribing french broadcast news? In *Interspeech*, pages 1665–1668.

Goldwater, S., Jurafsky, D., and Manning, D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, (52 (1)):181–200.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the french language. In *International Conference on Language Resources and Evaluation*, pages 114–118.

Hutin, M., Jatteau, A., Vasilescu, I., Lamel, L., and Adda-Decker, M. (2020). Ongoing Phonologization of Word-Final Voicing Alternations in Two Romance Languages: Romanian and French. In *Interspeech*, pages 4138–4142.

Jatteau, A., Vasilescu, I., Lamel, L., and Adda-Decker, M. (2019a). Final devoicing of fricatives in French: Studying variation in large-scale corpora with automatic alignment. In *International Congress of Phonetic Sciences*.

Jatteau, A., Vasilescu, I., Lamel, L., Adda-Decker, M., and Audibert, N. (2019b). "gra[f]e!" word-final devoicing of obstruents in standard french: An acoustic study based on large corpora. In *Interspeech*, pages 1726–1730.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: A review. *Int. J. Speech Technol.*, 15(2):99–117, jun.

Labov, W. and Rosenfelder, I. (2011). New tools and methods for very large scale measurements of very large corpora. In *New Tools and Methods for Very-Large-Scale Phonetics Research Workshop*.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89:30–65.

Labov, W. (1963). The social motivation of a sound change. *WORD*, 19(3):273–309.

Lamel, L., Gauvain, J., and Canseco-Rodriguez, L. (2004). Speaker diarization from speech transcripts. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2021). A review of speaker diarization: Recent advances with deep learning.

Pellissier-Tanon, T., Weikum, G., and Suchanek, F. M. (2020). YAGO 4: A Reason-able Knowledge Base. In *Extended Semantic Web Conference*.

Priva, U. C. and Gleason, U. C. E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language*, 96:413 – 448.

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley, CA.

Tian, L., Lai, C., and Moore, J. (2015). Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations. In Khiet Truong, et al., editors, *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech 2015*, pages 39–41, United States of America. IEEE, Institute of Electrical and Electronics Engineers. Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech 2015 ; Conference date: 14-04-2015 Through 15-04-2015.

Tranter, S. and Reynolds, D. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565.

Vasilescu, I. and Adda-Decker, M. (2006). Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech. In *Interspeech*.

Vasilescu, I., Hernandez, N., Vieru, B., and Lamel, L. (2018). Exploring temporal reduction in dialectal spanish: A large-scale study of lenition of voiced stops and coda-s. In *Proc. Interspeech 2018*, pages 2728–2732.

Vasilescu, I., Wu, Y., Jatteau, A., Adda-Decker, M., and Lamel, L. (2020). Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *Traitement Automatique des Langues*, 61(1).

Weikum, G., Dong, L., Razniewski, S., and Suchanek, F. M. (2021). Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. In *Foundations and Trends in Databases*.