spINAch: A Diachronic Corpus of French Broadcast Speech Controlled for Speakers' Age and Gender

Simon Devauchelle*†, David Doukhan†, Rémi Uro‡, Lucas Ondel Yang*, Valentin Pelloin†, Olympia Imbert-Brégégère†, Véronique Lefort†, Kévin Picard†, Emeline Seignobos†, Albert Rilliard*

*Université Paris Saclay, CNRS, LISN - Orsay, France;
†Institut National de l'Audiovisuel - Paris, France;
‡LIASD, Université Paris 8 - Saint-Denis, France.

simon.devauchelle@universite-paris-saclay.fr, r.uro@iut.univ-paris8.fr, lucas.ondel@cnrs.fr,
{ddoukhan,vpelloin,olimbertbregegere,vlefort,kpicard,eseignobos}@ina.fr, albert.rilliard@lisn.fr

Abstract

We present *spINAch*, a large diachronic corpus of French speech from radio and television archives, balanced by speakers' gender, age (20-95 years old), and spanning 60 years from 1955 to 2015. The dataset includes over 320 hours of recordings from more than two thousand speakers. The methodology for building the corpus is described, focusing on the quality of collected samples in acoustic terms. The data were automatically transcribed and phonetically aligned to allow studies at a phonemic level. More than 3 million oral vowels have been analyzed to propose their fundamental frequency and formants. The corpus, available to the community for research purposes, is valuable for describing the evolution of Parisian French through the representation of gender and age. The presented analyses also demonstrate that the diachronic nature of the corpus allows the observation of various phonetic phenomena, such as the evolution of voice pitch over time (which does not differ by gender in our data) and the neutralization of the /a/-/ɑ/ opposition in Parisian French during this period.

Keywords: Speech Corpus, Diachrony, Broadcast News, Parisian French, Gender and Age Bias evaluation

1. Introduction

Diachronic changes in speech may be studied longitudinally, focusing on a specific speaker who may represent a given population category, as in Harrington et al. (2000). Other longitudinal studies are more focused on individual changes in relation to specific and documented life events (Riverin-Coutlée and Harrington, 2022). Cross-sectional corpora, like the one described by Stuart-Smith (2020), allow for investigating changes at the population level. Stuart-Smith (2020) stratified their speakers according to social gender and age (middle-aged vs. younger) with two dates of recordings, allowing them to study language changes over four levels of date of birth, and introducing a distinction between real- and apparent-time differences. Real-time differences are found with the two dates of recording, and apparent-time differences are obtained by factoring in these recording dates with the actual ages of the speaker to consider the dates of birth. Studying speech variation across time raises the notable challenge of finding old recordings that can be compared to more recent ones. The scarcity of such resources explains why some studies compare only two groups of speakers – for example, one from the 1940s and one from the 1990s in Pemberton et al. (1998). Another important question when selecting a speech dataset is related to the speech style(s) it represents, with read vs. spontaneous speech being known to induce differences at various levels (Hollien et al., 1997).

The way we speak and its evolution over time is influenced by many factors, including contacts between populations (Mufwene, 2007), or the identification of social groups to shared reference from media outlets that can shape identity display (Stuart-Smith, 2006; Vigouroux, 2015). Gathering resources capable of some generalization over a population requires a large and complex dataset. While corpora such as $VoxCeleb$ (Nagrani et al., 2017) or $CommonVoice$ (Ardila et al., 2020) feature thousands of speakers and very large acoustic datasets (over 100k hours), they are mostly synchronic resources and thus not well suited for studying language evolution. Large resources in terms of speaker diversity are a rare feature within diachronic corpora described in the literature (e.g., Zou et al., 2012, is a relatively large resource, but features very few speakers). Typical diachronic datasets feature dozens of speakers (e.g., Pemberton et al., 1998; Stuart-Smith, 2020), and are generally based on read speech, with notable exceptions (Hollien et al., 1994; Barras et al., 2002). Two diachronic datasets based on broadcast archives were recently described for French (Suire and Barkat-Defradas, 2020; Uro et al., 2022), and feature hundreds of speakers, with some gen-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5805–5820
11–16 May 2026.

der balance, but have not been made available to the community due to authorship considerations, copyright restrictions, privacy concerns, etc.

In this paper, we present and analyze a new large-scale cross-sectional corpus of French, *spINAch*, which is made freely available to the research community. The complete corpus is freely available for research purposes at https://www.ina.fr/institut-national-audiovisuel/research/dataset-project#spINAch. The acoustics estimates presented in Section 2.5 are directly available at https://doi.org/10.5281/zenodo.18714702. This corpus comprises audio recordings of more than 2,000 speakers, recorded over a sixty-year-long time span (from the 1950s to the 2010s). The data is composed of excerpts from French radio and television archives from the *Institut National de l'Audiovisuel* (INA). Archivists prepared a list of potential speakers participating in broadcast shows (focusing on interviews and talk shows) in order to target known individuals. This allowed a stratification in terms of speakers' Age (between 20 and 95 years old) and Gender, for seven time Periods (a 10-year time span was selected between 1955 and 2015). The recordings, totaling more than 320 hours of speech, were automatically transcribed and forced aligned in order to allow the extraction of acoustic analyses (formants and fundamental frequency, f_o). This dataset (including the audio recordings, their automatic and manual transcriptions, acoustic analyses, with anonymous demographic information about the speakers) is made available to the research community. We present in section 2 the methods used to gather and analyze this large dataset, with details on its composition, the acoustic measurements made, and quality evaluations. Section 3 proposes some preliminary diachronic analysis of the changes that can be observed across this time span for French spoken in national media outlets.

2. Corpus description

The corpus was collected in two iterations (the first being described in Uro et al., 2022) with an identical methodology, except for improvements of state-of-the-art diarization and music detection methods. An evaluation of the extraction methods of the two phases applied on a subsample is given in section 2.6.2 to ascertain that the resulting acoustic segments propose comparable linguistic information, if obtained through different processes.

2.1. Archive selection

An essential step in building a gender- and age-balanced cross-sectional corpus over sixty years was to spot in the archives' metadata potential target speakers that match the Age and Gender criteria: female and male speakers spread across four age groups (20-34, 35-49, 50-64, and over 64 years old), in equivalent number at seven time Periods separated by 10 year time-steps (1955-1956, 1965-1956, 1975-1976, 1985-1986, 1995-1996, 2005-2006, 2015-2016). A target of 30 speakers per Age, Gender, and Period category was set, without duplicates across categories. The construction of such a balanced corpus was only possible thanks to the expertise of INA's archivists, who parsed the television and radio databases to identify potential target speakers. For each period, they selected media with reasonable acoustic quality, such as studio-recorded talk shows free of background noise featuring interactive conversations, and verified that participants had enough speaking time. Achieving this requires archivists to review or listen to the collections. By cross-referencing the speaker's birth date with the date of the first broadcast, they estimated the participant's age. The compilation of this corpus involved a back-and-forth process with the archivists' identification work, which helped us fill in the missing speaker categories. Based on INA's documentation databases, archivists identified about 10,000 individuals who matched these characteristics. Difficulties in completing some profiles could not always be overcome, especially for women, and for younger or older persons, from the earliest periods (see Table 2). This bias of female representation in the media is known and well documented (Coulomb-Gully, 2011; Doukhan et al., 2018b).

2.2. Sound-Track extraction

Automated signal-processing routines were applied to obtain single-track WAV files sampled at 16 kHz from archives, and to discard recordings having undesirable properties. A first decompression step was performed using `ffmpeg`, leading to up to 5 uncompressed tracks (stereo, mono, or Dolby) from heterogeneous archives encoded with various codecs. Audio tracks were inspected for the presence of speaking clock: a method used from 1970 to 1990 to embed time-code information in archives using a dedicated audio track (Vallet and Carrive, 2014). The speaking clock was spotted using a 1000 Hz beep detector, along with hard-coded rules describing characteristics of its temporal patterns, leading to the exclusion of the corresponding audio tracks. A signal bandwidth estimator, based on the cumulative sum of the long-term spectrogram, was used to discard recordings with bandwidth below 8 kHz, often corresponding to undesirable archive encoding or transcoding strategies that may result in biased acoustic parameter extraction. Lastly, we used autocorrelation to detect time delays between

the remaining audio tracks. When a time delay was detected, we kept only the first track; otherwise, we mixed the remaining tracks.

2.3. Manual speaker identification

The next step, and the more time-consuming one, was to identify whether and when each targeted speaker actually speaks within the raw audio archives. This was a manual process, supported by the voice activity detection (VAD) and the diarization of each archive. Speaker identification was realized by four authors of this study, resulting in a total involvement of about 40 days. Preprocessed audiovisual archives were presented to annotators using ELAN (Sloetjes and Wittenburg, 2008), displaying the cluster identifiers obtained during the diarization and cleaning process (see section 2.4), synchronized with the archives' audio and video tracks. Annotators were provided with a shared spreadsheet containing a list of archive identifiers and target speakers. The list was enriched with details about the target speaker (gender, age, occupation) to support the identification process. Annotators reported the target speaker's cluster identifier in the spreadsheet.

Several criteria were defined along the identification process to reject speakers having undesirable properties, resulting in a manual rejection rate of about 11%: bad acoustic quality of speaker utterances (telephone speech, outdoor recordings, large amounts of background noise or music, strong audio effects), diarization under-segmentation errors resulting in several speakers sharing the same cluster identifier, use of foreign language or strong non-French speaking accent, dubbed speaker, homonym speaker having different age and occupation than the target, retrospective show broadcasting the voice of the target speaker over several decades resulting in incorrect speaker age estimation, etc.

2.4. Data extraction, cleaning, and transcription

Once a target speaker is identified, all segments obtained from *pyannote* [v3.1] (Bredin, 2023) diarization corresponding to their voice are extracted, excluding overlapping voice segments. The audio excerpts are submitted to a cleaning procedure because the primary objective of the corpus is studying speech's acoustic characteristics – a process sensitive to the presence of noise or background music. On top of *pyannote* predictions, we also applied *InaSpeechSegmenter* (ISS v0.8; Doukhan et al., 2018a) as a voice activity detector and merged its outputs with those from the diarization in order to better identify spoken segments. Outputs from *pyannote* overlapping (even marginally) non-speech events detected by ISS were discarded. To avoid any acoustic bias from telephone-quality speech segments, we use the *LIUM SpkDiarization* [v8.4.1] (Meignier and Merlin, 2010) to detect and remove them. Some cleaning was already done when applying *pyannote* and ISS, as it removes what is considered noise or music to keep only speech, but background music is still possible. We use a music detection model and apply a threshold of 0.8 to the ratio of the detected duration to the segment's total duration. The music segmentation model[1] described in Pelloin et al. (2026) uses *music2vec* (Li et al., 2022) embeddings and classifies each frame. It obtains a frame-level F1-Score of 89.7% on *OpenBMAT* (Meléndez-Catalán et al., 2019) and 92.0% on *Seyerlehner* (Seyerlehner et al., 2007), two music detection datasets of TV broadcast content. At the end of the cleaning process, after removing all parts containing noise, music, or overlapping speech, we obtained the diarized speech segments for each target speaker.

These speech excerpts were then fed to *Whisper* [large-v3] (Radford et al., 2022) in order to obtain a lexical transcription. This transcription was used to force-align its phonetic transcription to the speech signal using the *Montreal Forced Aligner* (MFA, version 3.0; McAuliffe et al., 2017). An evaluation of the transcription accuracy, both in terms of word error rate and of phone error rate, is given in section 2.6.1. Without applying any cleaning (i.e., using raw segments from *pyannote* and ISS), the total number of phones from the MFA output exceeded 4.6M oral vowels. Vowels with predicted duration over 200 milliseconds were removed (about 5.1%). After removing unvoiced vowels using the method described in the next section 2.5, the music detection model reduces again this vowel set by 8.2%. This leads to a set of 3,016,134 million vowels available in the clean version of the corpus.

2.5. Acoustic measurements

Several acoustic parameters have been extracted and are provided with the corpus to support phonetic analyses of the speech productions it contains. First, the speech's f_o was estimated for each segment with a 10 ms time step, using two different pitch detection algorithms for robustness (Vaysse et al., 2022): the autocorrelation algorithm implemented in *Praat* (Boersma, 1993; Boersma and Weenink, 2025), and the *REAPER* estimator (Talkin, 2015). Frames that any of the two algorithms annotated as unvoiced were deemed unvoiced, and frames where they differ by more than a 20% gross-error difference in their f_o prediction

[1] https://hf.co/ina-foss/ssl-music-detection-music2vec

were also marked as unvoiced – because possibly unreliable. For the remaining frames (about 79% of the total number of frames), the value estimated by $Praat$ was retained.

Then, the first five formants were estimated along the signal using *Praat*'s implementation of the Burg algorithm, with the same 10 ms time step, and adapting the ceiling parameter for each speaker and vowel category following the strategy proposed by Escudero et al. (2009). The strategy consists of estimating formants for a set of ceilings, and using the one that minimizes the formants' variance for a given vowel category and a specific speaker. In our case, we used a set of twenty ceilings above and below the reference ceiling recommended for female and male speakers in $Praat$ documentation (respectively 5.5 kHz and 5 kHz), spaced by steps of approximately 50 Hz (for details, see *Praat*'s documentation[2]) above or below the reference ceiling. The best ceiling that was kept minimizes the sum of variances observed for the first three formants of all the vowels of a given category for a specific speaker. We used the first three formants, in place of the first two in Escudero et al. (2009) because the third formant is relevant for rounding, which is an important feature of the French vocalic system (e.g., Ménard et al., 2009).

For each formant, the median of all values observed along the middle third of each vowel was kept. Formants are expressed in Hertz and converted to a Bark scale using the equation in Traunmüller (1990). For f_o, the median of all valid values observed along the vowel was considered. The f_o values are expressed in Hertz and in semitones (relative to 1Hz).

2.6. Quality evaluations

2.6.1. Evaluation of automatic transcription and forced alignment

A crucial aspect of this corpus construction lies in its fully automatic transcription and subsequent phonetic alignment, which enable the study of the acoustic characteristics of vowels over time. To evaluate this process, one hour of speech was randomly sampled from the total corpus, representing 245 speakers spread across the seven periods. These samples were manually transcribed by four L1 French speakers to reflect their full content of speech. These human-made transcriptions were then submitted to the same MFA-based forced-alignment to obtain a phonetic version. Primary concerns related to the transcription quality were related to the possibility of $Whisper$ adding lexical items that were not actually pronounced, a risk that may be increased in older recordings – acoustically uncommon – and therefore less likely to have been included in $Whisper$'s training corpus. The automatic transcription quality is evaluated using word error rate (WER) and phone error rate (PER).

After the text normalization, the WER is 11.7 using *Whisper* large-v3. This score is one point higher than that reported for *CommonVoice* 15 (Ardila et al., 2020) on the French subset. The phone-level evaluation is a crucial component for the acoustic analysis of the vowels detailed in section 3, given the nature of our dataset (recorded speech from interviews in which speakers may repeat themselves and exhibit disfluencies or hesitations). The results indicate that the PER (substitutions, deletions, and insertions summed and divided by the total number of phones) reached 7.74, with a phone-level precision of 93.7%, discarding the risk of major insertions. When focusing only on vowels, the PER degraded by about three points, reaching 10.26. The most common errors made by *Whisper* were deletions. They represent 64.39% of phone-level errors. On the evaluation subset, 5.45% of vowels were deleted during the automatic alignment process, compared to the manual transcriptions. The first three most deleted vowels are the /œ/ (58%), /ø/ (22%) followed by the /ə/ (9%). Hesitations, expressed in French by words like "heu" or "euh" (phonetized by /œ/ or /ø/), represent 28% of the deleted vowels. Then, the next 5% are derived from the conjunction "et" (/e/), followed by the words "de" and "que" (4.9% and 3.5%) – these words are likely used in repetitions and hesitations and were only transcribed once by $Whisper$. Among all vowel classes, the vowel /ə/ is the most frequently inserted and is mostly derived from the negation word "ne" (23% of /ə/ insertions), which is generally not produced in French (Abeillé and Godard, 2021). $Whisper$ tends to add it, resulting in more formal lexical predictions.

2.6.2. Comparison of extraction methods

Given the size of the corpus, it was produced in two phases (in 2021 and 2025), separated by four years. The processing for the first version prioritized speech quality when identifying speakers, at the risk of missing potential targets and thereby lowering recall. In addition, more up-to-date software has since been released, so the automated processing of archives differs slightly between the two phases, mainly in the algorithms employed (for details on the first version, see Uro et al., 2022). This notably explains why there are more speakers in the years 1965, 1985, and 2005, as the newer version was more efficient.

In order to evaluate if these two processing methods introduced a qualitative difference in terms of the distribution of the acoustic characteristics of phonemes, an evaluation was necessary, as both

[2] https://www.fon.hum.uva.nl/praat/manual/FormantPath.html

processing methods end up with comparable, but not identical, speech segments out of the same original recording. The main differences introduced by the two processings are (i) different diarization of the archive, and (ii) possibly different transcription and alignment, as the cleaning process (removing background music, noises, etc) was performed with different algorithms. Our evaluation thus focuses on verifying that the phonetic characteristics of a given speaker (in our case, the distribution of their formants for oral vowels) are comparable across extraction methods. That is, the /a/s (and other vowels) extracted with the two methods shall have comparable acoustic characteristics for the same speaker. The question here is to verify, for a subset of speakers from the first phase (Uro et al., 2022), if the formants measured on vowels detected using the new processing are equivalent to those measured on vowels detected by the initial method. Even if the two methods give slightly different sets of vowels (in terms of the number of vowels observed for a given speaker), on a sufficiently long dataset, the formants of each vowel category shall have comparable distributions within each speaker, across processing methods, if one thinks the extraction method in itself does not bias the phonetic content.

To evaluate the possibility of a bias linked to the processing chain, eight speakers were randomly drawn from the eight categories of the Periods (1955-1956, 1975-1976, 1995-1996, 2015-2016) and Gender (Female, Male) processed at the first phase, for a total of 64 individuals (8 speakers from 8 categories). The new extraction method was applied to the 64 archives of the initial corpus to extract and transcribe the target speakers' voices a second time. From these two sets of extracted vowel segments, thanks to the initial (therefore Method [1]) and newer software (therefore Method [2]), the same phoneticization pipeline and the same f_o and formant detection process (see section 2.5) were applied. An equivalent number of vowels, approximately 90,000, is obtained by both Methods. Details regarding the number of occurrences obtained for each oral vowel are presented in Table 1.

A linear mixed model was then fit to the value of each of the first three formants, with the extraction Method ([1]/[2]) as the main predictor, controlled for Gender and Vowel category, and using Periods and Speakers (nested within gender and period) as random variables. This is formalized in formula 1, that follows R's *lme4* syntax (R Core Team, 2024; Bates et al., 2015), and where z_i are the standardized values of either of the first three formants in Bark, M is the *Method* ([1] or [2]) used for processing the archives, G being the *Gender* of a given speaker, V the *Vowel* class (12 levels, see Table 1), P the *Period* (4 levels), and S the index of the *Speaker* producing a given phone:

$$z_i \sim M + G + V + (1|P/G/S) \quad (1)$$

The tables summarizing the factors (fixed and random) of the three regression models (one for each formant) are given in Appendix A (Section 8.1).

The models showed there were no significant differences introduced in the distribution of these formants by the extraction *Method* (for F_1: $\chi^2_{(1)} = 0.084, p = 0.772$; for F_2: $\chi^2_{(1)} = 0.96, p = 0.327$; for F_3: $\chi^2_{(1)} = 0.6147, p = 0.433$), while the other parameters had, obviously, major effects on the formants values — primarily the Vowel categories, but also the gender and a large inter-speaker variability (cf. Tables 4 & 5). This result confirms that the extraction method employed here is robust to the evolution of diarization, speech-to-text, forced alignment, and music detection algorithms. This is fortunate given the surge of use of comparable frameworks in phonetic studies e.g., Ballier and Méli, 2024; Coats, 2025; Christodoulidou et al., 2025. We believe this comparison of the two methods supports that the extracted vowels from all Periods of the corpus shall give coherent information, regardless of how they were segmented. We consider the data extracted using these two methods to be comparable in terms of phonetic distribution, which justifies prioritizing improvements to the processing pipeline by incorporating state-of-the-art advances in the second phase.

Phone	Method [1]	Method [2]
[i]	13,338	14,194
[e]	13,132	13,814
[ɛ]	13,483	14,317
[a]	18,140	19,313
[ɑ]	769	828
[ɔ]	5,930	6,332
[o]	2,960	3,172
[u]	4,146	4,511
[y]	4,721	4,961
[ø]	2,652	2,799
[ə]	8,705	9,167
[œ]	1,165	1,257
Total	89,141	94,665

Table 1: Number of occurrences for each vowel category for the two segment sets obtained by the two extraction Methods on the same 64 speakers.

2.7. Summary of corpus features

Table 2 presents the distribution of the data in terms of speaker count and recording duration across the seven periods, with columns presenting gender

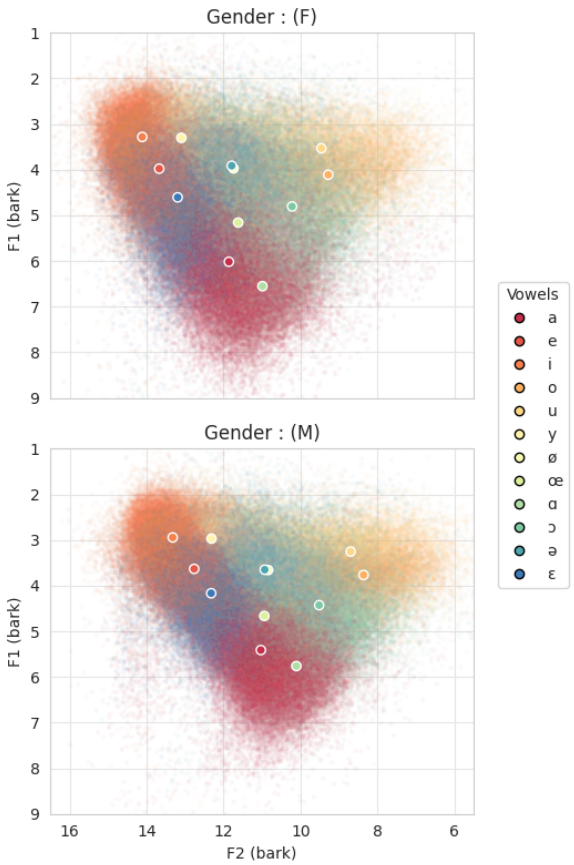

Figure 1: Plot of vowels' F_1 and F_2 (in Bark) values obtained from a sample of 240 speakers in the *spINAch* corpus, balanced in terms of period and gender. Median values by vowel category (encircled in white) and for each vowel (color points) are shown for both genders.

and rows presenting age categories. A sample of oral vowel formants estimated using the procedure described above is shown in Figure 1, resulting in well-known vocalic triangles. Cleaned oral vowels' distribution is reported in the table in Appendix B 6 [3]. Details about the whole phone dataset and its categories (including nasal vowels and consonants) are given in Appendix C (7).

3. Data Analysis

This section presents a series of preliminary analyses of the corpus data, highlighting its versatility and showing how diachronic data reveal new insights into language description.

[3]Acoustic estimates from the aligned oral vowels are hosted at *Zenodo* on this url https://doi.org/10.5281/zenodo.18714702.

3.1. Does voice pitch change with time?

Our voices are important parts of our personalities, as they index aspects of each individual, from their gender to their health (Laver, 1968; Eckert, 2019; Podesva and Callier, 2015). An important component of voice quality is related to its perceived pitch: a lower or higher voice being a central component of gender perception through vocal cues (Leung et al., 2018; Simpson and Weirich, 2020), but also to a series of interactive functions related to Ohala's *Frequency Code* (Ohala, 1994). The construction of an individual's voice pitch is mediated by cultural components, notably those related to culturally variable representations of gender (van Bezooijen, 1995; Ohara, 2001). As cultural values evolve with time, the role of women in society has undergone important changes since the end of the Second World War. As voice pitch is (negatively) related to social power in interaction (Ohala, 1994; Spencer-Oatey, 1996; Goudbeek and Scherer, 2010), one may hypothesize that female voice pitch decreases over time. A pitch decrease was claimed by some publications (e.g., Berg et al., 2017, albeit without diachronic evidences), but it is controversial in the literature (e.g., Hollien et al., 1997; Pemberton et al., 1998).

The corpus presented here provides insight into potential changes in vocal characteristics for both genders among the subset of French personalities who are invited to radio and television shows. Thanks to the acoustics measurements detailed in section 2.5, we have a set of f_o measurements, one for each oral vowel annotated as voiced (when both pitch detection algorithms returned coherent measures). There are 3,016,134 f_o measurements, which are the median values observed on the voiced frames of each corresponding vowel. The vowels come from 2,109 speakers, female or male, distributed uniquely across seven Periods as shown in Table 2. The speakers' ages range from 20 to 95.

3.1.1. Methods

We tried to evaluate a potential evolution of f_o across Periods, possibly linked to the speaker's gender, as a potentially different effect could be expected for female and male speakers, but controlling for changes linked to the speaker's age (Berg et al., 2017; Gisladottir et al., 2023). Linear mixed-effect regression models were fitted (following Gries, 2021; Crawley, 2013, and using R's *lmer* library; R Core Team, 2024; Bates et al., 2015) to the median f_o values of each vowel, expressed in semitones and standardized to avoid numerical problems. The models took as predictors three fixed factors: the speaker's *Age* (in years, centered around 50 years and divided by 30), their

	1955/6		1965/6		1975/6		1985/6		1995/6		2005/6		2015/6		Total
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	
20-34	16 0.32	34 1.42	40 2.85	29 9.81	16 0.44	16 1.41	29 3.96	52 7.88	28 3.59	27 2.94	52 9.97	53 10.59	25 3.71	30 3.76	447 62.65
35-49	21 0.85	70 2.94	37 3.41	37 8.07	24 0.92	37 1.87	48 8.15	63 12.24	33 4.32	44 6.43	57 11.28	58 11.22	36 4.02	53 7.47	618 83.19
50-64	18 0.62	52 2.78	33 4.91	34 11.25	27 2.76	41 2.33	45 6.69	60 13.98	29 5.96	47 6.94	56 11.73	62 12.02	26 3.22	50 5.53	580 90.72
≥65	21 1.82	17 1.11	29 3.49	32 9.25	18 2.28	26 4.55	21 3.97	62 15.21	31 7.82	37 6.13	48 13.88	59 13.20	33 5.14	30 5.55	464 93.4
Total	76 3.61	173 8.25	139 14.66	132 38.38	85 6.4	120 10.16	143 22.77	237 49.31	121 21.69	155 22.44	213 46.86	232 47.03	120 16.09	163 22.31	2109 329.96

Table 2: Number of speakers (on top of the cell) and duration of recordings in hours (at the bottom of the cell) in each category of *Age* (rows) by *Period* and *Gender* (columns) in the *spINAch* corpus.

Gender (two levels: "F" or "M"), and the *Period* of time corresponding to their recording (seven levels: 1955-56, 1965-66, 1975-76, 1985-86, 1995-96, 2005-06, 2015-16). The models also controlled for variation associated with two random factors: the *Speaker* (2,109 were considered) and the *Vowel* category (12 levels). As vowel production is speaker-specific, the factor *Vowel* was nested in *Speaker*, itself nested in *Gender* and in *Period*. The *spINAch* corpus is cross-sectional, so each speaker belongs to a specific *Gender* and a specific *Period*. *Gender* was nested in *Period* as gender representation in society may evolve across time. Double interactions between each pair of the three random factors were also kept in the model, while the three-way interaction was discarded during a model simplification process (Crawley, 2013), as not significant. The model considered here follows equation 2 (based on *lmer*'s syntax), where zf_o stands for the standardized f_o, A for *Age*, G for *Gender*, P for *Period*, S for *Speaker*, and V for *Vowel* category. The square in equation 2 encodes the two-way interactions among the A, G, and P factors. The ANOVA table for this model is presented in Table 3.

$$zf_o \sim (A + G + P)^2 + (1|P/G/S/V) \qquad (2)$$

Factors	χ^2	**Df**	$Pr(>\chi^2)$
A	6.9508	1	0.0083780 **
G	14.9094	1	0.0001128 ***
P	0.0246	6	0.9999997
A:G	57.6536	1	3.126e-14 ***
A:P	22.8561	6	0.0008461 ***
G:P	0.0593	6	0.9999958

Table 3: Analysis of Deviance Table (Type II Wald χ^2 tests) obtained for the model fitted to the f_o measurements; the A, G, and P factors correspond to the *Age*, *Gender*, and *Period* factors described in the text.

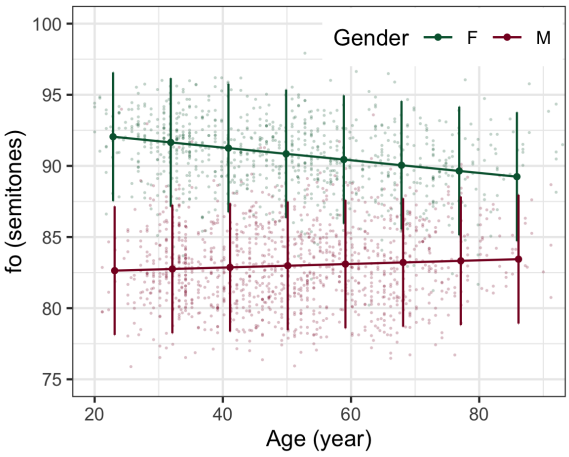

Figure 2: f_o values (y-axis) estimated by the model for Age (x-axis) across Genders (colors), plotted over points that represent each speaker's mean f_o.

3.1.2. Results

Results showed that the speaker's *Gender* has an (expected) major effect on f_o values, with the mean f_o difference of 7.7 semitones across genders, while, as main factors, *Age* showed only a limited effect, and *Period* had none. *Age* is nonetheless, and as described in the literature, fundamental to explain f_o changes, but conditioning on the speaker's *Gender* (cf. the A:G line). Female voices have their pitch lowered with *Age* (a mean lowering of about 2.8 semitones in 60 years), while the reverse tendency is observed for male voices (a more modest rise of about 0.8 semitones in 60 years; see Figure 2).

As for *Period*, this factor has a limited impact on f_o values. Its effect as a main factor is not significant, but it exhibits a significant interaction with *Age*. This interaction (not plotted for space reasons) is linked with the evolution of f_o across speakers of different *Ages* at a given *Period*. This effect (controlled for Gender, as discussed above) consists of an increase of f_o with age for the 1955-56 and 1965-66 *Periods*, while the tendency is reversed starting with the 1985-86 *Period*, when pitch tends to diminish with age. This means that across older

Periods, as people get older, they tend to increase their pitch, controlling for their gender tendencies, whereas in newer periods, older individuals tend to decrease their pitch. The effect size is comparatively limited with respect to the *Age:Gender* interaction, but is still interesting, as potentially linked to varying social conditions in the population. A possible explanation may relate to improved life expectancy in more recent periods, linked to better health conditions.

3.2. Evolution of French vocalic system

One documented change in Parisian French during the twentieth century was an evolution of its vocalic system. Among other variations, the opposition between /a/ and /ɑ/ is not productive in the main variant of French spoken in France (for an overview of French diachrony, see Abeillé and Godard, 2021). This vowel shift has already been observed on a series of speech corpora from different periods over a century (Cęcelewski et al., 2024), but on a dataset featuring only male speakers (because of the difficulty in finding female speakers in the archives). We focus here on these two vowels (/a/ an /ɑ/), and not on the complete vocalic system of French, in order to give a simple example of diachronic changes and the importance of controlling for gender – one of the key features of the *spINAch* corpus.

3.2.1. Methods

These two vowels were annotated by the MFA algorithm, which distinguishes between the two possible variants in its phonetic dictionary based on its acoustic model for French. In the corpus, we get a sample of 623,003 and 28,202 occurrences of /a/ and /ɑ/ respectively, which first shows that they are clearly used unequally in French. We fit two linear mixed-effect regression models, one on each of the first two formants (F_1 and F_2) expressed in bark and standardized. For F_1, after a simplification procedure, the model that was kept to describe the variation of this formant, controlling for variations linked to individual *Speaker* (S), *Gender* (G), *Period* (P), *Age* (A), and *Vowel* category (V) corresponds to the Equation 3.

$$zF_1 \sim (A+G+P+V)^2 + A{:}P{:}V + (1|P/G/S/V) \quad (3)$$

For F_2, a complex interaction between the speaker's age and the period was found. Following (Stuart-Smith, 2006), we thus estimated each speaker's birth date, and evaluated the same model but using *apparent time* in place of *Period* and *Age*. After simplification, the model corresponds to Equation 4, where AT refers to the *Apparent Time* when a given speaker starts learning their phonological system.

$$zF_2 \sim AT * (G+V) + (1|G/S/V) \quad (4)$$

3.2.2. Results

The model fitted to the first formant (see Table 8 in Appendix D) expects effects for the speaker's gender ($\chi^2_{(1)} = 18.9, p < 1.0\text{e}{-4}$) and for the vowel category ($\chi^2_{(1)} = 2618.8, p < 1.0\text{e}{-5}$). It also shows interactions between the Gender with the vowel category ($\chi^2_{(1)} = 30.6, p < 1.0\text{e}{-5}$) and between the time Period with the vowel category ($\chi^2_{(6)} = 15.9, p < 0.05$) and a triple interaction between Period with vowel and the speaker's age ($\chi^2_{(6)} = 13.8, p < 0.05$). The largest diachronic changes along F_1 are dependent on the vowel category, and consist of a F_1 decrease from 1955 to 1985, but the differences across the two vowel categories were mostly kept unchanged over time. An increase of the first formant is known to follow the jaw aperture (Erickson, 2002). Apart from F_1 obvious relation to vowel aperture, increased jaw aperture is also related to vocal effort (e.g., Rilliard et al., 2018). The observed diachronic changes in vocalic characteristics along this dimension may thus be linked to an evolution in recording practices in media outlets over this period, with a more declamatory style and microphones placed farther from the mouth in more ancient recordings (see e.g. Boula de Mareüil et al., 2012; Devauchelle et al., 2024). In terms of the two vocalic categories, the differences across the /a/ and /ɑ/ along F_1 do not change with time: it seems the vowels tagged as/ɑ/ had always larger jaw openings.

For the model fitted on the second formant (see Table 9 in Appendix D), we observed an expected effect of vowel category ($\chi^2_{(1)} = 5377.1, p < 1.0\text{e}{-4}$) as both phones differ along the antero-posterior dimension. There is also an interaction between the Apparent Time and the Vowel category ($\chi^2_{(1)} = 204.4, p < 1.0\text{e}{-4}$) that is represented in Figure 3. It shows that the two vowels' F_2 values converge over the twentieth century, so they are statistically comparable for people born around the middle of the century. We also observe a significant interaction between Apparent Time and the speaker's gender ($\chi^2_{(1)} = 11.6, p < 1.0\text{e}{-3}$), that is independent of vowel category. Because of a longer vocal tract, males tend to have lower F_2 values than females, but male speakers born more recently tend to amplify this difference, once controlled for vowel: they display a lowering F_2 trend over Apparent Time that is not observed for female speakers (whose mean F_2 values are almost flat with time).

This continued lowering of F_2 for these two vowel categories along the xx$^{\text{th}}$ century is visible in the analysis proposed by Cęcelewski et al. 2024 – who

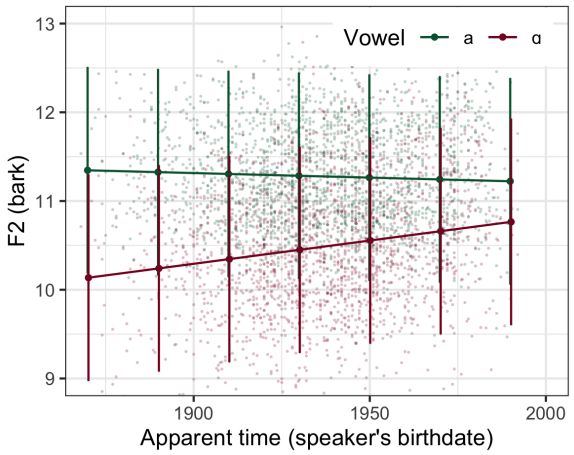

Figure 3: F_2 values (y-axis) estimated by the model for Apparent Time (x-axis) for both vowel categories (/a/ and /ɑ/; colors), plotted over points that represent each speaker's mean F_2.

worked on male-only datasets. The interpretation of a gender-specific tendency supported by our dataset may thus differ from theirs. Lowering F_2 is a correlate of a posteriorized articulation, an articulatory setting that tends to lower a voice's pitch, which may be a marker of masculinity (van Bezooijen, 1995) – while the contrary, an anteriorization strategy, was linked by other works to a phonostyle of seduction used by French females (Léon, 1993; Rilliard et al., 2018). So, during the first half of the century, we observe in our dataset a convergence of the two phonetic categories, with the/ɑ/ vowel progressively anteriorized until it mixes with /a/; this evolution was not gender-dependent. Once the vocalic system has only one category of open vowel, it has more space for sociophonetic variations – and during the second half of the century, Parisian French-speaking males appear to use this available space to reinforce their voice's masculinity by posteriorizing their articulation of the now single open vowel. Such a gender-specific tendency was not observed in our diachronic data for female speakers (who kept a more anterior articulation, compared to males). Working on a subset of the *spINAch* corpus, Elie et al. (2024) had already observed gender-exaggerating articulatory tendencies in female and male speakers, who are using lips and larynx positions to respectively shorten or lengthen their vocal tract. Meanwhile, in Elie et al. (2024)'s work, the tendency did not evolve diachronically. Here, the extra vocalic space left by the fusion of /a/ and /ɑ/ appears to have been used for indexical means by male speakers.

4. Conclusion

In this paper, we present the *spINAch* corpus, featuring more than 320 hours of speech extracted from radio and television broadcasts over a 60-year period, from speakers selected to balance gender and age distributions. This dataset comprises more than two thousand speakers, represented by speech segments of varying duration, with a median speaker duration of over six minutes (390 s). The speaker's birth dates range from 1870 to 1990. The corpus, which is made available for research purposes[4], contains the audio samples, their automatic transcription, and the phonetic alignment. The speakers' identities are not disclosed, with only non-identifying demographic information retained (age at the time of recording, with a five-year precision, and gender). All audio segments of a given speaker resulting from the diarization process have been assigned a random ID to make it difficult to reconstruct the speaker's argument for potential authorship purposes. The acoustic analyses (f_o and formants) presented in this paper focused on oral vowels. More than 3 million vowels have been analyzed, and these measurements are presented in a separate file to allow interested researchers to study this aspect of the corpus directly[5].

The two rapid analyses of this dataset presented in section 3, beyond their specific findings, showed that the *spINAch* corpus contains reliable data, as we were able to reproduce several known characteristics of f_o changes with age or vocalic variation in French. These data enable a variety of phonetic investigations of changes in French as it is spoken in France's national media outlets over this 60-year period. We hope our efforts to gather this dataset will enable the community to conduct more research on the diachrony of the French language.

5. Acknowledgments

This work was partly funded by ANR "Gender Equality Monitor" (GEM) grant ANR-19-CE38-0012. We are especially grateful to Pascal Flard at INA for his valuable assistance in recovering part of the archives needed for this work.

6. Bibliographical References

[4]Available at https://www.ina.fr/institut-national-audiovisuel/research/dataset-project#spINAch

[5]Available at https://doi.org/10.5281/zenodo.18714702

Anne Abeillé and Danièle Godard. 2021. *La grande grammaire du français*. Actes Sud Imprimerie nationale éditions, Arles [Paris].

Nicolas Ballier and Adrien Méli. 2024. Investigating acoustic correlates of whisper scoring for l2 speech using forced alignment with the italian component of the isle corpus. page 20–32.

Claude Barras, Alexandre Allauzen, Lori Lamel, and Jean-Luc Gauvain. 2002. Transcribing audio-video archives. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–13–I–16, Orlando, FL, USA. IEEE.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Martin Berg, Michael Fuchs, Kerstin Wirkner, Markus Loeffler, Christoph Engel, and Thomas Berger. 2017. The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors. *Journal of Voice*, 31(2):257.e13–257.e24.

Paul Boersma. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the institute of phonetic sciences*, 17:97–110.

Paul Boersma and David Weenink. 2025. Praat: doing phonetics by computer [computer program]. version 6.4.45.

Philippe Boula de Mareüil, Albert Rilliard, and Alexandre Allauzen. 2012. A diachronic study of initial stress and other prosodic features in the french news announcer style: Corpus-based measurements and perceptual experiments. *Language and Speech*, 55(2):263–293.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023*, pages 1983–1987.

Polychronia Christodoulidou, James Tanner, Jane Stuart-Smith, Michael McAuliffe, Mridhula Murali, Amy Smith, Lauren Taylor, Joanne Cleland, and Anja Kuschmann. 2025. A semi-automatic pipeline for transcribing and segmenting child speech. In *Interspeech 2025*, pages 4278–4282.

Steven Coats. 2025. *An automatic pipeline for processing streamed content: New horizons for corpus linguistics and phonetics*, page 257–274. De Gruyter.

Marlène Coulomb-Gully. 2011. Genre et médias : vers un état des lieux. *Sciences de la société*, 83:3–13.

Michael J. Crawley. 2013. *The R book*, second edition edition. Wiley, Chichester, West Sussex, UK.

Juliusz Cęcelewski, Cédric Gendrot, Martine Adda-Decker, and Philippe Boula de Mareüil. 2024. Étude en temps réel de la fusion des /a/ ~/ɑ/ en français depuis 1925. In *Actes des 35èmes Journées d'Études sur la Parole*, page 71–81, Toulouse, France. ATALA and AFPC.

Simon Devauchelle, Albert Rilliard, David Doukhan, and Lucas Ondel Yang. 2024. Variation of Perceived Voice Pitch Across Time Periods, Gender, and Age in French Media Archives. In Valentina De Iacovo, Bianca Maria De Paolis, and Daniela Mereu, editors, *The voice in the media and new technologies*, volume 12 of *Studi Associazione Italiana Scienze della Voce*, pages 47–71. Officinaventuno.

David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018a. An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE.

David Doukhan, Géraldine Poels, Zohra Rezgui, and Jean Carrive. 2018b. Describing gender equality in french audiovisual streams with a deep learning approach. *VIEW Journal of European Television History and Culture*, 7(14):103–122.

Penelope Eckert. 2019. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776.

Benjamin Elie, David Doukhan, Rémi Uro, Lucas Ondel-Yang, Albert Rilliard, and Simon Devauchelle. 2024. Articulatory Configurations across Genders and Periods in French Radio and TV archives. In *Interspeech 2024*, pages 3085–3089.

Donna Erickson. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, 59(2–3):134–149.

Paola Escudero, Paul Boersma, Andréia Schurt Rauber, and Ricardo A. H. Bion. 2009. A cross-dialect acoustic description of vowels: Brazilian and european portuguese. *The Journal of the Acoustical Society of America*, 126(3):1379–1393.

Rosa S. Gisladottir, Agnar Helgason, Bjarni V. Halldorsson, Hannes Helgason, Michal Borsky, Yu-Ren Chien, Jon Gudnason, Sigurjon A. Gudjonsson, Scott Moisik, Dan Dediu, Gudmar Thorleifsson, Vinicius Tragante, Mariana Bustamante, Gudrun A. Jonsdottir, Lilja Stefansdottir, Gudrun Rutsdottir, Sigurdur H. Magnusson, Marteinn Hardarson, Egil Ferkingstad, Gisli H. Halldorsson, Solvi Rognvaldsson, Astros Skuladottir, Erna V. Ivarsdottir, Gudmundur Norddahl, Gudmundur Thorgeirsson, Ingileif Jonsdottir, Magnus O. Ulfarsson, Hilma Holm, Hreinn Stefansson, Unnur Thorsteinsdottir, Daniel F. Gudbjartsson, Patrick Sulem, and Kari Stefansson. 2023. Sequence variants affecting voice pitch in humans. *Science Advances*, 9(23):eabq2969.

Martijn Goudbeek and Klaus Scherer. 2010. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–1336.

Stefan Thomas Gries. 2021. *Statistics for linguistics with R: a practical introduction*, 3rd revised edition edition. De Gruyter Mouton textbook. de Gruyter Mouton, Berlin Boston.

Jonathan Harrington, Sallyanne Palethorpe, and Catherine Watson. 2000. Monophthongal vowel changes in received pronunciation: an acoustic analysis of the queen's christmas broadcasts. *Journal of the International Phonetic Association*, 30(1–2):63–78.

Harry Hollien, Rachel Green, and Karen Massey. 1994. Longitudinal research on adolescent voice change in males. *The Journal of the Acoustical Society of America*, 96(5):2646–2654.

Harry Hollien, Patricia A. Hollien, and Gea De Jong. 1997. Effects of three parameters on speaking fundamental frequency. *The Journal of the Acoustical Society of America*, 102(5):2984–2992.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

John D. M. Laver. 1968. Voice quality and indexical information. *British Journal of Disorders of Communication*, 3(1):43–54.

Yeptain Leung, Jennifer Oates, and Siew Pang Chan. 2018. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 61(2):266–297.

Y. Li, R. Yuan, G. Zhang, Y. MA, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, E. Benetos, N. Gyenge, R. Liu, and J. Fu. 2022. Lv-49: Mapmusic2vec: A simple and effective baseline for self-supervised music audio representation learning. In *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*.

Pierre Léon. 1993. *Précis de phonostylistique. Parole et expressivité*. Nathan Université, Paris.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, page 498–502. ISCA.

Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*.

Salikoko S Mufwene. 2007. Population movements and contacts in language evolution. *Journal of language contact*, 1(1):63–92.

Lucie Ménard, Sophie Dupont, Shari R. Baum, and Jérôme Aubin. 2009. Production and perception of french vowels by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America*, 126(3):1406–1414.

John J. Ohala. 1994. *The frequency code underlies the sound-symbolic use of voice pitch*, 1 edition, page 325–347. Cambridge University Press.

Yumiko Ohara. 2001. *Finding one's voice in Japanese: A study of the pitch levels of L2 users*. DE GRUYTER MOUTON, Berlin, New York.

Valentin Pelloin, Lina Bekkali, Reda Dehak, and David Doukhan. 2026. Data selection effects on self-supervised learning of audio representations for french audiovisual broadcasts. In *Fifteenth International Conference on Language Resources and Evaluation (LREC 2026)*, Palma, Mallorca, Spain. European Language Resources Association.

Cecilia Pemberton, Paul McCormack, and Alison Russell. 1998. Have women's voices lowered across time? a cross sectional study of australian women's voices. *Journal of Voice*, 12(2):208–213.

Robert J. Podesva and Patrick Callier. 2015. Voice quality and identity. *Annual Review of Applied Linguistics*, 35:173–194.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Albert Rilliard, Christophe d'Alessandro, and Marc Evrard. 2018. Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis. *The Journal of the Acoustical Society of America*, 143(1):109–122.

Josiane Riverin-Coutlée and Jonathan Harrington. 2022. Phonetic change over the career: a case study. *Linguistics Vanguard*, 8(1):41–52.

Adrian P. Simpson and Melanie Weirich. 2020. *Phonetic Correlates of Sex, Gender and Sexual Orientation*. Oxford University Press.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-elan and iso dcr. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Helen Spencer-Oatey. 1996. Reconsidering power and distance. *Journal of Pragmatics*, 26(1):1–24.

Jane Stuart-Smith. 2006. *The Influence of the Media*, 0 edition, page 140–148. Routledge.

Jane Stuart-Smith. 2020. Changing perspectives on /s/ and gender over time in glasgow. *Linguistics Vanguard*, 6(s1):20180064.

Alexandre Suire and Melissa Barkat-Defradas. 2020. Evolution of human pitch: Preliminary analyses in the french population using ina audiovisual archives of vox pops. In *2020 IASA-FIAT/IFTA Joint Conference*.

David Talkin. 2015. Reaper: Robust epoch and pitch estimator.

Hartmut Traunmüller. 1990. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88:97–100.

Rémi Uro, David Doukhan, Albert Rilliard, Laetitia Larcher, Anissa-Claire Adgharouamane, Marie Tahon, and Antoine Laurent. 2022. A semi-automatic approach to create large gender- and age-balanced speaker corpora: Usefulness of speaker diarization & identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3271–3280, Marseille, France. European Language Resources Association.

Félicien Vallet and Jean Carrive. 2014. Quand l'horloge parlante a beaucoup à raconter sur l'évolution des techniques d'archivage audiovisuel. In *Journées d'étude sur la parole*.

Reneé van Bezooijen. 1995. Sociocultural aspects of pitch differences between japanese and dutch women. *Language and Speech*, 38(3):253–265.

Robin Vaysse, Corine Astésano, and Jérôme Farinas. 2022. Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *The Journal of the Acoustical Society of America*, 152(5):3091–3101.

Cécile B. Vigouroux. 2015. Genre, heteroglossic performances, and new identity: Stand-up comedy in modern french society. *Language in Society*, 44(2):243–272.

Yu Zou, Yan Wang, and Wei He. 2012. Diachronic contrastive analysis on read speech in broadcast news: Evidence from pitch and duration. In *2012 8th International Symposium on Chinese Spoken Language Processing*, page 291–295, Kowloon Tong, China. IEEE.

7. Language Resource References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Blai Meléndez-Catalán, Emilio Molina, and Emilia Gómez. 2019. Open broadcast media audio from tv: A dataset of tv broadcast audio with relative music loudness annotations. *Transactions of the International Society for Music Information Retrieval*, 2(1):43–51.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, page 2616–2620. ISCA.

Klaus Seyerlehner, Tim Pohle, Markus Schedl, and Gerhard Widmer. 2007. Automatic music detection in television productions. In *Proc. of the 10th International Conference on Digital Audio Effects (DAFx'07)*. SCRIME/LaBRI Bordeaux.

8. Supplementary Materials

8.1. A : ANOVA Tables

ANOVA table for models presented in section 2.6.2, fitted for the first three formants (expressed in Bark and standardized): Table 4 presents the results for the three fixed factors (*Method*, *Gender*, and *Vowel*), while Table 5 presents the effect of the random factors, obtained using single-term deletion with the *lmerTest* library (Kuznetsova et al., 2017).

8.2. B : Cleaned Vowels Summary Table

See the caption of the table 6.

8.3. C: Phone Summary Table

See the caption of the table 7.

8.4. D: ANOVA Tables

ANOVA tables for models presented in section 3.2 for the analysis of diachronic changes in formants F_1 (Table 8) and F_2 (Table 9).

Model	Factors	χ^2	Df	$Pr(>\chi^2)$
F_1	Method	0.08	1	0.772
	Gender	17.37	1	0.002 **
	Vowel	202660	11	0.000 ***
F_2	Method	0.96	1	0.327
	Gender	68.24	1	0.000 ***
	Vowel	287610	11	0.000 ***
F_3	Method	0.61	1	0.433
	Gender	49.72	1	0.000 ***
	Vowel	69210	11	0.000 ***

Table 4: Analysis of Deviance Table (Type II Wald χ^2 tests) obtained for the fixed factors of the models fitted respectively to formants F_1, F_2, and F_3 obtained on vowels segmented with two different processing *Methods*, controlled for *Gender* and *Vowel* category (see section 3).

Model	Deletion	npar	loglik	AIC	LRT	Df	$Pr(>\chi^2)$
F_1	<none>	18	-182288	364613	-	-	-
	$(1\|S{:}(G{:}P))$	17	-194498	389030	24419.6	1	0.00000 ***
	$(1\|G{:}P)$	17	-182288	364611	0.0	1	0.90799
	$(1\|P)$	17	-182290	364614	2.8	1	0.09284
F_2	<none>	18	-166592	333220	-	-	-
	$(1\|S{:}(G{:}P))$	17	-173107	346248	13030.3	1	0.00000 ***
	$(1\|G{:}P)$	17	-166592	333218	0.0	1	0.8749
	$(1\|P)$	17	-166592	333219	0.1	1	0.3931
F_3	<none>	18	-187973	375982	-	-	-
	$(1\|S{:}(G{:}P))$	17	-209152	418338	42358.0	1	0.0000 ***
	$(1\|G{:}P)$	17	-187973	375980	1.0	1	0.4542
	$(1\|P)$	17	-187973	375980	0.0	1	1.0000

Table 5: ANOVA-like table for random effects obtained by single term deletions of the *Speaker* (S), *Gender* (G), and *Period* (P) factors, for the models fitted on each formant (F_1, F_2, F_3). Values of the likelihood ratio test (LRT) are reported for each deleted term, compared to the full model (see section 3).

Phoneme			Cleaned
Vowels	Oral	[i]	443,795
		[y]	148,757
		[e]	443,329
		[ø]	89,547
		[ə]	285,534
		[ɛ]	473,529
		[œ]	36,875
		[a]	623,003
		[u]	149,453
		[ɔ]	191,682
		[o]	102,608
		[ɑ]	28,202

Table 6: Summary table of **cleaned** oral vowel returned by MFA after phonetic transcription and forced alignment of the **automatic** transcriptions ($Whisper$ [large-v3]) (see section 2).

Phoneme			Automatic	Manual
Vowels	Oral	[i]	706,217	2,047
		[y]	238,829	766
		[e]	675,676	2,118
		[ø]	127,524	435
		[ə]	512,990	1,581
		[ɛ]	710,354	2,169
		[œ]	52,340	375
		[a]	923,020	2,754
		[u]	232,540	672
		[o]	147,186	416
		[ɔ]	262,739	786
		[ɑ]	45,190	140
	Nasal	[ɑ̃]	421,404	1,267
		[ɔ̃]	256,564	770
		[ɛ̃]	184,927	551
Consonants	Plosive	[p]	462,424	1,361
		[t]	627,664	1,847
		[k]	450,392	1,350
		[b]	122,218	349
		[d]	539,279	1,738
		[g]	58,156	172
		[c]	122,142	344
		[ɟ]	4274	19
	Fricative	[f]	175,399	562
		[s]	767,359	2,358
		[ʃ]	60,638	149
		[v]	272,070	804
		[z]	115,033	332
		[ʒ]	216,210	672
		[ʁ]	912,787	2,781
	Approximant	[ɥ]	59,972	171
		[w]	124,831	419
		[j]	175,202	503
		[l]	657,357	1,895
		[ʎ]	56,563	168
	Nasal	[m]	429,348	1,279
		[n]	275,343	825
		[ɲ]	39496	120
		[ŋ]	849	3
		[mʲ]	6,990	21
	Affricates	[t͡ʃ]	800	5
		[d͡ʒ]	439	0
		[t͡s]	302	1

Table 7: Summary table of the number of phones returned by MFA after phonetic transcription and forced alignment of the **automatic** ($Whisper$ [large-v3]) and **manual** transcriptions, without any cleaning at this stage (see section 2). Note that MFA uses a fine-grained phonetic transcription that includes some phonetic phenomena such as palatalization and non-standard French transcriptions; we kept its default choices.

Factors	χ^2	**Df**	$Pr(>\chi^2)$
A	0.3767	1	0.53938
G	18.9247	1	1.360e-05 ***
P	2.1955	6	0.90087
V	2618.7972	1	< 2.2e-16 ***
$A:G$	6.5716	1	0.01036 *
$A:P$	9.4499	6	0.14981
$A:V$	0.3280	1	0.56685
$G:P$	0.2253	6	0.99978
$G:V$	30.6383	1	3.109e-08 ***
$P:V$	15.8983	6	0.01431 *
$A:P:V$	13.8443	6	0.03142 *

Table 8: Analysis of Deviance Table (Type II Wald χ^2 tests) obtained for the fixed factors of the model fitted to F_1 obtained on /a/ and /ɑ/ *Vowels* (V), across *Age* (A), *Gender* (G) and *Period* (P) (see section 3.2 for details).

Factors	χ^2	**Df**	$Pr(>\chi^2)$
AT	5.3243	1	0.021030 *
G	0.5006	1	0.479247
V	5377.1017	1	< 2.2e-16 ***
$AT:G$	11.5903	1	0.000663 ***
$AT:V$	204.4467	1	< 2.2e-16 ***

Table 9: Analysis of Deviance Table (Type II Wald χ^2 tests) obtained for the fixed factors of the model fitted to F_2 obtained on /a/ and /ɑ/ *Vowels* (V), along *Apparent Time* (AT), controlled for *Gender* (G) (see section 3.2 for details).