DOI: 10.17469/O2112AISV000004

SIMON DEVAUCHELLE, ALBERT RILLIARD, DAVID DOUKHAN,
LUCAS ONDEL YANG

Variation of Perceived Voice Pitch Across Time Periods, Gender, and Age in French Media Archives

Have female voices lowered with time? A diachronic corpus based on French television or radio shows, constructed around four Periods (1955-56, 1975-76, 1995-96, 2015-16), grouping 1028 cisgender female and male speakers of four age categories. The speakers' voices were automatically segmented at the phone level, and acoustically analyzed. The fundamental frequency (F_0) and the formants were estimated; from the formants, the vocal tract length was calculated. Linear mixed regressions were fitted to these acoustic cues to estimate the role of the speaker's Age, Gender, and Period. The models showed few gender-specific changes in pitch-related parameters across Periods. Mean F_0 decreases with age for females, while it increases for males. The first formant decreases for open vowels over time Periods. These observations do not support a lowering of female voices over time. Results are discussed in the light of previous studies and changes in the structure of the French population.

Keywords: Voice pitch, Gender, Diachrony, Audiovisual archives, Fundamental frequency, Vocal Tract Length, Formants.

1. *Introduction*

Beyond the features relevant to the linguistic message *stricto sensu*, our voices carry various information relative to individual and social characteristics, thus participating in the indexicalization of one's social representation (Eckert, 2008). Our social roles and identities vary across contexts as we play different *characters* depending on interlocutors and social roles (Sadanobu, 2015); these complex identities are constructed during infancy and adolescence (Drager, 2015). Audiovisual media shape one's behavior through their prescriptive use of language and the display of social types (Stuart-Smith, 2017). The present work proposes a series of measurements related to the voice's perceived pitch in a diachronic perspective (here, we'll consistently use "pitch" to refer to the perceived tone – high or low – of a voice, not to fundamental frequency or other acoustic parameters). We view the voice pitch as an essential constituent of one's distinctive vocal characteristics, which typically participates in constructing gender display (Cartei et al., 2019, 2022).

Attributing a gender category to an interlocutor is important for spoken interactions in many cultures. It notably allows the use of adequate terms of address or honorifics (Tawilapakul, 2022), which are embedded in cultural systems and the localized language varieties (for terms of address in varieties of Spanish, see,

e.g., Rebollo Couto & Lopes, 2011). Gender categories may be attributed rapidly when cues are not ambiguous (Doukhan et al., 2023). This attribution is linked to a complex system of visual and acoustic cues (Richoz et al., 2017). Among the acoustic cues, various aspects were studied, and it has been shown that a combination of factors shall be used to decide gender (Merritt & Bent, 2022). However, the perceived pitch of a speaker's voice constitutes a fundamental aspect of gender indexicality in its complex realities (Leung et al., 2018; Pépiot & Arnold, 2021; Weirich & Simpson, 2018).

Individuals construct their voices during their lifespan, and the vocal characteristics are notably tuned to target the gender representation each wants to display in a given context (Holmes, 1998; Podesva, 2007, 2011). This vocal display is trained during childhood, even before puberty (Guzman et al., 2014), and of course, there is a significant vocal evolution during adolescence (Markova et al., 2016; Yamauchi et al., 2015). Later, during adulthood, we adapt our voices to the expectations of the sociocultural context (Ohara, 2001). The socially constructed characteristics of voices are thus part of the cues produced and perceived to interact smoothly within our society, and displaying more prototypical vocal characteristics impacts the speech perception process (Johnson, 2006). As cultural constructions, our voices are not purely intrinsic characteristics of our phonatory system but are tuned to how we use them.

As well as being an important indicator of gender, the voice pitch plays a varied role in vocal communication. A fundamental aspect of the voice's pitch during interactions was captured by Ohala's *Frequency Code* (Ohala, 1983, 1994), which describes the symbolic relation between the pitch of vocalizations and its behavioral interpretation. It is based on a systematic review of literature (Morton, 1977) that shows, in mammal and bird calls, a strong relation between low/descending pitch associated with hostile, aggressive calls vs. high/rising pitch associated with submissive, appeasement calls. The symbolic use of this vocal difference is based on the relation between an animal's size and the pitch of its voice; larger animals tend to have lower voices and to be more potent (Ohala, 1983, 1994). In human communication, this symbolic code, among others, is used for various behaviors (Gussenhoven, 2004) and may explain, e.g., cross-linguistic trends like the prevalence of raising pitch for questions. Such interpretations of voice pitch also resonate with the social judgments given to higher and lower voices regarding psychological powerlessness/power, intersecting with the speaker's gender – with a preference for higher-pitched voices for femininity that vary across cultures (van Bezooijen, 1995). Low voice pitch, a sexual dimorphism correlate, was described as an important feature for males' reproductive success and attractiveness (Apicella et al., 2007; Quené et al., 2021). The reverse holds for high-pitched female voices being judged as more sexy or attractive, with large variations linked to behavioral and sociocultural factors (Barkat-Defradas et al., 2021).

The social contexts within which a speaker is evolving thus create strong expectations of their vocal characteristics. In light of the preceding paragraphs,

these expectations may be divergent when female speakers behave in public arenas, typically in the political field, where a higher voice carries "feminine" attributes linked to powerlessness ("short, weak, dependent, modest" in the terms of van Bezooijen, 1995), and a lower voice is associated to "masculine" characteristics and power, credibility, etc. (Puts et al., 2006). As Coulomb-Gully (2022) puts it, female voices in the (French) political arena (a typical place for the expression of patriarchal power) are criticized for entering there and expressing themself: their voices are "*de trop*" – always too much of something (too high, too loud, etc.). Such mismatches in sociocultural expectations led some female politicians to follow voice coaching to deepen their voice; similar social impositions led others to modify their accents to fit the perceived expectations linked with a given position (Riverin-Coutlée & Harrington, 2022). These impositions also apply differently in different places, where prototypically feminine features may be relevant for being successful. Cinema and TV actresses also receive varying demands for their voices, depending on the producers' cultural background. French producers have different criteria than companies from the USA, looking after voice talents for dubbing that match the original voices (Le Fèvre-Berthelot, 2015).

Within this context that puts the voice's pitch as a central cue indexing gender, there are growing claims that the female voice's pitch has lowered over time, and most of them (e.g., Ellis et al., 2023; Krahé & Papakonstantinou, 2020) base their claim on the same reference (Pemberton et al., 1998). These claims are repeated over several media platforms and sometimes twisted in the process. A newspaper article (Nebelsztein, 2018), for example, cherry-picked two scientific papers (Berg et al., 2017; Pemberton et al., 1998) presenting studies supporting a lowering of the fundamental frequency of females to conclude female voices have lowered while male voices haven't – and it is not isolated (e.g., Robson, 2018; Taylor, 2018). There are several problems with this claim, firstly because it does not consider other studies with contrasting conclusions (e.g., Hollien et al., 1997, or the more recent work by Suire & Barkat-Defradas, 2020); secondly because the study by Berg et al. (2017) does not present diachronic evidence and uses a relatively peculiar production protocol that may influence their measure; third, because these works test the pitch of Australian (Pemberton et al., 1998) and German (Berg et al., 2017) females and the voice pitch is culturally encoded, so it is difficult to generalize to other populations; finally, the diachronic work of Pemberton et al. (1998) finds a difference amounting to a lowering in the mean pitch of 1.8st (which is arguably limited) and does not test male speakers, nor female of different age, so it is more complex to know if there are some other factors (like vocal effort – e.g., Berg et al., 2017; Titze & Sundberg, 1992) that may be involved in the described difference.

To provide referential values for the case of voices spoken and perceived in France, we present here a diachronic study that measured two correlates of voice pitch – the voice fundamental frequency (F_0) and an estimation of the vocal tract length (VTL) based on its resonances – within the speakers of a cross-sectional dataset based on 1028 cis-gender female and male speakers between 20 and 80 years old broadcasted

in French national television and radio channels along 60 years spanning between 1955 and 2015. The effects of the speakers' gender, age, and broadcasting periods are evaluated to observe potential variations. The paper first presents the corpus construction process, its evaluation, the acoustic measurements, and the statistical analysis; it then presents the changes observed for the three controlled factors and discusses them in the light of the literature.

2. *Methods*

2.1 Corpus

2.1.1 Speaker characteristics and identification process

This work is based on the diachronic corpus collected by Uro et al. (2022), composed of extracts from broadcasted television or radio shows from the French National Institute of Audiovisual (INA) archives. Programs broadcasted through regional channels were excluded from analyses to keep only voice representations at the national level. Each type of show has a specific stylistic feature. As speech style impacts voice characteristics, studio-based talk show interviews were preferred to maximize sound recording quality (to avoid environmental noises) and gather data from interactive discussions where the participants shall have reasonable speaking time and use question/answer interactions. Note that on a corpus of such a scale (more than 850 different archives entries were used), the topic in each show could not be controlled, but many domains are represented, from news and politics to arts, science, education, etc. The shows were selected for being broadcasted at four different time *Periods* spaced by 20 years (1955-1956, 1975-1976, 1995-1996, 2015-2016), and potential speakers were identified by INA's archivists to reach a target of thirty speakers of both genders in four age categories (20-35, 36-50, 51-65, over 65). Thirty speakers for each of the 32 categories of four time Periods * four Age categories * two Genders amounts to a theoretical target of 960 speakers. The documentalists identified more than 6,000 potential target speakers within INA archives.

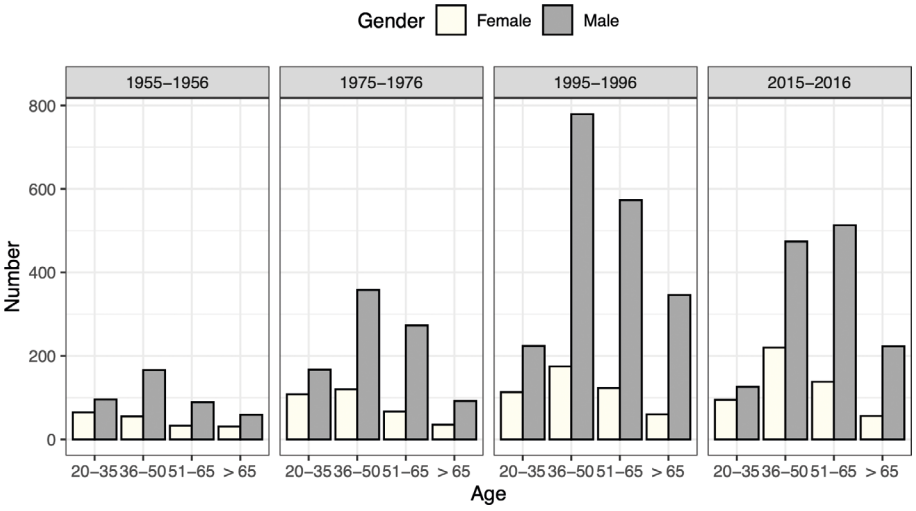
A semi-automatic process, through diarization and speaker identification, allowed the effective identification of the targets in actual recordings (Uro et al., 2022) and led to a dataset comprising 1028 speakers with a minimal amount of three minutes of recordings per speaker; these speakers are spread across time Periods, Age categories, and Gender according to Table 1. It is worth emphasizing the solid and consistent bias of female representation in the media (Coulomb-Gully, 2011; Doukhan et al., 2018): even trying to identify an even number of speakers of each gender, the dataset has about 50 % more males than females. This is explained by the difficult task of finding female speakers, especially in the youngest and oldest age categories, in the 1950s and the 1970s: the archivists managed to find a sufficient number of targets, but during the identification stage, many of the females were given minor roles, often only appearing on stage while males spoke on their behalf. The process led to identifying more males in some categories than the 30-speaker

initial goal because only the male target(s) speak in some archives featuring several target speakers.

Table 1 - *Number of speakers for each of the time Periods (by row), Age category (column), and Gender (Female or Male, in columns), plus marginal sums*

	20-35		*36-50*		*51-65*		*> 65*		*Row sum*		
	F	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*Both*
1955-56	18	41	21	72	18	51	19	15	76	179	255
1975-76	18	18	23	42	28	37	18	25	87	122	209
1995-96	31	30	33	45	29	48	29	34	122	157	279
2015-16	31	31	30	53	29	49	31	31	121	164	285
Column sum	98	120	107	212	104	185	97	105	406	622	1028

Figure 1 - *Number of potential target speakers identified within INA databases across Periods (individual plots), age categories (x-axis), and gender (white or grey bars)*



The representational bias in the media also applies to age categories, with the youngest and oldest categories underrepresented. This explains the fewer targets in these age categories (Table 1), as illustrated in Figure 1, representing the number of potential targets identified by the archivists within INA databases. The overall aging of the French population may be observed with the growing proportion of the 51-65 y. o. category over the Periods, and the decrease of the 20-35 one. The difference in the number of targets between time Periods is linked to the smaller size of the oldest archives, with fewer channels and fewer shows recorded at that time.

For each target speaker identified in the corpus, a minimum of three minutes of speech was selected, removing passages where non-speech events exceed a given signal-to-noise ratio threshold to keep only relatively clean speech. A similar process also removed overlapping speech events. This cleaning procedure was essential

to keep only qualitative speech recordings; it removed about 2/3 of the spoken segments (Uro et al., 2022).

For each speaker, the selection process ended with passages of speech of more than two seconds (the different passages amounted to more than 3 minutes of speech), but not necessarily corresponding to complete speech turns, as the noise removal procedure may delete parts of the turns, e.g., where music was found in the background. These passages of speech are mainly rid off pauses, which were removed at the diarization stage. Table 2 presents the total duration of speech considered for each category of period, age, and gender.

Table 2 - *Total duration (in minutes) of speech passages selected for all speakers belonging to the same time Period (by row), Age category (column), and Gender (Female or Male, in columns), plus marginal sums*

	20-35		*36-50*		*51-65*		*> 65*		*Row sum*		
	F	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*Both*
1955-56	21	98	52	200	39	168	102	161	214	527	741
1975-76	38	86	43	120	172	130	147	272	400	608	1008
1995-96	238	194	253	390	369	402	454	339	1314	1325	2639
2015-16	287	240	175	439	211	325	291	342	964	1346	2310
Column sum	584	618	523	1149	791	1025	994	1014	2892	3806	6698

2.1.2 Automatic transcription, phone segmentation, and evaluation

The entire speech corpus was automatically transcribed using Whisper (Radford et al., 2023). Then, the output transcripts were phonetized and force-aligned on the speech signals to extract vocalic excerpts, using pre-trained models for French from Montreal Forced Aligner (MFA; McAuliffe et al., 2017) and its French phonetic dictionary and language models (McAuliffe & Sonderegger, 2022c, 2022b, 2022a, 2022d). Timing indications proposed by the ASR predictions were also used to strengthen the alignments. Only oral vowels were selected by taking the middle frame of the detected vowel, resulting in the selection of over 1,300,000 frames. Twelve French oral vowels were selected: /a/, /ɑ/, /e/, /ɛ/, /œ/, /ə/, /ø/, /i/, /o/, /ɔ/, /u/ and /y/. Frames corresponding to vowels more than 200 milliseconds long were discarded from the study, preserving 96 % of the detected vowels. Automatic misalignment or poor automatic transcription can result in longer durations. We also removed frames corresponding to unvoiced phones using the voiced/unvoiced decision described in the next section.

An evaluation of the automatic transcription was conducted to evaluate the quality of the pipeline. While Whisper has been tested on French material (Radford et al., 2023), the evaluation corpora were based on read speech, potentially differing from the more spontaneous productions of our audiovisual corpus. A subset of 81 speech segments (about 20 minutes) was randomly selected across all periods, gender, and age categories. Four L1 French speakers manually transcribed the segments, which were then phonetized and forced-aligned by MFA. Our main concern was that the ASR system would make up words that had not been pronounced,

especially for older materials where sound artifacts are more common. Excerpts from old audio archives are less likely to have been used for Whisper training as they are less common. A further assumption was that Whisper tended to "smooth" transcriptions by removing hesitations and repetitions. To get a clearer idea of the system efficiency for this application, two scores were used corresponding to two analysis levels: the word error rate (WER) and the phone error rate (PER). Results are presented in section 3.1. In the following analysis, the central 25ms frame of each vowel was targeted for the acoustic analysis.

2.2 Acoustic analysis

The parameters targeted in this paper are known correlates of perceived pitch (Laver, 1980; Ohala, 1994) and are related to anatomical differences between females and males (Gisladottir et al., 2023; Sataloff, 2017; Simpson, 2009; Titze, 1989): from each vowel, segmented as described above, we estimated the voice fundamental frequency (F_0) and the first four formants. We also estimated the Vocal Tract Length (VTL) from the formants used as proxies of the vocal tract resonances (Titze et al., 2015).

For F_0 estimations, it has been shown that pitch detection algorithms perform differently on difficult signals (Vaysse et al., 2022). We used three algorithms to robustly estimate the fundamental frequency: Praat's AC algorithm (Boersma & Weenink, 2022) with an F_0 range of 65-650 Hz, REAPER (Talkin, 2015) with the default setting and with Hilbert transform (thus estimating two voicing estimates), and FCN-F0 (Ardaillon & Roebel, 2019) with its default pitch range (30-1000Hz) and Viterbi smoothing. We used F_0 values estimated by Praat and FCN-F0, plus the voiced/unvoiced decision produced by Praat and REAPER. Only the frames classified as voiced by Praat and REAPER were considered; among these frames, those who received F_0 values from Praat and FCN-F0 that do not differ by more than a 20 % gross-error-rate threshold (e.g., Jouvet & Laprie, 2017) were considered valid. This strategy considers only 52.2 % of valid frames in the original signal, less than if considering only the Praat algorithm output (61.1 %). On these frames, the value returned by Praat was considered.

The first four formants (F_1, F_2, F_3, F_4) were estimated by applying Praat's implementation of the Burg algorithm (Boersma & Weenink, 2022). As this work is focused on an estimate of acoustic differences linked with a possibly varying expression of gender through vocal cues, it seems dubious to use the speaker's gender to change the algorithm parameters – and typically, the frequency ceiling: Praat's recommendation is to set the ceiling to 5500Hz for female, and 5000Hz for males, but this proved problematic, e.g. for /u/ vowels produced by females (Gendrot & Adda-Decker, 2005). We thus followed here the strategy proposed by Escudero et al. (2009), which consists of extracting, for each vowel, several potential sets of formants with different ceilings and then deciding an optimal ceiling for each speaker and each vowel category. The best ceiling minimizes the sum of the variances of each of the first three formants (using the logarithm of their values in

Hertz) for this speaker and this vowel; three formants were considered here (only two in Escudero et al., 2009) as in French, the third formant is relevant for rounding (e.g., Ménard et al., 2009). The set of ceilings that was used is based on the default ceiling for each gender (5.5 and 5 kHz for females and males), plus twenty ceilings above and twenty below this default value, spaced by percentual of the default ceiling value; this was done using Praat's *FormantPath* function, that implements part of Escudero et al. (2009) algorithm, applying the different ceiling according to equation 1 (from Praat help), where the frequency ceiling (*FC*) at the i^{th} step above or below the middle-frequency ceiling (MFC) equals MFC multiplied by a factor determined by plus or minus (if above or below) *i* times the *step* size.

$$FC_i = MFC * e^{(\pm i * step)} \quad (1)$$

In our case, *i* varies from 1 to 20, and the *step* factor was set to 0.01, so to have steps of approximately 50Hz below or above the middle ceiling frequency – for ceilings ranging from above 4000Hz up to 6100 Hz for males, and from 4500 up to 6700Hz for females. Formants were estimated frame by frame; those on frames classified as unvoiced during the F_0 estimation process were discarded. The F_0 and the formant values estimated at the middle of each vowel are considered for the remainder of this work; F_0 was expressed in semitones relative to 1Hz and formants in Bark to fit their perception.

We used the first four formants to estimate the vocal tract length (VTL). This was done by using equations 2 and 3 from Lammert & Narayanan (2015), where F_1, F_2, F_3, and F_4 are the first four formant values (in Hz), and C=34000cm/s is the speed of sound. One estimation of VTL was done on each vowel of the corpus.

$$\phi = \frac{0.089\,F_1}{1} + \frac{0.102\,F_2}{3} + \frac{0.121\,F_3}{5} + \frac{0.669\,F_4}{7} \quad (2)$$

$$VTL = \frac{c}{4*\phi} \quad (3)$$

2.3 Statistical models

On each vowel, the measurements of the F_0, the first three formants, or the VTL were used as the dependent variable (DV) of linear mixed models fit with R's *lme4* library (Bates et al., 2015; R Core Team, 2023), with as fixed factors the speaker's Age (in years, averaged over the different shows from which their speech was extracted, if applicable), the speaker's Gender (Female, Male), and the time Period (1955-56, 1975-76, 1995-96, 2015-16). For models based on the F_0 and the VTL dependent variables, the Speaker and the Vowel (the 12 oral vowels considered here) factors were used as random factors in the model, with Vowel nested within Speaker, to account for the individual variability in vowel performances (see eq. 4 for the maximal model based on these measurements, where *DV* is either F_0 or VTL). For the models fitting formants, the Vowel factor was used as a fixed factor to allow describing formant variations linked to vowels, and the Speaker factor was

kept as a random factor (see eq. 5 for the maximal models based on formants, where F_i stands for formant formants F_1, F_2 or F_3).

$$DV \sim Age * Gender * Period + (Speaker/Vowel) \quad (4)$$

$$F_i \sim Age * Gender * Period * Vowel + (Speaker) \quad (5)$$

A model simplification procedure was then applied to each model described by equations 4 and 5, using R's *step()* function (Kuznetsova et al., 2017). For the model based on VTL, the simplification procedure deleted the triple and double interactions, leaving only the main fixed factors (see Table 3). For the model based on F_0, the triple interaction and the double interaction between Age and Period were deleted, leading to a model with three fixed factors and two double interactions: between Gender and Age, and between Gender and Period (see Table 4). For the models based on formants, no simplification was done.

Table 3 - *Output of the backward reduction process for the model fitted on VTL, for Random factors (first part: number of eliminated factors, number of parameters, log. Likelihood, Akaike Information Criterion, Likelihood Ratio Test, degrees of freedom, and p-value), and for Fixed factors (second part: number of eliminated factors, Sum of squares, mean square, numerator and denominator degrees of freedom, F value, and p-value)*

Random	*Eliminated*	*npar*	*logLik*	*AIC*	*LRT*	*Df*	*p*
<none>		19	-1670409	3340857			
(1 | Vow:Spk)	0	18	-2044955	4089946	749092	1	< 2.2e-16
(1 | Spk)	0	18	-1670770	3341575	721	1	< 2.2e-16
Fixed	*Eliminated*	*Sum Sq*	*Mean Sq*	*NDF*	*DenDF*	*F value*	*p*
Age:Gen.:Per.	1	6.20	2.07	3	999.05	1.9621	0.11801
Age:Period	2	0.94	0.31	3	1002.36	0.2970	0.82756
Gen.:Per	3	2.31	0.77	3	1002.88	0.7319	0.53308
Age:Gender	4	2.75	2.75	1	1002.03	2.6074	0.10668
Age	0	6.71	6.71	1	1003.95	6.3746	0.01173
Gender	0	2369.13	2369.13	1	1001.65	2250.1529	< 2e-16
Period	0	100.53	33.51	3	1006.18	31.8271	< 2e-16

Table 4 - *Output of the backward reduction process for the model fitted on F_0 for Random factors (first part: number of eliminated factors, number of parameters, log. Likelihood, Akaike Information Criterion, Likelihood Ratio Test, degrees of freedom, and p-value), and for Fixed factors (second part: number of eliminated factors, Sum of squares, mean square, numerator and denominator degrees of freedom, F value, and p-value)*

Random	*Eliminated*	*npar*	*logLik*	*AIC*	*LRT*	*Df*	*p*
<none>		19	-3085128	6170293			
(1 | Vow:Spk)	0	18	-3097069	6194174	23883	1	< 2.2e-16
(1 | Spk)	0	18	-3094811	6189658	19367	1	< 2.2e-16

Fixed	*Eliminated*	*Sum Sq*	*Mean Sq*	*NDF*	*DenDF*	*F value*	*p*
Age:Gen.:Per.	1	18.430	6.143	3	998.47	0.4760	0.69909
Age:Period	2	71.255	23.752	3	1001.64	1.8401	0.13817
Age:Gender	0	246.772	246.772	1	1004.04	19.1181	1.357e-05
Gender:Period	0	105.208	35.069	3	1004.40	2.7169	0.04356

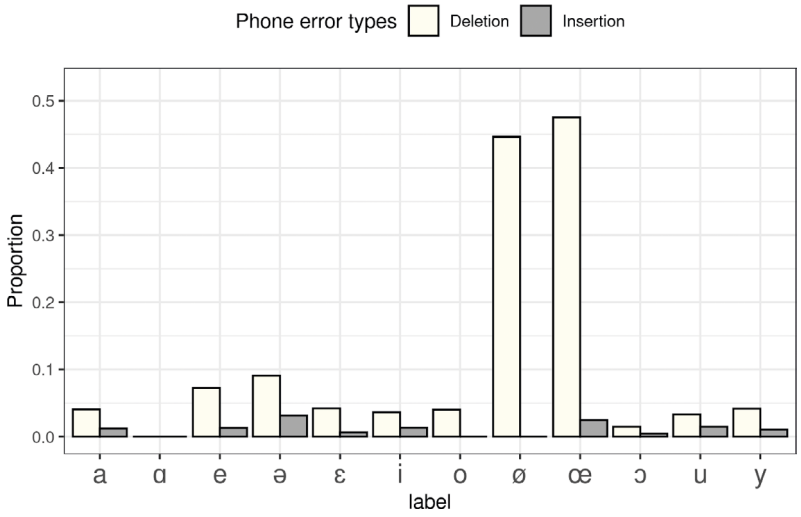
3. *Results*

3.1 Evaluation of the automatic transcription and phonetic alignment

On the subset of 81 speech segments, and compared to the human transcription, Whisper reached a 12.8 WER, using Whisper large v2 and after text normalization. The WER on our dataset is thus slightly higher than the WER on FLEURS and on Common Voice described in the OpenAI paper (Radford et al., 2023), but it is still a close performance. As we are here interested in the evolution of some phonetic characteristics of French spoken in the media, the PER appears to be more adapted to evaluate the efficiency of the ASR system. Some difficult-to-decide transcription differences may not lead to pronunciation differences (e.g., for French: plurals where spelling differs but not pronunciation) – and indeed, the PER was 7.9 % when all phones were considered and 9.8 % for oral vowels only.

Let's emphasize that the precision at the phone level is 97.4 %; in other words, Whisper produces few hallucinations in its predictions, at least once phonetized. 5.4 % of vowels in the evaluation subset were deleted during the automatic alignment process compared to the manual transcription – showing a reduced effect of Whisper in smoothing its outputs. Indeed, most of the errors at the phone level are deletions (68.2 %). As shown in Figure 2, the most frequently deleted vowels were /œ/ (46.7 % deletion), /ø/ (44.6 %), /ə/ (8.6 %), and /e/ (7.2 %). One-quarter of deleted oral vowels are derived from the French word for hesitation, "*euh*" (pronounced /œ/ or /ø/), followed by the function words "et" (6.5 %) and "de" (4.7 %). As shown in Figure 2, the vowel class the most "hallucinated" is the /ə/: 33 % of its insertions come from the negation word "*ne*" which is generally not produced in casual conversations but added by Whisper, producing more formal sentences in its outputs.

Figure 2 - *Proportion of vowels deleted (white) or inserted (grey) during the automatic transcription process*



Similar to Barras et al. (2002) findings for another ASR system evaluated for the transcription of French media archives, we do not notice differences in the phone or word error rates of Whisper transcriptions over time. Varying performances linked to gender may be related to the specific characteristics of the selected speakers (speaking style, recording conditions, etc.) and not to gender itself.

3.2 Variations of VTL with Age, Gender, and Period

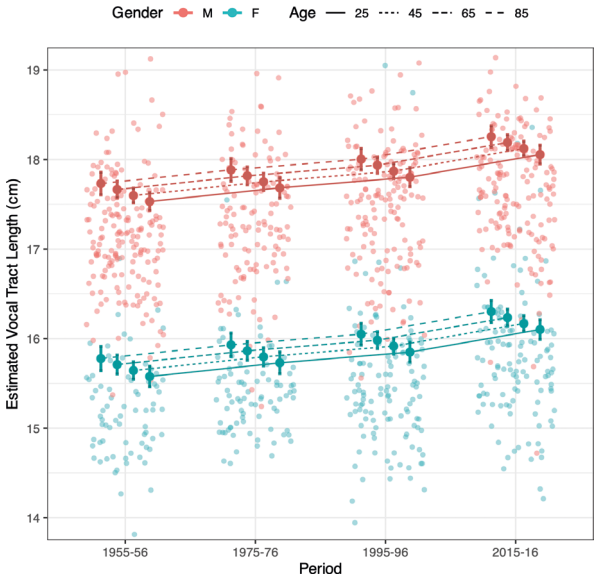
The mean tendencies of the linear mixed model fitted on values of VTL are displayed in Figure 3 for the three fixed factors together (let's recall there were no significant interactions between them). The main effect in the model is –of course– related to the speaker's Gender and reflects the mean difference of about two centimeters observed between adult females and males in vocal tract length, corresponding to several descriptions in the literature (De Pinto & Hollien, 1982; Hollien et al., 1994; Vorperian et al., 2009, 2011, 2019). The two other factors have much smaller effect sizes if they show significant and continuous trends. There is a tendency for vocal tract resonances to decrease with the speakers' Age, that amount, in terms of estimated VTL, to a lengthening of about 0.25cm between 25 and 85 years old. Comparatively, the effect linked to Period is twice as large, with an increase of about 0.5 cm of estimated VTL between the 1955s and 2015s.

3.3 Variations of F_0 with Age, Gender, and Period

The model fitted to the speakers' estimated F_0 values is based on two interactions: one between the Gender and Age factors, and the second between the Gender and Period factors. As for VTL, the Gender * Age interaction (see Figure 4) corresponds to the description in the literature of F_0 changes with these two factors, with Sataloff

et al. (2017) describing a decreasing trend of F_0 with age for females (descending pitch with age) and an increase for males (rising with age; see also Gisladottir et al., 2023). The trend of F0 change with age observed in our dataset corresponds to a decrease of 1.7st for 60 years or 0.3st for 10 years of age in female voices. For male speakers, the mean F_0 value tends to increase at a slower pace, with a rise of 1.3st for 60 years corresponding to a rate of 0.2st for 10 years.

Figure 3 - *Mean VTL (with confidence intervals) estimated by the model based on speaker Age (values for 25, 45, 65, and 85 y. o. are plotted), Gender, and time Period; the values estimated for each speaker are plotted as individual lightly colored points in the background, to give a better idea of the observed variation*

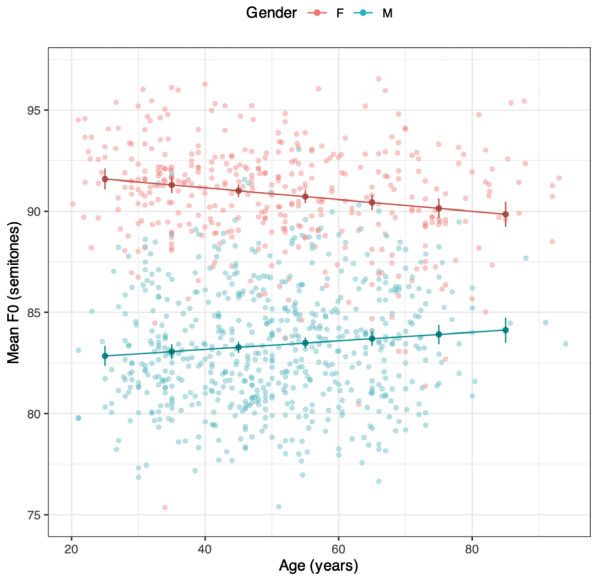


Note that the study by Leung et al. (2022) describes a downward trend of F_0, independent of gender, but the authors do not include the interaction between these two factors in their model and have a population sample with about twice as many females as males. The study by Berg et al. (2017), with a population between age 40 and 79 years old, observed a downward trend of F_0 for females but no significant change with age on this measurement for male speakers.

The interaction between Gender and Period is represented in Figure 5. A post-hoc comparison (with Bonferroni adjustment) between periods, fixing gender, showed that the only significant difference is an increase of F_0, observed between 1995-96 and 2015-16 for male speakers ($\chi^2 = 8.44$, $p = 0.044$). This limited importance of the interaction (the observed differences are smaller than one semitone), as well as the fact that the changes across periods do not show a clear trend but rather differences from one period to the other, not always in the same direction, may lead to conclude these differences may be linked to others, uncontrolled, factors (e.g., different characteristics of the shows, like the position of the microphones, the

ambient noise, etc.), than a result that can be interpreted as a gender-specific trend over time. We can nonetheless observe that the cloud of points corresponding in Figure 5 to females from the years 1955-1956 seems to be somehow higher than the other clouds, without this leading to a significant difference in their mean.

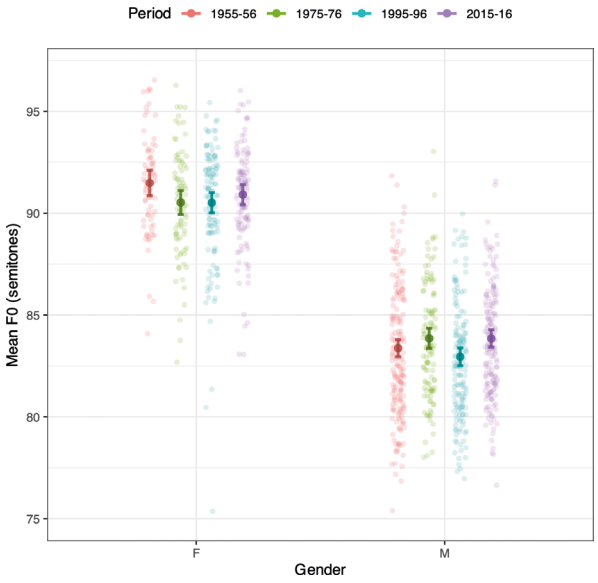
Figure 4 - *Mean F0 (with confidence intervals) estimated by the model based on speaker Age (x-axis), and Gender (colors and separate lines); the mean values estimated for each speaker are plotted as lightly colored points in the background to give a better idea of the observed variation*



3.4 Variations of Formants with Age, Gender, and Period

None of the three models based on formants could be simplified: it shows there are some complex interactions between the four fixed factors that will be difficult to fully understand, notably because they relate to age and phone. We are not aware of theoretical reasons for such differences (Hermes et al., 2023), apart from potential changes in the phonetic implementation of the vowel inventory (Cecelewski et al., 2023) – but such changes are not sufficiently described to provide stable hypotheses. In the remainder of the description, the Age factor will thus be overlooked to simplify the interpretation and understanding of the other factors (it was set to 40 years).

Figure 5 - *Mean F0 (with confidence intervals) estimated by the model based on speaker Gender (x-axis), and time Period (colored points for each gender); the mean values estimated for each speaker are plotted as lightly colored points in the background, to give a better idea of the observed variation*



The shapes of the vocalic spaces (not shown) show some differences across Gender and Periods. The larger effect (after the obvious effect of vowels, which fit the expected distribution of oral vowels in French across the (F_1, F_2) plane (e.g., Apostol et al., 2004) is related to Gender, with higher formant values for female speakers, as expected. For the interaction of the factor Period with the other factors, the vowels tend to be shifted toward lower F_1 values with more recent time Periods. This change is shown in Figure 6: it is possible to observe higher values for this formant during the 1955-56 Period compared to the latter ones, especially for open vowels /a/ and /ɑ/. This is less marked for closed vowels, especially for males. Higher formants for the same vowel are typically observed with a larger mandible opening (Erickson, 2002), which may be linked to higher vocal effort (Rilliard et al., 2018).

Variations of the F_2 across time Periods are plotted in Figure 7. The effect of Period is restricted to a few vowels, /ɑ/ and /ø/. For /ɑ/, we observe a rising of F_2 and thus an anteriorization tendency with time corresponding to the loss of the phonemic distinction between /ɑ/ and /a/ during the second half of the twentieth century (Cecelewski et al., 2023). The effect on /ø/ goes in the other direction, with a lowering of F_2. It is unclear which articulatory change may produce this difference if a centralization of the articulation is possible; we are unaware of a description of such a change in the literature. These two changes in the second formant do not seem gender specific, as they are observed for both genders, with comparable time pace and amplitude.

Figure 6 - *Mean F_1 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*

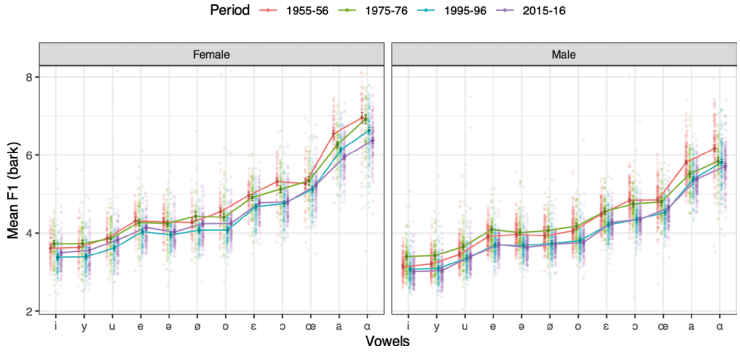


Figure 7 - *Mean F_2 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*

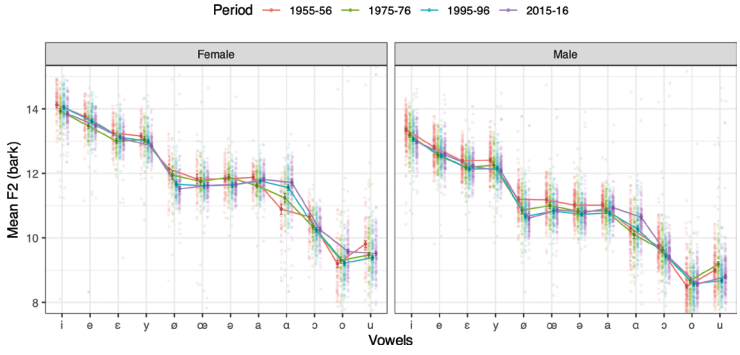
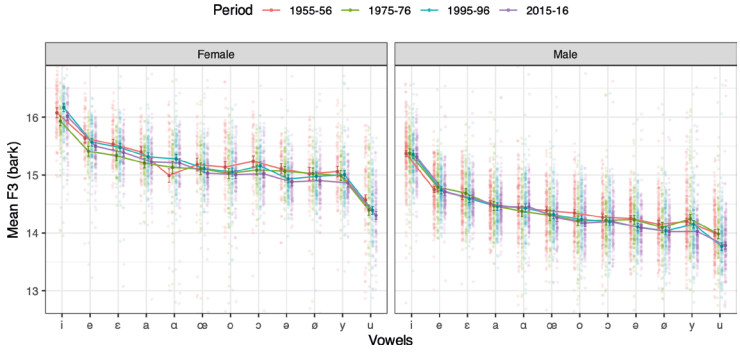


Figure 8 - *Mean F_3 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*



The changes in F_3 are presented in Figure 8. There are few interpretable tendencies for this formant, as the main change related to Period applies to the /ɑ/ produced by females in 1955-56. Still, it is difficult to interpret this change as something other than a bias potentially linked to formant estimation.

4. *Discussion & Conclusions*

This work presented the use of an extensive (111 hours) corpus of spontaneous speech (Uro et al., 2022) produced mainly in studio conditions in French national media archives, spanning 60 years in four periods. It is somehow balanced for gender and age – with limits related to strong gender and age bias in the representation of speakers featured in audiovisual media. An automatic transcription was made and evaluated. The evaluation showed a robust quality of the Whisper ASR system for French and media speech (targeting here interventions during talk show interviews), with a slightly higher word error rate than the one presented in their evaluation (Radford et al., 2023) on another speech genre. It also shows Whisper tends to remove the hesitations and repetitions typical of spontaneous speech productions: this shall not be a problem for our goal, but it should be carefully considered for studies interested in such phenomena.

We have presented an analysis of some measurements related to the perceived pitch of a voice: its fundamental frequency and formants, measured on oral vowels only. A preceding work (Rilliard et al., 2023) on the same corpus extracted similar measures but without phonemic segmentation but reached a comparable conclusion if limited by the absence of information on phones. The formants were used to estimate the vocal tract length that is related to its resonances. These source and resonance characteristics are cited in the literature as important correlates of perceived pitch (Ohala, 1983, 1994) and constitute major acoustic cues for the perceptual attribution of a gender to a voice (Guzman et al., 2014; Leung et al., 2018). In relation to the display of gender and a possible secular trend supporting an evolution tendency in France – and especially, in our case, through the national media, our data analysis does not support the view of a change (toward a lowering) in voice pitch specific to female voices, as in Hollien et al. (1997) for another linguistic and cultural context. let's recall these authors found an effect of style on pitch, but not a diachronic tendency. This is significant in our case, and we would like to restate that the corpus was based on talk shows featuring interviews – thus our results shall be restrained to this kind of material.

For the VTL, the model captured an increase in its estimated length across Periods, which corresponds to a lower pitch – but this lowering is observed for both genders; this impedes an interpretation as a gender-specific change. For F_0, the changes captured by the models in relation to Periods are minimal and do not follow a time trend but rather tend to vary from one period to the other in different directions, with amplitude below the threshold of statistical significance in most cases (but in one case for male speakers). Finally, for formants, the main

finding related to Period is linked to F_1, which has higher values for open vowels in 1955-56 compared to more recent Periods – and this observation is once again not gender specific if it may contribute to a lower pitch. Other findings are more related to changes in the vocalic system of French (loss of the /ɑ a/ opposition) than to changes in the voice pitch.

More robust variations in the acoustic parameters were nonetheless observed, which cues to a lowering of pitch. The most striking one is the change in F_0 with the speaker's age, with a descending trend for females and a rising trend for males. This leads to a remarkable reduction of about 3 st in the mean difference between male and female mean F_0 values from 20 to 80 years of age; about one-third of the difference between young adults is lost when comparing the older part of the population. These gender-specific changes correspond to several descriptions in the literature, either on acoustic measurements (Gisladottir et al., 2023) or based on physiological considerations (Sataloff et al., 2017). They are, however, not observed systematically (Berg et al., 2017). It is known that the voice production task influences F_0 changes (Hollien et al., 1997) thus, the controlled production task proposed by Berg et al. (2017) may explain some of the observed discrepancies. Let's recall our dataset proposes spontaneous productions during broadcasted talk shows that are clearly a different style of voice than the excerpts proposed in the literature just cited hereabove (which, with the exception of Hollien et al. 1997, do not include a diachronic dimension). A more comparable dataset was described by Suire & Barkat-Defradas (2020) and is based on vox pop interviews in French media archives. Unfortunately, this corpus was not controlled for age (for obvious reasons) thus, it cannot be used as a comparable reference here – if this study also does not observe trends of pitch lowering for female voices. In our study, a trend of VTL lengthening was also observed with age, which was gender-independent but supported the same interpretation as for F_0 of lowered pitch voices for older females. This trend does not depend on Periods.

The dataset presented here does not support the claim that the pitch of female voices has lowered since the 1950s, independently of other controlled factors – and typically their age (a conclusion similar to Hollien et al., 1997). In the same age range, females participating in talk shows on national French media tend to have comparable pitches in 1955 and in 2015. Let's also note that the mean pitch of voices within a given gender category varies by about one octave across individuals (see Figure 5). This must be highlighted, as the overall means (about 210Hz for females and 130 Hz for males in this dataset) are not necessarily representative of what female or male voices necessarily are: the mean observed for one speaker varies between about 130 and 260Hz for females, and 90 and 200Hz for males. There are thus no clear-cut F_0 differences that separate both gender voices. There are, nonetheless, notable differences in the distribution of individuals participating in television and radio shows in these two periods. As shown in Figure 1, based on a selection of 6,000 individuals from INA databases, there were more females between 20 and 35 years old and fewer females between 51 and 65 years old in the

1955-56 period than in 2015-16 (this also applies to males). This may be explained by the aging of the population in France during this period (INSEE, 2023). In the more recent Period (2015-16), there are more chances to observe a female speaker in the media (the proportion of females in the shows has increased since the 1950s; see Doukhan et al., 2018), but the chances of this individual being above 50 years old also increased. Not controlling for the population's age and sampling speakers randomly from TV and radio shows would thus lead to an age difference for the observed population. It is possible that auditors in the 2010s were exposed to female voices that were, on average, lower than the voices presented in a random sample in the 1955-56 period just because these 2010s females are, on average, older than in the 1950s. This may induce an erroneous (on the basis of the dataset presented here) conclusion about lower female voices in recent periods.

An important feature of this work is that it also looks at male speakers (unlike Pemberton et al., 1998), and this allowed us to describe trends across periods that are not gender specific but may be interpreted as features correlated with a lowering of voices across this time period. This is the case of VTL or of the first formant for open vowels, with lower resonances in more recent periods. To explain such a tendency, the changes in the studio settings across these periods may give an important clue, with the microphones' positions being more distant from the speaker's mouth in older archives, while they are closer in recent shows: this impacts the speaking style, and typically the vocal effort, with more projected voices in older archives (Boula de Mareüil et al., 2012). Producing higher vocal effort is known to increase the voice's F_0 (Alku et al., 2024; Berg et al., 2017; Titze & Sundberg, 1992), and it impacts the first formant, typically for the open vowels (Rilliard et al., 2018). Thus, changes in voice pitch observed in the media that are not gender- or age-dependent may be explained by situational variables and changes in sound recording practices.

Has the pitch of female voices changed with time? This dataset does not support this claim; meanwhile, this dataset certainly presents a series of possible biases that limit the reach of the conclusions one may draw from it. For example, it is a speech genre (speech presented in talk shows from the French audiovisual media) that features mostly individuals from high socioeconomic and prestige status (for example, almost all the speakers have their Wikipedia page). Thus, the dataset does not represent the general French population (a difference with studies like Berg et al., 2017; Gisladottir et al., 2023). The dataset is more adequate to describe the type of voices that are presented, with prescriptive characteristics, to the French population through the national audiovisual media. More specifically, the absence of a global trend does not mean some individuals cannot have lowered their voices: reports exist that some female personalities have decided or felt compelled to lower their voices (Coulomb-Gully, 2022). Typically, they are females with social positions of power (in the political sector or in the economy). It shall be possible to label the speakers of the current dataset in terms of professional occupation and to check if this allows observing a pitch lowering for a subpart of the (female) population;

this is an ongoing work that shall be presented soon. A problem with this approach is that females with political or economic responsibilities will constitute a very reduced or even absent population group in the 1950s.

The proposed dataset is also worth other types of analyses than trying to find a secular trend in voice pitch. An obvious body of research is linked with modeling potential articulatory changes in a diachronic perspective, as the convergence of the two /a/ of French (Cecelewski et al., 2023).

Acknowledgements

The authors wish to thank INA's documentalists, Laetitia Larcher and Anissa-Claire Adgharouamane, for their comprehensive research within the archives and their competent advice for selecting speakers within relevant shows. This work has been partially funded by the French National Research Agency (project Gender Equality Monitor – ANR-19-CE38-0012).

References

Alku, P., Kodali, M., Laaksonen, L. & Kadiri, S.R. (2024). AVID: A speech database for machine learning studies on vocal intensity. In *Speech Communication*, *157*, 103039. https://doi.org/10.1016/j.specom.2024.103039

Apicella, C.L., Feinberg, D.R. & Marlowe, F.W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. In *Biology Letters*, *3*(6), 682–684. https://doi.org/10.1098/rsbl.2007.0410

Apostol, L., Perrier, P. & Bailly, G. (2004). A model of acoustic interspeaker variability based on the concept of formant–cavity affiliation. In *The Journal of the Acoustical Society of America*, *115*(1), 337–351. https://doi.org/10.1121/1.1631946

Ardaillon, L., Roebel, A. (2019). Fully-Convolutional Network for Pitch Estimation of Speech Signals. In *Proceedings of Interspeech 2019*, Graz, Austria, 15-19 September 2018, 2005–2009.

Barkat-Defradas, M., Raymond, M. & Suire, A. (2021). Vocal Preferences in Humans: A Systematic Review. In Weiss, B., Trouvain, J., Barkat-Defradas, M. & Ohala, J.J. (Eds.), *Voice Attractiveness*. Singapore: Springer, 55–80.

Barras, C., Allauzen, A., Lamel, L. & Gauvain, J.-L. (2002). Transcribing audio-video archives. In *Proceedings of IEEE Int. Conference on Acoustics Speech and Signal Processing*, Orlando, Florida, USA, 13-17 may 2002, I-13-I–16.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Berg, M., Fuchs, M., Wirkner, K., Loeffler, M., Engel, C. & Berger, T. (2017). The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors. In *Journal of Voice*, *31*(2), 257.e13-257.e24.

Boersma, P., Weenink, D. (2022). PRAAT: Doing phonetics by computer [Computer program]. Version 6.2.08. http://www.praat.org/

Boula de Mareüil, P., Rilliard, A. & Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. In *Language and Speech*, *55*(2), 263–293.

Cartei, V., Banerjee, R., Hardouin, L. & Reby, D. (2019). The role of sex-related voice variation in children's gender-role stereotype attributions. In *British Journal of Developmental Psychology*, *37*(3), 396–409.

Cartei, V., Reby, D., Garnham, A., Oakhill, J. & Banerjee, R. (2022). Peer audience effects on children's vocal masculinity and femininity. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200397.

Cecelewski, J., Gendrot, C., Adda-Decker, M. & Boula de Mareüil, P. (2023). A Diachronic Study of Vowel Harmony in French Broadcast Speech since 1940. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 August 2023, 798-802.

Coulomb-Gully, M. (2011). Genre et médias: Vers un état des lieux. In *Sciences de la Société*, *83*, 3–13.

Coulomb-Gully, M. (2022). *Sexisme sur la voix publique: Femmes, éloquence et politique*. Avignon: Éditions de l'Aube.

De Pinto, O., Hollien, H. (1982). Speaking fundamental frequency characteristics of Australian women: Then and now. In *Journal of Phonetics*, *10*(4), 367–375.

Doukhan, D., Devauchelle, S., Girard-Monneron, L., Chávez Ruz, M., Chaddouk, V., Wagner, I. & Rilliard, A. (2023). Voice Passing: A Non-Binary Voice Gender Prediction System for evaluating Transgender voice transition. In *Proceedings of Interspeech 2023*, Dublin, Ireland, 20-24 August 2023, 5207–5211.

Doukhan, D., Poels, G., Rezgui, Z. & Carrive, J. (2018). Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. In *VIEW Journal of European Television History and Culture*, *7*(14), 103.

Drager, K.K. (2015). *Linguistic variation, identity construction and cognition*. Berlin: Language Science Press.

Eckert, P. (2008). Variation and the indexical field. In *Journal of Sociolinguistics*, *12*(4), 453–476.

Ellis, S., Goetze, S. & Christensen, H. (2023). Moving Towards Non-Binary Gender Identification Via Analysis of System Errors in Binary Gender Classification. In *Proceedings of IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, 4-10 June 2023, 1–5.

Erickson, D. (2002). Articulation of Extreme Formant Patterns for Emphasized Vowels. In *Phonetica*, *59*(2–3), 134–149.

Escudero, P., Boersma, P., Rauber, A.S. & Bion, R.A.H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. In *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393.

Gendrot, C., Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech 2005*, Lisboa, Portugal, 4-8 September 2005, 2453-2456.

Gisladottir, R.S., Helgason, A., Halldorsson, B.V., Helgason, H., Borsky, M., Chien, Y.R., Gudnason, J., Gudjonsson, S.A., Moisik, S., Dediu, D., Thorleifsson, G., Tragante, V., Bustamante, M., Jonsdottir, G.A., Stefansdottir, L., Rutsdottir, G., Magnusson, S.H., Hardarson, M., Ferkingstad, E., Halldorsson, G.H., Rognvaldsson, S., Skuladottir, A., Ivarsdottir, E.V., Norddahl, G., Thorgeirsson, G., Jonsdottir, I., Ulfarsson, M.O., Holm, H., Stefansson, H., Thorsteinsdottir, U., Gudbjartsson, D.F., Sulem, P., Stefansson, K. (2023). Sequence variants affecting voice pitch in humans. In *Science Advances*, *9*(23), eabq2969.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation* (1st ed.). Cambridge: Cambridge University Press.

Guzman, M., Muñoz, D., Vivero, M., Marín, N., Ramírez, M., Rivera, M.T., Vidal, C., Gerhard, J. & González, C. (2014). Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice. In *International Journal of Pediatric Otorhinolaryngology*, *78*(10), 1592–1598.

Hermes, A., Audibert, N. & Bourbon, A. (2023). Age-related vowel variation in French. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 august 2023, 2045–2049.

Hollien, H., Green, R. & Massey, K. (1994). Longitudinal research on adolescent voice change in males. In *The Journal of the Acoustical Society of America*, *96*(5), 2646–2654.

Hollien, H., Hollien, P.A. & De Jong, G. (1997). Effects of three parameters on speaking fundamental frequency. In *The Journal of the Acoustical Society of America*, *102*(5), 2984–2992.

Holmes, J. (1998). Signalling gender identity through speech. In *Moderna Språk*, *92*(2), 122–128.

INSEE. (2023, July 3). *Pyramide des âges par état matrimonial 2020 – France métropolitaine Séries de 1901 à 2020*. https://www.insee.fr/fr/statistiques/2418118

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. In *Journal of Phonetics*, *34*(4), 485–499.

Jouvet, D., Laprie, Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *Proceeedings of the 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 28 August – 2 September 2017, 1614–1618.

Krahé, B. & Papakonstantinou, L. (2020). Speaking like a Man: Women's Pitch as a Cue for Gender Stereotyping. In *Sex Roles*, *82*(1–2), 94–101.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. In *Journal of Statistical Software*, *82*(13), 1–26.

Lammert, A.C., Narayanan, S.S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. In *PLOS ONE*, *10*(7), e0132193.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Le Fèvre-Berthelot, A. (2015). Doublage au féminin: Transposer le genre. In *Genre En Séries*, *2*, 50–72.

Leung, Y., Oates, J. & Chan, S.P. (2018). Voice, Articulation, and Prosody Contribute to Listener Perceptions of Speaker Gender: A Systematic Review and Meta-Analysis. In *Journal of Speech, Language, and Hearing Research*, *61*(2), 266–297.

Leung, Y., Oates, J., Papp, V. & Chan, S.-P. (2022). Speaking Fundamental Frequencies of Adult Speakers of Australian English and Effects of Sex, Age, and Geographical Location. In *Journal of Voice*, *36*(3), 434.e1-434.e15.

Markova, D., Richer, L., Pangelinan, M., Schwartz, D.H., Leonard, G., Perron, M., Pike, G.B., Veillette, S., Chakravarty, M.M., Pausova, Z. & Paus, T. (2016). Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. In *Hormones and Behavior*, *81*, 84–96.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, 20-24 August 2017, 498–502.

McAuliffe, M., Sonderegger, M. (2022a). French language model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022b). French MFA acoustic model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022c). French MFA dictionary v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022d). French MFA G2P model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

Ménard, L., Dupont, S., Baum, S.R. & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. In *The Journal of the Acoustical Society of America*, *126*(3), 1406–1414.

Merritt, B., Bent, T. (2022). Revisiting the acoustics of speaker gender perception: A gender expansive perspective. In *The Journal of the Acoustical Society of America*, *151*(1), 484–499.

Morton, E.S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. In *The American Naturalist*, *111*(981), 855–869.

Nebelsztein, M. (2018, August 26). *Bonne nouvelle: La voix des femmes est plus grave qu'avant*. terra femina. https://www.terrafemina.com/article/pourquoi-la-voix-des-femmes-est-plus-grave-qu-avant_a343484/1

Ohala, J.J. (1983). Cross-Language Use of Pitch: An Ethological View. In *Phonetica*, *40*(1), 1–18.

Ohala, J.J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In Hinton, L., Nichols, J. & Ohala, J.J. (Eds.), *Sound Symbolism* (1st ed.). Cambridge: Cambridge University Press, 325–347.

Ohara, Y. (2001). Finding one's voice in Japanese: A study of the pitch levels of L2 users. In Pavlenko, A., Blackledge, A., Piller, I. & Teutsch-Dwyer, M. (Eds.), *Multilingualism, Second Language Learning, and Gender*. Berlin: De Gruyter Mouton, 231-256.

Pemberton, C., McCormack, P. & Russell, A. (1998). Have women's voices lowered across time? A cross sectional study of Australian women's voices. In *Journal of Voice*, *12*(2), 208–213.

Pépiot, E., Arnold, A. (2021). Cross-Gender Differences in English/French Bilingual Speakers: A Multiparametric Study. In *Perceptual and Motor Skills*, *128*(1), 153–177.

Podesva, R.J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. In *Journal of Sociolinguistics*, *11*(4), 478–504.

Podesva, R.J. (2011). The California Vowel Shift and Gay Identity. In *American Speech*, *86*(1), 32–51.

Puts, D.A., Gaulin, S.J.C. & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. In *Evolution and Human Behavior*, *27*(4), 283–296.

Quené, H., Boomsma, G. & Van Erning, R. (2021). Attractiveness of Male Speakers: Effects of Pitch and Tempo. In Weiss, B., Trouvain, J., Barkat-Defradas, M. & Ohala, J.J. (Eds.), *Voice Attractiveness*. Singapore: Springer, 153–164.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202), Hawaii, USA, 23-29 July 2023. PMLR, 28492–28518.

Rebollo Couto, L., Lopes, C.R. (Eds.). (2011). *As formas de tratamento em português e em espanhol: Variação, mudança e funções conversacionais = Las formas de tratamiento en español y en portugués: variación, cambio y funciones conversacionales*. Niteroi: Editora da UFF.

Richoz, A.-R., Quinn, P.C., Hillairet De Boisferon, A., Berger, C., Loevenbruck, H., Lewkowicz, D.J., Lee, K., Dole, M., Caldara, R. & Pascalis, O. (2017). Audio-Visual Perception of Gender by Infants Emerges Earlier for Adult-Directed Speech. In *PLOS ONE*, *12*(1), e0169325.

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

Rilliard, A., Doukhan, D., Uro, R. & Devauchelle, S. (2023). Evolution of voices in French audiovisual media across genders and age in a diachronic perspective. In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 August 2023, 753–757.

Riverin-Coutlée, J., Harrington, J. (2022). Phonetic change over the career: A case study. In *Linguistics Vanguard*, *8*(1), 41–52.

Robson, D. (2018). *The reasons why women's voices are deeper today*. BBC. https://www.bbc.com/worklife/article/20180612-the-reasons-why-womens-voices-are-deeper-today

Sadanobu, T. (2015). "Characters" in Japanese Communication and Language: An Overview. In *Acta Linguistica Asiatica*, *5*(2), 9–28.

Sataloff, R.T. (2017). *Clinical assessment of voice* (Second edition). San Diego: Plural Publishing, Inc.

Sataloff, R.T., Kost, K.M. & Linville, S.E. (2017). Chapter 13. The Effects of Age on the Voice. In Sataloff, R.T. (Ed.), *Clinical assessment of voice* (Second edition). San Diego: Plural Publishing, Inc, 221–240.

Simpson, A.P. (2009). Phonetic differences between male and female speech. In *Language and Linguistics Compass*, *3*(2), 621–640.

Stuart-Smith, J. (2017). Bridging the gap(s): The role of style in language change linked to the broadcast media. In Thogersen, J., Coupland, N. & Mortensen, J. (Eds.), *Language (De)standardisation in Late Modern Europe: Style and media studies*. Nashville: Novus Press, 51–84.

Suire, A., Barkat-Defradas, M. (2020). Evolution of human pitch: Preliminary analyses in the French population using INA audiovisual archives of Vox Pops. In *Proceedings of 2020 IASA-FIAT/IFTA Joint Conference*, online, Ireland, 26-29 October 2020.

Talkin, D. (2015). REAPER: Robust epoch and pitch estimator. [Computer Program] https://github.com/google/REAPER

Tawilapakul, U. (2022). Face threatening and speaker presuppositions: The case of feminine polite particles in Thai. In *Journal of Pragmatics*, *195*, 69–90.

Taylor, V. (2018). *Women's Voices Are Getting Lower Over Time, Thanks To Sexism*. Romper. https://www.romper.com/p/womens-voices-are-getting-lower-over-time-thanks-to-sexism-study-says-9406100

Titze, I.R. (1989). Physiologic and acoustic differences between male and female voices. In *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707.

Titze, I.R., Baken, R.J., Bozeman, K.W., Granqvist, S., Henrich, N., Herbst, C.T., Howard, D.M., Hunter, E.J., Kaelin, D., Kent, R.D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D.G., Noé, H., Scherer, R.C., Smith, J.R., Story, B.H., Švec, J.G., Ternström, S., Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. In *Journal Acoust. Soc. America*, *137*(5), 3005–3007.

Titze, I.R., Sundberg, J. (1992). Vocal intensity in speakers and singers. In *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946.

Uro, R., Doukhan, D., Rilliard, A., Larcher, L., Adgharouamane, A.-C., Tahon, M. & Laurent, A. (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*, Marseille, France, 20-25 June 2022. European Language Resources Association, 3271–3280.

van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. In *Language and Speech*, *38*(3), 253–265.

Vaysse, R., Astésano, C. & Farinas, J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. In *The Journal of the Acoustical Society of America*, *152*(5), 3091–3101.

Vorperian, H.K., Kent, R.D., Lee, Y. & Bolt, D.M. (2019). Corner vowels in males and females ages 4 to 20 years: Fundamental and F1–F4 formant frequencies. In *The Journal of the Acoustical Society of America*, *146*(5), 3255–3274.

Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J. & Gentry, L.R. (2009). Anatomic development of the oral and phar-

yngeal portions of the vocal tract: An imaging study. In *The Journal of the Acoustical Society of America*, *125*(3), 1666–1678.

Vorperian, H.K., Wang, S., Schimek, E.M., Durtschi, R.B., Kent, R.D., Gentry, L.R. & Chung, M.K. (2011). Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imaging Study. In *Journal of Speech, Language, and Hearing Research*, *54*(4), 995–1010.

Weirich, M., Simpson, A.P. (2018). Gender identity is indexed and perceived in speech. In *PLOS ONE*, *13*(12), e0209226.

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K.-I., Nito, T., Tayama, N. & Yamasoba, T. (2015). Quantitative Analysis of Digital Videokymography: A Preliminary Study on Age- and Gender-Related Difference of Vocal Fold Vibration in Normal Speakers. In *Journal of Voice*, *29*(1), 109–119.