EVOLUTION OF VOICES IN FRENCH AUDIOVISUAL MEDIA ACROSS GENDERS AND AGE IN A DIACHRONIC PERSPECTIVE.

Albert Rilliard[1,2], David Doukhan[3], Rémi Uro[1,3], Simon Devauchelle[3]

[1] Université Paris Saclay, CNRS, LISN, Orsay, France; [2] Federal University of Rio de Janeiro, Brazil; [3] French National Institute of Audiovisual, Paris, France

albert.rilliard@lisn.upsaclay.fr, {ddoukhan,ruro,sdevauchelle}@ina.fr

ABSTRACT

We present a diachronic acoustic analysis of the voice of 1023 speakers from French media archives. The speakers are spread across 32 categories based on four periods (years 1955/56, 1975/76, 1995/96, 2015/16), four age groups (20-35; 36-50; 51-65, >65), and two genders. The fundamental frequency (F_0) and the first four formants (F1-4) were estimated. Procedures used to ensure the quality of these estimations on heterogeneous data are described. From each speaker's F_0 distribution, the base-F_0 value was calculated to estimate the register. Average vocal tract length was estimated from formant frequencies. Base-F_0 and vocal tract length were fit by linear mixed models to evaluate how they may have changed across time periods and genders, corrected for age effects. Results show an effect of the period with a tendency to lower voices, independently of gender. A lowering of pitch is observed with age for female but not male speakers.

Keywords: Gender, Diachrony, Vocal Tract Resonance, Vocal register, Broadcast speech

1. INTRODUCTION

Vocal characteristics are an important part of identity, with our voices indicating our gender and other social constructs [1]. Voice obviously changes with age, a lot during our development until adulthood [2], but also later with aging voices [3]. Social constructs, and the acoustic characteristics linked to them, may change with cultures – as was described by van Bezooijen for Japanese and Dutch female voice [4]: it is thus essential to study these characteristics in varied cultural and language contexts (and it typically fits this ICPhS theme). Some studies also claimed that voices may have changed across time (for populations of comparable age); this was claimed for female voices, for example in Australia [5].

We present here a first acoustic analysis of a corpus of voices extracted from French broadcast archives in a diachronic perspective, trying to balance the selection of speakers according to their age and gender. A total of 1023 speakers were selected, and samples of their vocal production were extracted from the original archive. This corpus aims to extract acoustic cues from a large sample, representative of female and male voices presented in French media, to describe them and their possible changes with time and age. Some studies have described diachronic changes of voices across genders, but these studies are few, may have varying conclusions or population sampling, and address populations from different cultural backgrounds. [5] addresses young adult women's voices, with two points in time – while [6] studied changes in male voices over five time periods, focusing on media anchors: comparatively few speakers were included (four for the first three periods), with possible longitudinal effects. For French, [7] selected short samples of voices of anonymous individuals from media archives (mean duration of 5 s.), comparing F_0 of males and females across 70 years. They report complex patterns of F_0 changes across time, with possible opposite tendencies for genders after the 1960s, but don't control for age, a factor known to induce changes in voice F_0 [8]. Vocal characteristics are also affected by vocal effort [9], which has substantial effects on F_0 [10] and on formants [11].

Long-term acoustic measures give reliable information on voice characteristics. [12] showed the long-term average spectrum stabilizes for articulatory changes with about 10 seconds of voiced speech. [13] compared several measures linked to register and found the base-F_0 [14] was faster to stabilize (with less than ten s. of voiced speech) than mean or median F_0. A study [15] linked to the estimation of the vocal tract length (VTL) from formants showed the method proposed in [16] was the most rapidly stable, even on a reduced dataset. The literature reports that gender is perceived notably through two acoustic cues – F_0 and supraglottal resonances [17, 18].

The remaining parts of the paper present estimates of F_0 and VTL made on a corpus collected following the semi-automatic strategy proposed by [19] to gather speech samples from speakers across 60 years period, found in TV and Radio archives of the French National Institute of Audiovisual (INA) (speakers belong to different categories of age and gender). The acoustic measures related to the speakers' voice pitch are fitted by linear mixed models evaluating potential changes across time for both genders and controlling for age.

2. METHODS

2.1. INA's diachronic corpus

INA's diachronic corpus contains speakers of both genders, spread across four age categories (20-35, 36-50, 51-65, over 65 years old) and four time periods (1955-56, 1975-76, 1995-96, and 2015-16). With an initial goal of gathering at least 30 individuals for each of these 32 categories of age, gender, and time period, samples of voice for 1023 speakers were collected (see Table 1). Some of these categories being much less present in media (typically women), finding target speakers from all age categories in the earliest periods was challenging. It was thus mandatory to increase the targeted period for the 1955-56 and 1975-76 periods by considering the years between 1954-1957 and 1974-1977 (note the additional years only represent a small part of the selected speakers so that these periods will be referred to with the originally targeted years for simplicity). The collection was done thanks to INA's archivists identifying specific women and men within the four age categories and featuring in programs from the four given targeted time periods. Following the method described in [19], the programs featuring target speakers were submitted to a diarization process, and the ID corresponding to each target speaker was then hand-picked. These speakers' samples were then selected to keep samples with minimal noise or background music and remove silences.

These archive extracts were submitted to procedures to discard excerpts with adverse characteristics. `LIUM_SpkDiarization` [20] was used to reject segments associated with a telephone quality. Speakers with less than 10 seconds of valid pitch estimates were rejected (the filtering protocol is detailed in section 2.3). Table 1 presents the distribution of the speakers. Voice samples from 1023 speakers were obtained from 878 radio or TV programs (the voice of 85 persons was collected from more than one program). The median duration of valid acoustic features per speaker was 125 seconds. The amount of unique speakers in the 32 categories varies between 15 and 74.

	20-35		36-50		51-65		>65	
	F	M	F	M	F	M	F	M
1954-57	17	41	22	74	18	50	17	15
1974-77	18	17	23	41	28	39	20	26
1995-96	32	31	32	46	29	47	29	35
2015-16	30	31	30	52	28	48	29	31

Table 1: Number of speakers per time period (rows), age group, and gender in the corpus

2.2. Voicing, F_0 and Formant estimation

To estimate robust F_0 and formant measurements from widely different materials (due to heterogeneity in recording and archival conditions), the voicing decision, and then F_0 and formant estimations, were made by concurrent algorithms. The `Spleeter` source separation framework was used to separate voice from other phenomena (music or noise) [21]. The acoustic features were then estimated twice, from both the original and speech-separated signals. For voicing and F_0 estimation, [22] showed good performances of `Praat` auto-correlation (ac) algorithm; it also shows that these estimations could be improved in noisy conditions by combining `REAPER`'s voicing estimation with `FCN-F0`'s F_0 estimation (based on neural-network). We thus used `Praat` [23] for estimating voicing and F_0 (ac algorithm with a 65-650 Hz F_0 range) and for estimating the first four formants (using the recommended settings for female and male voices: 5 formants for a respective ceiling frequency of 5.5 or 5kHz). `REAPER` algorithm [24] was used with two distinct settings (default and Hilbert transform) to obtain two other voicing estimates. `FCN-F0` was used with default pitch range (30-1000 Hz), and Viterbi smoothing to estimate F_0 [25].

2.3. Acoustic features filtering

The frames detected as voiced by the six estimations (`Praat` and `REAPER` with two sets of parameters on the raw and separated signals) were kept. From these frames, those where the four F_0 estimations (by `Praat` and `FCN-F0` on raw and separated signals) were below a 20% gross error rate threshold were kept. This strategy allowed us to keep 52.2% of the signal's original frames instead of 61.1% obtained with `Praat`'s `ac` algorithm. From these frames, the formant estimates made on both signals (raw and source-separated) were compared to keep only frames with four formant estimates, with variations below a 20% gross-error-rate threshold (keeping about 96% of the preceding frames) and 50.2% of the signal's original frames.

2.4. Long-term features

For each speaker, the valid samples (cf. above section) were grouped in chunks of at least 10s; i.e., for a 36s extract, three chunks of 12s were used. From each of these chunks, two long-term estimates were made: the speaker's base-F_0 (defined as the seventh decile of the F_0 distribution on the chunk [14, 13]), and the speaker's vocal tract length (equal to the median of the length estimations made on each frame based on the first four formants, using the equation proposed in [16, 15]). Let's make it clear the estimation of the "vocal tract length" is a way to evaluate the tendency one speaker has to produce higher or lower resonances (formants) as a result of their articulatory habits; the relation to actual vocal tract length is undoubtedly more complex [26]. For each speaker, these two long-term estimates (base-F_0 and vocal tract length: VTL) were estimated for each chunk of voiced samples. The median number of chunks was 14 by speaker, ranging from 1 to 175.

2.5. Statistical analysis

The base-F_0 and the estimated VTL were fitted with two linear mixed models that took as fixed effects the actual age (in years) of the speaker, the time period of the recording (four levels: 1955-56, 1975-76, 1995-96, 2015-16), and the speaker's gender – and as random effects the speaker for which the measure was done, and the media program from which the corresponding speaker's chunk was extracted (the program factor was nested in the speaker factor). Following [27], a maximal model was fit (using R's `lme4` library [28]) that included all the interactions between the fixed factors. This model was then submitted to a simplification procedure to remove non-significant terms and reach a minimal adequate model, one for F_0 and one for VTL. These two models are used here to describe the results of the variation of the base-F_0 and VTL across age, gender, and time period.

3. RESULTS

3.1. Base-F_0

The minimal model fitted on base-F_0 was based on the three main fixed factors (`Age`, `Period` and `Gender`), with the `Age:Period` and `Age:Gender` interactions, plus the original random structure (`Program` nested in `Speaker`). An estimation of the conditional and marginal coefficients of determination (using [29]) for this model showed the model explained about 90% of the variance, with fixed factors responsible for 58% of this. Figure 1 presents the effect of the `Age:Gender` interaction on base-F_0: there is a major difference in F_0 across genders (about 9 st), and F_0 decreased with age for female voices. Figure 2 presents the `Age:Period` interaction: while Base-F_0 was increasing with Age in the 1950s, it showed a tendency to decrease with age at later periods (1995-96 and 2015-16). Note that this second interaction has a much smaller effect size than the `Age:Gender` one.

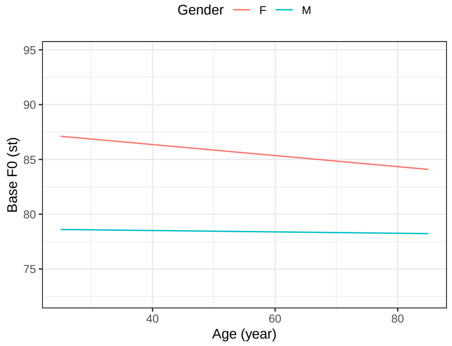

Figure 1: Fit of base-F_0 (in semitones) from the Speaker Age and Gender.

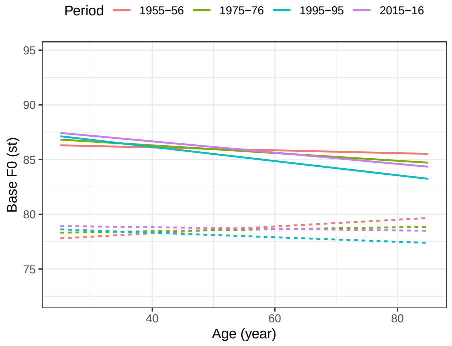

Figure 2: Fit of base-F_0 (semitones) for speaker Age by Period and separated by Gender (males: dashed lines) to show the relative effect of the Period slopes on voices of each gender.

3.2. Vocal tract length

The minimal model fitted on VTL was based on the three fixed factors without interaction and keeping the original random structure. The conditional and marginal coefficient of determination for this model showed a much smaller part of the variance was explained by the factors, with the complete model explaining about 38% and fixed factors 31%. The Gender had the largest effect size, with a mean difference of estimated VTL of 1.7 cm between females and males. Period was the next important factor, with a modest but significant increase in VTL of 0.13 and 0.2 cm for the 1995-96 and 2015-16 periods, compared to 1955-56. Age showed a significant but small (0.03 cm for ten years) increasing slope of estimated VTL with age.

4. DISCUSSION & CONCLUSION

This study intends to bring discussions and gather remarks and advice from the community by presenting preliminary results on a complex topic related to the perception of voice quality, and importantly voice pitch, across gender, age, and time periods in France. As the material used to extract the stimuli could not, for obvious reasons, be of homogeneous quality (recording conditions, signal quality, etc.), there are a series of limitations related to this approach. A problem is linked with the estimation of vocal tract resonances, as highlighted, e.g., by [26]. The quality of archive acoustic quality further complicates this: we had to implement a series of checks to assert we didn't process voices with telephone quality or other deterioration linked to, e.g., compression.

The question of the formant estimation algorithm's parameters is essential, as it may lead to varying results. The data presented here are based on the default parameters recommended by Praat for each gender, but this is a potential bias for estimating VTL in speakers with non-standard characteristics. We also tested the parameters recommended by [16] (estimate six formants for a 5.5kHz ceiling frequency): this option led to very different results – and a complex interaction between Gender, Age, and Period with some cases of female speakers having longer estimated VTL than male ones. It was thus not kept here, but questions remain on the estimation of VTL. While distinct settings are generally recommended for the analysis of male and female voices (frequency range for F_0 estimation, frequency ceiling for formats), the use of these a priori settings for the investigation of gender characteristics is questionable since voices, recording, and archival strategies may have changed over the last 60 years – and because it may hide some non-standard features.

Conversely, measures of F_0 seem much more robust and show trends that confirm the literature, with an effect of age on female voice [3, 30], and a clear gender difference for gender representation in French media. One finding of this study is linked to a change in the evolution of base-F_0 with age across time periods: while it tends to increase in the 1950s for males or stay almost constant for females, steeper slopes were observed in the two later periods (1995-96, and 2015-16). This evolution is not dependent on gender (the interaction was not significant), but the slopes are different for each gender, as shown by figure 2. This decreasing tendency raises questions on the use of voice in public displays (varying social use [31] or changing health conditions?), with lowered voices for older speakers across time periods.

5. ACKNOWLEDGEMENTS

This work has been funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012). The authors are deeply indebted to INA's archivists Laëtitia Larcher and Anissa-Claire Adgharouamane and who selected the target speakers and also wished to thank Luc Ardaillon and Jean-Hugues Chenot for their valuable advice.

6. REFERENCES

[1] D. C. Sergeant and G. F. Welch, "Gender differences in long-term average spectra of children's singing voices," *Journal of Voice*, vol. 23, no. 3, pp. 319–336, May 2009.

[2] M. Fouquet, K. Pisanski, N. Mathevon, and D. Reby, "Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood," *Royal Society Open Science*, vol. 3, no. 10, p. 160395, Oct 2016.

[3] A. Russell, L. Penny, and C. Pemberton, "Speaking fundamental frequency changes over time in women: A longitudinal study," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 101–109, Feb 1995.

[4] R. van Bezooijen, "Sociocultural aspects of pitch differences between japanese and dutch women," *Language and Speech*, vol. 38, no. 3, pp. 253–265, 1995.

[5] C. Pemberton, P. McCormack, and A. Russell, "Have women's voices lowered across time? a cross sectional study of australian women's voices," *Journal of Voice*, vol. 12, no. 2, pp. 208–213, Jan 1998.

[6] Y. Zou, Y. Wang, and W. He, "Diachronic contrastive analysis on read speech in broadcast news: Evidence from pitch and duration," in *2012 8th International Symposium on Chinese Spoken*

Language Processing. IEEE, 2012, pp. 291–295.

[7] A. Suire and M. Barkat-Defradas, "Evolution of human pitch: Preliminary analyses in the french population using ina audiovisual archives of vox pops," in *2020 IASA-FIAT/IFTA Joint Conference*, 2020.

[8] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, Aug 2011.

[9] M. Berg, M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, and T. Berger, "The speaking voice in the general population: Normative data and associations to sociodemographic and lifestyle factors," *Journal of Voice*, vol. 31, no. 2, pp. 257.e13–257.e24, Mar 2017.

[10] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, May 1992.

[11] A. Rilliard, C. d'Alessandro, and M. Evrard, "Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 109–122, Jan 2018.

[12] A. Löfqvist and B. Mandersson, "Long-time average spectrum of speech and voice analysis," *Folia Phoniatrica et Logopaedica*, vol. 39, no. 5, pp. 221–229, 1987.

[13] P. Arantes and A. Eriksson, "Temporal stability of long term measures of fundamental frequency," in *Speech Prosody 2014*. ISCA, May 2014, pp. 1149–1152.

[14] H. Traunmüller and A. Eriksson, "The perceptual evaluation of f_0 excursions in speech as evidenced in liveliness estimations," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1905–1915, 1995.

[15] K. Johnson, "Vocal tract length normalization," *UC Berkeley Phonology Lab Annual Reports*, vol. 14, 2018.

[16] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLOS ONE*, vol. 10, no. 7, p. e0132193, Jul 2015.

[17] K. Pisanski and D. Rendall, "The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2201–2212, Apr 2011.

[18] Y. Leung, J. Oates, S.-P. Chan, and V. Papp, "Associations between speaking fundamental frequency, vowel formant frequencies, and listener perceptions of speaker gender and vocal femininity-masculinity," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 7, pp. 2600–2622, Jul 2021.

[19] R. Uro, D. Doukhan, A. Rilliard, L. Larcher, A.-C. Adgharouamane, M. Tahon, and A. Laurent, "A semi-automatic approach to create large gender- and age-balanced speaker corpora: Usefulness of speaker diarization & identification," in *13th Language Resources and Evaluation Conference*, 2022, pp. 3271–3280.

[20] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[21] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[22] R. Vaysse, C. Astésano, and J. Farinas, "Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 3091–3101, 2022.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1.09)," 2020. [Online]. Available: http://www.praat.org

[24] D. Talkin, "Reaper: Robust epoch and pitch estimator," 2015. [Online]. Available: https://github.com/google/REAPER

[25] L. Ardaillon and A. Roebel, "Fully-Convolutional Network for Pitch Estimation of Speech Signals," in *Proc. Interspeech 2019*, 2019, pp. 2005–2009.

[26] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.

[27] S. T. Gries, *Statistics for linguistics with R: a practical introduction*, 3rd ed., ser. De Gruyter Mouton textbook. Berlin Boston: de Gruyter Mouton, 2021.

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] K. Bartoń, *MuMIn: Multi-Model Inference*, 2022, r package version 1.47.1. [Online]. Available: https://CRAN.R-project.org/package=MuMIn

[30] A. Yamauchi, H. Yokonishi, H. Imagawa, K.-I. Sakakibara, T. Nito, N. Tayama, and T. Yamasoba, "Quantitative analysis of digital videokymography: A preliminary study on age- and gender-related difference of vocal fold vibration in normal speakers," *Journal of Voice*, vol. 29, no. 1, pp. 109–119, Jan 2015.

[31] P. Boula de Mareüil, A. Rilliard, and A. Allauzen, "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*, vol. 55, no. 2, pp. 263–293, 2012.