Interspeech 2024
1-5 September 2024, Kos, Greece

Do Speaker-dependent Vowel Characteristics depend on Speech Style?

Nicolas Audibert[1], *Cécile Fougeron*[1], *Christine Meunier*[2]

[1]Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), Paris, France
[2]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
nicolas.audibert@sorbonne-nouvelle.fr, cecile.fougeron@sorbonne-nouvelle.fr, christine.meunier@univ-amu.fr

Abstract

Based on the examination of 385K vowels pertaining to 6 oral and 3 nasal vowel categories produced by twenty-three French speakers in both read and spontaneous speech, the present study questions the interplay between intra-speaker style-dependent variability in vowel production and speaker-specific vowel properties. Acoustic properties of the speakers' vowels in the 12-D MFCC space are compared to that of other speakers in two styles. Results show that vowels do index speaker-distinctiveness better in read vs. spontaneous speech. Furthermore, in both speech styles, distinctions between speakers are the largest for the nasal vowels. Intra-speaker variability in vowel production is also examined between speech styles and is found to depend on the speaker and on the vowel category. However, for most speakers and most vowels, the variation between styles is smaller than the distinction between speakers in both styles. Implications of these results for speaker identification are discussed.

Index Terms: vowels, French, speaker discrimination, speaker identification, indexical properties, speech style

1. Introduction

Various phonemic classes have been tested for their relative potential to discriminate between speakers. Among those, vowels are often recognized as good candidates to index speaker-specific information, including physiological, anatomical, pathological condition, age or sex/gender attributes that might affect the vocal tract dimensions and articulatory displacements, as well as idiolectal and regiolectal specificities of talkers [e.g 1, 2, 3, 4]. Indeed, acoustic characteristics of single vowels or of the acoustic vowel system have proved useful when used for speaker discrimination or identification ([4-10]). In their analysis of the impact of phonemic content on voice comparison processes in French, [10] further demonstrated that both oral and nasal vowels were the most informative (along with nasals), and that nasal vowels outperformed the oral vowel class.

However, if vowels respect one of the requirements necessary for phoneme candidates to discriminate among speakers, i.e. high inter-speaker variability ([11, 12]), it is unclear whether the second attribute of low intra-speaker variability is fulfilled. Indeed, vowels have been extensively used as test cases in the phonetic literature on speech variation, and their acoustic properties have been shown to vary quite a lot according to speech conditions, style, interlocutor, rate, etc. (among many others see [14, 17] for studies on French). These variations affect mostly the spectral and durational characteristics of vowels (see e.g. [18, 19], but also amplitude [20]. Globally, these effects can be summarized by a general tendency for vowels to lose some of their contrastive acoustic properties in casual or fast speech, either by being centralized, more coarticulated, or more variable within their category as compared to clearer or slower style of speech (e.g. see [17]).

This within-speaker variability in pronunciation has been shown to have implications for speaker identification [21, 22]. Recently, [23] tested on a corpus produced by 20 Brazilian speakers how the spectral characteristics of vowel (among other phonetico-acoustic dimensions) discriminate between speakers in two speech styles: a spontaneous dialogue between familiar speakers (twins) vs. an interview with a non-familiar experimenter. Results confirmed that vowel formants (and especially higher formants F3 and F4) discriminate well among speakers as compared to other features, and that vowels did contrast speakers better in the dialogue condition than in the interview data.

Considering that understanding the indexical properties conveyed in the speech signal about the speaker identity is essential for both theoretically grounded questions about speech communication and social interactions, as well as for speaker recognition applications, the objective of the current paper is twofold. By examining the acoustic properties of 385,724 vowel tokens produced by twenty-three French speakers in both read and spontaneous speech (interviews), the present study aims to assess (a) whether and how style-dependent variability in vowel production depends on the talker, and (b) whether speaker-specific characteristics are best indexed in read or spontaneous speech, and by specific vowels.

2. Method

2.1.1. Speakers and speech material/style

Twenty-three French speakers were selected from the PTSVOX database [24]. This database is specifically designed to study the factors of inter- and intra-speaker variability in forensic voice comparison. These factors include speaking style (reading or spontaneous speech), recording equipment (microphone or telephone), sex, and various information about the speaker (smoking, health issues, etc.). For this study, we selected microphone recordings.

The selected cohort included 12 male and 11 female French speakers, between 18 and 24 years-of-age, all students in a police school. All were recorded on a minimum of two sessions in two speech tasks: while reading three short texts (Read speech) and during an interview about their studies and occupations (Spontaneous speech).

10.21437/Interspeech.2024-310

After a manual orthographic transcription of the recordings, the speech files were automatically aligned at the phoneme and word levels, and these alignments were manually checked and corrected.

The 9 French vowel categories have been selected for analysis: the 6 oral vowels: /i, y, u, E, O, a/ (with E and O representing archiphonemes for respectively /e, ɛ/ and /o, ɔ/), and the 3 nasal vowels: /ɛ̃, ɔ̃, ɑ̃/. A total of 385,724 vowel tokens were extracted: 311,064 in spontaneous speech and 74,660 in read speech. Table 1 summarizes the distribution of occurrences by speech style, vowel and speaker sex.

Table 1: *Number of vowel tokens for each vowel category and speech style (read or spontaneous speech), for male (M) and female (F) speakers.*

Vowel category	Read speech		Spontaneous speech	
	F	M	F	M
i	3,396	3,740	13,760	20,584
y	2,420	2,708	7,184	10,424
u	1,460	1,580	6,664	8,656
E (e/ɛ)	10,192	11,092	39,164	54,228
O (o/ɔ)	4,864	5,388	9,408	14,128
a	7,800	8,476	27,972	39,804
ɛ̃	1,256	1,384	5,348	7,844
ɔ̃	1,972	2,200	8,280	12,688
ɑ̃	2,272	2,460	10,436	14,492

2.1.2. Acoustic analysis

In order to characterize the acoustic properties of the vowels, we used MFCC features instead of the classical phonetic formant features. Following [25], motivations for choosing this multidimensional feature space typically used in speech technology are (a) MFCCs produce a more extensive representation of vowel quality than that obtained with a formant features (which target more specifically phonemic contrastive aspects of the vowels), (b) extraction of MFCC features requires less manual correction to avoid erroneous formant estimation, (c) MFCCs allows the inclusion of nasal vowel categories, for which poles and zeros need to be accounted for.

On each target vowel, 12 MFCCs (excluding coefficient 0 related to the overall level of energy) are extracted with a custom Praat script on a 15 ms frame, centered on the middle of the vowel, using a filter bank spaced by 100 Mel.

2.1.3. Between-speaker and within-speaker distances as proxies of speaker distinctiveness and style-dependent variability

In each speech style, each speaker is characterized by a metric, the between-speaker distance, capturing the acoustic distance between her/his vowels and the vowels of the other speakers of the same sex. As such the 'Between-Speaker_distance' is a proxy of the speaker's distinctiveness within the pool of speakers, based on her/his vowel production. Each speaker is also characterized by a within-speaker distance, which captures the acoustic distance between her/his vowels in read speech and her/his vowels in spontaneous speech. As such the 'Within-Speaker_distance' is a proxy of the variability of the talker's vowel system between the two styles.

These distance metrics have been computed according to the following procedure:

(a) In order to account for the variety of segmental contexts in which the vowels occur in the spontaneous condition and the discrepancy in the number of samples produced per speaker, tokens are sampled according to the distribution of vowel phonemic contexts. The following context categories are defined: consonants are coded with their place of articulation, vowels and glides as anterior and posterior, and pauses are considered as a category. Left and right contexts are considered all together with no coding of order, so that a [labial_velar] context is considered equivalent to a [velar_labial] one for instance. For each style and speaker, contexts in which less than 8 vowel tokens occur are excluded. As a result, 62 different contexts are considered for spontaneous speech, and 45 for read speech. The frequency of occurrence is computed for each speaker, context and speech style in order to account for the large variability of the frequency across contexts. Indeed, the frequency of contexts ranges from 0.007% to 19% in read speech and from 0.001% to 18% in spontaneous speech, with 32% to 35% of [dental_labial] and [dental_dental] in both speech styles.

(b) For each style and each speaker, a category centroid is computed for the 9 French vowels categories /i, y, u, E, O, a, ɛ̃, ɔ̃, ɑ̃/ in each of the defined contexts (e.g, a centroid representing the average of all the /a/s in a [labial_velar] context of speaker A in Read condition).

(c) Comparisons between speakers are thus done between vowels occurring in the same context. For each vowel and context, the Euclidean distance on the 12-dimension MFCC space is computed between the speaker's centroid and the centroid of all the other speakers of the same sex. Thus, each speaker is characterized in each speech style by a set of 'Between-Speaker_distances', i.e. 993 to 2182 distances corresponding to 9 vowels categories * N number of available contexts * K-1 (with K=12 for male, 11 for female) speakers.

(d) Comparisons between the two speech styles follow the same principle: each speaker is characterized by a set of 'Within-Speaker_distances' corresponding to the 9 vowels * 45 contexts shared between the two styles * 2 speech styles.

3. Results

Separate models were fitted for male and female speakers. In each model, distance values were log-converted to account for the asymmetric distribution of data. First a linear mixed model was built using the R package *lme4* [26] to test whether the *Between-Speaker distances* (assessing the distinctiveness of each speaker's vowel characteristics) vary according to STYLE (read vs. spontaneous), the TALKER and the 9 VOWEL CATEGORIES, taken in interaction and as fixed factor, with a random intercept accounting for the second talker included in the comparison (lmer(log(BETWEEN-SPK_DIST) ~ STYLE * TALKER1 * VOWEL + (1|CONTEXT) + (1| TALKER2).

A second model was fitted to test whether the *Within-speaker distances* (corresponding to the comparison between the 2 speech styles) vary according to vary according to the TALKER and the 9 VOWEL CATEGORIES, taken in interaction and as fixed factors, with contexts as random intercept (lmer(log(WITHIN-SPK_DIST) ~ STYLE * TALKER * VOWEL + (1|CONTEXT)). To account for the unbalanced contexts, observations were weighted by their frequency of occurrence in all linear mixed effects models. For each fitted model,

diagnostic plots of residuals were visually inspected to ensure that there were no obvious violations of homoscedasticity or normality of residuals. The significance of fixed effects and interactions is assessed by likelihood ratio tests in which the full linear mixed model is compared with the same model without the evaluated fixed factor [27]. Estimated marginal means are computed for each style, speaker and vowel and used in pairwise comparisons.

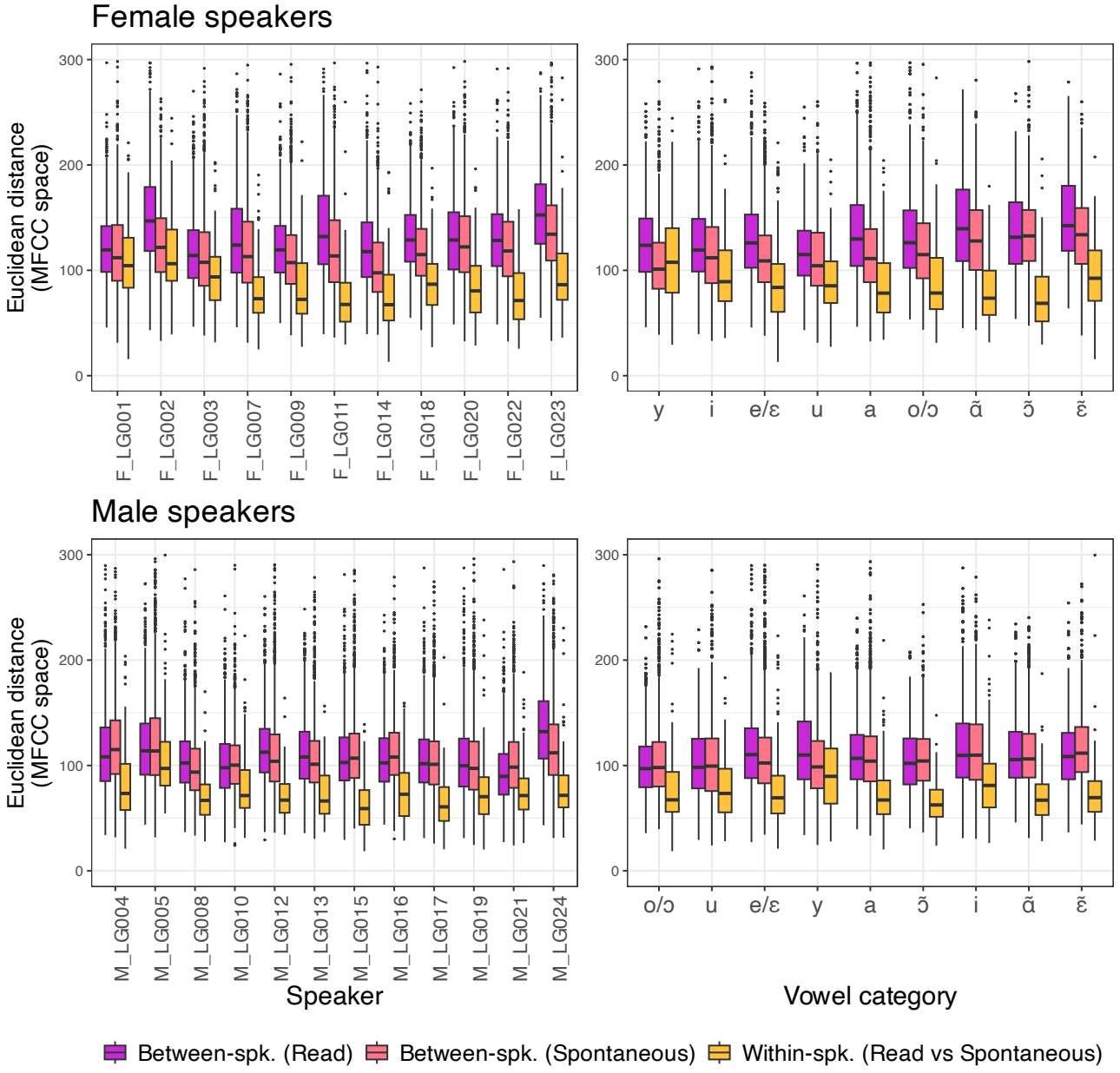

Figure 1: *Distribution of distances for the female (top panels) and male (bottom panels) speakers, by talker (left panels) and by vowel (right panels). Between speaker distance (i.e. speaker distinctiveness within the speaker pool) in Read (purple) and Spontaneous (pink) styles. In orange, within-speaker distance (i.e. distinction between read and spontaneous speech). On the right panels, vowel categories are ordered from left to right by ascending estimated marginal mean of between-speaker distances (averaged over Read and Spontaneous speech) for each speaker sex.*

Figure 1 presents the distribution of between- and within-speaker distances for the female and male speakers respectively. Vowel-wise distributions are displayed with all speakers pooled (on the right panels) and speaker-wise distributions with all vowel categories pooled (on the left panels), however interactions are also considered in the results presented in the following sections.

3.1.1. Vowel and talker-dependent effect on speaker's distinctiveness

As illustrated by the purple and pink boxes in the left panels of Figure 1, the between-speakers distances (indexing the talker's distinctiveness) are found to be affected by speech style with a significant interaction with the talker (F: $\chi^2(90) = 462.8$,

p<.001; M: $\chi^2(99) = 1181.6$, p<.001). For most talkers, between-speaker distances are larger in read speech than in spontaneous speech. This tendency is shown by all the female speakers and 7 out of the 12 male speakers. For the remaining male speakers, there is either no difference in distinctiveness across styles (M-LG010 and M-LG015), or a significant reverse tendency with more distinctiveness in spontaneous speech (M-LG004, M-LG016 and M-LG021).

The purple and pink boxes in the right panels of Figure 1 reflect the speaker-discriminant properties of individual vowels in the two speech styles. Tendencies appear to be dependent on sex and style. For the female speakers, the three nasal vowels /ɛ̃, ɔ̃, ɑ̃/ show the larger between-speakers distances in both read and spontaneous speech. For the male speakers, the same set is found in spontaneous speech, but vowels /ɑ̃, ɛ̃, i/ present the larger between-speakers distance in read speech.

3.1.2. Vowel- and talker-dependent effects on variability between speech styles

Both models fitted on within-speaker distances for male and female speakers show a significant interaction between the TALKER and VOWEL factors (F: $\chi^2(80) = 230.5$, p<.001; M: $\chi^2(88) = 173.6$, p<.001), showing that variability between read and spontaneous speech is both vowel- and talker-dependent. Comparison of likelihood ratios show that the TALKER effect is larger than the VOWEL effect, for both female (291 vs. 120) and male speakers (303 vs. 107).

In the left panels of Figure 1, the within-speaker distances (indexing the talker's variability between read and spontaneous speech) in the orange boxes vary according to the talkers. Nonetheless, for most talkers (except F-LG001 and M-LG005, for whom significant differences are found only in Read speech), the variability between styles is significantly smaller than the variability between-speakers in both Read and Spontaneous speech (p<.001 for all pairwise comparisons).

Also illustrated by the orange boxes in the right panels of Figure 1, the nine vowel categories are not equally variable according to speech style in our corpus. The vowels /y, ɛ̃/ for female speakers and the vowels /u, y/ for male speakers are the most variable between the read and spontaneous styles. For all vowels, the within-speaker variability according to style is smaller than the between-speakers variability in both Read and Spontaneous speech (p<.001 for all pairwise comparisons, except /y/ for female speakers in spontaneous speech, p=.032).

4. Discussion

In this study, we examined quite a large number of vowels produced by twenty-three French speakers in order to question how speaker-specific acoustic characteristics, which convey his/her distinctiveness in the speaker pool, are related to the talker itself, to the speech style and to the set of vowels considered.

First, the speaker's distinctiveness –in terms of acoustic distance between his/her vowels and that of the other speakers in the pool– is found to be greater in read speech than in spontaneous speech. Considering that the acoustic space of vowels is usually larger in read speech, with more peripheral acoustic realizations of the vowels [e.g. 15, 16, 17], it is not surprising that acoustic distances between speakers are larger. Overall larger acoustic distances are also found for female speakers which are also known to have a larger acoustic space than male speakers [4, 28, 29]. Nonetheless, style-related differences in the speaker's vowel distinctiveness cannot be reduced to a difference in acoustic space. Indeed, in [23], a mismatch in speech style was found to lower discriminatory performances in terms of speaker comparison. Interestingly, in this study, the discrimination power of vowels was worse in the interview condition (corresponding more or less to our spontaneous style) than in an interactive dialogue between familiar interlocutors. Although not reported in this study, we can assume that if the acoustic space is reduced in one of the two styles compared, this reduction is probably observed in the style for which the speakers were best discriminated, i.e. the interactive dialogue.

The second interesting result of this study is that speaker-specific characteristics are best indexed by specific vowels as found in [30] for instance. Indeed, the 9 vowel categories tested are not equally performing for contrasting speakers in the two speech styles. As found in [11], nasal French vowels are good candidates for speaker discrimination. Indeed, they combine the advantages of vowels, which are good candidates for estimating individual characteristics relating to the shape of the vocal tract [4-10], and those of nasal consonants, which relate both to the morphology of the nasal tract and to individual specificities in the timing of velar movements [31, 32]. Interestingly, among the nasal vowels which are the best candidates to index speaker specific characteristics, /ɔ̃/ and /ɑ̃/ show little variability according to style. In speaker verification or identification protocols, speech extracts containing these vowels should be favored if a choice is feasible.

This study also shows that style-dependent variability in vowel acoustics depends on the talker. This is not surprising since a large variety of speech adaptive strategies can be adopted by speakers [33, 34]. More interestingly, within-speaker variability according to the speech style is smaller than the variability between speakers for most of the speakers included in the pool.

5. Conclusion

To conclude, the fact that speaker's vowels are best contrasted in read speech in our study could have implications for applications in forensic contexts, where the recordings to be compared are usually acquired in different speech situations. Indeed, the questioned recording originated from a wiretap or crime scene is rather produced in a spontaneous speech style, while the reference recording is often a read version of the same content obtained later on. Nonetheless, the real implications of these results in terms of speaker identification or verification, both in automatic application but also by human listeners need to be tested in further studies. Indeed, [35] for instance, showed that performances of verification and identification systems are affected by variations in style of the speech material used, but only with a small loss in accuracy.

6. Acknowledgements

This work was supported by the research program ANR17-CE39-0016 (VoxCrim) and by the "Investissements d'Avenir" program ANR-10-LABX-0083 (Labex EFL). It contributes to the IdEx Université Paris Cité - ANR-18-IDEX-0001.

This research was partly funded by ChaSpeePro (CRSII5_202228) of the Swiss National Science Foundation.

7. References

[1] P. Ladefoged and D.E. Broadbent. “Information conveyed by vowels,” *The Journal of the Acoustical Society of America*, 29(1). pp. 98–104, 1957.

[2] P. French, P. Foulkes, P. Harrison, V. Hughes, E. san Segundo and L. Stevens. “The vocal tract as a biometric: output measures, interrelationships, and efficacy,” In *Proc. 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, Aug. 2015.

[3] R.D. Kent and C. Rountrey “What Acoustic Studies Tell Us About Vowels in Developing and Disordered Speech,” *American Journal of Speech-Language Pathology.* Aug 4;29(3), pp. 1749-1778, 2020.

[4] S. P. Whiteside. “Identification of a speaker’s sex : a study of vowels,” *Perceptual and Motor Skills,* 86(2), 579-584, 1998.

[5] J. J. Wolf. “Efficient acoustic parameters for speaker recognition,” *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[6] H. Cao and V. Dellwo. “The role of the first five formants in three vowels of mandarin for forensic voice analysis,*”* in *Proc. 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia, Aug. 2019, pp. 617-621.

[7] M. Antal and G.G. Toderean, “Speaker recognition and broad phonetic groups,” in *SPPRA*, pp. 155–159, 2006

[8] K. Amino, O. Takashi, K. Kamada, M. Hisanori, and A. Takayuki. “Effects of the phonological contents and transmission channels on forensic speaker recognition,” in *Forensic Speaker Recognition*, pp. 275–308, Springer, 2012.

[9] K.K. Paliwal, “Effectiveness of different vowel sounds in automatic speaker identification,” *Journal of Phonetics*, 12 (1), pp. 17-21, 1983

[10] J.P Eatock and J.S. Mason, “A quantitative assessment of the relative speaker discriminating properties of phonemes,” in *Acoustics, Speech, and Signal Processing- ICASSP- 94,* vol. 1, pp. I–133.1994

[11] M. Ajili, J-F Bonastre, W. Ben Kheder, S. Rossato, and J. Kahn. Phonetic content impact on Forensic Voice Comparison. In *Proc. 2016 IEEE Spoken Language Technology Workshop (SLT)*, San Diego (CA), USA, Dec. 2016, pp. 210-217

[12] F. Nolan. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press. 1983

[13] P. Foulkes and P. French. “Forensic Speaker Comparison: A Linguistic-Acoustic Perspective,” In: Solan L.M., Tiersma, P.M. (eds), *The Oxford Handbook of Language and Law*. Oxford University Press: Oxford. pp. 557–573, 2012.

[14] B. Harmegnies and D. Poch-Olivé. “A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish”. *Speech Communication*, 11, pp. 429–437, 1992.

[15] C. Gendrot and M. Adda-Decker. “Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German”. In *Proc. Interspeech 2005,* Lisbon (Portugal), Sept. 2005, pp. 2453–2456.

[16] C. Meunier and R. Espesser. “Vowel reduction in conversational speech in French: The role of lexical factors”. *Journal of Phonetics*, 39 (3), pp. 271-278, 2012.

[17] N. Audibert, C. Fougeron, C. Gendrot, and M. Adda-Decker. “Duration- vs. Style-Dependent Vowel Variation: a Multiparametric Investigation,” In *Proc. 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, Aug. 2015.

[18] S.-J. Moon and B. Lindblom. “Interaction between duration, context, and speaking style in English stressed vowels”. *The Journal of the Acoustical Society of America*, 96, pp. 40–55, 1994.

[19] R. Smiljanic and A. Bradlow. “Production and perception of clear speech in Croatian and English”. *The Journal of the Acoustical Society of America*, 118, pp. 1677–1688, 2005

[20] J. Wilson Black, J. Hay, L. Clark, and J. Brand. "The overlooked effect of amplitude on within-speaker vowel variation" *Linguistics Vanguard*, vol. 9, no. 1, pp. 173-189, 2024.

[21] J. Kahn, N. Audibert, J.-F. Bonastre, and S. Rossato. “ Inter and intra-speaker variability in french: an analysis of oral vowels and its implication for automatic speaker verification”. In *Proc. 17th International Congress of Phonetic Sciences (ICPhS)*, Hong-Kong, China, Aug. 2011, pp. 1002-1005.

[22] J. Dankovičová and F. Nolan. “Some acoustic effects of speaking style on utterances for automatic speaker verification”. *Journal of the International Phonetic Association*, 29(2), pp. 115–128, 1999

[23] J.C. Cavalcany, A. Eriksson and P.A. Barbosa. “On the speaker discriminatory power asymmetry regarding acoustic-phonetic parameters and the impact of speaking style”. *Front. Psychol.* 14:1101187, 2023.

[24] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, “PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire,” in *Proc. Journées d’Études sur la Parole*, Nancy, France, Jun. 2020, pp. 73–81.

[25] E. Ferragne and F. Pellegrino, “Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics,” *Journal of Phonetics*, vol. 38, no. 4, pp. 526–539, 2010.

[26] 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, Oct. 2015.

[27] B. Winter. *Statistics for linguists: An introduction using R*. Routledge, 2019.

[28] S.P. Whiteside. “Gender-specific fundamental and formant frequency patterns in a cross sectional study,” *The Journal of the Acoustical Society of America*, 110, pp. 464-478, 2001

[29] M. Weirich and A.P. Simpson. “Investigating the relationship between average speaker fundamental frequency and acoustic vowel space size,” *The Journal of the Acoustical Society of America*, 134(4), pp. 2965-2974, 2013

[30] G. de Jong, K. McDougall, T. Hudson, and F. Nolan. “The speaker discriminating power of sounds undergoing historical change: a formant-based study,” In *Proc. 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, Aug. 2007, pp 1813-1817.

[31] J. Dang and K. Honda. “Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation,” *The Journal of the Acoustical Society of America*, 100(5), pp. 3374-3383, 1996.

[32] K. Amino and T. Arai. “Speaker-dependent characteristics of the nasals,” *Forensic science international*, 185(1-3), pp. 21-28, 2009

[33] S.H. Ferguson and D. Kewley-Port. “Talker Differences in Clear and Conversational Speech:Acoustic Characteristics of Vowels”. *Journal of Speech,Language, and Hearing Research*, Vol. 50, pp. 1241–1255. 2007

[34] D. D'Alessandro, A. Bourbon, and C. Fougeron. “Effect of age on rate and coarticulation across different speech-tasks”. In *Proc. 12th International Seminar on Speech Production*, New Haven (CT), USA, Dec. 2020, pp 14-18.

[35] M. Grimaldi and F. Cummins. “Speech style and speaker recognition: a case study,”. In *Proc. Interspeech 2009*, Brighton, UK, Sept. 2009, pp. 920-923.