DURATION- VS. STYLE-DEPENDENT VOWEL VARIATION: A MULTIPARAMETRIC INVESTIGATION

Nicolas Audibert[1], Cécile Fougeron[1], Cédric Gendrot[1], Martine Adda-Decker[1&2]

[1] Laboratoire de Phonétique et Phonologie, UMR7018 CNRS/Univ. Paris 3 Sorbonne-Nouvelle
[2] LIMSI/CNRS
{nicolas.audibert, cecile.fougeron, cedric.gendrot, martine.adda-decker}@univ-paris3.fr

ABSTRACT

This study examines phonetic variation in the acoustic properties of the French /i,e,a,o,u/ sub-system as a function of vowel duration and speech style in order to better understand the interplay between these two well-known factors of vowel reduction. Over 1000k vowels extracted from three large corpora of continuous French including read speech (BREF), partly scripted journalistic speech (ESTER) and casual conversations (NCCFr) were split into duration classes (short, medium, long). Six metrics based on F1 & F2 frequencies were used to capture variation in multiple dimensions. Style and duration show comparable effects in terms of reduction in the acoustic working space (overall and in F1 or F2 dimensions), centralization of vowel categories and variability within vowel categories. However, interesting differences are found in terms of overlap between vowel categories: more neutralization of vowel contrasts independent from vowel duration is found in casual conversational speech.

Keywords: Style, vowel duration, French, reduction, acoustic metrics.

1. INTRODUCTION

Phonetic forms of segments are known to be modulated according to speech style and communication situation, and to be function of segment duration, two factors which are not independent from each other. Differences in speech style, a general term embedding very different aspects, can be conceived as differences along a continuum ranging from clear speech (where acoustic cues are made to be maximally informative for the ease of intelligibility and segmental distinctiveness) to casual interactional speech (where situational, performance, and communicative requirements may be met at the expense of clarity at the segmental level). This continuum can also be seen as a modulation between hyper- and hypo-articulated segmental outputs. Differences in phonetic forms according to speech styles have been reported along different acoustic dimensions, including variations in the acoustic properties of vowel segments with changes in both temporal and spectral properties. Casual speech, for instance, has been described to be prone to reduction in vowel duration and in vowel spectral properties [2, 8, 13, 17], while clear speech has been characterized by an increased duration of vowels [3, 4, 18, 21] as well as an increased size of the vowel acoustic space [12, 21, 24]. This link between temporal and spectral changes according to speech style is usually accounted for by the biomechanical trade-off between articulation time and articulatory precision. Undershoot is then explained by a lack of time and dynamic adjustment to reach articulatory, and thus acoustic, segmental targets [15, 18]. Different large corpora studies on vowel reduction in French [8, 17] have indeed reported a progressive acoustic vowel space reduction when moving from long to short vowel durations, within a given speech style.

However, one may wonder whether this temporal/spatial trade-off is the sole responsible of vowel reduction in casual speech. Several studies have shown interesting interactions between style and duration on acoustic reduction in French. In their study based on MFCC-based parameters, [19] have shown that differences in spectral reduction could be found between read and conversational speech only for short segments. Comparison between the patterns of acoustic space reduction observed according to vowel duration in [3] and [17] also suggests style-dependent interaction. Looking at variation between the same duration classes (long vs. mid vs. short vowels), the first study is based on scripted journalistic speech, while the later is based on vowels produced in a casual interaction between friends. If both studies report an overall reduction of the vowel acoustic space, vowels in casual speech ([17]) present a global centralization of most vowel categories, while journalistic speech ([3]) features most reduction along the F1 dimension, due to a massive reduction on F1 for /a/.

The present paper further explores the relationship between duration-dependent and style-dependent vowel reduction in three large corpora of continuous French, produced in distinct situations: laboratory read speech, radio broadcast news, casual face-to-face conversations. They therefore vary in many aspects: type of speakers (professional or not),

interactive/communicative goals, degree and type of planning (reading vs. scripted vs. spontaneous), and linguistic content (written press, broadcasted news, debate between friends).

A further original aspect of our study consists in looking at various possible dimensions of variation of a vowel system (here a subset of the French vowel system), capitalizing on the fact that reduction or expansion of a vowel space is "neither uniform nor simple" ([3], see also [4, 10]). For this, various metrics based on F1/F2 are used to describe and quantify variation in terms of working space reduction (overall, F1- and F2- dimensions), of centralization of vowel categories, of exemplar dispersion within vowel categories, and of neutralization of contrasts between categories.

2. SPEECH MATERIAL AND METRICS

In order to explore style- and duration-based vowel variation, naturally/un-controlled continuous speech produced by a rich number of speakers is studied. Productions extracted from three publicly available French corpora with various content are used. *Read speech* comes from the French BREF corpus [14] composed of read parts of the newspaper *Le Monde* produced by non-professional speakers in a laboratory setting. This reading task was considered as difficult for most of the speakers. *Prepared speech* corresponds to partly scripted speech samples originated from broadcasted news and political/societal debates found in various radio and TV programs within the ESTER corpus [6, 7]. Professional speakers produce the largest part of this quite careful speech material. Finally, *conversational speech* produced in a casual face-to-face setting comes from the NCCFr corpus [24], which includes partly free and partly guided discussions on societal topics between friends.

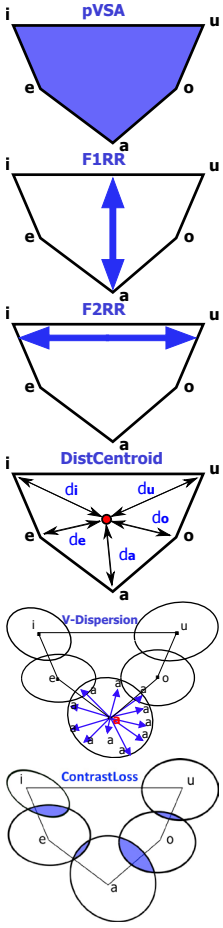
A sub-system of French vowels extracted from these corpora is investigated. It included exemplars of the oral vowels /i, e, ɛ, a, o, ɔ, u/. Vowels /e, ɛ/ and /o, ɔ/ were merged into single classes, hereafter labelled /e/ and /o/, to account for their frequent neutralization and regional variation in French.

A total of 1 143 941 vowels were included in the analysis. Vowels were subdivided into three duration classes: short (up to 50 ms), mid (up to 80 ms), and long (up to 300 ms), following the criteria applied in [8] and [17]. They are produced by about 50 different speakers in each corpus (74 for BREF, 61 for ESTER, 45 for NCCFr) with a balance between male and female speakers. Speakers were selected based on the amount of vowel exemplar available (at least 15 for the least frequent vowel, /u/, in each duration class).

While the ESTER and NCCFr corpora show a similar distribution of vowels into the three duration classes (52%-30%-18% of short, mid and long vowels for ESTER, 50%-29%-21% for NCCFr), short vowels are slightly under-represented in the BREF corpus (22% short, 44% mid, 34% long).

F1/F2 formant values were extracted using Praat [16] on a forced automatic alignment, taking the average of values extracted at 1/3, 1/2 and 2/3 of each vowel. A filter was applied to discard aberrant formant values according to the same criteria used in [8]. In the three corpora, for each speaker and each duration class, the acoustic metrics described in Figure 1 were computed from F1 and F2 frequencies.

Figure 1: Description of the six acoustic metrics used to quantify vowel space variation.



pVSA: *Vocalic Space Area (pentagon)*. Surface of vocalic shape defined by the /i,e,a,o,u/ pentagon in F1/F2 plane. Vowel positions are computed as weighted centroids of measured tokens.

F1RR: *First formant range ratio*. Designed to measure the ambitus in the F1 dimension (jaw/tongue height), see [22].

F2RR: *Second formant range ratio*. Designed to measure the ambitus in the F2 dimension (front-back tongue dimension and/or lip rounding), see [22].

DistCentroid: Vowel Distance to Centroid A common metric computing the average distance of the 5 vowels' centroids to the overall centroid of the speaker's vowel space.

V-Dispersion: Vowel Dispersion Area Averaged intra-category vowel dispersion: area of the 95% confidence ellipses.

ContrastLoss: Contrast between vowel categories (or system distinctiveness) Misclassification rate in a linear predictive model of vowel class membership based on F1 and F2 values (see [10]).

3. RESULTS

A linear mixed effect model, with corpus (BREF, ESTER, NCCFr) and duration class (short, intermediate, long) as fixed effects and speaker as a random effect, is used to predict variation on acoustic metrics.

Significant main effects are found for both duration and corpus predictors (all $p<10^{-5}$) on all the metrics. Statistical results are summarized in Table 1 and differences are illustrated in the boxplots of

Figure 2. In order to compare the influence of duration and style on metrics values, effects sizes (estimated by χ^2 values in the likelihood ratio tests, given in Table 1) are compared, indicating a larger effect of duration class for all metrics except ContrastLoss.

Corpus effect: All duration classes pooled, the corpus is found to be a significant predictor of acoustic variation for all the metrics studied with two- or three-ways distinctions between corpora. In particular, the NCCFr corpus is always distinct from the other two. In all duration classes the corresponding vowel space is more reduced, both overall (pVSA) and in specific dimensions (F1RR, F2RR), in the casual-speech corpus (NCCFr) than in the more formal speech corpora (ESTER and BREF not being distinguished when compared by duration classes). Vowel categories are also more centralized (DistCentroid) and vowel tokens within each vowel categories are more variable (V-Dispersion) in casual speech compared to more formal speech (ESTER and BREF). The measure of within-system distinctiveness, assessed as the rate of token misclassification (ContrastLoss), shows a similar pattern with more systematic differences along the casualness continuum. For each duration class, vowel tokens are more accurately classified (i.e. less misclassified, less overlapping in acoustic shape across vowel categories) in read speech (BREF) than in journalistic speech (ESTER), while misclassifications are the most frequent for vowels produced in the conversational speech corpus (NCCFr).

Figure 2: Boxplots for the 3 corpora NCCFr, ESTER, BREF (casual/prepared/read) and 3 duration classes (short, medium, long) for the 6 acoustic metrics: pVSA, DistCentroid, F1RR, F2RR, V-Dispersion, ContrastLoss.

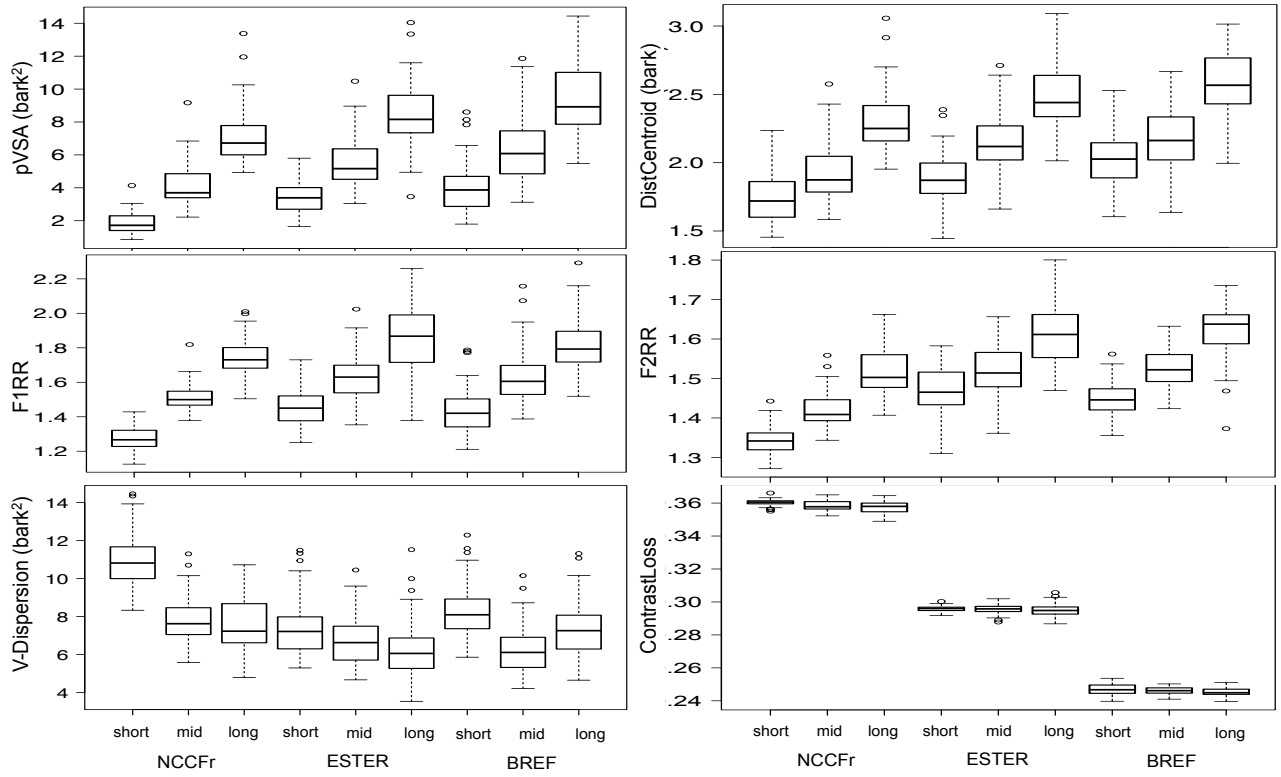


Table 1: Summary of main effects and interaction for linear mixed models with each of the 6 metrics as dependent variable. Model specification: corpus and duration class as fixed effect, speaker as random effect. Stars indicate significant differences ($p<10^{-4}$). χ^2 values with two degrees of freedom are reported as effect size estimates.

	Main effects				Interactions	
Metrics	Corpus effect	χ^2	Duration class effect	χ^2	Duration class effect per corpus	Corpus effect per duration class
pVSA	NCCFr<*ESTER<*BREF	68	short<*mid<*long	1009	in the 3 corpora: short<*mid<*long	in 3 duration classes: NCCFr<*ESTER=BREF
DistCentroid	NCCFr<*ESTER<*BREF	69	short<*mid<*long	1045		
F1RR	NCCFr<*ESTER=BREF	39	short<*mid<*long	950		
F2RR	NCCFr<*ESTER=BREF	119	short<*mid<*long	689		
V-Dispersion	NCCFr<*ESTER=BREF	64	short<*mid=long	332	in the 3 corpora: short<*mid=long	
ContrastLoss	NCCFr>*ESTER>*BREF	1290	short>*mid>*long	25	in the 3 corpora: short>*long	in 3 duration classes: NCCFr>*ESTER>*BREF

Duration-class effect: All corpora pooled, vowel duration is also a significant predictor of vowel variation on all metrics. A three-way distinction between short, mid and long vowels is found showing a larger acoustic space reduction (pVSA, F1RR, F2RR) and centralization (DistCentroid) from the long to the mid to the short vowel classes in the three corpora. Variability within vowel categories (V-Dispersion), however, significantly increases only for short vowels, as compared to the other duration classes, in the three corpora. A smaller effect of duration is found on distinctiveness between vowel categories (ContrastLoss), with a larger misclassification rate found for short vowels compared to long vowels only.

4. DISCUSSION

The five metrics measuring reduction and centralization of the vocalic system, as well as intra-class dispersion, show that similar reduction patterns are observed according to style and vowel duration. Vowels produced in the NCCFr casual speech corpus and vowels with shorter durations are more prone to reduction in the F1/F2 acoustic space, are less peripheral, have more unstable acoustic targets, and have more overlapping distributions reducing the accuracy of their classification into their respective vowel categories. However, interactions between style and duration factors show that variations differ in magnitude when vowels of the same duration class are compared across corpora. Effect size comparisons also show that style is a predominant predictor of variation as compared to vowel duration.

The effect of style appears mainly between two categories of corpora: the casual interactive speech corpus (NCCFr) and the more 'clear speech' corpora (BREF and ESTER). Some distinctions appear between BREF and ESTER when all vowel durations are pooled, but disappear when controlling for duration class (for pVSA and DistCentroid). To conclude that there is a difference between interactive vs. non-interactive speech situation is tempting. However, if we can ascertain that in the NCCFr corpus the interlocutors are conversing which each other, the content of the ESTER corpus would need a further comparison between journalistic monologues and broadcasted debates (even though such debates generally leave little place for interactions!).

The inclusion of a metric assessing the degree of distinctiveness between vowel categories (ContrastLoss) is particularly interesting for a multidimensional investigation of reduction. Indeed, it is the only metric showing a three-way distinction between the three corpora for each duration class. ContrastLoss is also found to be marginally affected by duration, as compared to the leading role played by this particular aspect in speaking style differences. Examination of misclassification scores between vowel categories shows that for all corpora and duration classes, overlap is largest for vowel /u/, which also has the largest intra-class dispersion values. However, inter-corpus variation is mainly linked to misclassification of /o/ and /i/ vowels.

In this study, we examined the effect of speech style and segmental duration on vowels produced in a natural environment rather than relying on instructed laboratory-induced variations as done in numerous studies (see e.g. [3, 11, 12, 18]). This large corpus-based approach makes the analysis of naturally occurring reduction phenomena feasible, but it does not allow for the control of all variables known to affect vowel variation. In the context of acoustic metrics computation, a strict control of segmental context would imply selecting only speakers with the same context distribution in each corpus and duration class. Examination of left and right segmental context distributions in each corpus and duration class indicates similar distributions between BREF and ESTER, but a different distribution in the NCCFr corpus (with an underrepresentation of alveolar contexts for vowels /a,e,i,o/ compared to the two other corpora). Although such differences might account for part of the variation measured between casual and formal speech, they are unlikely to explain all of it, particularly concerning the within-system distinctiveness (ContrastLoss) for which three-way distinctions are found between the three corpora.

5. CONCLUSION

Altogether, the present results outline the need for a multidimensional approach to quantify the various qualitative changes that a vocalic system can undergo when studying phonetic reduction. Style- and duration-dependent vowel variation was assessed in this study on over 1000k French vowels on dimensions linked to acoustic working space reduction, dispersion and distinctiveness. In each duration class, results indicate more reduction of the acoustic space, a larger degree of vowel centralization, and more intra-category dispersion in casual speech as compared to the other two more formal speech styles. More overlap between vowel categories is also found in casual speech and in journalistic speech compared to read speech. Duration-dependent variations modulate the vowel system in the same way for most dimensions, showing a three-way or two-way distinction between short and longer vowels, but to a smaller extent.

Acknowledgments: *This work was supported by the French ANR project Typaloc (ANR-12-BSH2-0003) and by the Labex EFL program (ANR-10-LABX-0083).*

7. REFERENCES

[1] Adank, P., Smits, R.,and van Hout R., 2004. A comparison of vowel normalization procedures for language variation research, *Journal of Acoustical Society of America,* 116, 5, pp. 3099–3107.

[2] Duez, D. 1995. On spontaneous French speech: Aspects of the reduction and contextual assimilation of voiced stops. *Journal of Phonetics*, 23 (4), 407–427.

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

[4] Ferguson, S. H., Kewley-Port, D. 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 112, 259–271.

[5] Fougeron, C., Audibert, N. 2011. Testing various metrics for the description of vowel distortion in dysarthria. *Proceedings of the 17th ICPhS*, 687-690.

[6] Galliano, S., Geoffrois, Gravier, G., E., Bonastre, J.-F., Mostefa, D., Choukri, K. 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. *Proceedings of European Conference on Speech Communication and Technology*, 139–142.

[7] Galliano, S., Gravier, G., Chaubard, L. 2009. The ESTER 2 evaluation campaign for the rich transcription of French broadcasts. *Proceedings of Interspeech*, Brighton (UK), 2583–2586.

[8] Gendrot, C., Adda, 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. *Proceedings of Interspeech*, Lisbon (Portugal), 2453–2456.

[9] Gordon, A. 2012. *The Effect of Speaking Style on Variability of Vowel Production for Native Monolingual English Speakers*. Undergraduate honors thesis, University of South Florida.

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

[11] Huet, K., Harmegnies, B., 2000. Contribution à la quantification du degré d'organisation des systèmes vocaliques. *Actes des XXIIIe Journées d'Etude sur la Parole*, 225–228.

[12] Johnson, K., Flemming, E., Wright, R. 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 69, 505–528.

[13] Johnson, K. 2004. Massive reduction in conversational American English *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*, Tokyo.

[14] Lamel, L., Gauvain, J., Eskenazi, M. 1991. BREF, a Large Vocabulary Spoken Corpus for French. *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Genova, Italy.

[15] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory, In Hardcastle & Marchal (eds): *Speech Production and Speech Modeling*, Dordrecht:Kluwer, 403–439.

[16] Boersma, P., Weenink, D. 2014. *Praat: doing phonetics by computer* [Computer program]. Version 5.4.04, retrieved 28 December 2014 from http://www.praat.org/

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

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

[19] Rouas, JL., Beppu, M., Adda-Decker, M. 2010. Comparison of Spectral Properties of Read, Prepared and Casual Speech in French. *International Conference on Language Resources and Evaluation (LREC 2010)*.

[20] Peterson, G. E., Barney, H. L. 1952. Control methods used in the study of the vowels, *Journal of the Acoustical Society of America*, 24, 175–184.

[21] Picheny et al 1986, Picheny, M. A., Durlach, N. I., Braida, L. D. 1986. Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29, 434–446.

[22] Sapir, S., Ramig, L., Spielman, J., and Fox, J. 2010. Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research*, 53(1): 114.

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

[24] Torreira, F., Adda-Decker, M., Ernestus, M., 2010. The Nijmegen corpus of casual French. *Speech Communication*, 52:201–212.