VOICING VARIATIONS IN FRENCH OBSTRUENTS: DISTRIBUTION AND ACOUSTIC QUANTIFICATION

Fanny Ivent, Martine Adda-Decker, Cécile Fougeron

Université Sorbonne Nouvelle - Paris 3; Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS
fanny.ivent@univ-paris3.fr; martine.adda-decker@univ-paris3.fr; cecile.fougeron@univ-paris3.fr

ABSTRACT

This study investigates phonetic variation of voicing correlates in French obstruents. A controlled set of words was extracted from a large broadcast news corpus providing 378 word-initial singleton obstruents /t, d, k, s/. An expert investigation, by eye and ear, of the productions revealed that phonologically voiced obstruents are phonetically voiced most of the time (96%), while a large number of phonologically voiceless obstruents are produced with a partial or complete phonetically voiced constriction, more frequently so for stops (74% /t/, 61% /k/) than for fricatives (30% /s/). In order to acoustically quantify our observations and establish reliable metrics for future larger-scale studies, 13 acoustic metrics were applied and tested. Out of these, three metrics were found to be particularly effective in accurately classifying our obstruents into the manually defined categories: an energy difference measure relative to the following vowel, an unvoiced/voiced frame ratio, and consonant duration.

Keywords: phonetic variation, voicing, obstruent, manual and automatic analysis, speech corpora.

1. INTRODUCTION

When dealing with speech variation, a major challenge consists in identifying the various sources of variation and in accurately describing how they impact speech production [15]. By modeling these factors, variation can be better explained and predicted to get a better understanding of speech production and perception processes, in healthy or pathological conditions for instance, or to improve automatic speech recognition systems. Studies on phonetic variation however are confronted to the fact that variation phenomena are diverse in nature, form and perceptual consequences. The availability of large datasets of natural speech recordings offers a good methodological test-bed for studying phonetic variation, but with the challenge of processing a large amount of data for which careful manual phonetic analyses become prohibitive in time and human labor. Speech corpora produced in more naturalistic conditions than laboratory-controlled settings are also known to offer a large inventory of pronunciation variants [9] calling for the need of defining relevant acoustic metrics able to localize and accurately describe the phonetic output.

The work proposed here is a first methodological step toward a larger-scale study on consonant variation in French in both healthy casual speech and pathological speech in motor disorders.

In contrast with Germanic languages, French is relatively poorly documented with respect to phonetic variation affecting obstruent singleton consonants. Indeed, most large-scale studies on continuous speech French corpora have dealt with vocalic variations [3,5,13,17], and studies on consonants have rather focused on variations linked to phonological processes, such as voice assimilation in consonant sequences [4,8,16]. Regressive C-to-C voice assimilation is a frequent process in French, but gradual voicing of voiceless consonant closures in word medial intervocalic context for instance, can be seen both in casual production of healthy speakers and can be quite pervasive in the speech of dysarthric speakers [11].

In this study, we focus on the variation affecting phonetic properties of voicing contrasts in French obstruents, with a twofold aim. First, on a controlled set of words selected in a natural corpus of continuous speech, we evaluate the frequency of occurrence of voicing alterations, i.e. phonetic voicing of voiceless obstruents and devoicing of voiced obstruent. Second, we test a selected set of automatable acoustic metrics adapted from the literature which can capture acoustic markers of vocal fold vibration and/or other temporal or energy cues known to be involved in the voice/voiceless contrast in French. The corresponding measurements are assessed for their potential to discriminate the cases of variation found in the data.

2. SPEECH MATERIAL

The productions used in this study were extracted from a sample of the French ETAPE corpus [7] including various broadcast programs (12 radio and 2 television programs), in which a fair amount of spontaneous speech produced by professional speakers (journalists or politicians) is available.

A set of minimally contrasting words were chosen in order to control for position in word and segmental context and to be able to make use of a fair amount of observable tokens. After examination of the lexical content of the recordings, the following words were selected: 'dans' (*in*) /dɑ̃/, 'temps' (*weather*) and 'tant' (*so much*) both /tɑ̃/, quand /kɑ̃/ (*when*), 'sans' (*without*) and 'cent' (*hundred*) both /sɑ̃/. As presented in Table 1, all together these words were produced 378 times, with about 100 occurrences per target consonants, by 7 to 39 different speakers, and they allowed for comparisons within a voiced/voiceless alveolar stop pair (/d/ vs. /t/), within an alveolar/velar voiceless stop pair (/t/ vs. /k/) and within a voiceless alveolar stop/fricative pair (/t/ vs. /s/). All consonants are word-initial and followed by the same vowel /ɑ̃/. Preceding contexts were not possible to control without reducing too much the amount of tokens. The test words occurred after a pause (10%), after a word ending by a vowel (48%) or by a consonant (42%).

Table 1: Description of the material in terms of number of tokens (N) and speakers (SPK).

	/dɑ̃/	**/tɑ̃/**		**/kɑ̃/**	**/sɑ̃/**	
	dans	temps	tant	quand	cent	sans
N	100	69	8	101	54	46
SPK	39	32	7	36	30	26

3. EXPERT CLASSIFICATION

3.1. Method and criteria

A manual annotation of the production was done by an expert phonetician (the first author) based on visual cues on both the signal and spectrogram, and on auditory impressions. Variations affecting the quality of the constriction for the stops and fricative were also annotated, but we will focus here only on variations linked to voicing. A categorical classification of the tokens was done according to whether the consonant deviated or not from a canonical production.

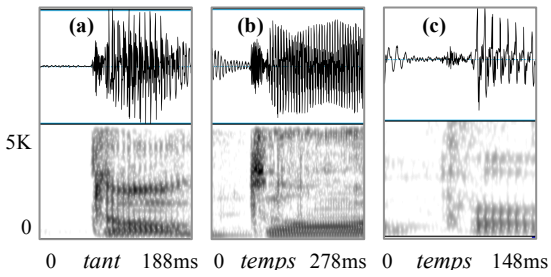
Note that in French, voiced stops are typically fully voiced, with vocal fold vibration throughout the full constriction period. Therefore, target voiced stops (e.g. /d/s) were categorized as phonetically voiced ([+v]) when produced with a periodic signal and a voiced bar throughout closure duration, and categorized as phonetically devoiced ([-v]) if signal periodicity was interrupted during closure, either completely or partially.

Voiceless stops (/t, k/) were considered as phonetically unvoiced ([-v]) if produced without any periodicity on the waveform and energy in the very low frequency band on the spectrogram. An example of such a rendition is given in Figure 1 (a). Voiceless stops were classified as phonetically voiced [+v], (see Figure 1b), if they presented periodicity in the signal and/or a voiced bar during the closing phase (which is assessed by a large drop in energy in the mid to high frequency band on the spectrogram -- note that the /ɑ̃/ context often entailed some nasality during C closure). Renditions with partial voicing during closure were also labelled [+v], even if this periodicity could be due to the voicing decay time of the preceding segments as shown in figure 1c. In this continuous speech corpus, consonants are quite short and this voice termination time often occupies a large portion of the consonant closure time (almost half of it in Figure 1c).

For the voiceless fricative (/s/), it was not always possible to judge periodicity in the noisy signal. The absence or presence of energy in the very low frequency band on the spectrogram (voiced bar) was thus a better criterion to classify phonetically unvoiced /s/ ([-v]) and voiced /s/ ([+v]), respectively. Again, partially voiced /s/ were considered as [+v].

Ambiguous cases were discussed between the authors and a forced classification was always done.

Figure 1: Examples of /t/ realizations. Phonetically [-v] in (a) and [+v] in (b) and (c)



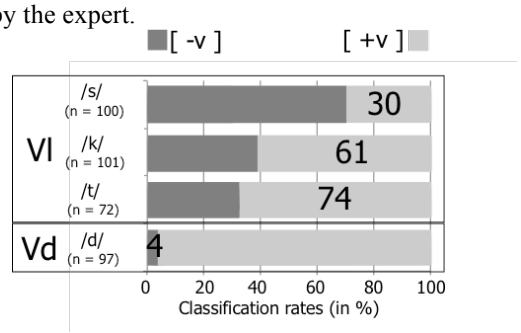
3.2 Classification results

Out of the 378 tokens, five of the /t/s had to be discarded due to the presence of external noise. For three of the /d/s, no acoustic cues for closure phase or release could be identified (nor perceived) and these cases were classified as deleted. The remaining 370 target consonants were classified as shown in Figure 2.

Surprisingly, a large number of underlyingly voiceless targets are found to be phonetically voiced ([+v]). Almost 3/4 of the /t/s (74%) showed partial or complete voicing during closure and were classified as [+v] according to our rather strict criteria. The alveolar stop is the most prone to voicing variation among the three voiceless obstruents. Nonetheless, more than half of the velar

voiceless stops /k/ (61%) features partial or complete voiced closures. For the fricative /s/, most of the renditions are phonetically voiceless, even though 30% present voicing cues during constriction.
If the phonetic voicing of underlying voiceless targets is quite frequent, the reverse is not true in our data. Only 4 renditions (4%) of /d/ have been found to be phonetically devoiced.

Figure 2: Distribution (in %) of the 370 target consonants classified in terms of phonetic voicing by the expert.



4. ACOUSTIC CLASSIFICATION

A manual description and classification of phonetic variants such as the one described above is not conceivable on larger-scale data. Automatable acoustic analysis has to be envisioned. Several acoustic parameters could be used to capture variation in phonetic voicing. To test these metrics, an acoustic study was carried out on the consonants showing the most voicing variation (the 270 stops (/t/, /k/), for which we also have a good stable reference (the voiced /d/). The fricative /s/ is discarded here as its acoustic characteristics are too far apart from those of the stops, adding difficulties in the comparisons of the voicing cues.

4.1 Acoustics metrics

Thirteen metrics have been applied to the manually segmented consonants in a semi-automated way using Praat [2]. Closure and burst phases were merged into a single interval.
First, consonant duration was measured, including both closure duration and burst (**target length).**
Second, a measure reporting the ratio of unvoiced frames over the total number of frames (**uv_ratio)** [6,10, see also 8,16 for the use of a voiced frame ratio] is computed with Voice Report function of PRAAT. **CV_uv** is a contextual extension of this metric corresponding to the difference between uv_ratios of the consonant and following vowel.
The other metrics have all been applied to a subregion in the middle of the obstruents. Values are then obtained as averages of 3 measurements taken in the middle and in two points at equal distance of ±7,5 ms from the middle). For the contextual metrics giving a measure on the consonant relative to the following vowel (CV_), the measurements in the vowel are made on a similar subregion in the vowel. Moreover, these metrics were computed both on the unfiltered signal and on a 0-500Hz low-pass filtered signal (_lf) in order to filter out potential effects of noisy incomplete closures.
Following [6,16], metrics based on signal energy were defined assuming that voicing into closure will increase the energy of the consonant. **Nrj** corresponds to the absolute intensity measured in dB while **nrj_lf** is limited to the low frequency band. **CV_nrj** gives the intensity difference between the obstruent and the following vowel [6] (same for **CV_nrj_lf** but in low frequencies**)**.
Following [1,6,10], Harmonics-to-noise ratios (**hnr** and **hnr_lf**) were obtained using the harmonicity object in PRAAT with the cross-correlational method (settings: time step 0.01, minimum pitch 75 Hz, silence threshold: 0.001, number of period window: 4.5). **CV_hnr** and **CV_hnr_lf,** give the hnr difference between the obstruent and the following vowel in the two (full and low) frequency bands.
Finally, a simple fundamental frequency measure (**f0_bin**) based on a cross-correlation f0 detection (with a pitch ceiling at 400Hz) was used. Continuous f0 values extracted with the PRAAT pitch object were binarized as [+v] when f0 > 75 Hz and [-v] when f0 ≦ 75Hz. This measure was also applied to the filtered signal, as f0 detection algorithms may be sensitive to other sources of noise (**f0_bin_lf**).

4.2 Classification results and discussion

In order to test whether these metrics are good predictors of phonetic voicing, we examine their contributions in the classification of the test consonants into the four groups defined in the manual analysis: $Vl_{[+v]}$, $Vl_{[-v]}$, $Vd_{[+v]}$, $Vd_{[-v]}$ (with 'Vd' and 'Vl' for underlyingly voiced/voiceless, and [+v], [-v] for phonetically voiced/unvoiced). A linear predictive analysis was done in R [14], with the 13 metrics entered as potential predictors of group membership. A stepwise forward variable selection using the Wilk's Lambda criterion, to the "greedy.wilk" function [18], was performed to identify which metrics are relevant to discriminate between classes. Three of the 13 metrics stand out as good predictors: **CV_nrj_lf**, **uv_ratio** and **target_length**. Table 2 gives the mean values of the four groups using these metrics.
A discriminant model based on these three metrics accurately classified 81% of the consonants into

their manually pre-defined groups (82% without cross-validation). From the recall scores (% of consonants accurately classified) given in Table 3, we can see that canonically produced underlying voiced consonants ($Vd_{[+v]}$) are best discriminated by the model (92% accuracy). On the other hand, canonically realized voiceless stops ($Vl_{[-v]}$) are poorly classified (66%) and 34% of them have been mixed-up with phonetically voiced variants ($Vl_{[+v]}$). Non-canonical consonants ($Vd_{[-v]}$ and $Vl_{[+v]}$) show interesting results. While the four phonetically devoiced /d/s ($Vd_{[-v]}$) were misclassified, 3 of them were indeed recognized as [-v] and assigned to the $Vl_{[-v]}$ group. For the phonetically voiced realization of /t/ and /k/ classification was rather performing with 83% of the $Vl_{[+v]}$ correctly predicted as such. Results show that the few wrongly classified $Vl_{[+v]}$ have been considered either as phonetically voiceless variants (9 cases $Vl_{[-v]}$) or as voiced /d/s (11 cases $Vd_{[+v]}$).

Table 2: Mean values on the predictive metrics for each class.

Variables	$Vl_{[-v]}$	$Vl_{[+v]}$	$Vd_{[-v]}$	$Vd_{[+v]}$
cv_nrj_bf	-22	-17	-25	-4
uv_ratio	68	35	60	6
target_length	88	96	96	61

Table 3: Classification summary of the cross-validated model (P=precision, R=recall).

Cross-validated counts

Predicted	$Vl_{[-v]}$	$Vl_{[+v]}$	$Vd_{[-v]}$	$Vd_{[+v]}$	#	*P*
$Vl_{[-v]}$	38	9	3	1	51	*75%*
$Vl_{[+v]}$	20	95	1	6	122	*78%*
$Vd_{[-v]}$	0	0	0	0	0	*0%*
$Vd_{[+v]}$	0	11	0	86	97	*89%*
#	58	115	4	93	35	
R	*66%*	*83%*	*0%*	*92%*		

The precision of the classification is another aspect to consider before applying this classification model to a larger dataset. Precision indicates the rate of false alarms, i.e. the % of consonants included in a group by the discriminant function, which do indeed belong to this group. Low precision would be problematic because it would mean that based on these three metrics, the model would return classes with more errors than correctly classified consonants. This is not the case here. Except for the predicted $Vd_{[-v]}$ group where no tokens were classified, relatively high precision scores are found in all predicted groups (75 to 89%).

5. DISCUSSION AND CONCLUSION

While C-to-C voice assimilation is known to be frequent in French, both word medially or across word boundary, variation in the voicing of consonants in a prevocalic context had not been systematically studied before. Our observation of about 100 exemplars of each of the consonants /d, t, k, s/ produced in a controlled set of words in a /_ɑ̃/ context showed that variation in phonetic voicing is quite pervasive in natural continuous French. Surprisingly, 68% of the voiceless stops, and 30% of the /s/, are phonetically voiced (fully or partially). Devoicing of voiced stops, on the contrary, seems to be quite rare. Even though, an effect of the right context has to be tested in our data (recall that our word initial consonants are preceded by either a pause, a word-final consonant or vowel), we can tentatively interpret the phonetic voicing of voiceless consonant as a coarticulatory anticipation of the vocal fold setting for the upcoming vowel. The preferred direction of voice assimilation in CC context argues in favor of this interpretation. Indeed the most frequent cases of C-to-C voice assimilation are regressive ([12], progressive assimilation is only found consonant+liquid clusters, with a devoicing of the liquid after voiceless C). Moreover, it seems that phonetic voicing of voiceless consonants is preferably triggered by a following vowel than by a following consonant since, in CC sequences, voiced stops are more frequently devoiced by a following voiceless C than the reverse [16].

In order to better understand the conditioning of voice modification in French consonants, an examination of a larger set of consonants in various context and word position is planned. To this aim, we tested in this study a set of acoustic metrics for their potential to discriminate phonetic voicing on stops. The relative measure of energy between the consonant and the vowel in the 0-500Hz frequency band (CV_nrj_lf), the proportion of unvoiced frames in the consonant (uv_ratio), and the duration of the consonant (target_length) were found to be good indicators of group membership. Overall, these metrics were able to predict phonetic voicing [+v] with a global (Vl/Vd mixed) accuracy rate of 87% and a good precision (83%). Voicelessness [-v], however, is not well predicted by these metrics, with a poorer discrimination (61%) but relatively good precision (75%). A closer look at the misclassified tokens is now needed to understand why some have wrongly been assigned to a [+v] or [-v] category based on these metrics.

6. ACKNOWLEDGEMENTS

This work was supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

7. REFERENCES

[1] Bárkányi, Z., Kiss, Z. 2010. Is /v/ different? *Proc. Twenty years of theoretical linguistics in Budapest* 25.

[2] 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/

[3] Bürki, A., Fougeron, C., Gendrot, C., Frauenfelder, U.H. 2011. Phonetic reduction versus phonological deletion of French schwa: Some methodological issues. *Journal of Phonetics* 39, 279–288.

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

[5] 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. *Proc. Eurospeech* Lisbon, 2453-2456.

[6] Gradoville, M.S. 2011. Validity in measurements of fricative voicing: Evidence from Argentine Spanish, *Proc. 5th Conference on Laboratory Approaches to Romance Phonology* Provo, 59–74.

[7] Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., Galibert, O. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, International Conference on Language Resources, Evaluation and Corpora. *Presented at the LREC - Eighth international conference on Language Resources and Evaluation* - Istanbul.

[8] Hallé, P., Adda-Decker, M. 2007. Voicing assimilation in journalistic speech, *Proc. 16th ICPhS* Saarbrücken, 493–496.

[9] Johnson, K. 2004. Massive reduction in conversational American English, *Proc. of the Workshop on Spontaneous Speech: Data and Analysis*. 29–54.

[10] Kiss, Z. 2013. Measuring voicing correlates of voicing in stops and fricatives. http://seas3.elte.hu/VLlxx/gkiss.html

[11] Kocjancic Antolík, T., Fougeron C. 2013. Consonant distortions in dysarthria due to Parkinson's disease, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia. *Proc. Interspeech* Lyon.

[12] Meunier, C. 1994. Les groupes de consonnes : problématique de la segmentation et variabilité acoustique. PhD thesis, Université de Provence.

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

[14] R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[15] Raymond, W.D., Dautricourt, R., Hume, E. 2006. Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. Language Variation and Change 18, 55–97.

[16] Snoeren, N.D., Hallé, P.A., Segui, J. 2006. A voice for the voiceless: Production and perception of assimilated stops in French. *Journal of Phonetics* 34, 241–268.

[17] Torreira, F., Ernestus, M. 2011. Vowel elision in casual French: The case of vowel /e/ in the word c'était. *Journal of Phonetics* 39, 50–58.

[18] Weihs, C., Ligges, U., Luebke, K. and Raabe, N. 2005. klaR Analyzing German Business Cycles. In Baier, D., Decker, R., Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, Springer-Verlag Berlin, 335-343.