EXPLORING CONSONANTAL VARIATION IN FRENCH-ARABIC CODE SWITCHING SPEECH: THE CASE OF GEMINATION

Djegdjiga Amazouz[1,2], Martine Adda-Decker[1,2], Lori Lamel[2], Jean-Luc Gauvain[2]

[1]LPP,UMR7018,CNRS-Sorbonne Nouvelle University, Paris, France

[2]LIMSI-CNRS, Paris Saclay University, Orsay, France
[amazouz,madda,lamel,gauvain]@limsi.fr

ABSTRACT

Code switching (CS) is the practice of moving back and forth between two languages in the same passage. This paper investigates consonantal variation in an 8-hour corpus of spontaneous code-switched French-Algerian Arabic speech. The study focuses on production variation in Arabic geminate consonants for which Arabic has a phonological opposition with simple consonants. This may influence bilinguals' production in French where this opposition does not exist. Experiments are realized with the help of automatic speech alignment authorizing simple and geminate pronunciation variants. The alignment system makes use of Arabic acoustic models which also cover all consonants of French, permitting investigation of simple/geminate variation in both languages. By associating the alignment results with acoustic measurements of consonant variation, this study shows that, the variation affects both French and Arabic simple consonants in code-switched speech with an average geminate variation of the simple consonants of 21.2% for French and 22.4% for Arabic.

Keywords: Code-switching, consonant variation, gemination, automatic speech alignment, French, Arabic

1. INTRODUCTION

Code-switching (CS) consists of switching from one language to another in the same speech turn. A large body of research has been devoted to CS [6, 13, 14, 21], especially addressing lexical, syntactical and morphological levels of CS. However, the acoustic-phonetic level has been less widely explored. Previous studies in phonetics, phonology and prosody that focused on the influence of one language on the production of the other one during switching often present diverging results across studies. Bullock & Toribio's [6] results suggest that speakers tend to stick to the spoken languages' standards (e.g. a stop is produced with a typical English burst when speaking English and as a typical Spanish stop when switching to Spanish) with only minor mutual interaction. However, more recent studies on voiced and voiceless stops in CS speech showed phonetic convergence between pairs of stops belonging to one or the other of the two CS languages [5, 8, 20], thus questioning earlier results that promoted the idea of clear language system separations. More recently, researchers in automatic speech recognition have turned attention to CS speech in order to test their system's automatic transcription and language identification abilities [24, 25, 26].

This short review of CS studies shows that the CS term covers a multi-faceted reality [4, 14, 10, 22], not only with respect to the various speech types and communication situations, but also regarding the addressed language pairs, studied linguistic levels and methodological approaches.

This paper presents a study on French (FR) Algerian Arabic (AA) code-switched speech. We focus on geminate consonants in CS speech making use of a recent methodological approach [19]. Gemination is the process of consonant doubling [7]. Geminate consonants are mainly characterized by their acoustic long duration compared their simple counterparts [15].

In this study, we investigate potential consonantal gemination in CS speech in AA and in French, the latter may arise due to the influence of AA for the bilingual speakers. To assess the influence of CS on geminate production we compare the French part of the CS speech with a French monolingual spontaneous speech corpus with the aim of disentangling variation due to spontaneous speech and variation due to CS. The proposed work is realized with the help of automatic speech alignment using simple/geminate consonants as a variation paradigm in the pronunciation dictionary of the speech alignment system. This experiment is followed by acoustic analyses of consonantal duration in order to support the alignment results.

2. FRENCH AND ALGERIAN ARABIC CONSONANTAL SYSTEMS

French and Algerian Arabic are phonologically distant languages. As, opposed to French, AA has a very rich consonantal system and a reduced vowel system. French is generally considered to comprise 21 consonants [9] with 6 stops and 6 fricatives, each being either voiced or voiceless, four nasals, three glides and two liquids /l/ and /ʁ/. There are no geminates in the French phonological system. However, gemination of French consonants may occur in word contact situations such as "il l'aime", *he loves her*, (which is different from "il aime" *he loves*) and within words with double letter consonants "il mourrait" *he would die* to distinguish from "il mourait" *he died*. However, such gemination processes in French are considered to be marginal.

AA, a North African Arabic dialect with 27 consonants, has two more consonants than Modern Standard Arabic (MSA): /p/ and /v/ [3]. AA and French share 18 consonants /p, b, t, d, k, g, m, n, f, v, s, z, ʃ, ʒ, ʁ/ɣ, l, w, j/. The AA consonantal system covers all French consonants except the labial-palatal glide /ɥ/ and the palatal and velar nasals /ɳ,ŋ/, the latter borrowed from English such as *camping* and *parking*. The AA phonological system also features dental fricatives /θ, ð/ and emphatic counterparts of /t, d, s/, not present in French.

All AA consonants have a geminate counterpart. The geminate consonants may appear in words that form minimal pairs with respect to their simple consonant counterparts /bat^{ʕ}al/ *hero*, batt^{ʕ}al *break a habit*. Often, they are just in phonetic opposition with the simple consonants: /kabbar/ *rise*, /t^{ʕ}lla/ *round*, /ssitta/ *six*. As in French, the AA dialect gives rise to gemination processes, which occur more frequently than the French ones. As a typical example, the coronal consonants are automatically geminated after the article أَل *the*, the article's consonant being assimilated to the following coronal. Beyond the phonological status of geminates and gemination of consonants in Arabic, they are orthographically marked by a diacritic ■ّ called *Shadda*. The explicit transcription of geminates helps to count them in speech providing information about the most produced consonants (see Section 4).

3. SPEECH MATERIAL AND ALIGNMENT

3.1. Speech material

These studies rely on two corpora, a French Algerian-Arabic CS speech corpus (FACST) from bilingual speakers, and a native French corpus of casual speech by monolinguals (NCCFr). The FACST corpus [3, 2] contains a total of 8 h of speech with CS in both French and Algerian Arabic from 20 speakers. The *Nijmegen Corpus of Casual French* (NCCFr) provides a reference baseline for consonant variation in French spontaneous speech. The corpus contains about 31h of conversational French of 46 native speakers raised in Central/Northern France [23].

Table 1: The most frequent AA geminates (75% of geminate tokens) in the FACST corpus. Occurrence counts for their simple counterparts, in AA/French CS data and in French NCCFr corpus.

Cons	Number of occurrences			
	AA		Fr	
	FACST			NCCFr
	geminate	simple	simple	simple
/l/	334	2820	7268	59227
/d/	141	950	4609	43497
/n/	138	2118	2867	27269
/s/	126	456	7072	77680

3.2. Automatic speech alignment

To study consonant variation in CS speech, we used the forced alignment paradigm which consists in automatically time-aligning the manually produced transcripts on the acoustic signal. Each word gets one or several pronunciations to handle potential production variants across repetitions and speakers. The acoustic models used to process the bilingual CS data consist of Arabic position-independent monophone acoustic models similar to those described in [11, 12, 17, 18]. A monophone setup was preferred as previous studies showed that large sets of context-dependent acoustic models (as typically used in speech recognition systems) tend to capture systematic coarticulatory variation: there is less of a need for the system to select different phone models than the canonical ones [1, 19]. The forced alignment system locates word and phone boundaries using the orthographic transcriptions and the best matching pronunciations chosen among the variants permitted in the dictionary. Hence, to study geminates and gemination variation, the alignment system is used with a pronunciation dictionary which allows each geminate to be replaced by its corresponding simple counterpart and vice versa.

4. GEMINATES AND GEMINATION AS CONSONANT VARIATION IN CS SPEECH

We start with a quantitative overview of geminates in the Arabic subset of the FACST corpus. Table 1 shows the frequency counts of the most frequent geminates in AA. For comparison, the frequency counts of their simple counterparts in both AA and French of the FACST CS speech, as well as those of the monolingual French NCCFr corpus are also shown. We did not include the rhotic geminate in this study, although it is among the top 5 most frequent geminates. These consonants have a distinct phonological class in AA and FR although their phonetic characteristics are the same [16]. It is interesting to note, that the geminates in Table 1 are the coronals /ll, dd, nn, ss/ which each have more than 120 tokens. So, we limit our investigations on these consonants, which are the most representative in our data. Experiment 1 investigates simple consonant gemination of AA and FR in CS speech as given by the forced alignment method, using the Arabic acoustic models. Table 2 lists the target consonants and competing variants, i.e., the simple and geminate form for each consonant, and provides examples for both languages. In experiment 2, the same protocol is applied to FR CS speech and FR monolingual speech, in order to compare productions in both settings (bilingual vs native). Experiment 3 addresses the question of whether, and if so, how often, geminates are simplified in spontaneous CS speech. In this case, the target consonants are the AA geminates /ll, dd, nn, ss/ and their simple counterparts are added as variants. This experiment is aimed at highlighting discrepancies between orthographic geminates as transcribed in the FACST corpus and the aligned variants chosen during the forced alignment.

Table 2: Competing variants for each target consonant and example lexical entries.

Targ.	Var.	Examples
[l]	[l, ll]	لِ /li/ (for) : [li], [lli] (AA) lu /lu/ (read) : [ly], [lly] (Fr)
[d]	[d, dd]	دَار /da:r/ (house) : [da:r],[dda:r](AA) dent /dã/ (tooth) : [dã],[ddã](Fr)
[n]	[n, nn]	نُور /nu:r/ (light) : [nu:r],[nnu:r](AA) ne /ne/ (light) : [nu:r],[nnu:r] (Fr)
[s]	[s, ss]	سَار /sa:r/ (walked)[sa:r],[ssa:r] (AA) sept /set/ (seven)[set],[sset] (Fr)

5. RESULTS AND ANALYSIS

This section reports results for the 3 experiments aiming at quantifying variation in the production of simple and geminate consonants. The figures provide the gemination (respectively simplification) variant rates for each experiment. These rates are supplemented by consonant duration results as obtained from the forced alignments.

5.1. Gemination in French-AA in CS speech

Hereafter, we present gemination variant rates measured on simple consonants occurring both in AA and FR CS speech. The overall gemination variant rates are similar in both languages: 22.4% for AA and 22.2% for FR, although there are differences across consonants as shown in Figure 1. For AA, the consonants /s,l/ have the highest variant rates of about 25%, whereas only /s/ has a similar value for FR. High gemination variant rates in AA are observed for /d,l/ with 21% and 25%. The consonant with the lowest gemination variant rate is the nasal /n/, with 19% and 16% of occurrences in AA and FR respectively. Using the consonant durations, a positive correlation (r= +0.67) is measured between the variant rates and the corresponding durations. Figure 2 shows that, with the exception of /s/ in AA, the consonants most frequently labeled as the geminate variant have a larger duration difference between the simple and geminate labelled forms (r= +0.55).

Figure 1: Expt 1: consonant (simple) gemination rates by target consonant in AA and FR CS speech

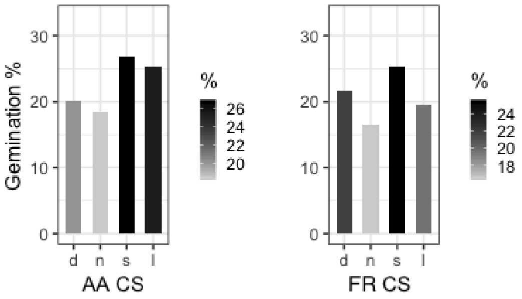
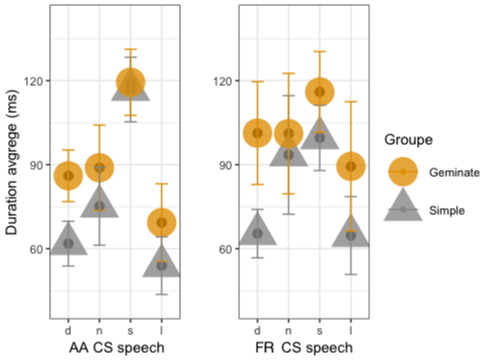


Figure 2: Average duration in (ms) of simple consonants in AA and FR CS speech. Circles: geminate variant selected ($C \rightarrow CC$); triangles: remains simple ($C \rightarrow C$). Error bars give standard deviation.



5.2. Gemination in French: comparing CS and monolingual speech

Comparing gemination variant rates between FR CS and FR monolingual consonants in Figure 3, it can be observed that the monolingual speech features higher rates than the CS speech ($\chi^2(2) = 8.01$, $p < 0.01$). This somewhat unexpected results may suggest that the phonological gemination contrast plays an important role in keeping the canonical pronunciations. The word and prosodic contexts of the gemination have not yet been examined. However, the high variation in monolingual speech suggests that these segments may have been stressed. The duration plots are shown in Figure 4.

Figure 3: Expt 2: Gemination rates of simple Cs in CS and monolingual French for each target C.

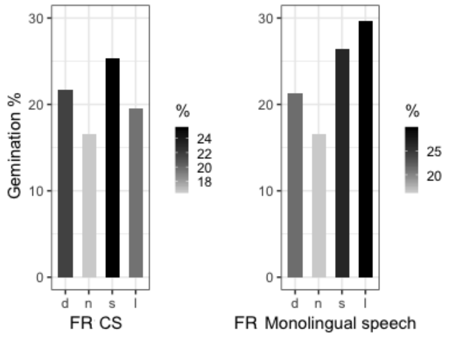
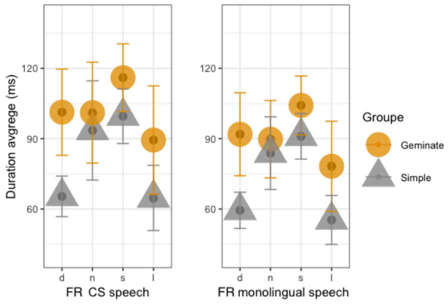


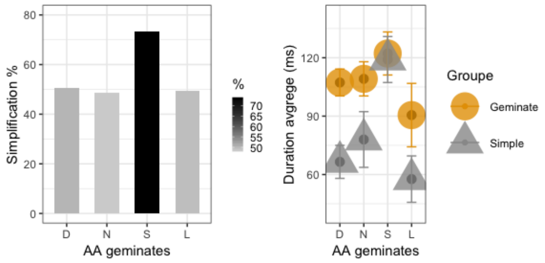
Figure 4: Average duration (ms) of French simple consonants in CS and monolingual speech. Circles: labelled as geminate variant ($C \to CC$); triangles: ($C \to C$) Error bars give standard deviation



5.3. AA geminate simplification in CS speech

Figure 5 shows the percentage of original AA geminates labelled as their simple consonant counterpart along with the corresponding duration plots. The simplification rates are seen to be higher than the gemination rates in the previous figures which were around 20%. Simplification is observed for all consonants, with rates ranging from 49% to 76%. The largest simplification rate of 76% is measured for /ss/ 'S'. The duration plot shows that the durations of geminate tokens aligned with their simple counterpart are significantly shorter than those which remained labelled as geminates.

Figure 5: Left: Simplification rates of AA geminates. Right: Average duration (ms) for simple C variants given by triangles ($CC \to C$); unchanged geminates given by circles ($CC \to CC$ variation)



6. DISCUSSION

Three points can be mentioned based on this study. First, the proposed method using automatic variant alignment can help us study the variation of simple and geminate consonants in large speech corpora. The duration analysis confirms that the aligned gemination and simplification variant labels are highly related to segment duration and that duration is a solid, although not unique, criterion to study variation in consonant gemination.

The study also shows that gemination of simple consonants, as revealed by our method, appears in both FR and AA CS speech. However, AA is the most affected by this variation because of the phonological distinction between simple consonants and geminates. In our data, the consonants most concerned by this gemination variation are /d, s, l/. By contrast, less gemination was observed for the nasal consonant /n/ in both languages and in both corpora. The FR monolingual speech also shows high gemination variant rates comparable to FR in CS.

Finally, the high simplification rates of geminate consonants (> 40%) suggest further investigations on a methodological level: acoustic models may be biased in favor of simple consonants due to their overwhelming presence in speech. On a linguistic level, geminates may feature other correlates than duration. Further studies are underway in an attempt to understand production differences by monolingual and bilingual speakers.

7. ACKNOWLEDGEMENTS

This research is supported by CNRS-France, SALSA project (Speech and Language technologies for Security Applications) under grant ANR 14-CE28-0021 and by French Labex EFL program ANR-10-LABX-0083.

8. REFERENCES

[1] Adda-Decker, M., Lamel, L. 1999. Pronunciation variants across system configuration, language and speaking style. *Speech Communication* 29(2-4), 83–98.

[2] Amazouz, D., Adda-Decker, M., Lamel, L. 2017. Addressing Code-Switching in French/Algerian Arabic Speech. *Proc. ISCA Interspeech 2017* 62–66.

[3] Amazouz, D., Adda-Decker, M., Lamel, L. 2018. The French-Algerian Code-Switching Triggered audio corpus (FACST). *Proc. Eleventh International Conference on Language Resources and Evaluation LREC 2018* 1468–1473.

[4] Auer, P. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

[5] Balukas, C., Koops, C. 2015. Spanish-english bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism* 19(4), 423–443.

[6] Bullock, B. E. 2012. *Phonetic reflexes of code-switching* chapter 10, 163–181. Cambridge: Cambridge University Press.

[7] Delattre, P. 1969. An acoustic and articulatory study of vowel reduction in four languages. *IRAL-International Review of Applied Linguistics in Language Teaching* 7(4), 295–326.

[8] Deuchar, M., Davies, P., Herring, J., Couto, M. C. P., Carter, D. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism* 93–111.

[9] Fougeron, C., Smith, S. 1999. Illustrations of the IPA: French. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet* 78–81.

[10] Gardner-Chloros, P. 2009. *Code-Switching*. Cambridge: Cambridge University Press.

[11] Gauvain, J.-L., Lamel, L., Adda, G. 2002. The LIMSI broadcast news transcription system. *Speech communication* 37(1-2), 89–108.

[12] Gelly, G., Gauvain, J.-L., Lamel, L., Laurent, A., Le, V. B., Messaoudi, A. 2016. Language recognition for dialects and closely related languages. *Odyssey, Bilbao, Spain*.

[13] Grosjean, F. 2008. *Studying bilinguals*. Oxford University Press, USA.

[14] Gumperz, J. J. 1982. *Discourse strategies* volume 1. Cambridge University Press.

[15] KHOUJA, M. K., ZRIGUI, M. 2005. Durée des consonnes géminées en parole arabe: mesures et comparaison. *TALN-RECITAL 2005, Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*.

[16] Lahrouchi, M. 2018. Not as you r: Adapting the french r into arabic and berber. *Rhotiques: l'invariant et ses avatars*.

[17] Lamel, L., Gauvain, J.-L., Adda, G., Adda-Decker, M., Canseco, L., Chen, L., Galibert, O., Messaoudi, A., Schwenk, H. 2004. Speech transcription in multiple languages. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* volume 3. IEEE iii–757.

[18] Lamel, L., Messaoudi, A., Gauvain, J.-L. 2009. Automatic speech-to-text transcription in Arabic. *ACM Transactions on Asian Language Information Processing (TALIP)* 8(4), 18.

[19] Mareüil, P. B. d., Adda-Decker, M. 2002. Studying pronunciation variants in French by using alignment techniques. *Seventh International Conference on Spoken Language Processing*.

[20] Piccinini, P., Arvaniti, A. 2015. Voice onset time in spanish–english spontaneous code-switching. *Journal of Phonetics* 52, 121–137.

[21] Poplack, S. 2012. What does the nonce borrowing hypothesis hypothesize? *Bilingualism: Language and Cognition* 15(3), 644–648.

[22] Sebba, M. 2009. *On the notions of congruence and convergence in code-switching* 40–57. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

[23] Torreira, F., Adda-Decker, M., Ernestus, M. 2010. The Nijmegen corpus of casual French. *Speech Communication* 52(3), 201–212.

[24] Vu, N. T., Adel, H., Schultz, T. 2013. An investigation of code-switching attitude dependent language modeling. *International Conference on Statistical Language and Speech Processing*. Springer Berlin Heidelberg 297–308.

[25] Vu, N. T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., Li, H. 2012. A first speech recognition system for Mandarin-English code-switch conversational speech. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE 4889–4892.

[26] Yılmaz, E., van den Heuvel, H., van Leeuwen, D. 2016. Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science* 81, 159–166.