VOWEL RAISING ACROSS SYRIA AND JORDAN IN THE DIVAL CORPUS

Sam Hellmuth[1], Rana Almbark[1], Chris Lucas[2] and Georgina Brown[3]

[1]University of York, [2]SOAS University of London, [3]Lancaster University
sam.hellmuth@york.ac.uk, rana.alhusseinalmbark@york.ac.uk, cl39@soas.ac.uk, g.brown5@lancaster.ac.uk

ABSTRACT

Levantine Arabic dialects display a pervasive pattern of front vowel raising (*imala*), whose distribution is subject to variation and change across the region. There are few acoustic descriptions of *imala*, but prior auditory transcriptions point to variation in the degree of raising, as well as in the categorical presence or absence of *imala*, between contexts and dialects. We report values of F1/F2 and F2-2F1 in two potential *imala* contexts (word-internal /a:/ and feminine suffix /-a/), in scripted speech produced by 123 speakers of dialects from across Syria and Jordan, from the Dialect Variation in the Levant [DiVaL] corpus, with comparison to earlier dialect descriptions. We find limited evidence of present day dialectal differences in the conditioning environments of *imala* of word-internal /a:/, but clear evidence of both gradient and categorical variation in the degree of movement in the front vowel diagonal affecting feminine suffix /-a/.

Keywords: dialectal variation, Arabic, vowel raising, speech corpora.

1. INTRODUCTION

Levantine Arabic dialects display varying degrees of a vowel raising/fronting pattern known in Arabic as *imala*. This name dates back to the very earliest descriptions of Arabic phonetics by mediaeval grammarians [1], and the typical phonological conditioning and regional distribution of the pattern is well-known [2-4]. However, even though *imala* has been described as involving lower F1 and higher F2 [3], there are very few studies which actually use or report results of acoustic analysis of *imala* [5-7]. This study seeks to help remedy the lack of acoustic descriptions of *imala* by exploiting a new corpus of speech data in dialects from across Jordan and Syria: the Dialect Variation in the Levant (DiVaL) [8].

Our aim is: i) to introduce the DiVaL corpus, ii) to demonstrate its potential through acoustic analysis of F1 and F2 in two *imala* conditioning contexts (word-medial /a:/ and word-final feminine suffix /-a/, and iii) to compare the results with prior dialectological descriptions based on auditory impression [9].

2. BACKGROUND TO THE STUDY

2.1. Prior descriptions of *imala*

The traditional term *imala* ([ʔimaːla] lit. ‘inclination’) covered a range of phenomena involving some degree of raising (and fronting) of /a/. Here we focus on general *imala* of word-medial long /a:/ and *imala* in the morphological context of word-final feminine suffix /-a/. One type of *imala* does not entail the other, but in all Arabic varieties that exhibit one or both types, *imala* is blocked by emphatic consonants (with secondary post-velar articulation) which trigger backing of /a/ ([tafxiːm] lit. ‘intensification’). Syrian dialects vary in both the presence and degree of *imala*: for word-medial /a:/ [9] gives Damascus *lābis* (no *imala*) vs. Aleppo *lēbis* vs. Soukhne (Homs region) *lībis* ‘wearing’; but for suffix /-a/: Dēr iz-Zōr (NE Syria) *wazza* (no *imala*) vs. Damascus *wazze* vs. Soukhne *wazzi* ‘goose’; *imala* of word-medial /a:/ is present in all non-*tafxi:m* environments in some dialects (e.g. Tartus) but in others only in non-*tafxi:m* environments historically adjacent to a high front vowel or palatal consonant (most of NW Syria).

Recent sociolinguistic studies show that both types of *imala* are subject to variation and change in Levantine dialects and beyond [7, 10, 11], so the incidence of *imala* in Syria may have changed since the 1980s data collection which informed [9]. In one of the best sociophonetic descriptions of *imala* to date, [5] measured F1 and F2 at five points in tokens of suffix /-a/ in an age-stratified sample of Palestinian Arabic speakers from Gaza City (using a within-speaker linear-normalised 'N-score' [(F2-F1) / F1]), and found significant differences between the youngest and oldest speakers at all measuring points. Crucially, [5] revealed variation in apparent time in the phonetic realisation of /-a/, which had not been detected in auditory analysis of the larger corpus from which the analysed data subset was drawn [12].

2.2. Dialect Variation in the Levant [DiVaL] corpus

A number of very large speech databases of Arabic dialects exist [13], but their size and data collection practices lead authors such as [5] to concede to frequent errors in the metadata. In addition, the level of granularity in the metadata typically cannot support analysis at the sub-regional level. The DiVaL corpus was created to facilitate reliable and detailed analysis tasks to be carried out at a fine-grained level.

Data was collected in 2021 using four production tasks: *st*: reading a scripted folk *story* twice; *rs*: ten scripted *read sentences* which target phonological and morphological variables such as phoneme /q/ and feminine suffix /-a/; *gs*: free translation of three *grammar sentences* from Standard Arabic to dialect, which target morpho-syntactic and lexical variables such as negation; *pd*: three *picture description* tasks, with visual prompts which target lexical items known to vary within and across dialects.

Text for the *st* and *rs* tasks were presented in pdf format in Arabic script using dialectal spelling conventions (see Table 1). Participants were asked to read in their own dialect not Standard Arabic. Text for the *gs* task was presented in Standard Arabic.

Participants were recruited to represent key major dialect regions in Syria and Jordan. All Jordanian data and most Syrian data was collected in Jordan; a few Syrian-origin speakers were recorded in the UK. Since most Syrians are displaced, their data is coded according to place of origin and not where recorded. In Jordan we worked with local fieldwork assistants to ensure the accuracy of the place of origin metadata. Full metadata will be published with the corpus [8].

Due to Covid restrictions, data was self-recorded remotely using the Awesome Voice Recorder (AVR) smartphone app [14]. We used AVR to obtain wav audio files containing full spectral information [cf. 15]. Other platforms were explored but these did not reliably yield wav files with full spectral information.

3. METHODS

This study uses data from the story (*st*) and read sentences (*rs*); sample scripted items are in Table 1.

The full corpus comprises 133 speakers: 52 from Jordan; 81 from Syria. We also identified subsets of speakers who can be grouped into linguistically meaningful dialect subgroups in Jordan or Syria with at least 14 speakers per subgroup. Table 2 shows the split of speakers in these subgroups by gender, and Figure 1 visualises speakers' place of origin as a map. The speakers were aged between 18-65 at time of recording with median age in the 26-35 age range. Level of education is included in the corpus metadata.

A romanised orthographic transcription was force aligned to the data using Prosody Lab Aligner [16]. Tokens of word-medial long /a:/, feminine suffix short /-a/ and word-medial short /-a-/ were identified using a Praat script, which also extracted the first and second formants (F1 and F2) in each identified vowel at the vowel midpoint. To remove erroneous values potentially arising from tracking or alignment errors, outliers were removed using the Modified Mahalanobis Distance method implemented in [17] using the joeyr package [18]. The remaining formant values were Lobanov normalised [19] using phonR [20]. We calculated linear F2-2F1 [F2-(2*F1)] which has been shown in prior dialect contact studies [21] to effectively capture direction and extent of movement in the front diagonal of the vowel space in a single metric. Values of F2-2F1 were explored in linear mixed-effects models (LMM) using lme4 [22].

Code	Target
rs08	الكل ظروفو صعبة كثيرهال أيام /l-kull ðˤuruːf-o-h sˤaʕb-a hal-ʔajjaːm/ <lkull DHuru:fuh Sa3ba hal2ayya:m> Everyone’s circumstances are hard these days.
st13	قالّو جحا ثلاثة ريال ومافي غيرها /qallu ʒuħa θalaːθa rjaːl w maːfi ɣayrha/ <qallu juHa thala:tha rya:l w ma:fi ghayrha> Juha told him: “Three riyals and nothing more”

Table 1: Arabic orthography, IPA, romanised transcription and translation of sample stimuli.

Code	Dialect area	F	M	Total
sy-neast	Al-Raqqa/Al-Hasakah	10	4	14
sy-nwest	Aleppo/Idlib	8	6	14
sy-homha	Homs/Hamah	8	8	16
sy-damas	Damascus + environs	12	6	18
sy-south	Daraa	8	6	14
jo-rural	Irbid/Al-Ramtha/Ajloun	9	7	16
jo-urban	Amman/Al-Salt	11	4	15
jo-south	Karak/Al-Tafilah/Ma'an	9	7	16
Total		76	48	123

Table 2: DiVaL subgroups included in this study.

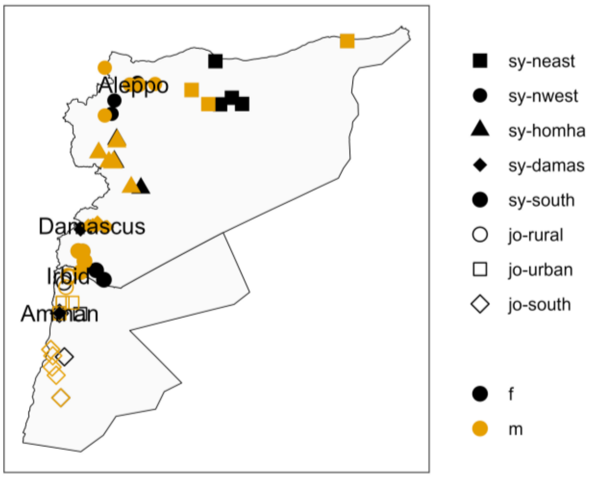

Figure 1: Place of origin of DiVaL speakers in this study.

4. RESULTS

4.1. Word-medial /a:/ in different environments

We examined F1 and F2 at the midpoint of four types of target words produced in the *rs* task: words expected to contain plain /aː/ such as <thala:tha> /θalaːθa/ 'three'; words where /aː/ is preceded by a palatal such as <2ayya:m> /ʔajjaːm/ 'days'; words followed by a palatal or front vowel such as <na:zil> /naːzil/ 'going down'; and words containing a trigger for tafxi:m (e.g. an emphatic coronal consonant) such as <Ta:la3> /tˤaːliʕ/ 'going up'. Target /aː/ in [tˤaːlaʕ] is expected to resist *imala* in all dialects due to the preceding emphatic [tˤ] and following post-velar [ʕ].

Our research question in this section is thus: to what extent do any patterns of (non-)overlap in the phonetic realisation of /aː/ in plain versus palatal versus emphatic contexts in the DiVaL corpus (collected in 2021) align with patterns reported in the earlier dialectological description by [9]. The corpus subset yields 2617 tokens for analysis (Table 3).

dialect	plain	prepal	follpal	emphatic	Total
sy-neast	102	38	90	56	286
sy-nwest	106	33	88	66	293
sy-homha	118	46	106	77	347
sy-damas	132	50	118	85	385
sy-south	98	39	93	67	297
jo-rural	127	41	102	71	341
jo-urban	107	39	102	72	320
jo-south	121	45	106	76	348
Total	911	331	805	570	2617

Table 3: Count of /aː/ tokens by dialect~wordtype.

Figure 2 visualises the F1/F2 vowel space of the different dialects, with varying degrees of overlap of formant values in /aː/ in plain versus pre-/post-palatal environments, and variation also in the degree of relative lowering/backing in emphatic environments.

The data were explored using a series of LMMs to predict values of F2-2F1 in the /aː/ data subset, with *wordtype* and dialect *subgroup* plus the interaction between them as fixed effects; the best fit model included fixed effects for *gender* and *task* (but not for *age* or *education*), with random intercepts for *speaker* and *word* and a random slope for *speaker* by *wordtype*. The model indicates a main effect of *task* (with larger F2-2F1 difference in *rs* across the board), but no effect of gender. There is a significant dialect *subgroup* by *wordtype* interaction with smaller F2-F1 values in emphatic environments in all dialects, as expected, being lower/backed in the vowel space.

Figure 3 visualises model predictions: the main differences are in the degree of difference between the emphatic versus all other environments, with a smaller degree of raising in Damascus and Daraa than in northern Syrian dialects, and which is similar to the degree of raising in northern Jordanian dialects. There is no indication of a different degree of vowel raising in plain versus palatal contexts, contra Behnstedt's observation in [9] that *imala* applied only in palatal contexts in northern Syria (equivalent to our sy-neast and sy-nwest).

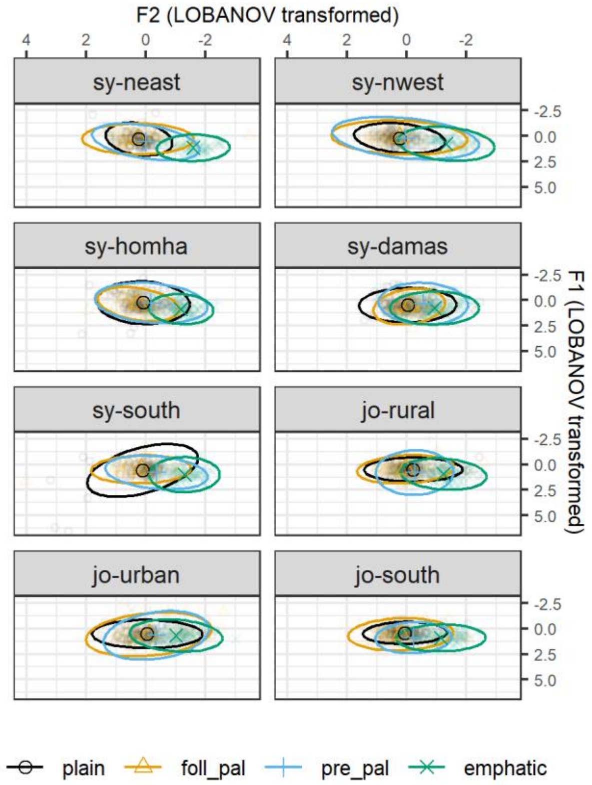

Figure 2: Mean normalised F1/F2 at the vowel midpoint in word-medial /aː/ by environment and by dialect group.

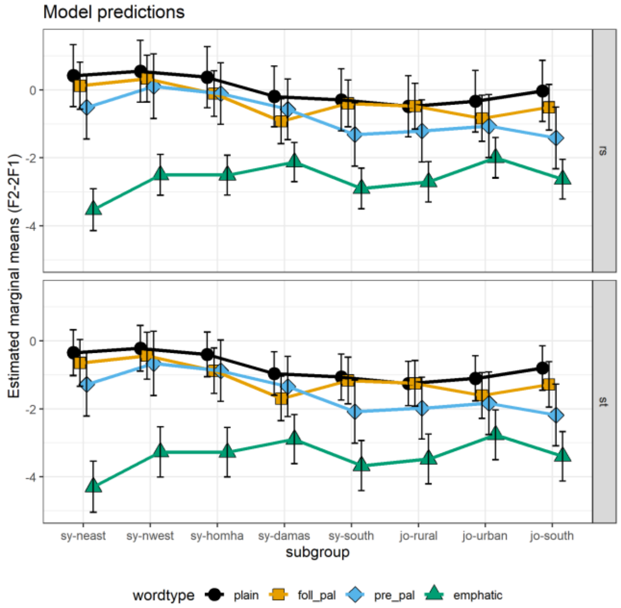

Figure 3: Estimated marginal mean + 95%CI of normalised F2-2F1 at the vowel midpoint in word-medial /aː/ by *wordtype*, dialect *subgroup* and *task*.

4.2. Feminine suffix /-a/ versus word-medial /-a-/

We examined F1/F2 at the midpoint of suffix /-a/ and word-medial /-a-/ in three target words produced at multiple points in the *rs* and st tasks: <madrasa> /madrasa/ ‘school’, <thala:tha> /θalaːθa/ ‘three’, and <madi:na> /madiːna/ ‘city’.

Our research question here is: to what extent do patterns of (non-)overlap in phonetic realisation of suffix /-a/ versus word-medial /-a-/ in DiVaL (collected in 2021) align with patterns reported in earlier dialectological descriptions by [9]. The corpus subset yields 3156 tokens for analysis (Table 4).

dialect	/-a/ suffix	medial /-a-/	Total
sy-neast	176	192	368
sy-nwest	172	180	352
sy-homha	204	213	417
sy-damas	225	234	459
sy-south	175	190	365
jo-rural	199	213	412
jo-urban	184	194	378
jo-south	195	210	405
Total	1530	1626	3156

Table 4: Short /a/ tokens by dialect~environment.

Figure 4 visualises the F1/F2 vowel space of each dialect, with varying degrees of overlap of formant values in suffix /-a/ versus word-medial /-a-/ between dialect sub-groups. The data were explored using a series of LMMs to predict values of F2-2F1 in the short [a] data subset, with *voweltype* and dialect *subgroup* plus the interaction between them as fixed effects; the best fit model included a fixed effect for *age* only (but not for *gender*, *task* or *education*), with a random slope for *speaker* by *voweltype*. There is a main effect of *age* (less F2-2F1 difference for the one male speaker in the oldest 56-65 age group, consistent with natural effects of aging on formant values). There is a significant dialect *subgroup* by *voweltype* interaction such that the degree of F2-2F1 difference between suffix /-a/ versus word-medial /-a-/ varies.

Figure 5 visualises the model predictions. There is no *imala* in the Al-Raqqa (sy-neast) data, and clear raising in suffix /-a/ in northern Syrian and southern Jordanian, but a reduced (intermediate) degree of raising in southern Syrian dialects (Damascus and Daraa). These patterns align closely with Behnstedt’s observations about *imala* of suffix /-a/ in Syrian varieties, both as a categorical (absent in Al-Raqqa) and gradient (raised to [i] in the Homs region but only to [e] in Damascus) phenomenon. The patterns also align with raising patterns in Jordan reported in [11], that is, with raising in all equivalent dialect groups to the three investigated here, but noting a higher degree of raising in southern Jordan (e.g. Karak) [4 fn4].

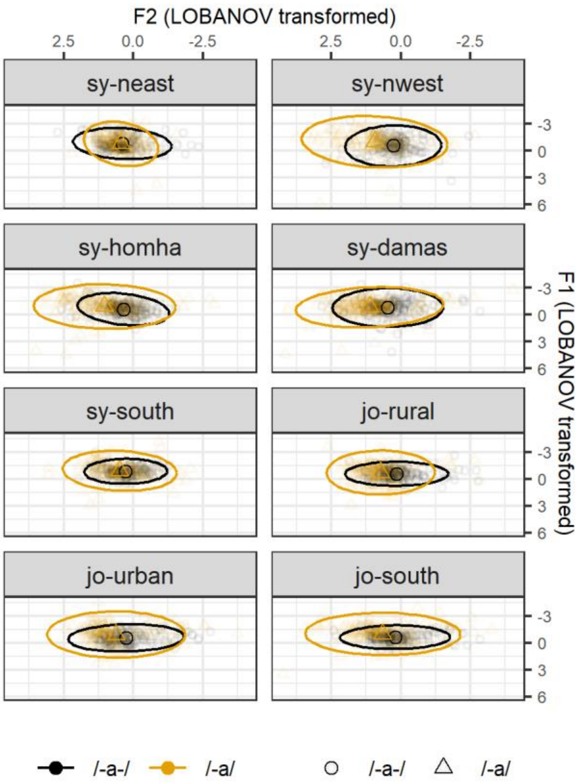

Figure 4: Mean normalised F1/F2 at the vowel midpoint in suffix /-a/ and word-medial /-a-/ by dialect group.

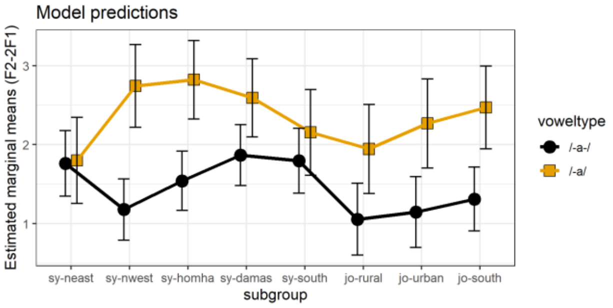

Figure 5: Estimated mean + 95%CI of norm. F2-2F1 at midpoint in suffix /-a/~word-medial /-a-/ by dialect.

5. CONCLUSION

We report quantitative investigation of F1/F2 in tokens of long and short /a/ in two contexts expected to display the vowel raising pattern known as *imala* (word-internal /a:/ and feminine suffix /-a/), from data newly collected in 2021 with speakers of a range of Syrian and Jordanian dialect areas. Observed patterns in the 2021 data, of categorical and/or gradient variation in the degree of raising in each context, mostly align with the earlier descriptions in [9], except for the apparent lack of sensitivity of *imala* to plain versus palatal contexts in NW Syria. This initial indication based on group-level acoustic analysis, can be further explored in future through auditory analysis at speaker level, thanks to the availability of the Dialect Variation in the Levant [DiVaL] corpus.

6. REFERENCES

[1] Owens, J. 2006. *A linguistic history of Arabic*. Oxford: Oxford University Press.

[2] Levin, A. 2011. ʾImāla, in *Encyclopedia of Arabic Language and Linguistics*, K. Versteegh, et al., Editors. Brill: Leiden. p. 311-315.

[3] Barkat-Defradas, M. 2008. Vowel raising, in *Encyclopedia of Arabic Language and Linguistics*, K. Versteegh, et al., Editors. Brill: Leiden. p. 678–82.

[4] Mitchell, T.F. 1993. *Pronouncing Arabic 2*. Oxford Clarendon Press.

[5] Cotter, W.M. 2016. A sociophonetic account of morphophonemic variation in Palestinian Arabic. *Proceedings of Meetings on Acoustics*, **26**(1): p. 060001.

[6] Kelly, N.E. 2017. A phonetic case study of a bidialectal speaker of Lebanese and Palestinian Arabic, in *Paper presented at Structural and Developmental Aspect of Bidialectalism, The Arctic University of Norway, Tromsø, Norway*.

[7] Abou Taha, Y. 2022. Contact-Induced Change in the Levantine: Evidence from Lebanese and Palestinian Arabic. PhD Dissertation, University of Ottawa.

[8] Almbark, R., S. Hellmuth, and G. Brown. (forthcoming). *Dialect Variation in the Levant*. UKDS. https://reshare.ukdataservice.ac.uk/856484/

[9] Behnstedt, P. 1997. *Sprachatlas von Syrien: Kartenband*. Otto Harrassowitz Verlag.

[10] Hennessey, A. 2011. The linguistic integration of the Palestinian refugees in Beirut: a model for analysis. MA Dissertation, American University of Beirut.

[11] Horesh, U., et al. 2022. Dialect contact and change in the Arabic feminine ending morpheme, in *Perspectives on Arabic Linguistics XXXIII. Papers selected from the Annual Symposium on Arabic Linguistics, Toronto, Canada, 2019*, A.-K. Ali and A. Hachimi, Editors. John Benjamins Publishing Company. p. 27-49.

[12] Cotter, W.M. 2013. Dialect contact and change in Gaza City. *Unpublished MA thesis. Colchester, UK: University of Essex*.

[13] Ali, A., et al. *The MGB-5 challenge: Recognition and dialect identification of dialectal arabic speech*. in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019. IEEE.

[14] Newkline. 2020. Awesome Voice Recorder [Android/iOS smartphone application]. Newkline Ltd.

[15] Zhao, L. and E. Chodroff. *The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects*. in *Proceedings of the Language Resources and Evaluation Conference*. 2022. ELRA.

[16] Gorman, K., J. Howell, and M. Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Journal of the Canadian Acoustical Association*, **39**(3): p. 192-193.

[17] Stanley, J.A. 2020. 6. The Absence of a Religiolect Among Latter-Day Saints in Southwest Washington. *Publication of the American Dialect Society*, **105**(1): p. 95-122.

[18] Stanley, J.A. 2021. joeyr: Functions for Vowel Data (R package version 0.6.2).

[19] Adank, P., R. Smits, and R. Van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, **116**(5): p. 3099-3107.

[20] McCloy, D.R. 2012. Vowel normalization and plotting with the phonR package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, **1**: p. 1-8.

[21] Dinkin, A.J. 2013. What's really happening to short-a before L in Philadelphia? *American Speech*, **88**(1): p. 7-31.

[22] Bates, D., et al. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1): p. 1-48.