

A Corpus Phonetic Study of Contemporary Persian Vowels in Casual Speech

Taylor Jones
University of Pennsylvania
tayjones@sas.upenn.edu

Department of Linguistics
University of Pennsylvania

PLC42
March 24, 2018



- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods
- 6 Findings
- 7 Future Work

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods
- 6 Findings
- 7 Future Work

Contemporary Iranian Persian (CIP) has 6 phonemic vowels,
however their exact characterization has been a source of
controversy in the literature.

Contemporary Iranian Persian (CIP) has 6 phonemic vowels, however their exact characterization has been a source of controversy in the literature.

The goal of this project is to provide a **phonetic foundation** for the analysis of Persian, both for future sociolinguistic study to better inform phonological theorizing about CIP.

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods
- 6 Findings
- 7 Future Work

Contemporary Iranian Persian has a relatively straightforward 6 vowel system, but with two sources of controversy:

- Are historical length distinctions preserved? Are they relevant for the phonology?
- What is the low back vowel?

One Vowel Space

i

u

e

o

æ

ɒ

Figure 1: Persian Vowels, according to Lazard (1992), Toosarvandani (2004), and Miller (2013).

Another Vowel Space

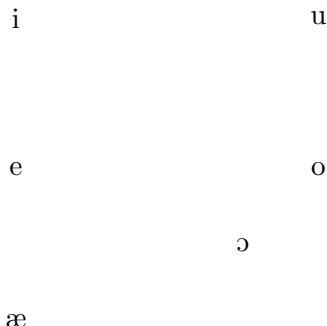


Figure 2: Persian Vowels, according to Ansarin (2004) and Aranow et al (2017).

A THIRD Vowel Space

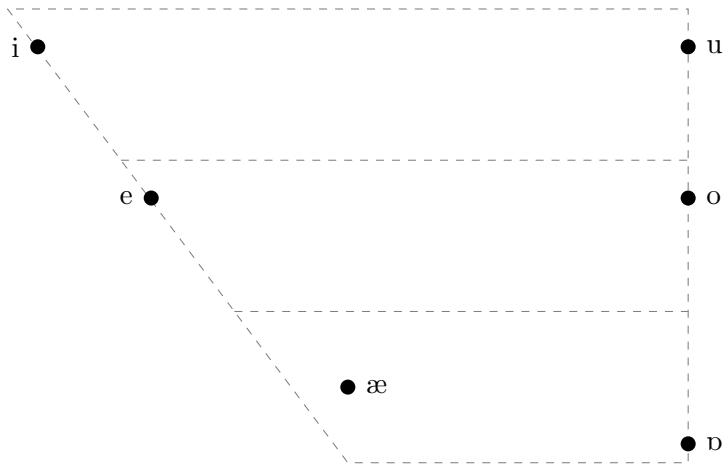


Figure 3: CIP Vowels according to the *JIPA*.

Middle Persian

i, ī

u, ū

e

o

æ

a, ā

Figure 4: Middle Persian Vowels

Importance

Why does this matter?

Importance

Why does this matter?

- Persian vowel assimilation (sometimes called “harmony”)

Importance

Why does this matter?

- Persian vowel assimilation (sometimes called “harmony”)
- Place or length?

Importance

Why does this matter?

- Persian vowel assimilation (sometimes called “harmony”)
- Place or length?
- if place, we need to know what the places are!

An example of “Assimilation”

- | | | | |
|----|---------------|---------------|---------------|
| 1. | <i>devist</i> | <i>divist</i> | ‘two hundred’ |
| 2. | <i>forush</i> | <i>furush</i> | ‘sale’ |
| | <i>fozul</i> | <i>fuzul</i> | ‘impertinent’ |
| | <i>sholuq</i> | <i>shuluq</i> | ‘crowded’ |
| 3. | <i>jahân</i> | <i>jâhân</i> | ‘world’ |
| | <i>maʔâsh</i> | <i>mâʔâsh</i> | ‘livelihood’ |

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work**
- 4 Corpus
- 5 Methods
- 6 Findings
- 7 Future Work

Length

The continued existence of a (historical) length contrast is still the subject of much debate in the Persian literature.

- Lazard (1992) claims length distinctions still obtain in speech.
- Toosarvandani (2004) and Rahbar (2009), among others, appeal to length distinctions to explain phonological processes of assimilation.

Evidence

None of the (phonological) works claiming length distinctions are relevant to the phonology empirically demonstrate the continued existence of such a distinction.

Phonetic Studies

There are only two phonetic studies of Persian Vowels.

Phonetic Studies

There are only two phonetic studies of Persian Vowels.

- Ansarin (2004): 12 female undergrads from Tabriz. Word lists.

Phonetic Studies

There are only two phonetic studies of Persian Vowels.

- Ansarin (2004): 12 female undergrads from Tabriz. Word lists.
- Aronow et al (2017): 2 Tehrani speakers. Word lists.

Ansarin 2004

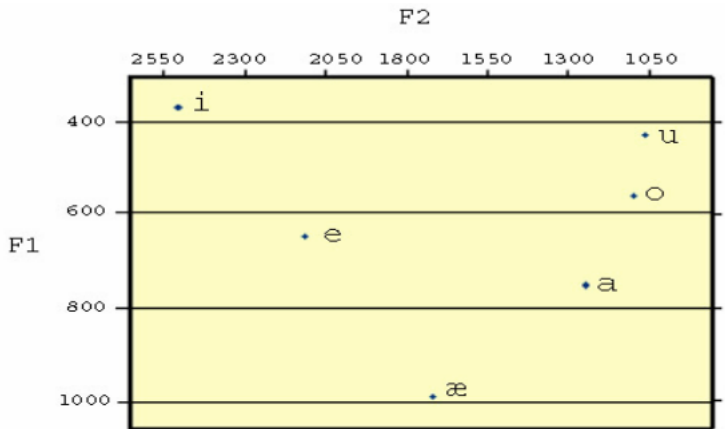


Figure 5: Persian vowels from Ansarin (2004)

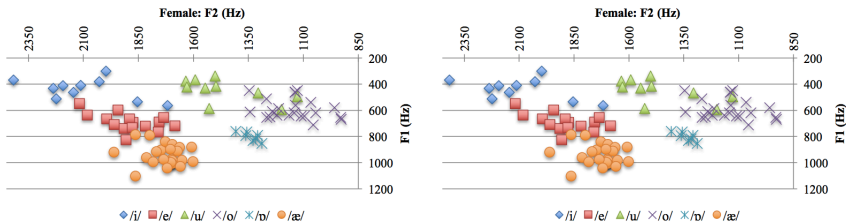


Figure 6: Persian vowels from Aronow (2017)

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus**
- 5 Methods
- 6 Findings
- 7 Future Work

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.
- 104 speakers.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.
- 104 speakers.
- Over 60 hours of speech.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.
- 104 speakers.
- Over 60 hours of speech.
- Recorded at 4KHz.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.
- 104 speakers.
- Over 60 hours of speech.
- Recorded at 4KHz.
- Transcribed by native speakers.

The Corpus

The present study makes use of the Linguistics Data Consortium (LDC) CALLFRIEND FARSI CORPUS.

- Casual telephone conversations between (and among) native Persian speakers.
- Multiple cities in Iran.
- 104 speakers.
- Over 60 hours of speech.
- Recorded at 4KHz.
- Transcribed by native speakers.

Cities



Figure 7: Cities represented in the Corpus

Population

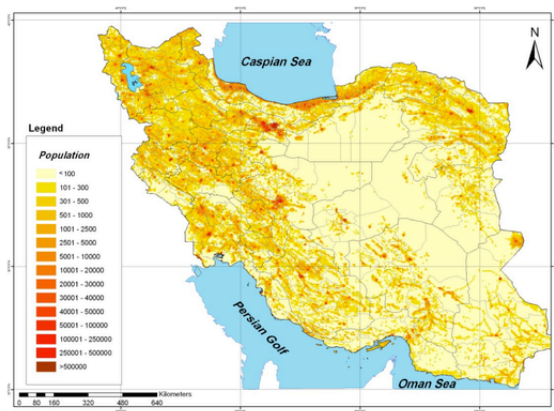


Figure 8: Population Density

Age and Gender

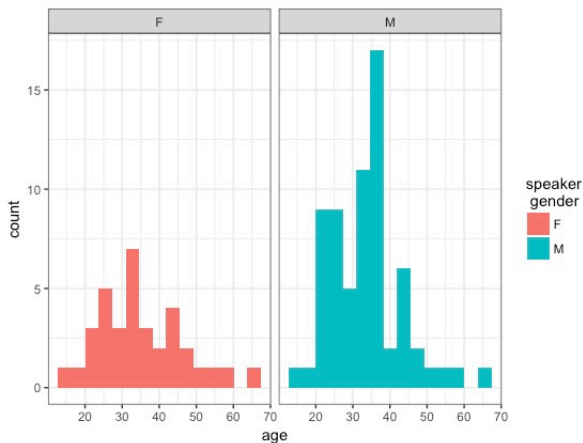


Figure 9: Distribution of age by gender

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods**
- 6 Findings
- 7 Future Work

Alignment

914-200 914-730 A: ANAN
 914-400 920-360 B: Re se Shabee do Shabee in e ke u hawa'n ruz tu Khune nist 'mahin sare kar e
 922-220 926-510 B: ke waazee Ne baDChehaye hamen protestooniha ta'tili dārand
 927-100 931-910 B: in Chon unja haat nan alibab barne harfe do ruz
 931-300 933-400 A: ANAN 'mahin namuz kareSH ro dāre na
 933-360 934-510 B: Are dare
 933-500 937-780 B: haw khune kar aikone haw unja kar aikone
 939-800 940-160 B: Kholāse in e
 941-450 944-660 B: dige be saikawti telefon intowri e
 944-470 946-190 A: Are too Agahi zabi buwan tu ruznāe
 946-800 947-250 B: ANAN
 947-800 948-720 A: ke nita ak'at telefon koton
 948-700 949-420 B: Chera
 949-820 953-660 A: ino ye dāneShgahi haat dārand Chiz neunehāye
 950-170 956-100 B: doonake mohtalef ro jan' alkonand
 950-800 950-610 B: ANAN ANAN
 950-590 957-880 A: ke yakSH farsi e
 950-400 960-960 A: wifan in in telephone ak'al'an dāre zabt alShw
 960-950 963-160 B: Meh' ne
 963-740 964-530 A: Are Are telefon ak'ro al'an dārand zabt alkonand
 964-500 968-940 B: Dada pas se harfe dige ro nuznāe
 967-140 971-110 A: (laugh) Are npe 'firas bud xogfite moVaze zabunetun bāSHin
 970-640 972-480 B: ba'le (laugh)
 971-100 976-820 A: (laugh) Na ingā haw Chī alShw goft hāh haw ke eidut AGHāye 'khalinton CHAR' ruz e da'vASH Shode BA joahuriKHAN
 970-800 980-900 B: Meh' Chera
 978-640 979-770 A: kA khuneneSHin Shodin
 980-910 983-300 A: (laugh) dowlat ro baband dige KHabar nadari eage
 983-340 987-610 B: na az kojā w'idut begu
 984-250 985-450 A: KHabar nadari' Nuuh vāsat begaw
 985-900 987-110 A: az da'
 987-500 989-930 A: sare budje tawafogh nakardand bā haw
 989-870 990-650 B: ANAN
 989-720 991-330 A: Aghe mebin al'an joahuriKHAN, aksariyat dārand toye wājes va toye sanā
 990-390 996-200 B: ANAN
 996-150 999-810 A: ba'd ra'is joahur bAHASHin jur dar jur dar naywand sare budje
 999-920 999-200 B: ANAN
 998-200 992-660 A: ba'd az zohre seShabee
 992-660 993-360 B: ANAN
 993-170 998-450 A: dowlat ta'tili Shod ya'ni Gheyr az Adewāyi ke hataan alibāyi sare kar bāSHand ke kārASHin
 994-750 995-330 B: (())
 998-410 999-820 B: ANAN
 998-920 912-240 A: wāzhe sādārsad e nadāike ye welyun nafar bikār Shodand
 912-250 914-510 B: ey wāy
 913-870 917-960 A: na Chizi nist wifani workhāsi begiri ya'ni waj- puli az daste kA newire
 917-900 918-920 B: ANAN
 918-360 920-500 A: woneteha umadān Khune workhāsi ejbāri dāria agiris
 920-500 921-710 B: ANAN
 921-500 924-940 A: ba'd ba'CHCHI dige wespe al'an se ruz
 923-800 925-120 B: akhāresh kojā KHANhad kashid
 925-870 927-750 A: DADA mojbūr and ye tawafogh konand chetwālan tā fardā pas fardā KHANhad kard
 926-610 928-270 B: Meh
 930-870 933-720 A: wāli al'an se ruz o nāw e ke dowlat dare dokkunesH ro baste
 934-800 934-330 B: ANAN
 934-120 937-740 A: Twāwan in Chiz wazalan parkhā indye dowlati haw ta'tili Shode
 937-920 938-670 B: 'wāb
 939-450 944-370 A: in edkreje neandunaw wālyāt o edkreje neandunaw komākhāye ejtemā'i haw darASH baste Shode
 944-360 945-640 B: 'wāb
 945-450 947-520 A: Are al'an CHAHār ruz e dowlat ta'tili e
 947-740 950-750 B: nan neandunaw 'huShāng az 'huShāng harf newāne ke
 950-540 951-840 A: Are
 951-430 954-350 A: kA ro az zohre se Shabee ferestānd Khune goftand Khodkhāfz
 954-190 954-800 B: ANAN

Figure 10: A typical transcription file

Alignment

```
861.740 864.530 A:      Are Are telefone mA ro al'An dArand zabt mikonand
864.560 868.040 B:      bAbA pas %e harfe dige ro nazanim
867.140 871.110 A:      {laugh} Are age ^irAn bud migoftam movAzebe zabunetun bASHin
870.640 872.480 B:      ba'le {laugh}
871.180 878.020 A:      {laugh} na injA hame CHi miSHe goft hAlA ham ke miduni AGHAYe ^kilinton
878.080 880.990 B:      %eh CHerA
878.640 879.770 A:      mA KHuneneSHin SHodim
880.810 883.300 A:      {laugh} dowlAt ro bastand dige KHabar nadAri mage
883.340 887.610 B:      na az kojA miduni begu
884.250 885.450 A:      KHabar nadAri %huh vAsat begam
885.990 887.100 A:      az da'-
```

Figure 11: A closer look

Alignment

The McGill Prosodylab-Aligner wrapper for HTK was used after some data processing:

Alignment

The McGill Prosodylab-Aligner wrapper for HTK was used after some data processing:

- split into utterances.

Alignment

The McGill Prosodylab-Aligner wrapper for HTK was used after some data processing:

- split into utterances.
- utterances with English words or code-switching were discarded.

Alignment

The McGill Prosodylab-Aligner wrapper for HTK was used after some data processing:

- split into utterances.
- utterances with English words or code-switching were discarded.
- {laugh}, {cough} etc. were discarded.

Alignment

The Farsi language model was trained on 21,002 sentences.

Alignment

Alignment was essentially perfect at the word level, and excellent at the phone level.

Alignment

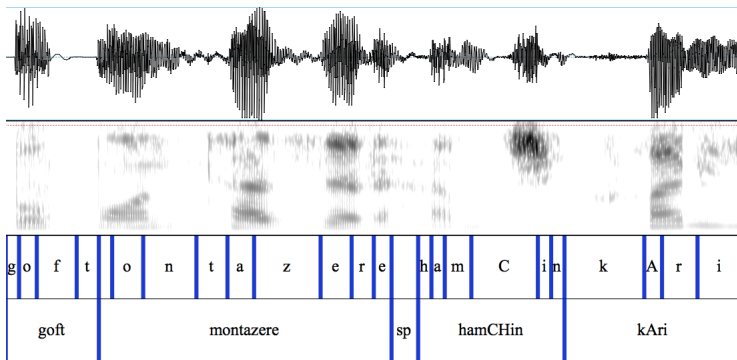


Figure 12: A sample alignment

Alignment

The end result was **70,711** vowels.

Extraction

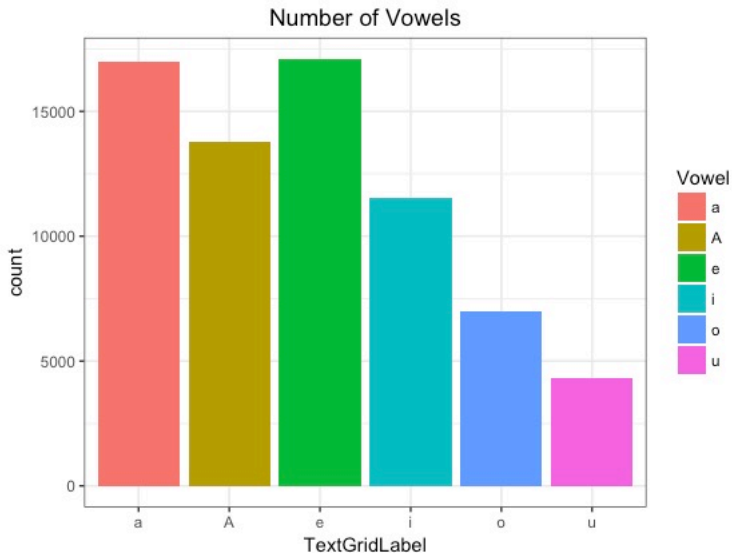
Two methods were used for vowel extraction:

- Praat scripts
- R scripts

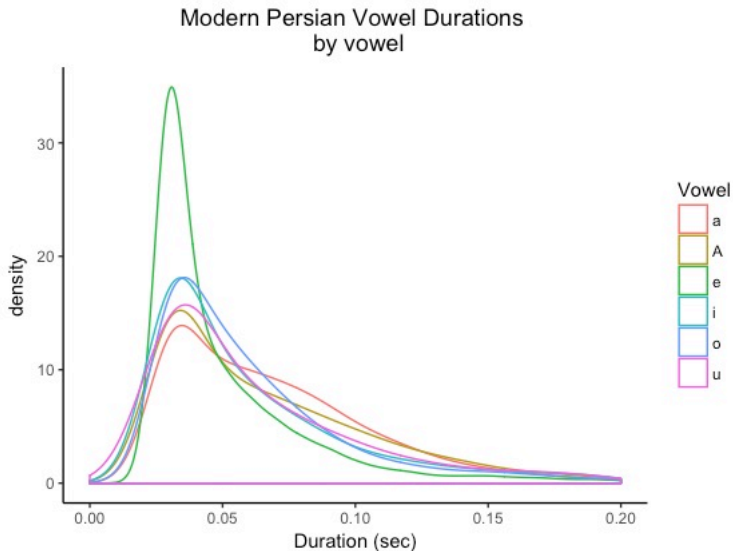
Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods
- 6 Findings**
- 7 Future Work

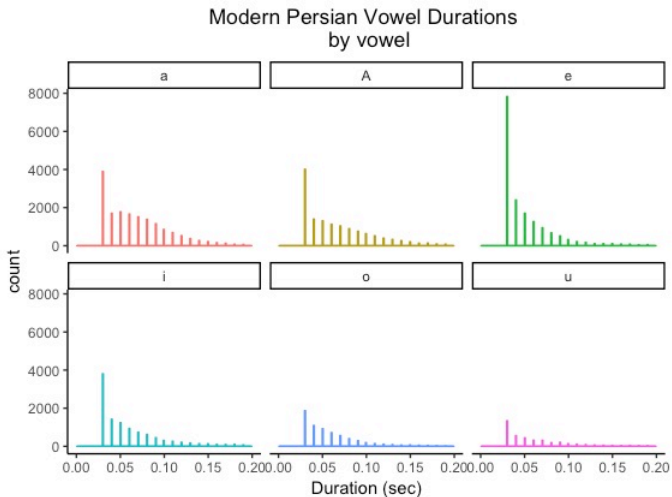
Summary



Summary



Summary



Vowel Reduction

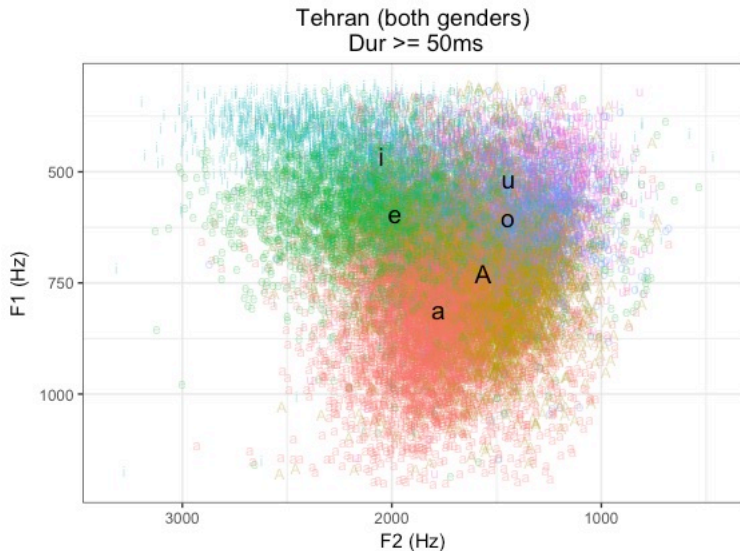
Persian has (normally) final stress. Stressed vowels are longer in duration.

Vowel Reduction

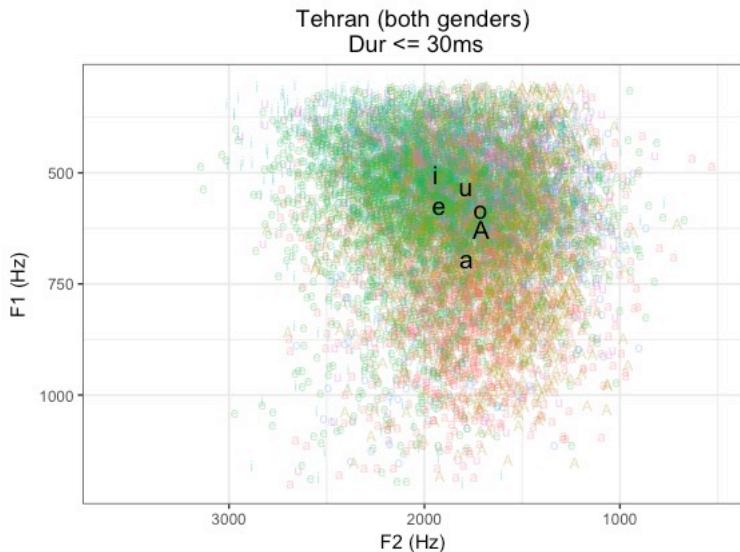
Persian has (normally) final stress. Stressed vowels are longer in duration.

One fact that emerged from the corpus is that Contemporary Iranian Persian vowels significantly reduce in unstressed position.

Tehrani Vowel Space



Reduction



Reduction

- *ruz* → [rz] 'day'
- *hich* → [htʃ] 'any'
- *chiz* → [tʃs] 'thing'
- *pul* → [p^hl̩] 'money'
- *miforusham* → [mifrʃæm] 'I am selling'
- *miforushin* → [mifrʃn] 'you (pl.) are selling'
- *pas* → [p^hs] 'so'

Assimilation? Reduction? Harmony?

More work is needed to tease apart what is happening with vowel assimilation.

Assimilation? Reduction? Harmony?

More work is needed to tease apart what is happening with vowel assimilation. So far, little phonetic evidence for the phenomenon.

A Wrinkle

The mysterious low back vowel may not be a single steady vowel for all speakers.

A Wrinkle

The mysterious low back vowel may not be a single steady vowel for all speakers. In fact, it looks (and sounds) like a diphthong for many speakers.

A Wrinkle

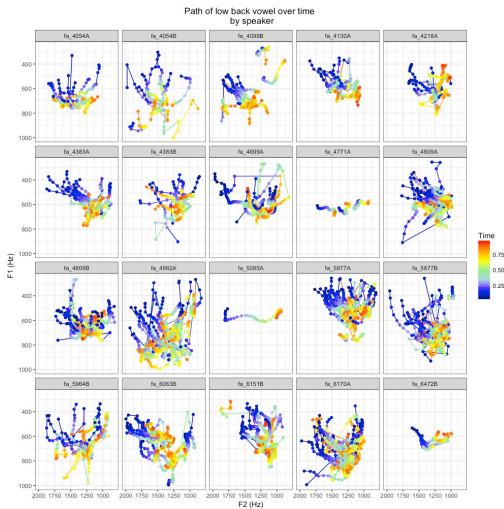


Figure 13: Diphthong? Off-glide?

A Wrinkle

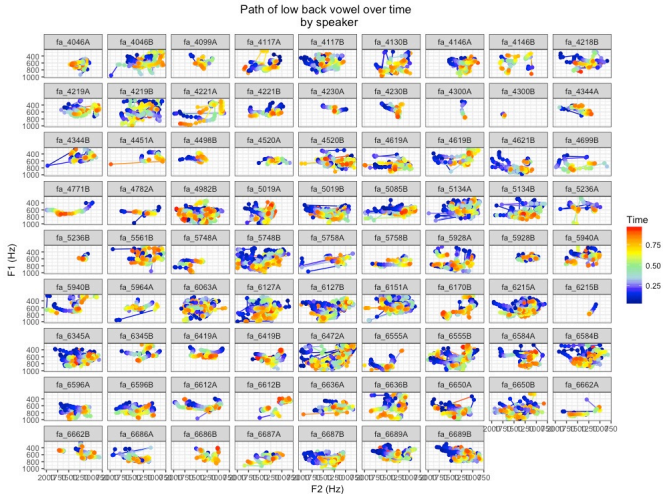
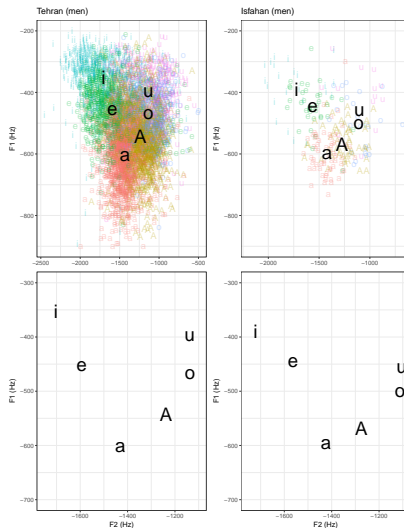


Figure 14: Other Speakers

Regional Variation



Gender Variation

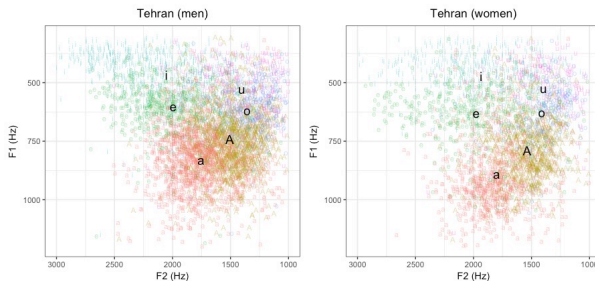


Figure 15: Vowel plots for Tehrani speakers by gender

Table of Contents

- 1 Goals
- 2 Background
- 3 Previous Work
- 4 Corpus
- 5 Methods
- 6 Findings
- 7 Future Work**

Future reseach

Big Question (1):

Future research

Big Question (1): How empirically real, and distinct from normal reduction processes, is the phenomenon of vowel assimilation in casual Contemporary Iranian Persian?

Future reseach

Big Question (2):

Future research

Big Question (2):

Assuming assimilation is empirically supported, do the phonetic patterns here inform our thinking about the phonological explanation?

Future reseach

Big Question (3):

Future research

Big Question (3): Given a corpus with such a good balance of age, gender, education, and location, is there evidence of sociolinguistic variation in the vowelspace of Contemporary Iranian Persian?

The End

Thank you!

Special thanks to Mark Liberman, Jianjing Kuang, Kyle Gorman, and the Linguistics Data Consortium.