Using forced alignment for segmental analysis

A Review

Erin Olson, Michael Wagner, Meghan Clayards
McGill University

Computational Field Workshop
McGill University, Montréal
28 May 2013

Table of Contents

What is the Prosodylab Aligner?

The Prosodylab Aligner (Gorman *et al.* 2011) is a tool for performing **forced alignment** on audio data

- Some details:
  - **Python** codebase
  - Compatible with **UNIX**-based systems (so far)
  - Based on the **Hidden Markov Model Toolkit** (HTK)
- It takes these files...
  - **.lab** files (transcripts of the audio files)
  - **.wav** files (the audio files themselves)
- ... and gives back **.TextGrid** files (readable in Praat (Boersma & Weenink 2013))
  - Both **words** and **segments** are aligned
  - No previously aligned data is necessary – just a transcript

Aligner demo

Demo

Training the Aligner

The Aligner is also capable of being **trained** on different data. This data can come from:

- a single **speaker**
- a single **dialect**
- a new **language**

Training the Aligner requires:

- At least **two hours** of transcribed training data
- A **phonetic dictionary**, such as the CMU Pronouncing Dictionary or Lexique

Training demo

Demo

Goals for this talk

- We've seen how good word alignment can be, but what about **segmental alignment**?
- How can we make this tool as useful as possible for **field linguists** in its present state?

Table of Contents

The studies

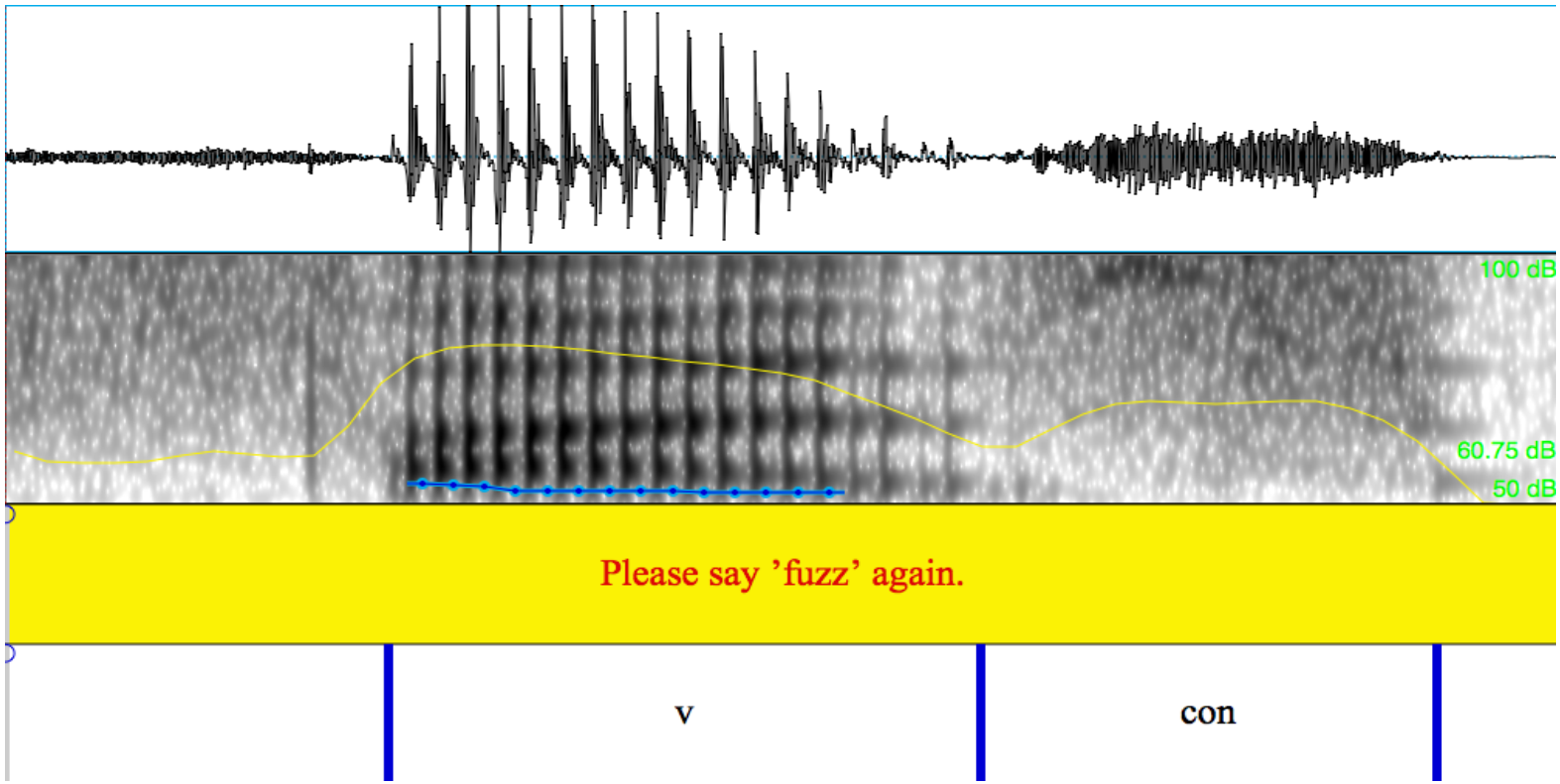
Goal: Hayes (2007) claims that vowel phonemes are "realized as extra short when a voiceless consonant follows" in English. Is this really the case?

- Two experiments performed comparing vowel length before **voiced** and **voiceless** obstruents to vowel length words before **sonorants**
  - One with **fricatives** (F): *fuss*, *fuzz*, *fun*
  - One with **stops** (S): *cot*, *cod*, *con*
- More details on experimental design:
  - All words were **monosyllabic** and spoken in a **carrier phrase** "Please say ___ again"
  - Experiment F had 6 (near) minimal triplets comparing [s] and [z] with [m] or [n]; 19 participants
  - Experiment S had 30 minimal triplets comparing stops with [m], [n], [ŋ], [l]; 27 participants
    - Participants only saw one word of each minimal triplet

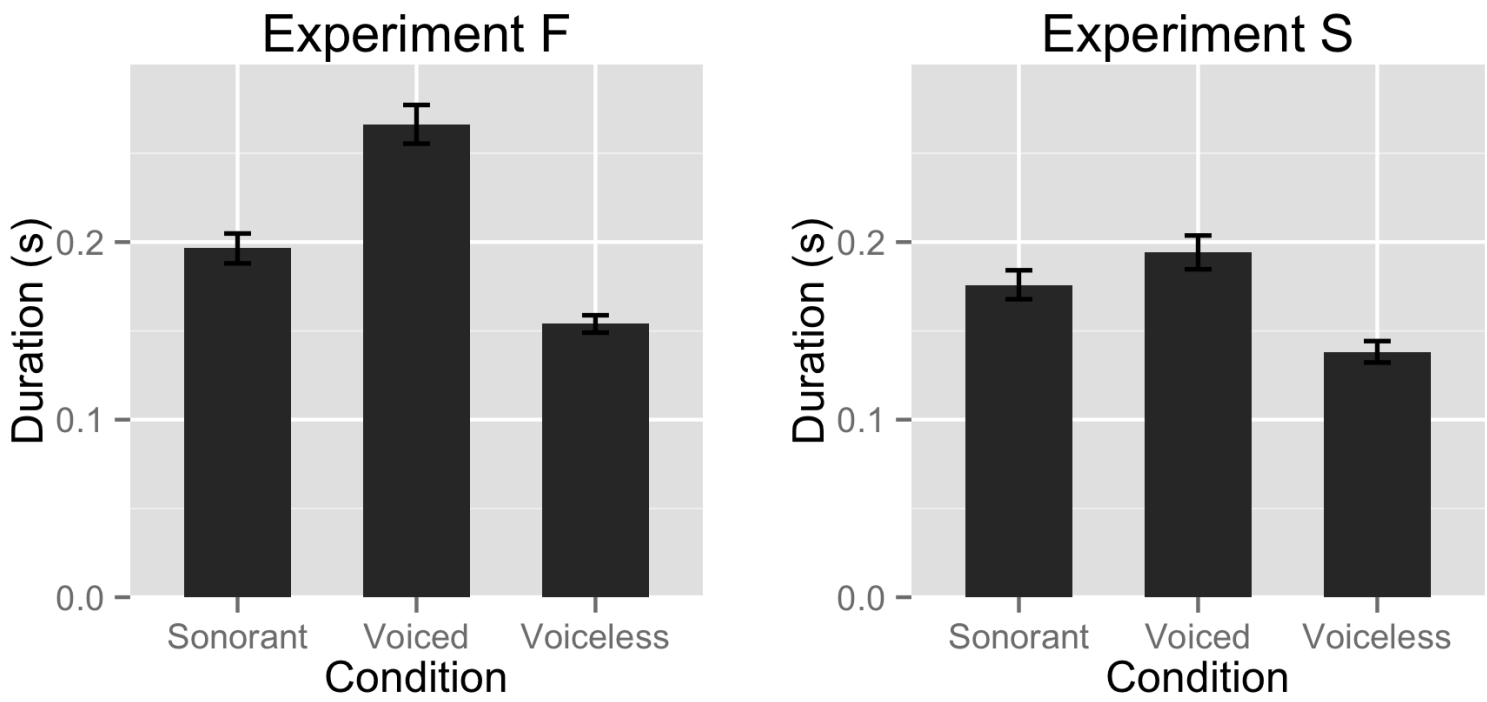
Human annotation

Two research assistants aligned the **vowel of interest** and the **following consonant** for both experiments

- For experiment S, stop consonants were split into closure and burst components



Results: Human annotation



Results of human annotation for experiment F and experiment S. Error bars represent 90% confidence intervals. All differences are significant, as found by a linear mixed model regression.

Results: Human annotation

- All **three** conditions are significantly different from one another in both experiments
- For experiment F, the **Sonorant** and **Voiceless** conditions were closest ($|t| = 3.628$)
- For experiment S, the **Sonorant** and **Voiced** conditions were closest ($|t| = 4.254$)

Alignment

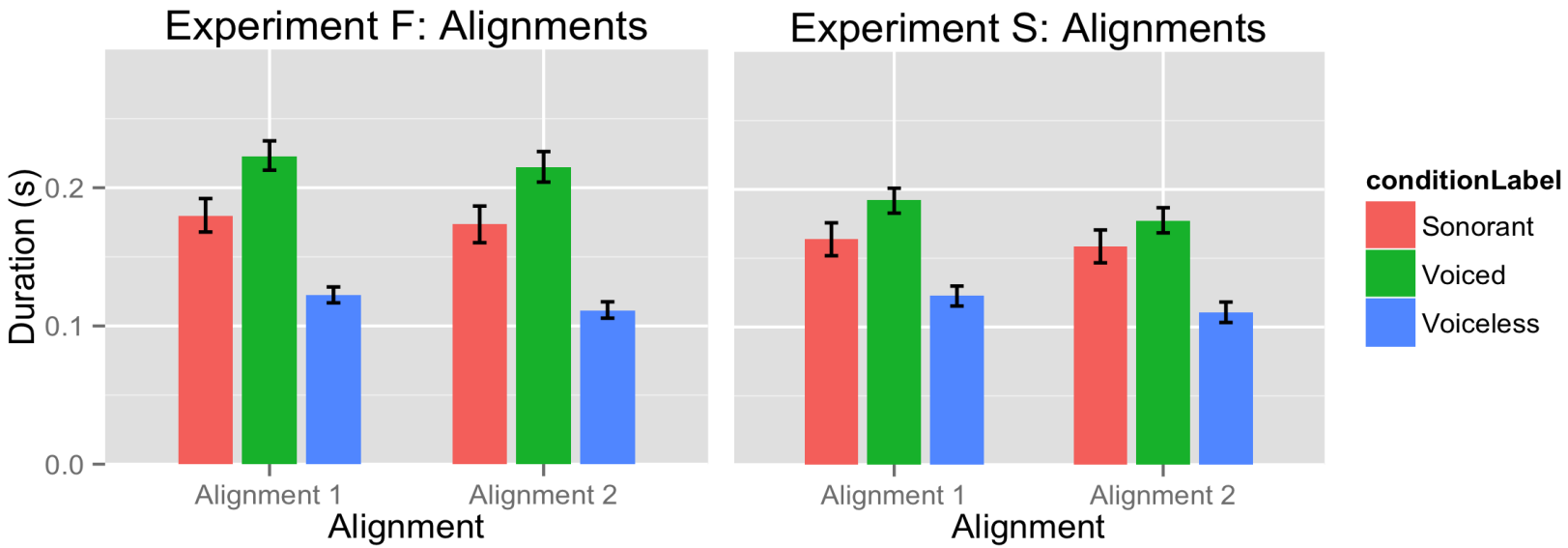
The training set:

- Around **four hours** of training data
- Previously collected through other Prosodylab experiments

Two alignments performed:

- One using the **CMU Pronouncing Dictionary** (Alignment 1, or A1)
- One using a modified version of the Pronouncing Dictionary, where stops are separated in **closures** and **bursts** (Alignment 2, or A2)

Results: Alignment



Results of aligned annotation for both experiments. Error bars represent 90% confidence intervals. All differences between conditions are significant, as found by a linear mixed model regression.

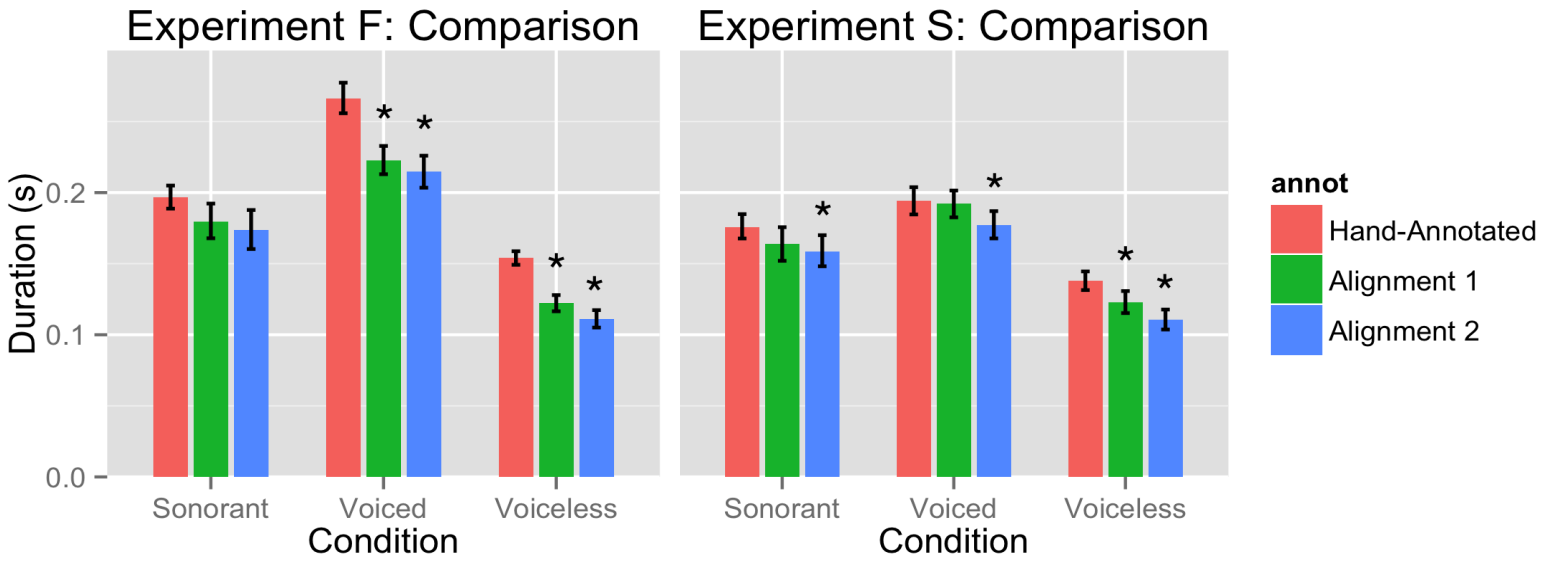
Results: Alignment

- All **three** conditions are still significantly different from another, in both experiments and both alignments.
- For experiment F, the **Sonorant** and **Voiceless** conditions were closest to one another ($|t| = 2.611$ for A1 and $|t| = 2.876$ for A2), just as in the hand-annotated data
- For experiment S, the **Sonorant** and **Voiced** conditions were closest to one another ($|t| = 3.192$ for A1 and $|t| = 2.147$ for A2), just as in the hand-annotated data

Take home message: the alignments give the same **qualitative** result as the hand-annotated data

Assessment: Duration

Are the measures of vowel duration **significantly different** from the human-annotated durations?



Results from all annotations, grouped by condition and annotation. Error bars represent 90% confidence intervals. Asterisks indicate significant difference from hand annotation, as measured by a mixed model linear regression.

Assessment: Duration

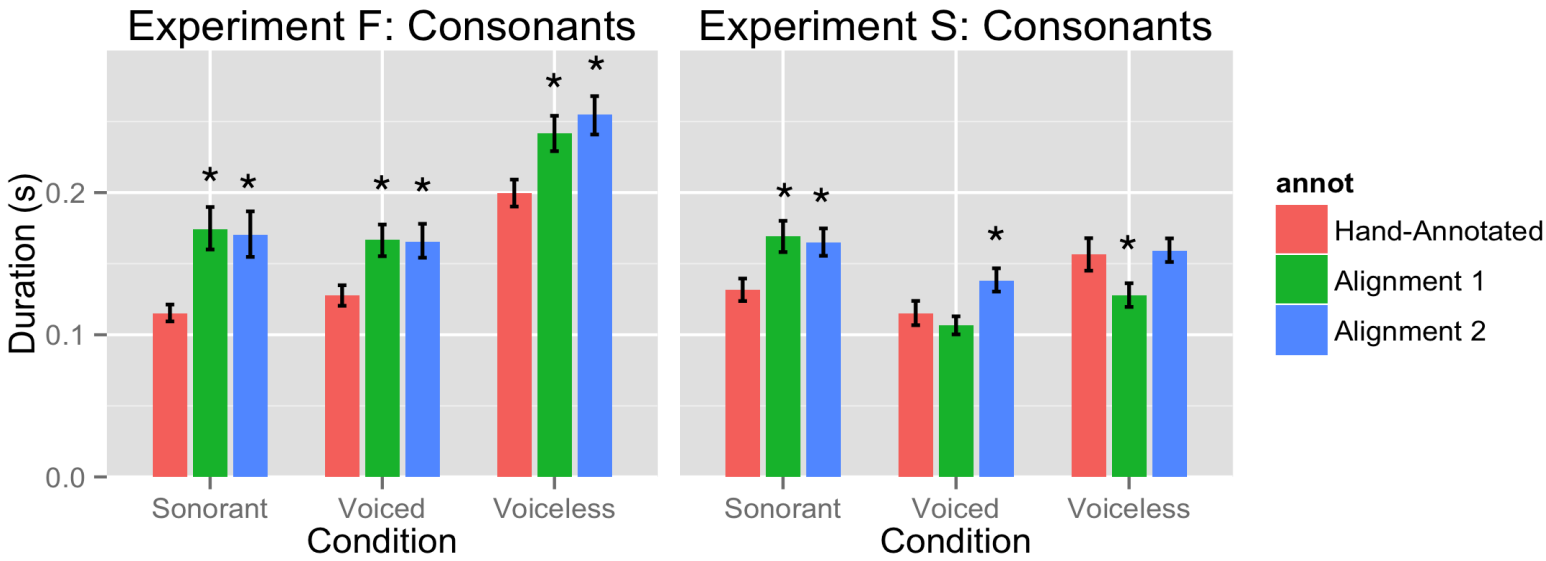
Durations as measured by the aligner are generally **significantly different** from durations as annotated by humans

- **Sonorants** seem to be the exception
- **A2** in experiment S is always significantly different from hand-annotation
  - Contrary to expectations – A2 mirrors human annotation style better

What is behind this consistent discrepancy?

Assessment: Consonant measures

How much of the vowel is being aligned with the consonant?



Consonant duration results from all annotations, grouped by condition and annotation. Error bars represent 90% confidence intervals. Asterisks indicate significant difference from hand annotation, as measured by a mixed model linear regression.

Assessment: Consonant measures

- In general, aligned consonant durations are **higher** than hand-annotated consonant durations
  - Implication is that part of the vowel is consistently being aligned as part of the consonant
- Of course, this could also be due to right-edge discrepancies as well, although a visual check of the alignments reveals that both are a factor

Assessment: Items

Durations measured for each individual item and condition for all alignments in all experiments. The alignment was deemed "bad" if the confidence intervals did **not overlap**.[1] Problematic items were then checked visually to see where the problem was

- In experiment F, particularly bad at:
  - the minimal triplet *bus* / *buzz* / *bun*, at both boundaries
  - in the environment [**r**_**s**], such as *grace, gross, rice*, at both boundaries
- In experiment S, particularly bad at:
  - vowels before [**ŋ**], such as in *ring, king*
  - the vowel [**ʌ**], such as in *luck, buck*, at both boundaries
  - vowels after [**ɹ**], such as in *root, rune, right, trait*
  - vowels after **glides** [j, w], such as in *white, mule*

[1] Broad, preliminary results only

Discussion

What makes the alignment models different from human-defined models?

- Consistently takes part of the vowel and counts it as part of **consonant**
- Seems to misalign segments that even humans have trouble with – particularly:
  - Rhotics
  - Glides
- Seems to misalign segments that aren't **frequent** in the training corpus, such as [ʌ] and [ŋ]

Why might this be?

- Aligner models look for the **first relevant cue**, and makes the boundary there
- No opportunity for **overlapping boundaries**, so a choice is forced

Table of Contents

Setbacks to current methods

Developed for use on **well-studied**, **well-established** (Indo-European) languages

- Requires a **consistent orthography** to have been developed and used extensively for the language
- Requires a **phonetic dictionary** to have been developed for the language
  - ... if the orthography doesn't match the phonetic content in the first place
- Assumes that words are relatively **short** and **unchanging**, with minimal morphological differences
  - Could be solved for these languages by making a **morpheme dictionary**, but that opens up a whole other set of assumptions

    What is the correct morphemic analysis? What are the accepted allomorphs of each morpheme? How much time can we spend making such a dictionary? etc.

The quick and dirty way: use what you have

Make a "phonetic" dictionary (and .lab files) from the transcribed data that you already have. Two methods

- Method 1: Transcriptions stored in a **spreadsheet**
- Method 2: Transcriptions stored as **TextGrids** or **ELAN** files

Method 2 has been used on the following Mi'gmaq data

- Two brief stories (around **5 minutes** total duration, or 314 utterances of varying length) elicited from a single speaker as part of the Field Methods in Linguistics course at McGill University
- Aligner models **trained** on these stories
- Both stories also **aligned** using the same models
- 25 utterances were also aligned **by hand** for comparison

Data collection demo

Demo

Broad assessment

Fairly ok at the **word** level. What about at the **segment** level?

- For each segment, beginning and end time were measured for both hand annotations and the alignment
- Differences between annotation types were calculated for each boundary
- Global results:
  - **Start**: average difference of 18 ms
  - **End**: average difference of 21 ms

Conclusion & Discussion

- The aligner gives the same **qualitative** results as human annotation with respect to duration
  - What about automated measurement tasks which rely on duration?
- The aligner gives **quantitatively different** results from human annotation
  - What can this tell us about segmental cues?
- The aligner is **trainable** on multiple languages, but not all of these languages have the resources necessary
  - What other tools would be useful for doing this sort of task?

Sources & Thanks

Boersma, Paul & Weenink, David (2011) *Praat: doing phonetics by computer* [Computer program]. Version 5.2.40, retrieved 11 September 2011 from http://www.praat.org/

Gorman, Kyle, Jonathan Howell & Michael Wagner (2011) "Prosodylab-Aligner: A tool for forced alignment in laboratory speech." *Proceedings of Acoustics Week in Canada, Quebec City*

Hayes, Bruce (2007) "About phonemes" *Introductory Phonology* Wiley/Blackwell

This research was supported by an FQRSC grant and a SSHRC grant to Michael Wagner, McGill University. Thanks go to the Prosodylab research assistants for their help in data collection, annotation, and assistance with interpreting data: Aron Hirsch, Lauren Mok, Elise McClay, and Thea Knowles. Thanks also go to Janine Metallic for her immeasurable assistance with the Mi'gmaq stories. Finally, special thanks go to Yasemin Boluk for her aid in developing the Python tools used in the demos.