Prosodic context in computational modeling of tone:

citation tones vs. running speech

RIKKER DOCKUM
YALE UNIVERSITY

BLS43, UC BERKELEY
FEBRUARY 5, 2017

BCS-1528386

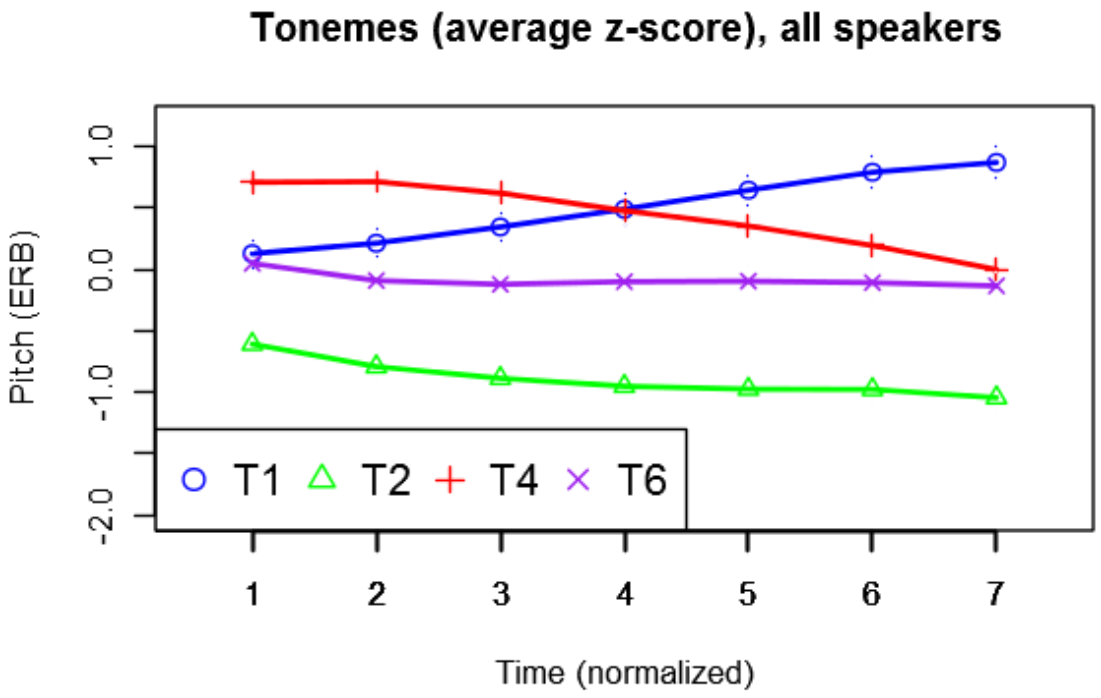
Yale MACMILLAN CENTER

Goals

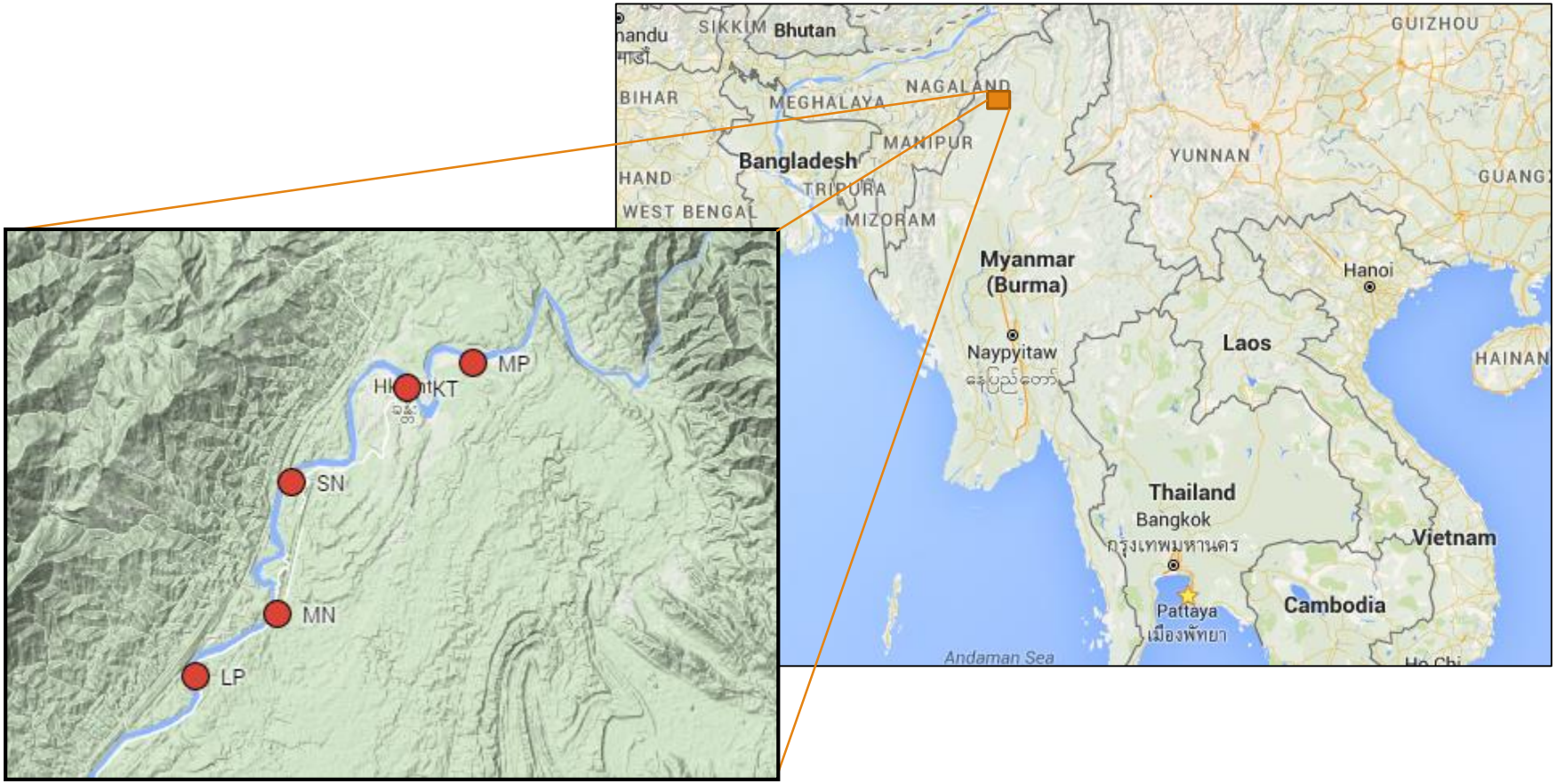
- **Explore computational modeling of tonal systems**
  - Expand on work by Shosted et al 2014, 2015
  - Use in identifying toneme categories, yes, but…
  - Computational modeling is a back-and-forth process
- **Identify the uses and limitations of the method**
  - Citation tones vs. running speech
  - Demonstrate how it can identify places where more human attention is needed
- **Model data gathered under very imperfect conditions**
  - Which, given limited time in field/limited access to speakers, can accelerate and improve results

Background

Tai Khamti tonemes



Data gathering locations: Upper Chindwin river valley



Data gathering

#	Form	Gloss	#	Form	Gloss	#	Form	Gloss	#	Form	Gloss
1	maː^{1}	dog	2	kʰaw^{2}	rice	3	paː^{4}	fish	4	kai^{6}	chicken
5	mi^{1}	bear	6	maː^{2}	horse	7	kʰaːi^{4}	buffalo	8	kʰaː^{6}	galangal
9	pʰaː^{1}	wall	10	ɔi^{2}	sugarcane	11	naːw^{4}	star	12	taw^{6}	turtle
13	sʰɤ^{1}	tiger	14	saːng^{2}	elephant	15	nɤn^{4}	moon	16	tʰo^{6}	bean/nut
Tone 1			Tone 2			Tone 4			Tone 6		

Frame questions:

1. Have you ever seen / eaten / etc _____?
2. What kind of _____ have you seen / eaten / etc?
3. Where have you seen ___ / Where can _____ be found / etc?

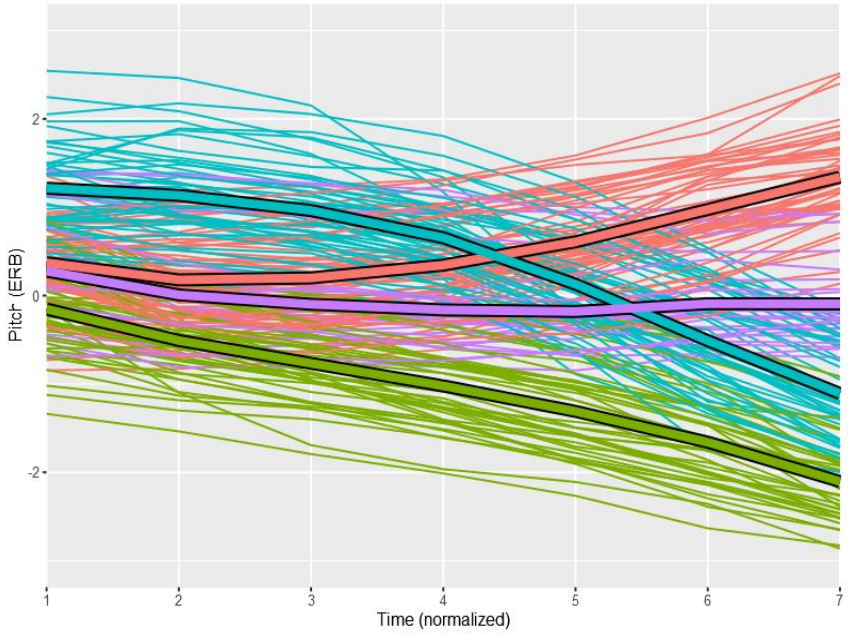
Corpora

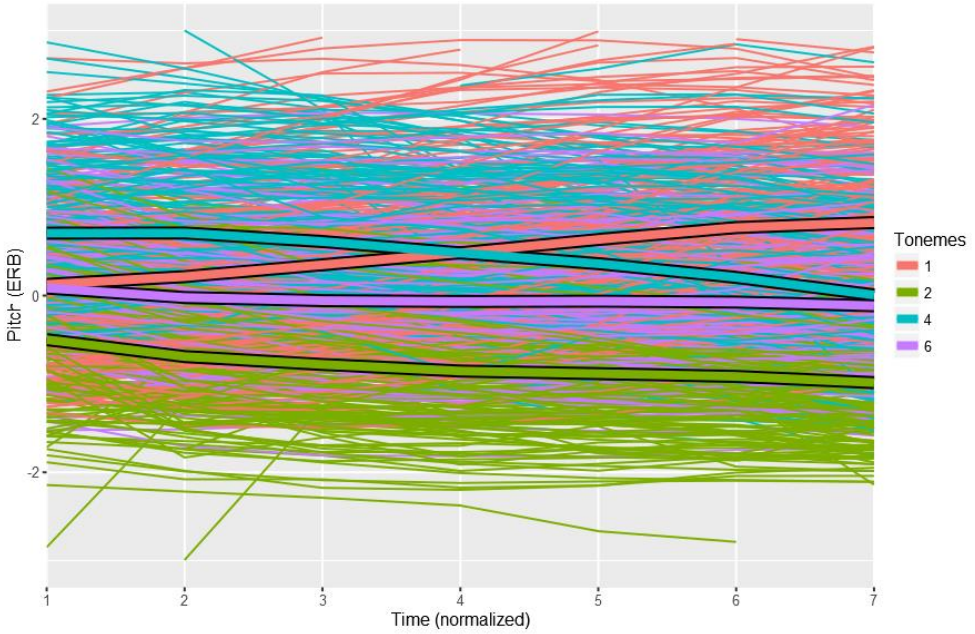
1. Question answering (stimuli response corpus)
   - 750 tokens
   - 16 types
   - 5 speakers (of 37 recorded)
   - Controlled for syllable shape (14 CV, 2 CVN)
2. Wordlist reading
   - 173 tokens
   - ~50 types
   - 1 speaker (of 5 recorded)
   - Controlled for syllable shape (all CV)

Corpora

ID	Sex	Age	Data type	Tone 1	Tone 2	Tone 4	Tone 6	Total
LP5	F	24	Wordlist reading	54 (31%)	43 (25%)	47 (27%)	29 (17%)	**173**
MN2	F	42	Stimuli responses	35 (28%)	18 (14%)	39 (31%)	32 (26%)	**124**
SN2	M	74	Stimuli responses	49 (31%)	37 (23%)	39 (25%)	34 (21%)	**159**
SN3	F	40	Stimuli responses	47 (31%)	34 (22%)	37 (24%)	34 (22%)	**152**
SN5	F	20	Stimuli responses	40 (24%)	43 (26%)	43 (26%)	40 (24%)	**166**
KT5	M	65	Stimuli responses	33 (22%)	34 (23%)	41 (28%)	41 (28%)	**149**
				204	**166**	**199**	**181**	**750**

Normalized pitch tracks





Wordlist reading

n = 173

(tones in isolation)

Stimuli responses

n = 750

(tones in situ)

Computational Modeling

Methods

- Principal component analysis (PCA)
  - e.g. Joliffe 1986, Johnson 2008
  - Dimensionality reduction
- K-means clustering
  - Hartigan & Wong 1979
  - One of many possible clustering techniques

- Traditionally-classified tones = “ground truth”

Principal Components Analysis

Principal Components Analysis

Dimensionality reduction

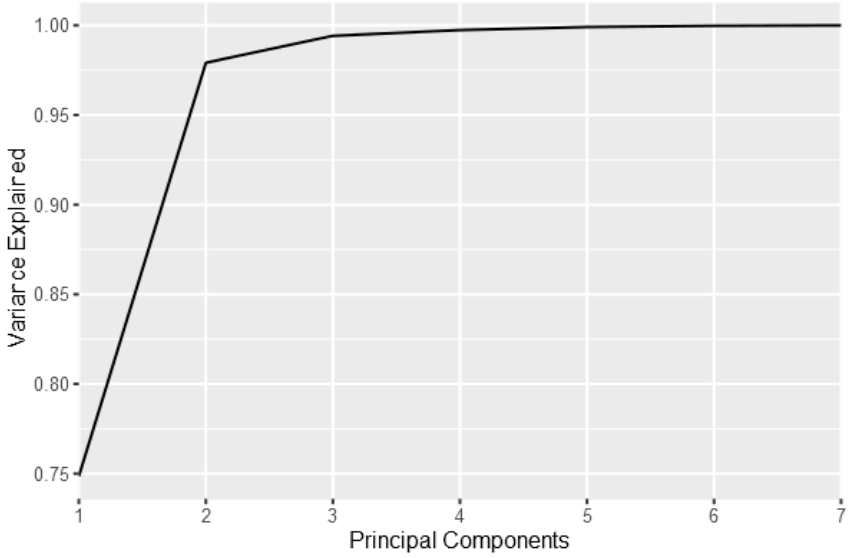
	pitchstep1	pitchstep2	pitchstep3	pitchstep4	pitchstep5	pitchstep6	pitchstep7
1	1.086929934	0.450587307	0.483235657	0.655530171	1.102906335	1.44515143	1.440590791
2	0.607606651	0.338903221	0.372242407	0.402931670	0.618526143	0.91845228	1.196864524
3	0.450851789	0.126137969	0.264717697	0.450897779	0.711481611	0.95368215	1.174989172
4	-0.504643927	-0.360536037	-0.137285978	0.371874477	0.739305016	1.14582241	1.290947897
5	-0.725597388	-0.463384357	-0.286780011	0.059922230	0.671011203	1.03863363	1.104717024
6	-0.768527739	-0.949704933	-0.517437858	-0.097089133	0.526519200	0.95343230	0.978498179
7	-1.213585513	-0.572240997	-0.244463835	0.151368409	0.583430711	0.97442031	1.011408001
8	0.408315294	0.215909355	0.307380727	0.393959592	0.541063253	0.71207016	1.011795175
9	0.156641030	-0.354174286	0.075335465	0.301133094	0.503122246	0.85323952	1.041414014
10	-0.362067897	-0.135754143	0.133953775	0.380156395	0.720650687	0.94268843	1.133367927
11	0.068417189	-0.184881004	0.008392662	0.331155047	0.667533277	0.83400051	1.039284555
12	0.182635555	0.507843073	0.644175868	0.962996379	1.370706611	1.50736590	1.402066941
13	-0.128511030	0.053684686	0.030938165	0.224180272	0.548967630	0.77603362	0.989532649
14	-0.560965397	-0.252739689	-0.032882953	0.237293309	0.684606730	0.91145627	1.000373531
15	-0.388062422	-0.096523341	0.275470168	0.644832694	0.878422042	0.91470442	0.872993163
16	1.032183889	1.205515178	1.401704796	1.628655543	1.951520197	2.04231035	2.043614874
17	0.651324715	0.988155328	1.115897178	1.289787063	1.548713170	1.67327114	1.680445314
18	0.202722234	0.611044823	1.049648083	1.282540384	1.530375017	1.66802413	2.022901045
19	0.583975264	0.925244672	0.962587753	1.145543656	1.426037247	1.47538417	1.450463737
20	-1.675185257	-1.093904639	-0.645080094	-0.322081240	0.284329104	0.85973581	0.977336656
21	-1.840604960	-1.355796751	-1.061998488	-0.762403217	-0.225977443	0.31304807	0.480204763
22	-1.303384780	-1.194278944	-0.921522656	-0.463218927	0.077550615	0.43397900	0.649980725
23	0.520170522	0.736512704	0.869630906	1.184537687	1.560095472	1.60206181	1.416973154
24	-0.330559382	0.055098408	0.133606921	0.326669008	0.809495879	1.10559538	1.267330260

From this…

	PCA1	PCA2
1	-2.4169825015	-0.97576080
2	-1.5956611265	-0.76044886
3	-1.4801442446	-0.95699041
4	-0.8704152590	-1.81333082
5	-0.4375612131	-1.83175525
6	0.0510062678	-1.95933833
7	-0.1936571278	-1.95211841
8	-1.2898404271	-0.68502753
9	-0.8907753308	-1.14810871
10	-0.9961026222	-1.40490895
11	-0.9705493451	-1.16477252
12	-2.4371604947	-1.33307986
13	-0.8710089612	-1.07370114
14	-0.6867905593	-1.48647940
15	-1.1525733951	-1.21832554
16	-4.2136089546	-1.22648444
17	-3.3320837262	-1.08861056
18	-3.0901902518	-1.62464215
19	-2.9727961740	-0.93688689
20	0.7094457043	-2.33669529
21	1.7740247948	-1.98265422
22	1.1184813592	-1.86693098
23	-2.9468046171	-1.12894111
24	-1.1895248665	-1.49867652

…to this

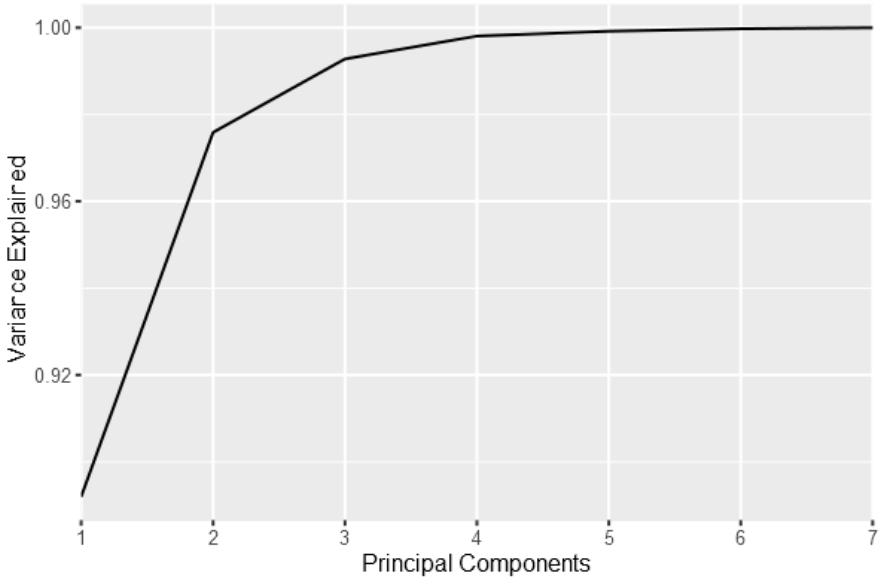
Principal Components Analysis



Wordlist corpus

Speaker = LP5

Measures = pitch (7 steps)



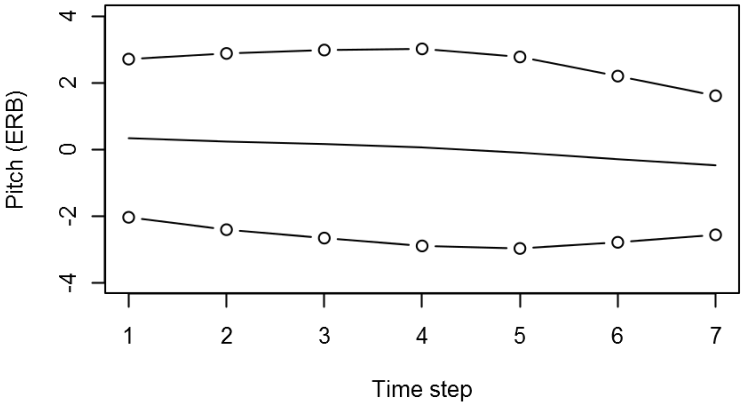
Stimuli corpus

Speakers = KT5, MN2, SN2, SN3, SN5

Measures = pitch (7 steps)

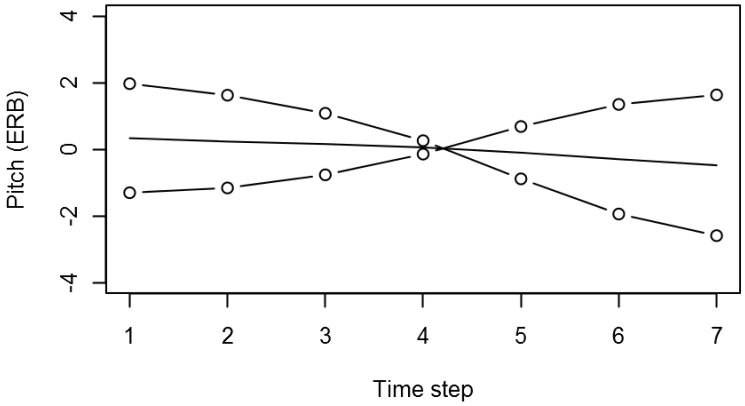
Identifying the PCs

PC1



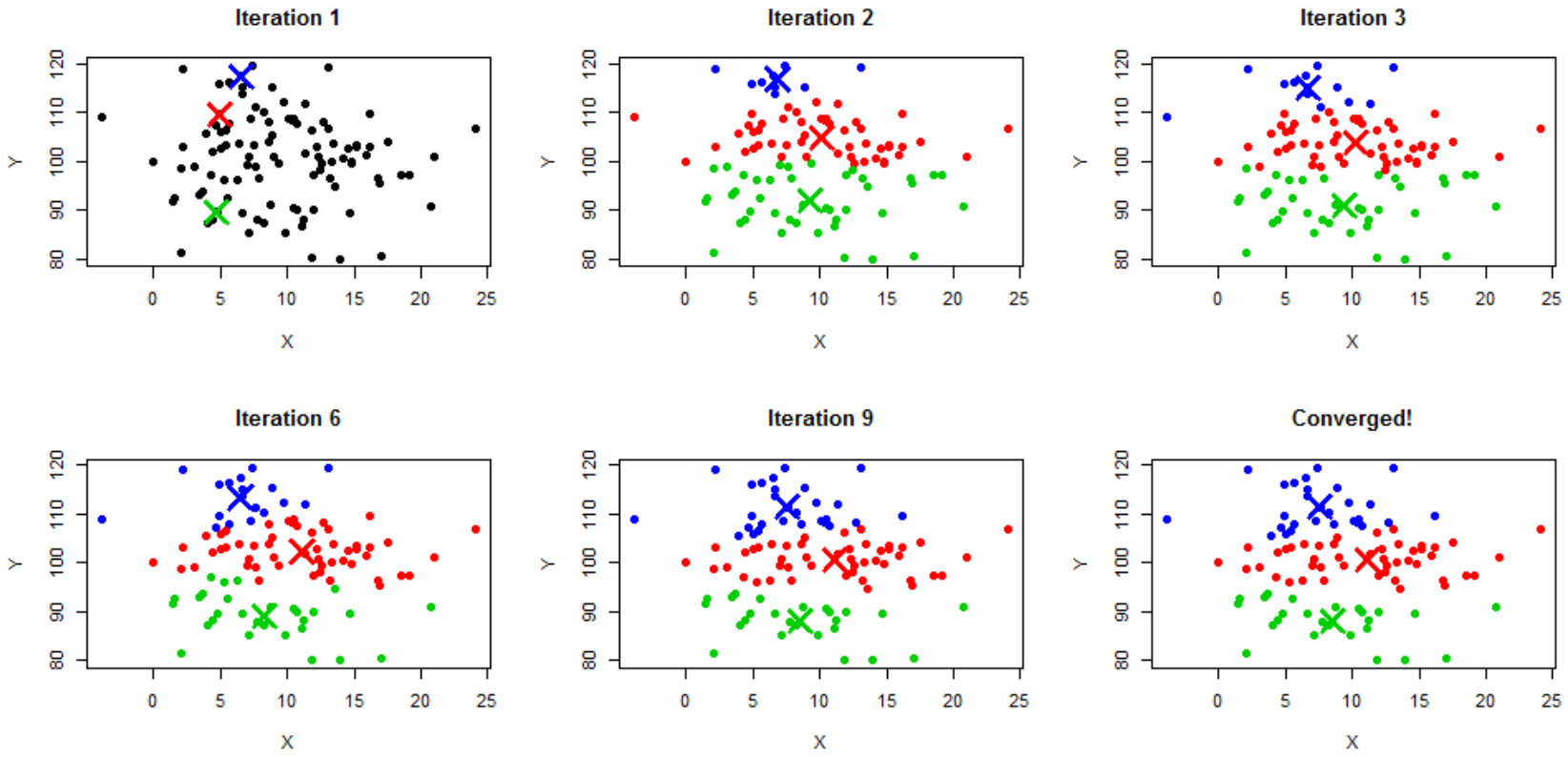
Pitch Height

PC2



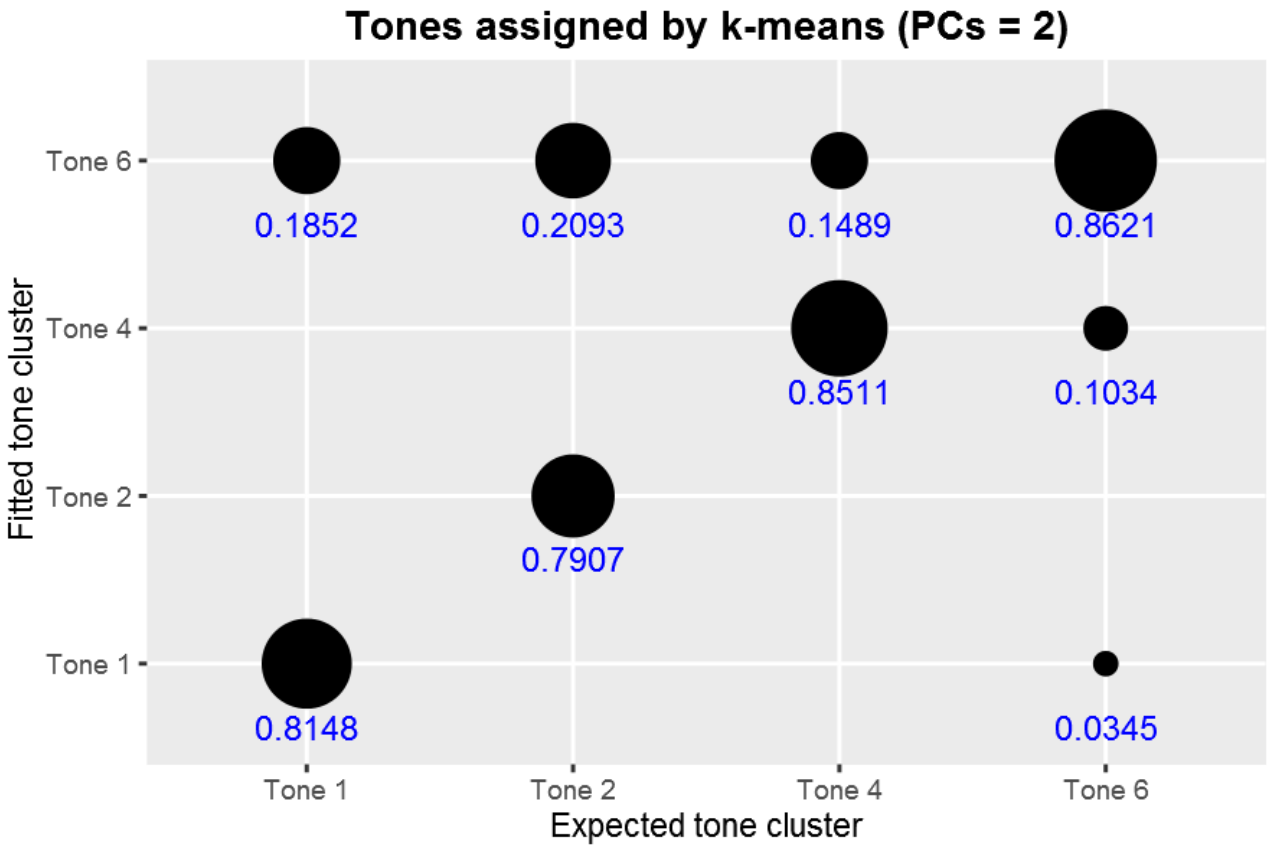
Pitch Slope

k-means clustering



[image from http://learnbymarketing.com/methods/k-means-clustering/]

Wordlist corpus (pitch only)

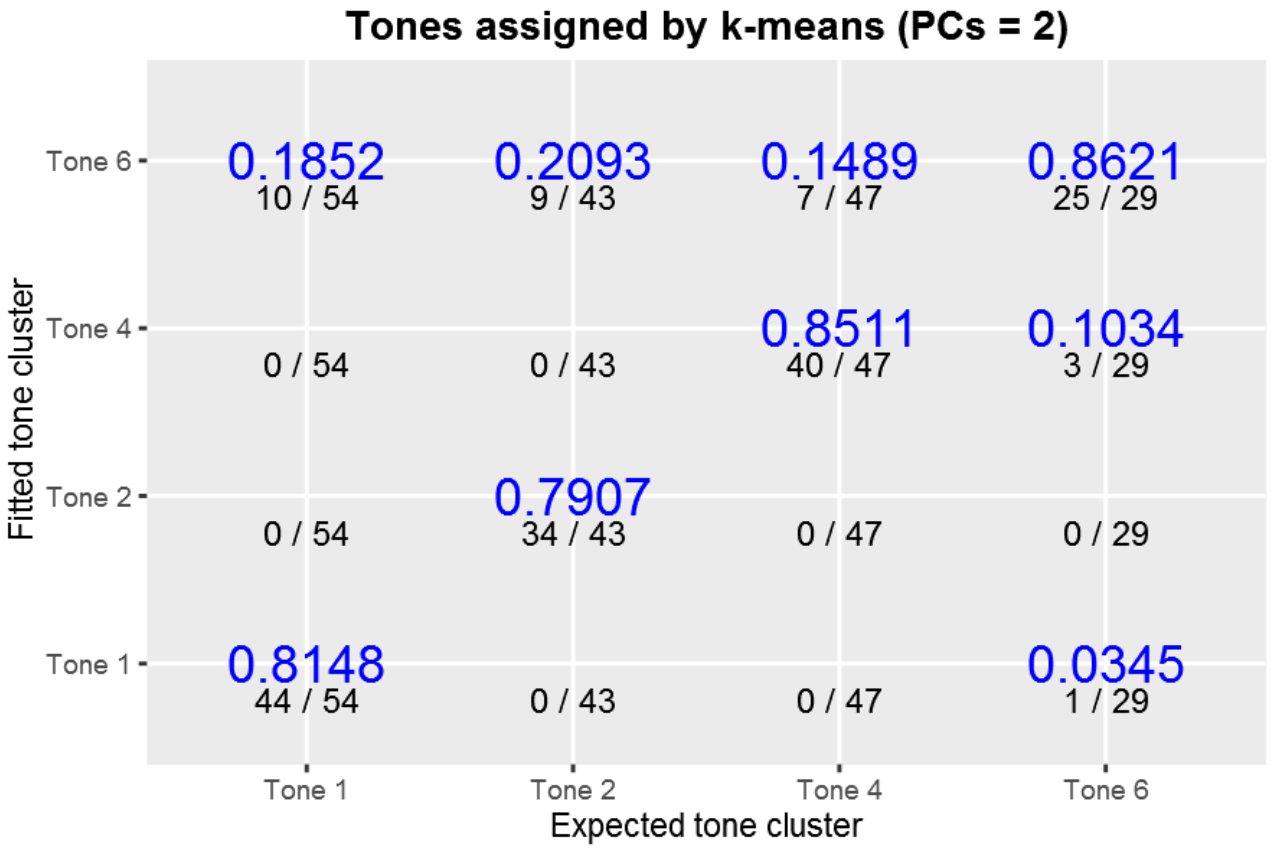


corpus = wordlist
measures = pitch
PCs = 2
***k* = 4**

	Precision	Recall	F-score
T1	0.9778	0.8148	0.8889
T2	1.0000	0.7907	0.8831
T4	0.9302	0.8511	0.8889
T6	0.4902	0.8621	0.6250
Overall	**0.8266**	**0.8266**	**0.8266**

T1 = rising
T2 = low falling
T4 = high falling
T6 = mid level

Wordlist corpus (pitch only)

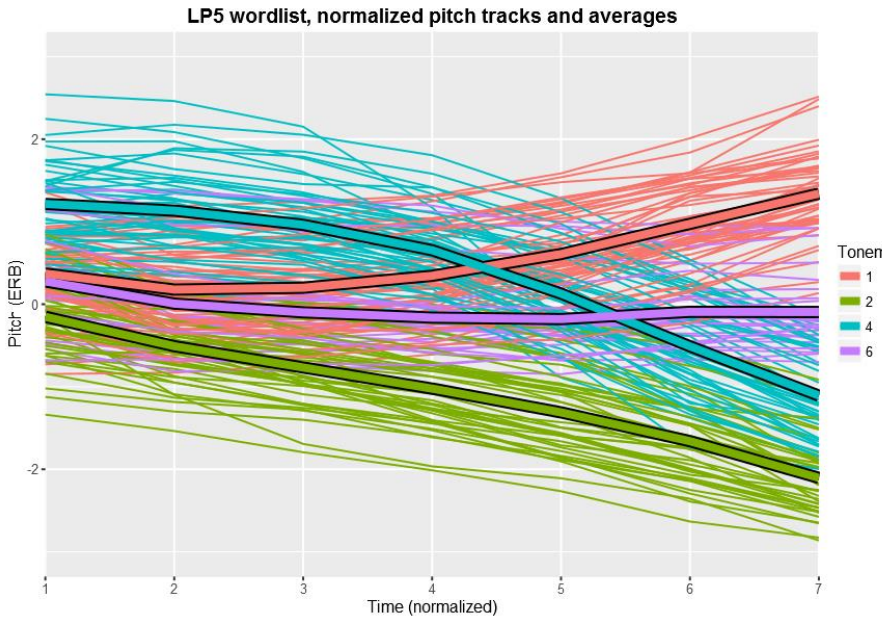


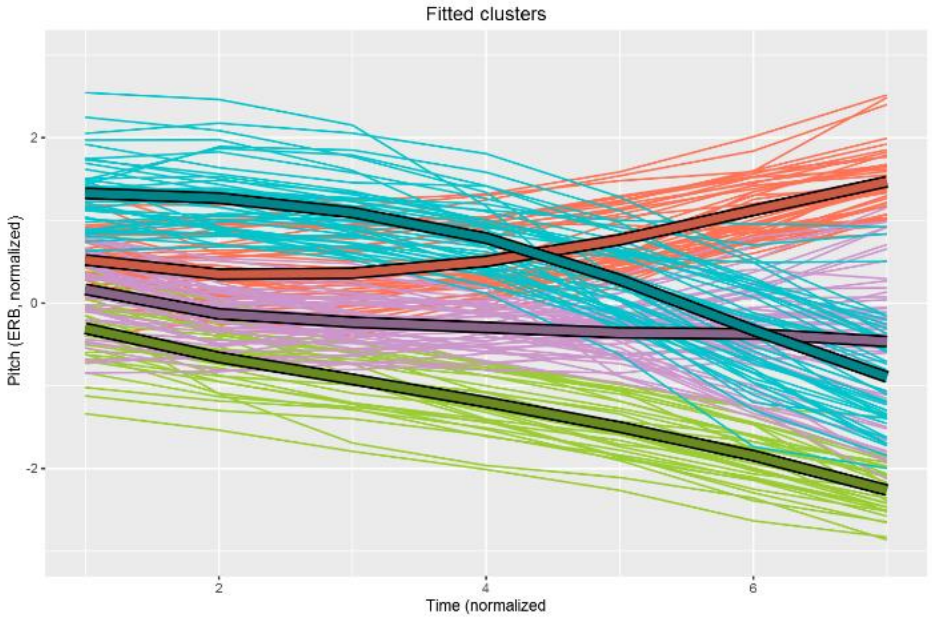
corpus = wordlist
measures = pitch
PCs = 2
***k* = 4**

	Precision	Recall	F-score
T1	0.9778	0.8148	0.8889
T2	1.0000	0.7907	0.8831
T4	0.9302	0.8511	0.8889
T6	0.4902	0.8621	0.6250
Overall	**0.8266**	**0.8266**	**0.8266**

T1 = rising
T2 = low falling
T4 = high falling
T6 = mid level

Wordlist corpus





Expected clusters

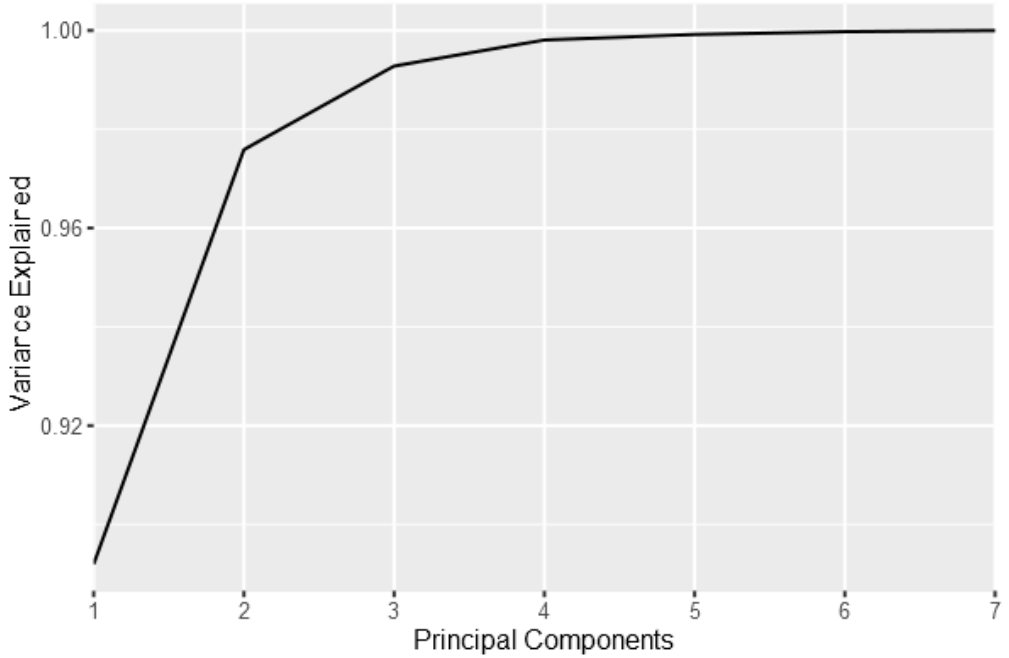
+ normalized pitch tracks

Fitted clusters

+ normalized pitch tracks

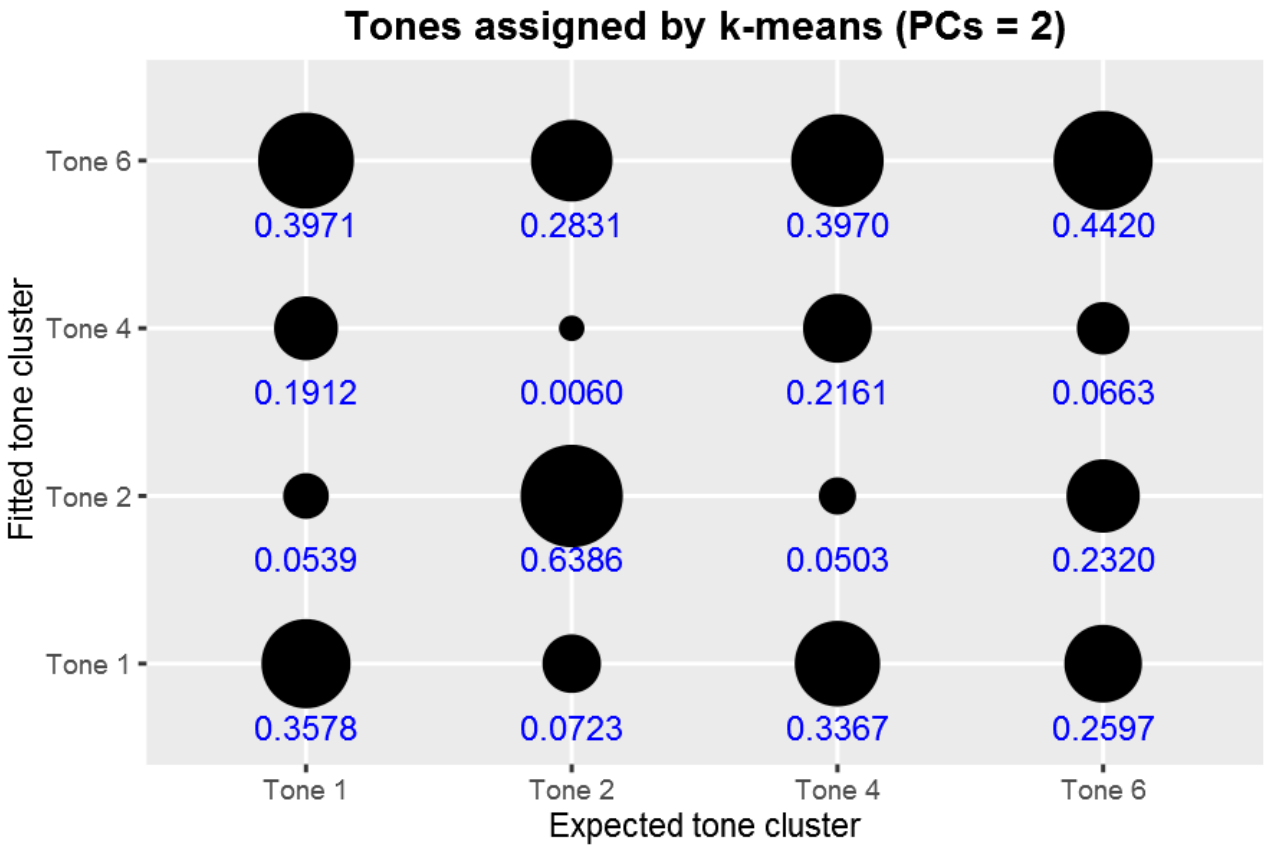
Q&A corpus

Q&A corpus (pitch only)



Speakers = KT5, MN2, SN2, SN3, SN5
Corpus = stimuli
Measures = pitch (x7)

Q&A corpus (pitch only)

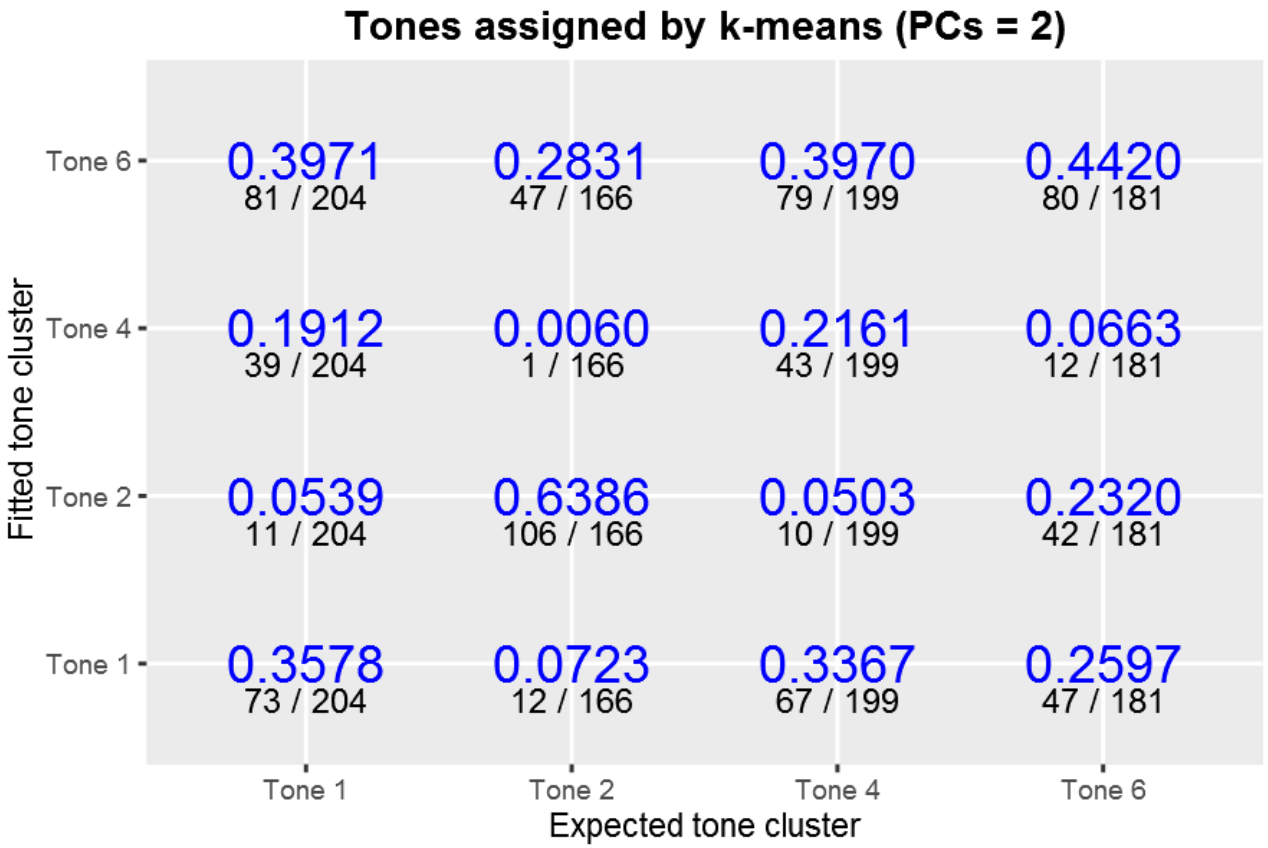


corpus = stimuli
measures = pitch
PCs = 2
***k* = 4**

	Precision	Recall	F-score
T1	0.3668	0.3579	0.3623
T2	0.6272	0.6386	0.6328
T4	0.4526	0.2161	0.2925
T6	0.2788	0.4420	0.3419
Overall	**0.4027**	**0.4027**	**0.4027**

T1 = rising
T2 = low falling
T4 = high falling
T6 = mid level

Q&A corpus (pitch only)

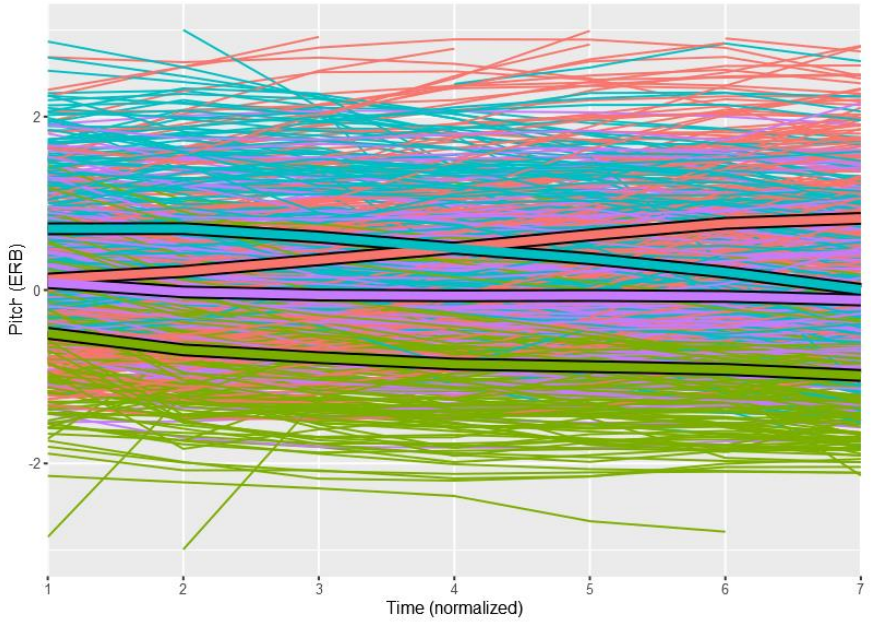


corpus = stimuli
measures = pitch
PCs = 2
***k* = 4**

	Precision	Recall	F-score
T1	0.3668	0.3579	0.3623
T2	0.6272	0.6386	0.6328
T4	0.4526	0.2161	0.2925
T6	0.2788	0.4420	0.3419
Overall	**0.4027**	**0.4027**	**0.4027**

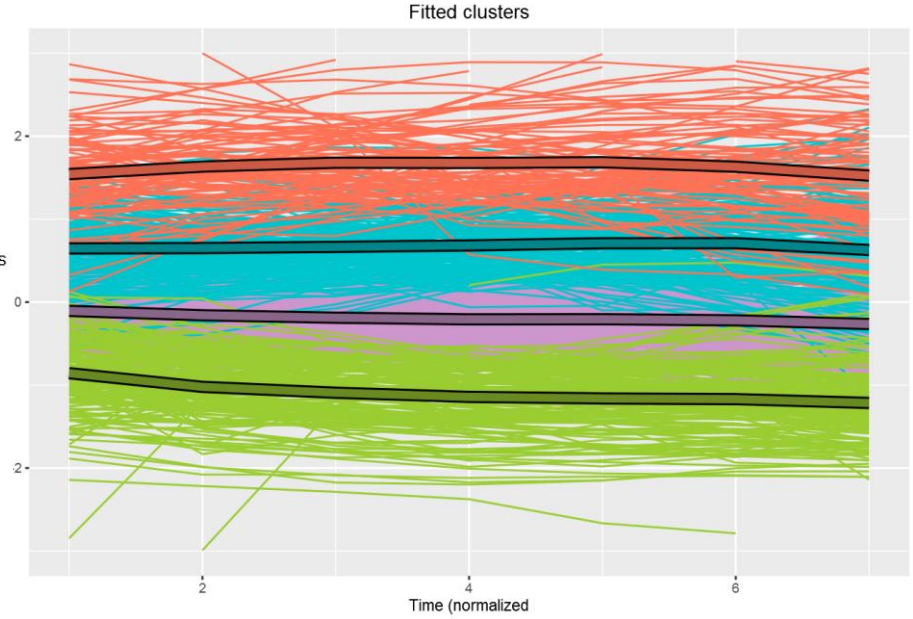
T1 = rising
T2 = low falling
T4 = high falling
T6 = mid level

Q&A corpus (pitch only)



Expected clusters

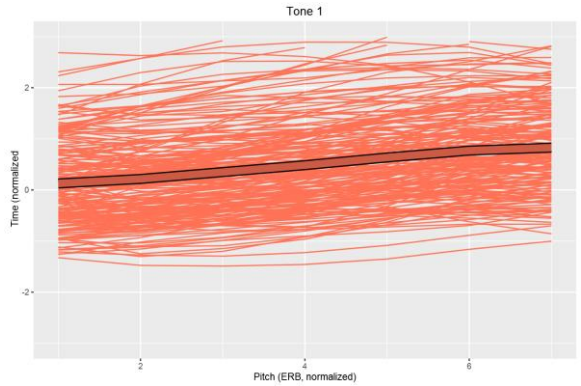
+ normalized pitch tracks

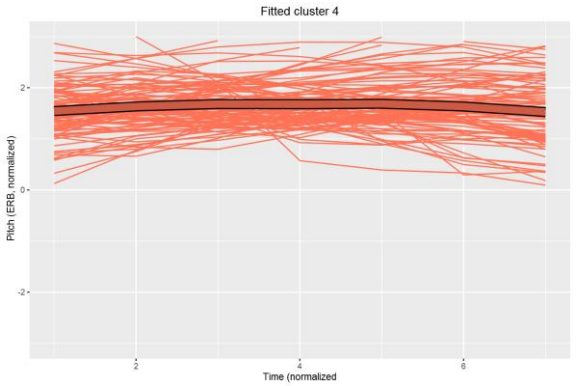


Fitted clusters

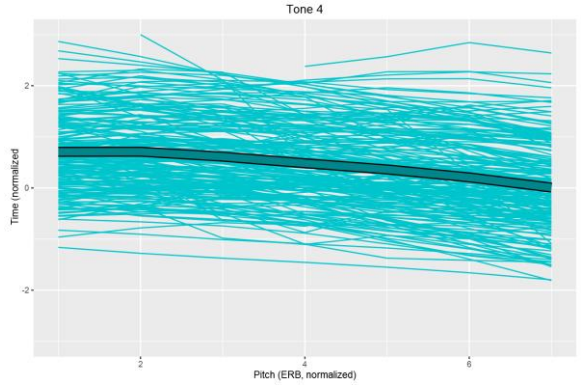
+ normalized pitch tracks

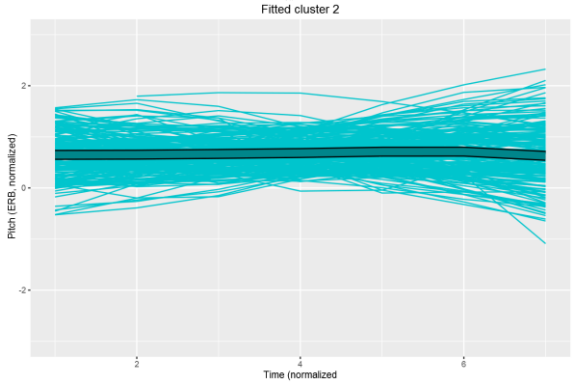
Q&A corpus (pitch only)





Tone 1





Tone 4

Expected

Fitted

Adding phonation measures

Adding phonation measures

Adding new columns to our input to PCA

	pitchstep1	pitchstep2	pitchstep3	pitchstep4	pitchstep5	pitchstep6	pitchstep7
1	1.086929934	0.450587307	0.483235657	0.655530171	1.102906335	1.44515143	1.440590791
2	0.607606651	0.338903221	0.372242407	0.402931670	0.618526143	0.91845228	1.196864524
3	0.450851789	0.126137969	0.264717697	0.450897779	0.711481611	0.95368215	1.174989172
4	-0.504643927	-0.360536037	-0.137285978	0.371874477	0.739305016	1.14582241	1.290947897
5	-0.725597388	-0.463384357	-0.286780011	0.059922230	0.671011203	1.03863363	1.104717024
6	-0.768527739	-0.949704933	-0.517437858	-0.097089133	0.526519200	0.95343230	0.978498179
7	-1.213585513	-0.572240997	-0.244463835	0.151368409	0.583430711	0.97442031	1.011408001
8	0.408315294	0.215909355	0.307380727	0.393959592	0.541063253	0.71207016	1.011795175
9	0.156641030	-0.354174286	0.075335465	0.301133094	0.503122246	0.85323952	1.041414014
10	-0.362067897	-0.135754143	0.133953775	0.380156395	0.720650687	0.94268843	1.133367927
11	0.068417189	-0.184881004	0.008392662	0.331155047	0.667533277	0.83400051	1.039284555
12	0.182635555	0.507843073	0.644175868	0.962996379	1.370706611	1.50736590	1.402066941
13	-0.128511030	0.053684686	0.030938165	0.224180272	0.548967630	0.77603362	0.989532649
14	-0.560965397	-0.252739689	-0.032882953	0.237293309	0.684606730	0.91145627	1.000373531
15	-0.388062422	-0.096523341	0.275470168	0.644832694	0.878422042	0.91470442	0.872993163
16	1.032183889	1.205515178	1.401704796	1.628655543	1.951520197	2.04231035	2.043614874
17	0.651324715	0.988155328	1.115897178	1.289787063	1.548713170	1.67327114	1.680445314
18	0.202722234	0.611044823	1.049648083	1.282540384	1.530375017	1.66802413	2.022901045
19	0.583975264	0.925244672	0.962587753	1.145543656	1.426037247	1.47538417	1.450463737
20	-1.675185257	-1.093904639	-0.645080094	-0.322081240	0.284329104	0.85973581	0.977336656
21	-1.840604960	-1.355796751	-1.061998488	-0.762403217	-0.225977443	0.31304807	0.480204763
22	-1.303384780	-1.194278944	-0.921522656	-0.463218927	0.077550615	0.43397900	0.649980725
23	0.520170522	0.736512704	0.869630906	1.184537687	1.560095472	1.60206181	1.416973154
24	-0.330559382	0.055098408	0.133606921	0.326669008	0.809495879	1.10559538	1.267330260

From this…

	PCA1	PCA2
1	-2.4169825015	-0.97576080
2	-1.5956611265	-0.76044886
3	-1.4801442446	-0.95699041
4	-0.8704152590	-1.81333082
5	-0.4375612131	-1.83175525
6	0.0510062678	-1.95933833
7	-0.1936571278	-1.95211841
8	-1.2898404271	-0.68502753
9	-0.8907753308	-1.14810871
10	-0.9961026222	-1.40490895
11	-0.9705493451	-1.16477252
12	-2.4371604947	-1.33307986
13	-0.8710089612	-1.07370114
14	-0.6867905593	-1.48647940
15	-1.1525733951	-1.21832554
16	-4.2136089546	-1.22648444
17	-3.3320837262	-1.08861056
18	-3.0901902518	-1.62464215
19	-2.9727961740	-0.93688689
20	0.7094457043	-2.33669529
21	1.7740247948	-1.98265422
22	1.1184813592	-1.86693098
23	-2.9468046171	-1.12894111
24	-1.1895248665	-1.49867652

…to this

Adding phonation measures

Adding new columns to our input to PCA

	pitchstep1	pitchstep2	pitchstep3	pitchstep4	pitchstep5	pitchstep6	pitchstep7	H1.H2step1	H1.H2step2	H1.H2step3
2	0.18596266	0.41522682	0.38338457	0.50580031	0.44801254	0.26450686	0.2873861	-0.96585201	-1.26525328	0.02025606
3	0.39895189	0.56594677	0.61996066	0.55462509	0.43433216	0.36569444	0.3980084	-1.29228627	-1.12152687	0.19173676
4	-0.45772242	-0.22421263	-0.06217098	0.18195289	0.56948480	0.60816723	0.6017988	0.15676950	-0.18678774	-0.03895307
5	-0.57919469	-0.43720186	-0.16689659	0.12982596	0.37772373	0.49919600	0.6515670	-0.21081010	-0.24577735	0.14883522
6	-0.58391206	-0.38531080	-0.04683953	0.09232287	0.23761785	0.34399454	0.6152433	-0.12252953	-0.43036685	-0.47587133
7	0.16921599	0.67586149	1.00065238	1.51154350	1.69882308	1.49125882	1.4837110	0.10292305	1.93668139	2.70972438
13	0.02132646	0.41593442	0.74544269	1.01338928	1.51720435	1.76345104	1.7860944	-0.19465930	1.47401594	2.46134708
14	0.03925246	0.28148939	0.66265285	1.18863956	1.54126293	1.62428864	1.9160579	-0.29463743	0.95725316	2.30366513
15	0.07180231	0.51193289	0.97187643	1.23510565	1.38134410	1.57216170	1.8368061	-0.47110449	0.33806611	2.17401970
16	0.56736199	1.17165703	1.63702553	2.34415923	2.77603442	3.00506271	3.7268203	0.92278768	1.90249815	3.04550415
17	0.48858191	1.03390984	1.72075884	2.37387866	2.69348045	3.19658791	3.9445269	1.43133393	1.87132554	2.92050012
18	0.89333222	1.28133587	1.66367867	2.26302048	2.64701436	2.87745786	3.4088695	0.86633829	0.42650348	1.95941795
19	0.03642204	0.51924482	0.63647145	0.85040416	1.01975773	1.16410924	1.3308682	0.34976368	-0.55991819	2.58509668
20	0.45697554	0.98201877	1.17684613	1.23274696	1.40422335	1.61013653	1.7063709	1.30322518	0.45789562	1.99350711
21	0.08807724	0.48598736	0.81360868	1.02423923	1.18840369	1.35280402	1.4473873	0.11029283	1.23642674	1.94931978
22	0.50768726	0.86148998	1.04829781	1.17778961	1.34336928	1.34997360	1.2615229	0.72706511	0.77846541	1.87279950
23	0.33951304	0.73129058	1.03485331	1.23864367	1.19783843	1.17849721	1.3999777	-0.08320312	-0.06357128	1.12302620
24	0.68647557	0.98862309	1.11174643	1.16528858	1.20161232	1.37285284	1.5631987	0.14613820	-0.60921732	0.81675691
25	0.83247816	0.87729317	0.99192525	1.18934716	1.17330811	0.94923305	0.9857927	-0.48518548	-0.31279532	2.02471107
26	0.22488095	0.51830134	0.80653262	1.04145763	1.20491448	1.28534563	1.1700059	0.80825814	1.35349650	2.95643956
27	0.22417335	0.33833369	0.51948068	0.66902130	0.83837487	0.94097766	1.0617423	1.10907065	1.10138727	2.25583995
28	0.54353927	0.48952539	0.71855368	0.97447098	1.26010771	1.35539857	1.3693148	1.15971830	1.57183018	1.82306132
29	0.22228640	0.59259991	0.62349868	0.67161585	0.69850486	0.81455215	0.8529987	1.35512726	0.61670656	0.49035401
30	0.02368514	0.14114765	0.36168467	0.47891131	0.65510506	0.72751668	0.7485090	0.42399462	0.48605759	0.97174184
31	0.14657262	0.31050121	0.58292931	0.71430805	0.46971244	0.41121705	0.6326976	0.66443765	0.37996410	0.22105908
32	0.82634557	0.99923717	1.11127470	1.22708612	1.44361338	1.65518741	1.7311371	-0.30398294	-0.81089844	1.20572455

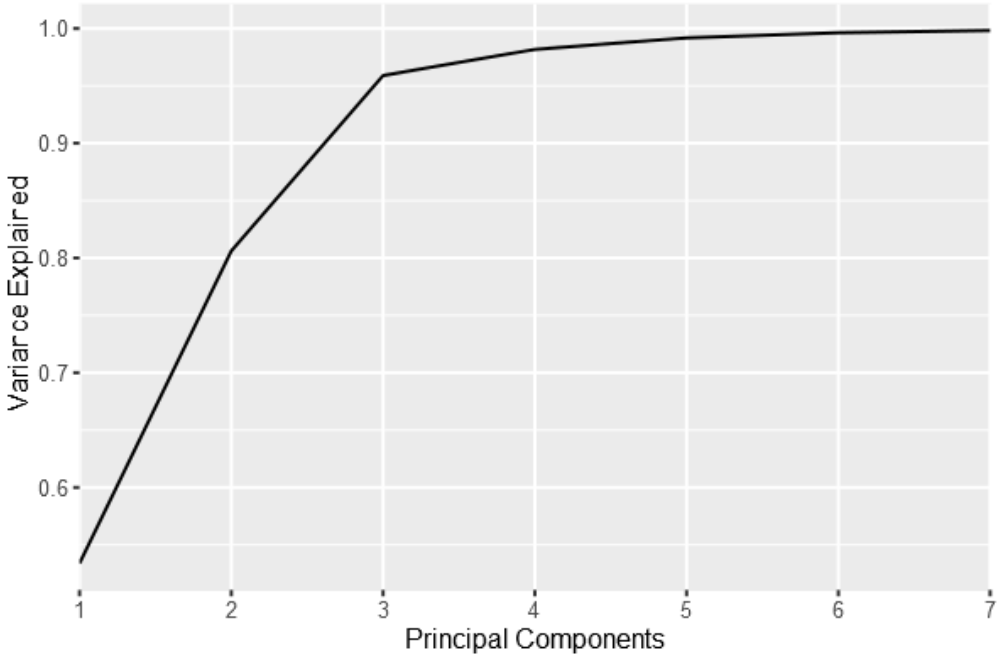
From this…

	PCA1	PCA2
1	-2.4169825015	-0.97576080
2	-1.5956611265	-0.76044886
3	-1.4801442446	-0.95699041
4	-0.8704152590	-1.81333082
5	-0.4375612131	-1.83175525
6	0.0510062678	-1.95933833
7	-0.1936571278	-1.95211841
8	-1.2898404271	-0.68502753
9	-0.8907753308	-1.14810871
10	-0.9961026222	.40490895
11	-0.9705493457	-1.16477252
12	-2.4371604947	-1.33307986
13	-0.8710089612	-1.07370114
14	-0.6867905593	-1.48647940
15	-1.1525733951	-1.21832554
16	-4.2136089546	-1.22648444
17	-3.3320837262	-1.08861056
18	-3.0901902518	-1.62464215
19	-2.9727961740	-0.93688689
20	0.7094457043	-2.33669529
21	1.7740247948	-1.98265422
22	1.1184813592	-1.86693098
23	-2.9468046171	-1.12894111
24	-1.1895248665	-1.49867652

?

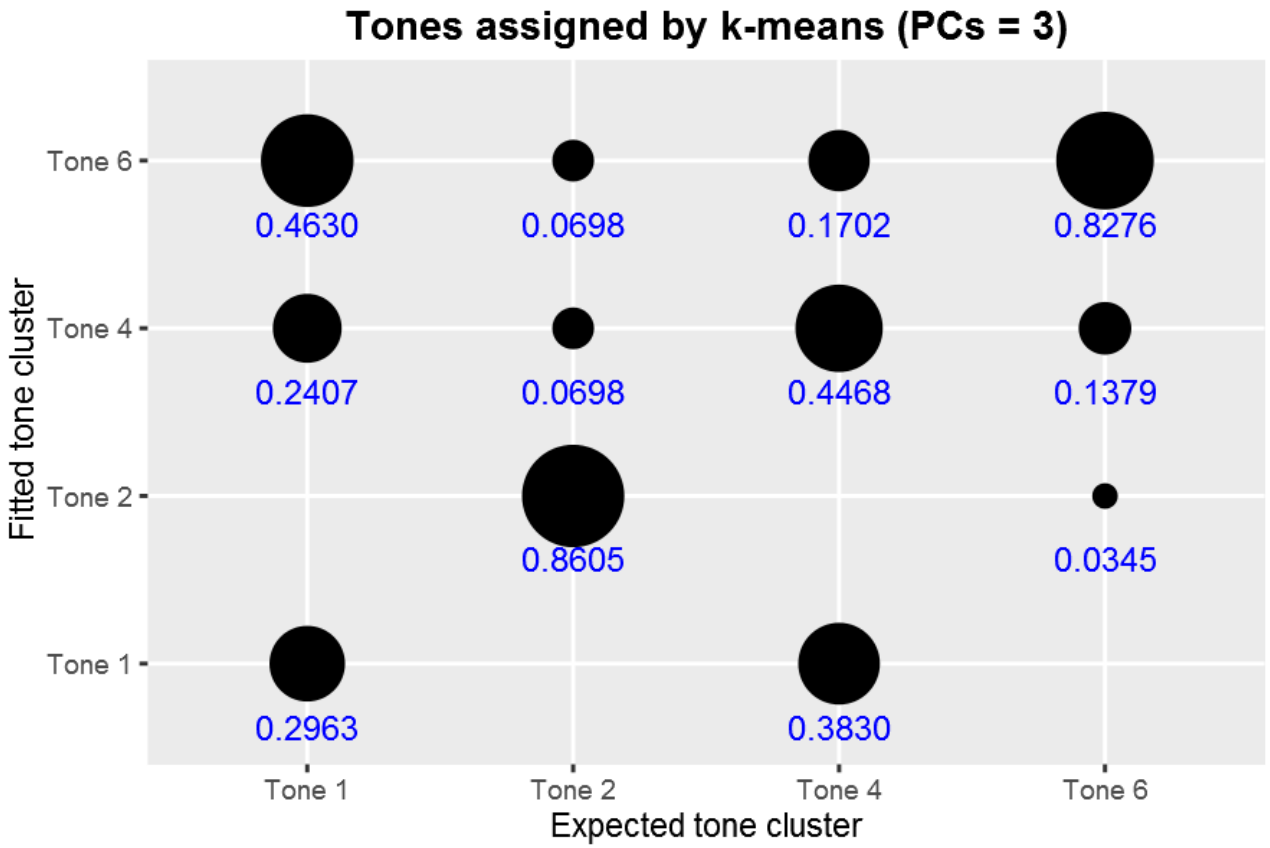
…to this

Adding phonation measures



Speaker = LP5
Corpus = wordlist
Measures = pitch (x7), spectral tilt (x3)

Wordlist corpus (pitch + phon)



corpus = wordlist
measures = pitch, spec. tilt
PCs = 3
***k* = 4**

	Precision	Recall	F-score
T1	0.4706	0.2963	0.3636
T2	0.9737	0.8605	0.9136
T4	0.5122	0.4468	0.4773
T6	0.4	0.8276	0.5393
Overall	**0.5665**	**0.5665**	**0.5665**

T1 = rising
T2 = low falling
T4 = high falling
T6 = mid level

Wordlist corpus

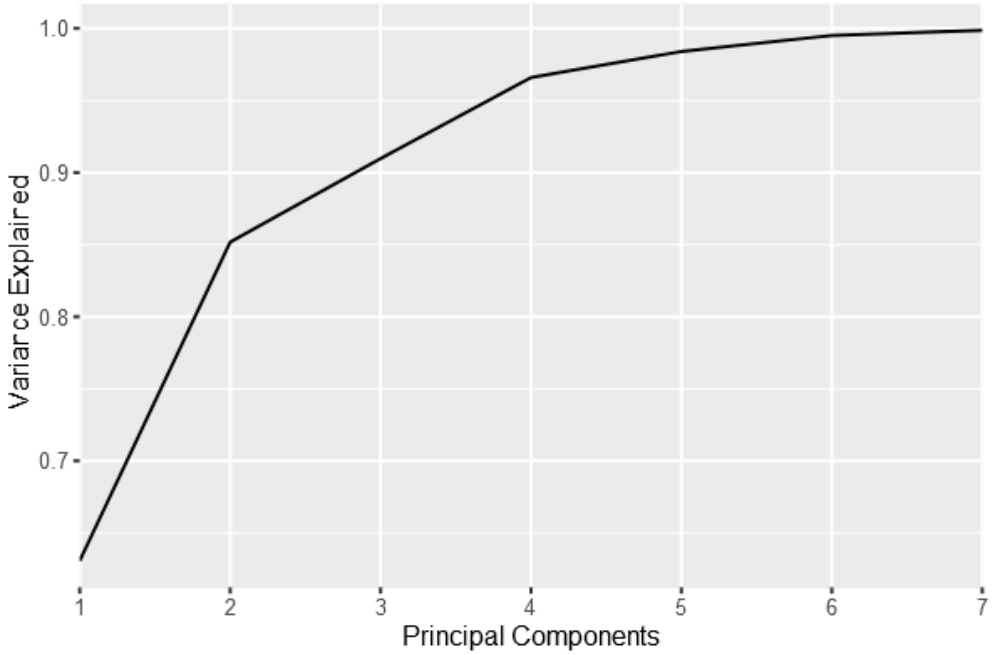
	Precision	Recall	F-score
T1	0.9778	0.8148	0.8889
T2	1.0000	**0.7907**	**0.8831**
T4	0.9302	0.8511	0.8889
T6	0.4902	0.8621	0.6250
Overall	**0.8266**	**0.8266**	**0.8266**

Pitch only

	Precision	Recall	F-score
T1	0.4706	0.2963	0.3636
T2	0.9737	**0.8605**	**0.9136**
T4	0.5122	0.4468	0.4773
T6	0.4	0.8276	0.5393
Overall	**0.5665**	**0.5665**	**0.5665**

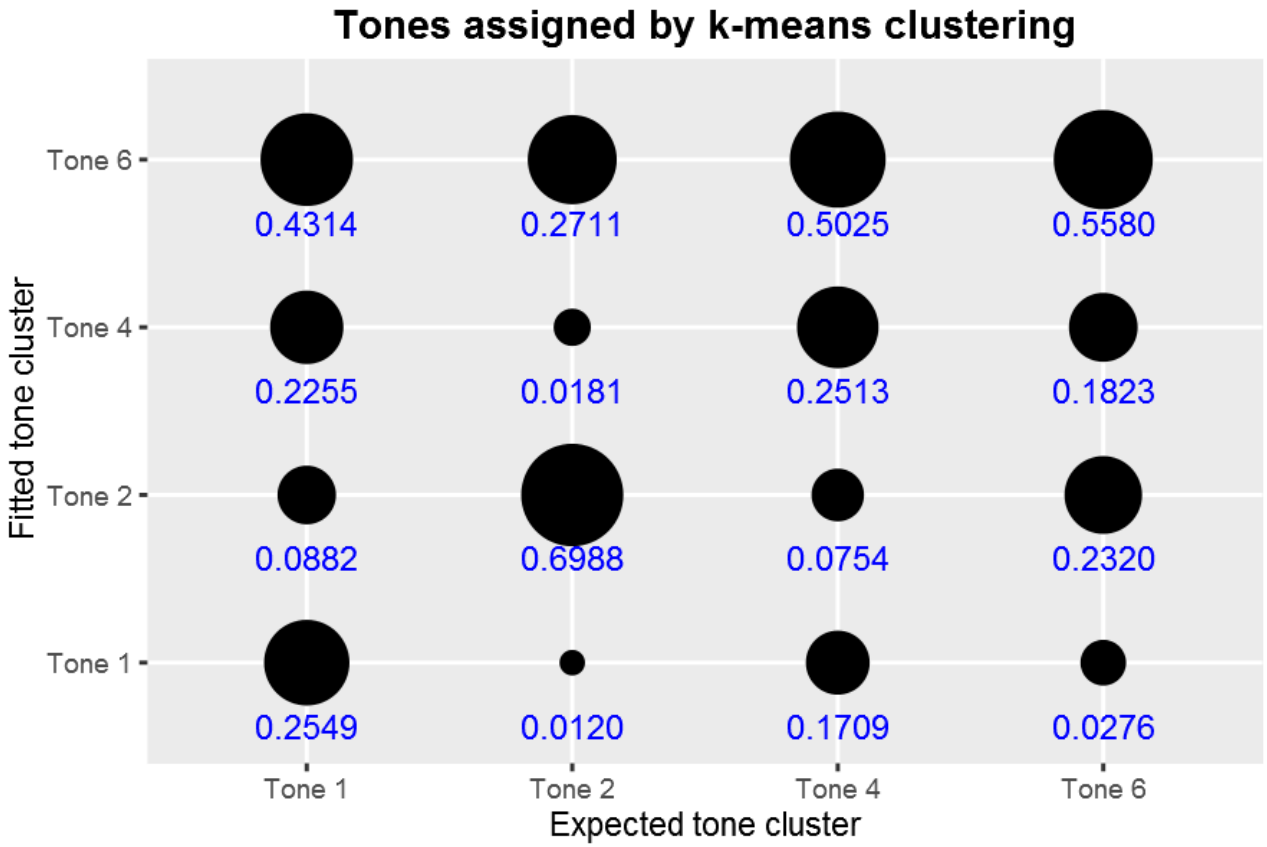
Pitch + spectral tilt

Q&A corpus (pitch + phon)



Speakers = KT5, MN2, SN2, SN3, SN5
Corpus = stimuli
Measures = pitch (x7), spectral tilt (x3)

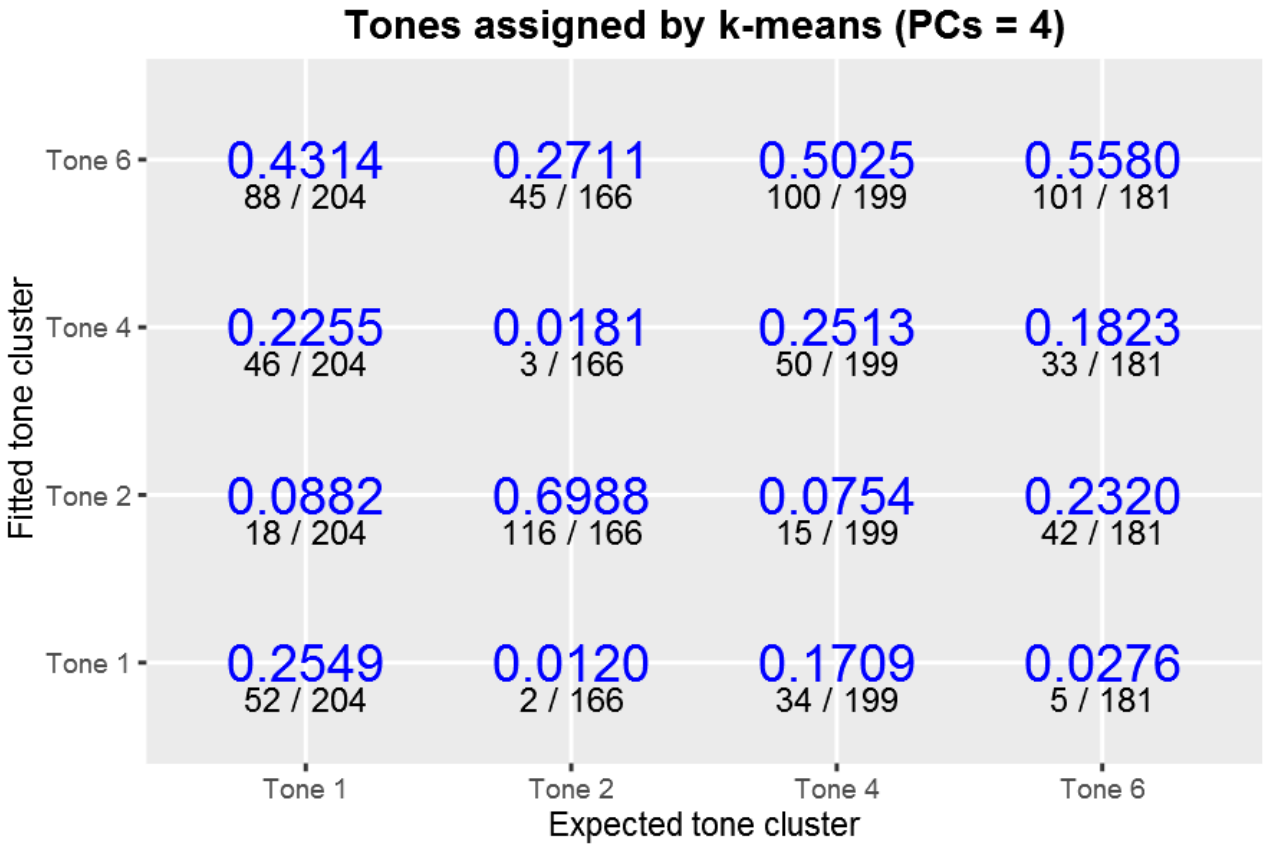
Q&A corpus (pitch + phon)



corpus = stimuli
measures = pitch, spec. tilt
PCs = 4
***k* = 4**

	Precision	Recall	F-score
T1	0.5591	0.2549	0.3502
T2	0.6073	0.6988	0.6499
T4	0.3788	0.2513	0.3021
T6	0.3024	0.5580	0.3922
Overall	**0.4253**	**0.4253**	**0.4253**

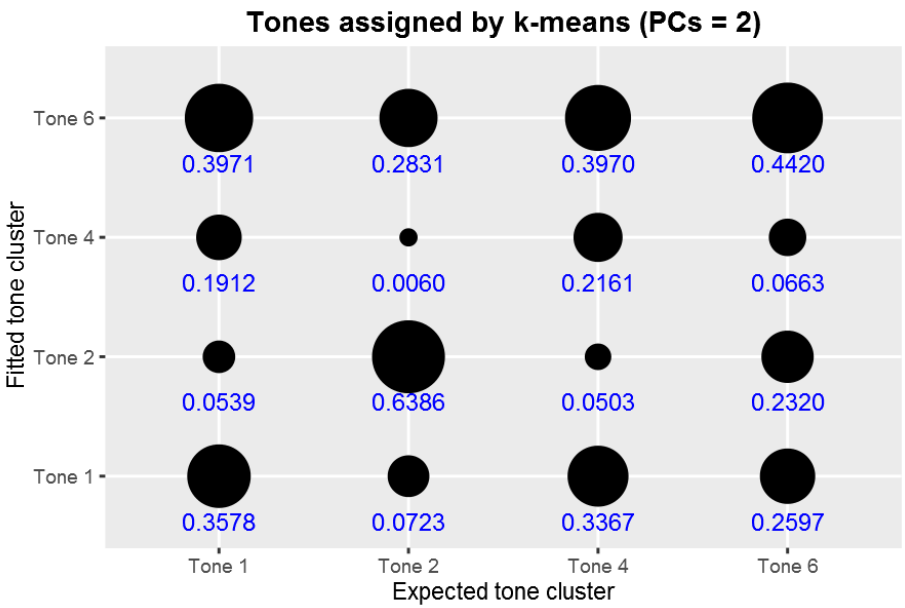
Q&A corpus (pitch + phon)



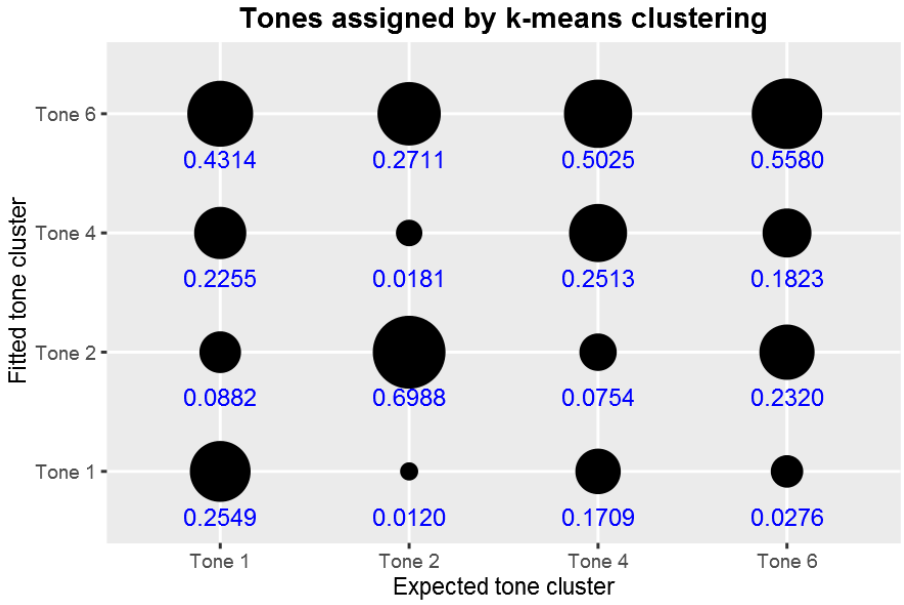
corpus = stimuli
measures = pitch, spec. tilt
PCs = 4
***k* = 4**

	Precision	Recall	F-score
T1	0.5591	0.2549	0.3502
T2	0.6073	0.6988	0.6499
T4	0.3788	0.2513	0.3021
T6	0.3024	0.5580	0.3922
Overall	**0.4253**	**0.4253**	**0.4253**

Q&A corpus



Pitch only



Pitch + spectral tilt

Conclusions

- There's a significant performance gap between citation tones and running speech
  - Excellent results on wordlist corpus (citation tones)
  - Poor results on stimuli corpus (tones in context)
- Generally verifies tonemes identified by linguist / native speakers
  - Thus has potential to tag new lexical material when access to a native speaker is limited, or even to help correct mistakes in analysis
- Other observations
  - Phonation possibly a cue in tone 2
  - Consistent with historical analysis
  - Combining pitch-only analysis and pitch+phonation analysis helped a case in which non-modal phonation acts as a phonetic cue

Future directions

- Identifying the limits of the method
  - Is the in-context data salvageable?
- Testing other clustering algorithms
- Testing performance on additional syllable structures
- Implementing as a classifier
  - Use wordlist data as a training corpus for new data
  - Determine how much data is “enough” data to converge at the same approximate centers; then use *k*-means as a classifier step for further input
  - (The question of how minimum sufficient dataset size is important in fieldwork, where time and access to native speakers can be limited resources)

Thank you!

Research supported by:

NSF grant BCS-1528386
and the Yale MacMillan Center

References

Bennett, Ryan. 2015. All measures extractor. Script for the Praat software program.

Boersma, Paul and David Weenink. 2016. Praat: doing phonetics by computer [Computer program]. Version 5.3.7.1. <http://www.praat.org>

Chao, Yuen-Ren. 1930. A system of tone-letters. *Le Maître Phonétique* 45: 24–27.

Charrad, Malika, Nadia Ghazzali, Veronique Boiteau, and Azam Niknafs. *NbClust*. Package for R <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf>

Dockum, Rikker. 2015. Tonal evidence and language classification: The place of Khamti in Southwestern Tai. Qualifying Paper, Yale University. Unpublished MS.

Dockum, Rikker. 2016. Chindwin Khamti tone: computational modeling and language documentation. Qualifying Paper, Yale University. Unpublished MS.

Halkidi, M., M. Vazirgiannis, and I. Batistakis. 2000. Quality scheme assessment in the clustering process. *Proceedings of PKDD*, Lyon, France.

Harris, Jimmy G. 1976. Notes on Khamti Shan. *Tai Linguistics in Honor of Fang-Kuei Li*. 113–141.

Hartigan, J. A. and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society,* Series C 28 (1): 100–108.

Inglis, Douglas. 2014. This here thing: Specifying Morphemes an3, nai1, and mai2 in Tai Khamti Reference-point Constructions. PhD Dissertation. University of Alberta.

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Wiley-Blackwell.

Jolliffe, I. T. 1986. Principal Component Analysis. Springer-Verlag. <doi:10.1007/b98835>

Ketchen, David J. and Christopher L. Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* 17(6): 441-458.

Kirk, P. L., Ladefoged, J., & Ladefoged, P. 1993. Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In A. Mattina & T. Montler (Eds.), *American Indian linguistics and ethnography in honor of Laurence C. Thompson*. Missoula, MT: University of Montana Press.

Krzanowski, W. J. and Y. T. Lai. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44(1): 23-34.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the World, Nineteenth edition.* Dallas, Texas: SIL International. <http://www.ethnologue.com.>

Morey, Stephen. 2005. Tonal Change in the Tai Languages of Northeast India. *Linguistics of the Tibeto-Burman Area* 28.2: 139–202.

Needham, J. F. 1894. Outline Grammar of the Tai (Khamti) Language.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Robinson, William. 1849. Notes on the Languages Spoken by the Various Tribes Inhabiting the Valley of Assam and Its Mountain Confines [Section: The Khamti]. *Journal of the Royal Asiatic Society of Bengal 18.1*: 183–237, 310–349.

Shosted, Ryan, Marissa Barlaz, and Di Wu. 2015. Modeling Iu Mien tone with eigenpitch representations. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0728. <http://www.icphs2015.info/pdfs/Papers/ICPHS1041.pdf>

SIL International. 2002. Mainlaind Southeast comparative wordlist.

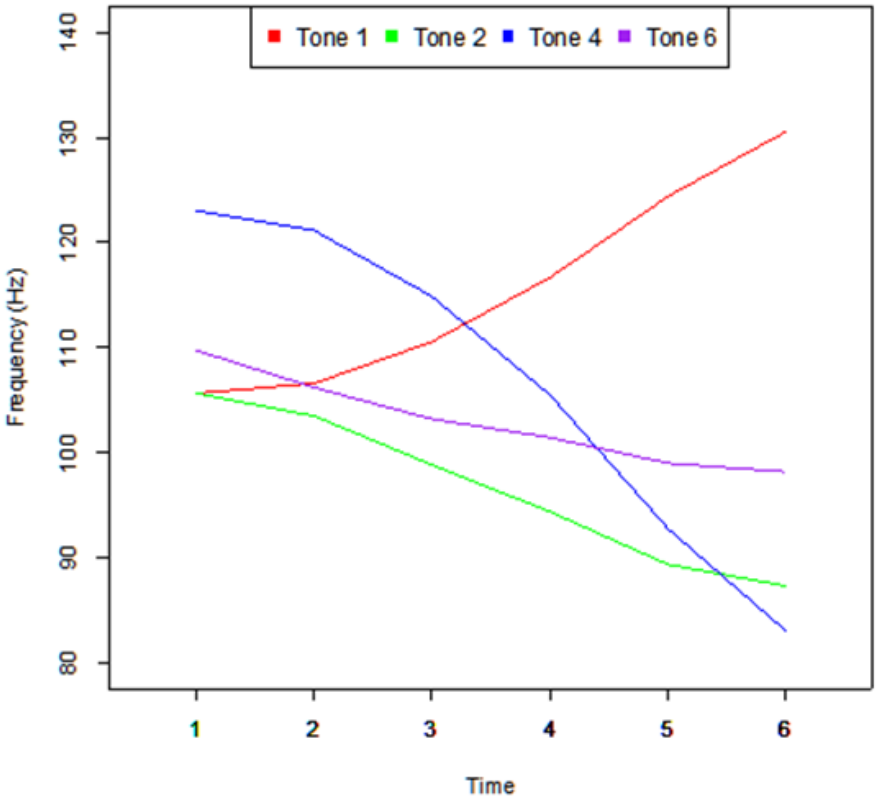
Scott, A. J. and M. J. Symons. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* 27(2):387-397.

Stanford, James N. 2008. A sociotonetic analysis of Sui dialect contact. *Language Variation and Change* 20(3):409-50.

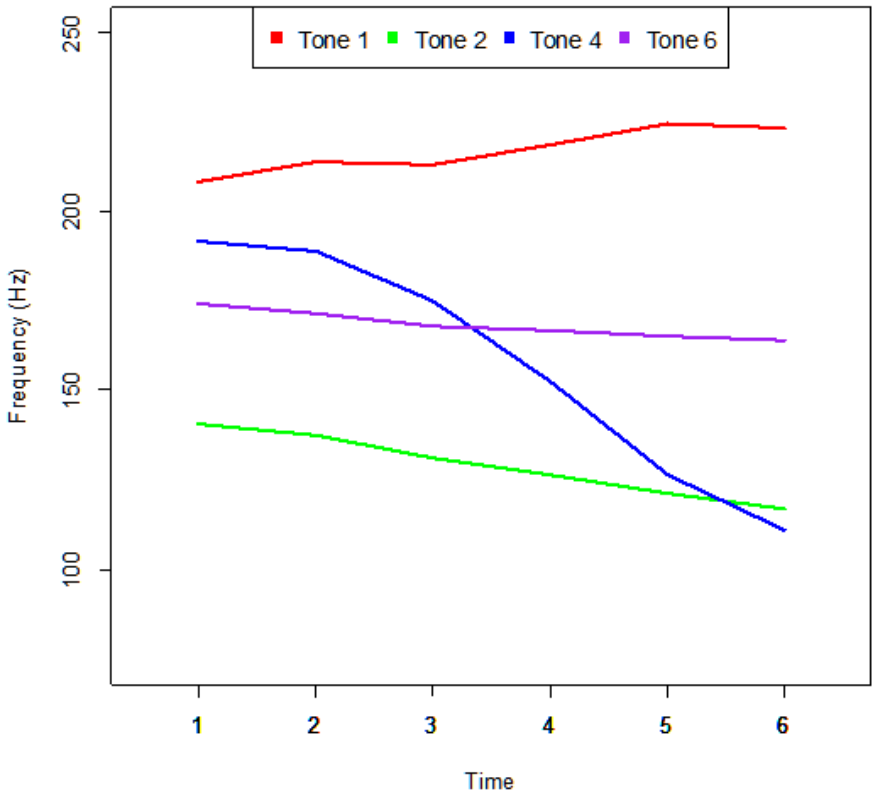
Bonus slides

Synchronic variation

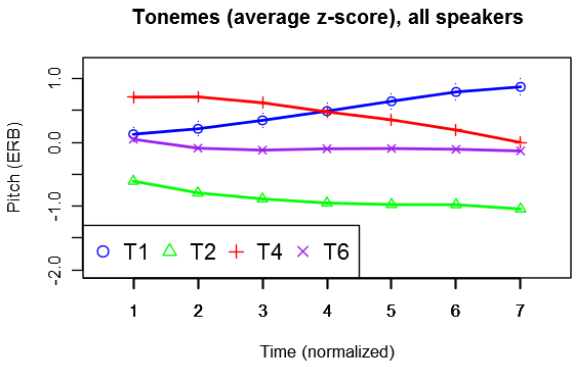
SAM, 38M

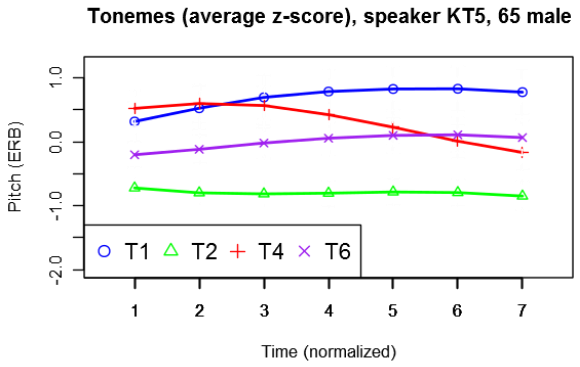


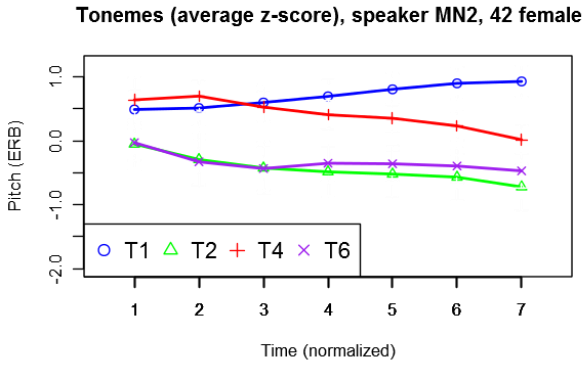
LSAT, 75M

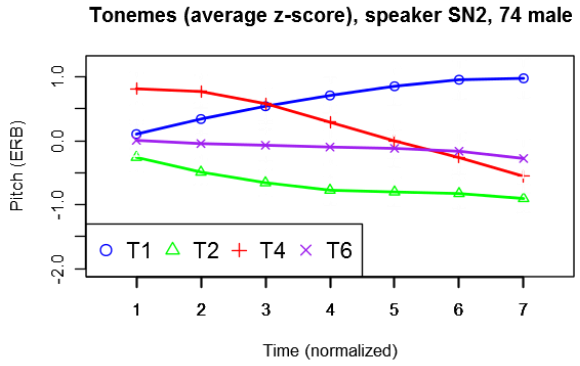


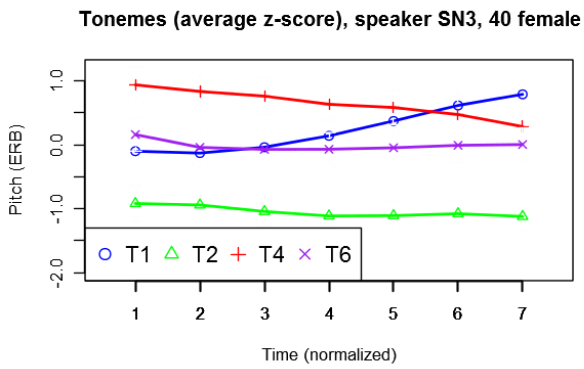
Additional speaker variation

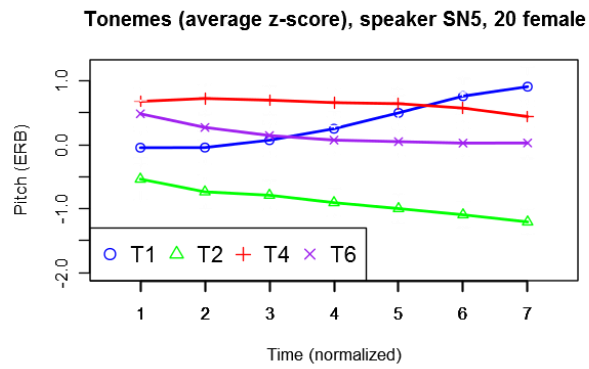










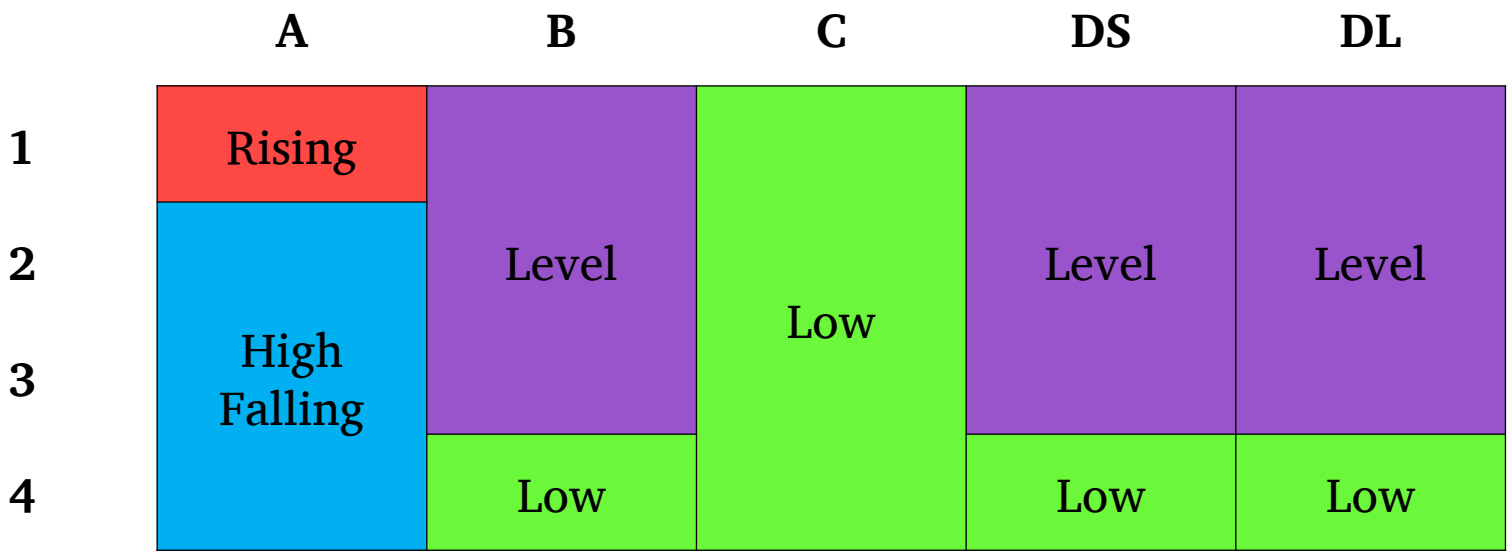


Background

Proto-Tai consonants

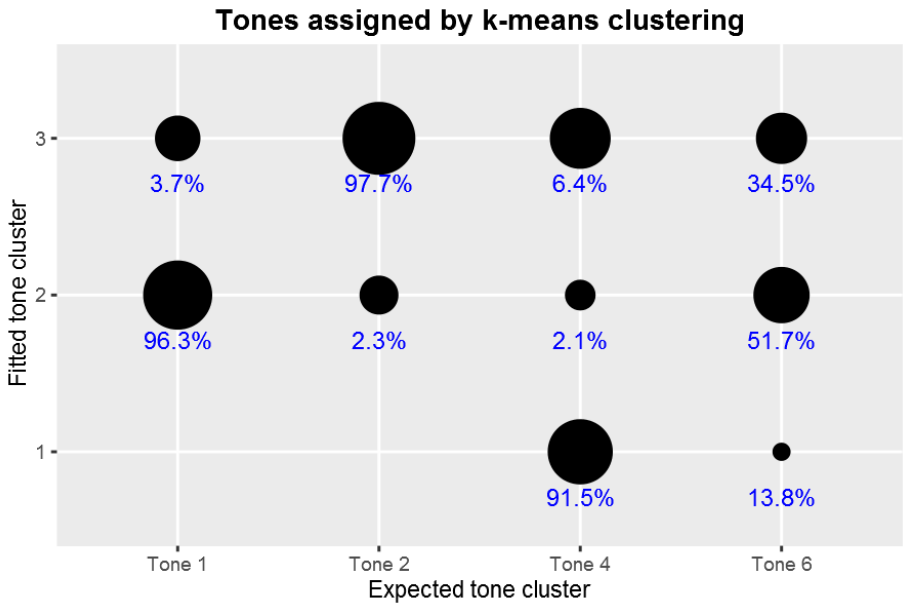
	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

Background

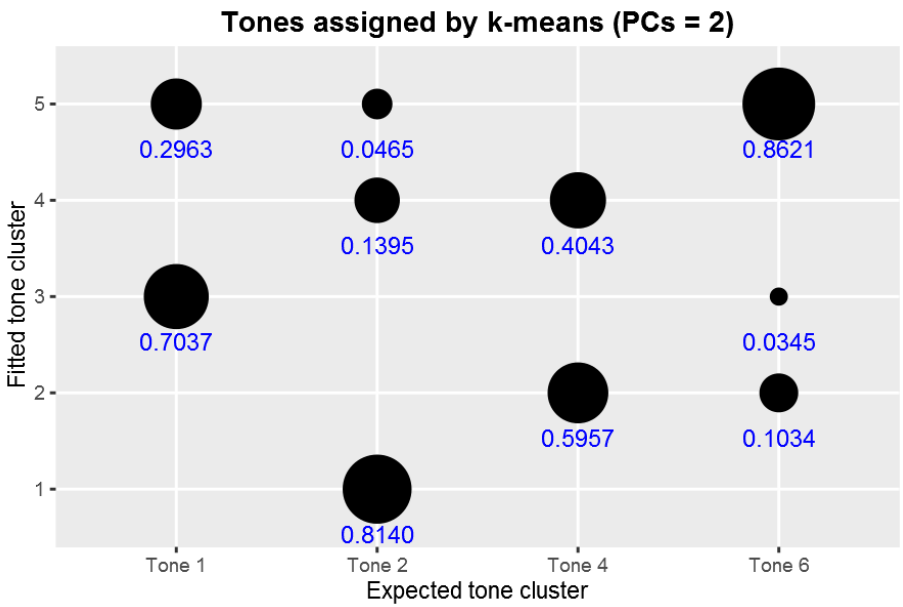


Chindwin Khamti tone splits

Testing other *k*-values



k = 3



k = 5