Adjusting to the New Normal(ization): Adapting Atlas of North American English Benchmarks to Lobanov-Normalized Data

Aaron J. Dinkin
San Diego State University

Adjusting to the New Normal(ization): Adapting Atlas of North American English Benchmarks to Lobanov-Normalized Data

Abstract

The Atlas of North American English (Labov et al. 2006) defines criteria for participation in certain dialect features in terms of formant benchmarks; e.g., a speaker is considered to have a raised TRAP vowel if their mean normalized F1 of TRAP is less than 700 Hz. Other researchers often compare their own findings to these benchmarks; but the majority of recent research in North American sociophonetics uses the Lobanov (1971) method to normalize formant measurements, producing values that are formally incomparable with Atlas benchmarks. This paper proposes a method of transforming benchmarks into Lobanov-comparable values. Benchmarks are expressed as z-scores relative to the entire Atlas corpus of normalized formant measurements, whose mean F1 is 650.7 Hz (s.d. 150.0 Hz) and mean F2 is 1595.5 Hz (s.d. 435.2 Hz). Thus, for example, a benchmark of 700 Hz in F1 is converted to 0.329 in Lobanov terms. This method is evaluated by comparing the effectiveness of the Lobanov-transformed benchmarks at distinguishing dialect regions to that of the original Atlas benchmarks. Fourteen such benchmarks are evaluated against three isogloss parameters; in 76% of cases, the Lobanov-transformed benchmarks are at least as effective as the original Atlas benchmarks at characterizing the Atlas' dialect regions. Therefore, this transformation can be recommended for researchers who want to compare Lobanov-normalized data to Atlas benchmarks.

Adjusting to the New Normal(ization): Adapting *Atlas of North American English* Benchmarks to Lobanov-Normalized Data

Aaron J. Dinkin

1 Introduction

The *Atlas of North American English* (Labov et al. 2006) is the landmark modern study of the dialect geography of the United States and Canada. In the years since its publication, sociophonetic research into North American vowel variation and change has been conducted in the context of the dialect regions and sound changes it documented. For some dialect features, *ANAE* defines criteria for participation in terms of specific normalized mean vowel formant measurements. For example, the Inland North dialect region is defined by the presence of the Northern Cities Shift, involving (among other features) the raising of the TRAP vowel and the fronting of the LOT vowel; a speaker is identified as participating in the fronting of LOT if their mean normalized F2 of LOT is greater than 1450 Hz, and the raising of TRAP if its mean F1 is less than 700 Hz (Labov et al. 2006:192–6).

Research following *ANAE* has frequently adopted the same formant cutoffs, in order to ensure comparability with *ANAE*'s regions and definitions. For instance, Roeder (2009) uses the 700-Hz cutoff for the raising of TRAP in order to diagnose her speakers in Michigan as participants in the Northern Cities Shift; and Brumbaugh and Koops (2017) use several of *ANAE*'s formant benchmarks in order to situate the speakers in their Albuquerque data set in relation to other, better-studied dialect regions. Although, as Stanley (2020) notes, the values of these benchmarks are somewhat arbitrary, they have emerged as *de facto* standards of comparison.

Changing methodological trends, however, call into question the applicability of *ANAE*'s formant benchmarks as a comparison point for future research. In order for formant measurements to be meaningfully compared between speakers, their values must be normalized. However, formant benchmarks chosen on the basis of data normalized by one method may not be meaningful when applied to data that is normalized differently. The goal of this paper is to propose a method for *ANAE* benchmarks to be applied to data normalized using the Lobanov (1971) method, which is currently the most common normalization method in North American sociophonetics but differs from the normalization presupposed by *ANAE* benchmarks.

2 Log-mean and Lobanov Normalization

In *ANAE*, formant measurements are normalized by a calculation based on Nearey (1978)'s log-mean normalization method. This calculation cancels out acoustic differences between speakers by multiplying each speaker's measured formant values by a speaker-specific uniform scaling factor applied to both F1 and F2. The scaling factor is chosen such that, after normalization, the geometric mean of a speaker's F1 and F2 measurements is a standardized value: approximately 989 Hz, the geometric mean of the unnormalized F1 and F2 measurements of the first 345 speakers in *ANAE*'s corpus (Labov et al. 2006:39–40).

normalization methods used, 2015–21	NWAV abstracts	*Am. Sp.*	*LVC*
Lobanov	36	14	15
log-mean (Nearey or *ANAE*)	7	4	5
Watt & Fabricius (2002)	3	1	6
Bark	2	4	3
other	2	0	1
normalization method not specified	19	1	0

Table 1. Normalization methods used by recent papers in selected sociolinguistics venues.

Although some new dialectological research in the years following the publication of *ANAE* has used *ANAE*'s normalization methodology, the Lobanov (1971) normalization method appears

to be much more widely used in recent work on North American English sociophonetics. The Lobanov method involves measuring F1 and F2 each in terms of *z*-score: for each formant, the arithmetic mean is set equal to zero, and formant values are measured in units of that speaker's overall standard deviation for that formant. Table 1 shows a survey of three major venues for sociolinguistic research—*American Speech*, *Language Variation and Change*, and the abstracts of the annual NWAV conferences—demonstrating that the Lobanov method is used in nearly two thirds of papers between 2015 and 2021 that mention using a specific formant normalization method.

Unlike *ANAE*'s method, the Lobanov method normalizes F1 and F2 independently of each other, each formant according to its own mean and standard deviation. Therefore, unlike *ANAE* normalization, the ratio of F1 to F2 is not preserved. Figure 1 illustrates outlines of the vowel spaces of two *ANAE* speakers[1], demonstrating this phenomenon: on the left, Gordon H. from Hammond, Indiana; and on the right, Ernest P. from Las Vegas. The top panels display formants normalized according to the *ANAE* procedure; they show that Ernest's F1 range is much "shorter" in the height dimension than Gordon's is, meaning that Ernest has a higher ratio of F2 range to F1 range. In the Lobanov-normalized outlines, shown in the bottom panels, the difference between the two speakers in relative height of the vowel space is eliminated.

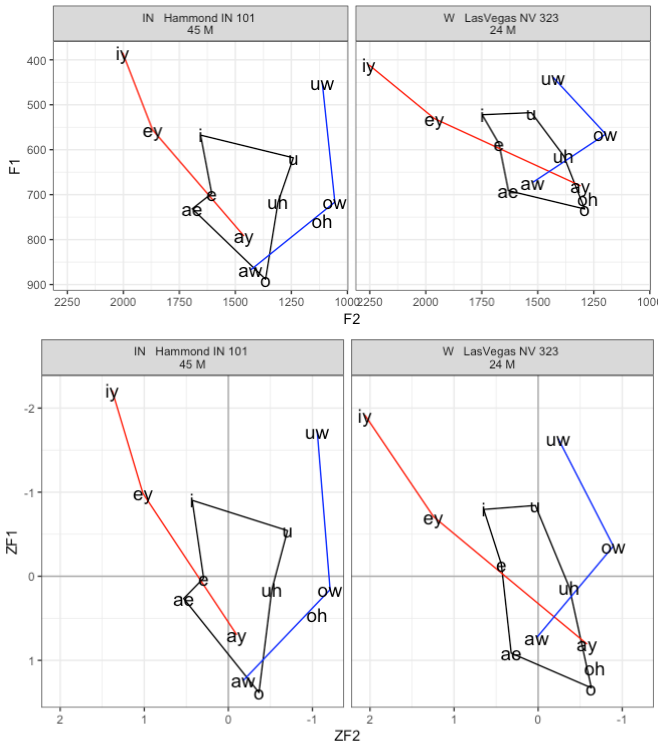

Figure 1: Vowel space outlines for two *ANAE* speakers: Gordon H. from Hammond, Indiana (left) and Ernest P. from Las Vegas (right). Top plots: *ANAE* normalization; bottom plots: Lobanov.

Barreda (2021, see also Barreda and Nearey 2017) contends that Lobanov normalization is not appropriate for sociophonetic research, inasmuch as it erases quantitative differences between speakers that may represent perceptible differences in vowel quality. Rankinen and de Jong (2021) make a similar point, comparing log-mean and Lobanov normalizations of the same data set and finding that Lobanov normalization erases potentially socially meaningful variation. However, the comparison in Figure 1 indicates that Lobanov normalization may be more effective than log-mean normalization for comparing speakers' participation in regional chain shifts such as the Northern

[1] Throughout this paper, Fruehwald (2010)'s packaging of the *ANAE* data is used as the source of *ANAE* formant measurements, in both original log-mean normalization and Lobanov normalization. Fruehwald (2010) also created the R script that produces vowel plot outlines such as those shown in Figure 1.

Cities Shift.

As noted above, one of *ANAE*'s criteria for participation in the Northern Cities Shift is the raising of mean TRAP to F1 < 700 Hz. In the ANAE normalization shown on the top of Figure 1, Gordon's mean TRAP is greater than 700 Hz, and Ernest's is less. The 700-Hz cutoff would therefore suggest that Ernest exhibits the NCS feature of TRAP-raising to a greater degree than Gordon does. But a glance at Figure 1 demonstrates that this is not the case: although Gordon's TRAP does not clear 700 Hz, it is substantially higher than low vowels such as LOT, close to the middle of his F1 range; and Ernest's TRAP is near the low edge of his vowel space. Under Lobanov normalization, Gordon's TRAP indeed appears much higher than Ernest's. Thus, the 700-Hz cutoff for NCS TRAP-raising, defined in terms of *ANAE*'s log-mean normalization, misclassifies Ernest and Gordon. So it appears that, in at least some cases, Lobanov normalization can be more accurate for classifying participation in regional dialect patterns than log-mean normalization.

The benchmarks defined in *ANAE* cannot be applied directly to Lobanov-normalized formants, since *ANAE* benchmarks are denominated in hertz and Lobanov normalization measures formants in units of *z*-score. However, the FAVE suite (Rosenfelder et al. 2014), a software package that automatically extracts formant measurements from audio recordings, translates Lobanov-normalized formants back into values denominated in hertz. It does this by setting a standardized value for each formant's mean and standard deviation and applying those to all speakers: mean F1 is set to 650 Hz, with a standard deviation of 150 Hz, and mean F2 is set to 1700 Hz, with a standard deviation of 420 Hz (for convenience, these values are repeated on the right side of Table 2 below). Because of this, some researchers (e.g., Gordon and Strelluf 2017) have applied *ANAE*'s formant benchmarks to the resulting values. However, it is not at all clear that this use of *ANAE* benchmarks is appropriate; the two methods of normalization do not necessarily produce comparable results, and there is no guarantee that, for instance, an F2 value like "1450 Hz" represents a similar degree of fronting in *ANAE* normalization as it does in FAVE-rescaled Lobanov normalization.

Since *ANAE* defined widely-used benchmarks for North American English, but the Lobanov method is the most prevalent normalization technique in North American sociophonetics, it is desirable to be able to apply the former to the latter. The purpose of this paper, therefore, is to convert the *ANAE* benchmarks to values that can be applied to Lobanov-normalized data in a principled and justified way, and to test the effectiveness of these translated benchmarks for diagnosing participation in the dialect features that the original benchmarks were chosen to diagnose.

3 Translating *ANAE* Benchmarks to Lobanov Normalization

Since the benchmarks are defined based on the regional features present in the normalized measurements of the *ANAE* speakers, and Lobanov normalization is based on *z*-scores, we will translate the benchmarks from *ANAE*-normalized units to Lobanov units using *z*-scores calculated on the basis of the entire *ANAE* corpus. Brumbaugh and Koops (2017) used a similar method to calculate Lobanov-normalized benchmarks based on means and standard deviations of their own New Mexico data set, but the formants of a data set specifically from one region may have different distributional properties than those of North America at large. Since the benchmarks were originally defined in the context of the *ANAE* data set, it makes the most sense to define the Lobanov-translated versions of the benchmarks in terms of the *ANAE* data as well.

	ANAE corpus	FAVE output
F1 mean	650.7 Hz	650 Hz
F1 s.d.	150.0 Hz	150 Hz
F2 mean	1595.5 Hz	1700 Hz
F2 s.d.	435.2 Hz	420 Hz

Table 2. Formant means and standard deviations of the log-mean normalized 132,051-token *ANAE* data set, and values used by FAVE to scale Lobanov-normalized outputs.

Fruehwald (2010)'s compilation of the *ANAE* corpus contains F1 and F2 measurements of 132,051 vowel tokens, representing the vowel spaces of 435 speakers. The mean *ANAE*-normalized F1 of these 132,051 tokens is 650.7 Hz, with a standard deviation of 150.0 Hz; the mean F2 is

1595.5 Hz, with a standard deviation of 435.2 Hz. For convenient reference, these values are shown on the left side of Table 2. We will apply these values to translate formant benchmarks into *z*-scores that can be applied to Lobanov-normalized data. For example, the benchmark for the Northern Cities Shift's raising of TRAP is F1 < 700 Hz. This differs by 49.3 Hz from the overall mean F1 of the entire corpus; 49.3 Hz is 32.9% of the corpus-wide standard deviation of F1; therefore the Lobanov-normalized benchmark for TRAP raising is F1 < 0.329. Similarly, the benchmark for NCS LOT-fronting is F2 > 1450 Hz; and that translates into a Lobanov-normalized benchmark of F2 > –0.334, based on the corpus's mean value and standard deviation of F2.

These Lobanov-normalized benchmarks are *not* simply equivalent restatings of the original *ANAE* benchmarks. It is not the case that a speaker who has TRAP F1 less than 700 Hz in *ANAE*-normalized measurements will therefore also have TRAP F1 less than 0.329 in Lobanov-normalized measurements; the two normalization methods may disagree on which speakers satisfy any given benchmark. What this conversion indicates, however, is that, for example, a value of 0.329 in Lobanov-normalized F1 represents roughly the same degree of vowel height, *compared to the overall range of vowel measurements in the corpus*, as a value of 700 Hz in *ANAE*-normalized F1.

The *ANAE* corpus's mean and standard deviation of normalized F1 are virtually the same as the values used by the FAVE suite to rescale Lobanov-normalized *z*-scores back into hertz, as shown in Table 2. This indicates that FAVE's rescaling produces a range of F1 values generally comparable to those produced by *ANAE* normalization. The F2 values, however, are very different: FAVE sets mean F2 at 1700 Hz for every speaker, while the mean of *ANAE*'s normalized F2 measurements is 1595.5 Hz. This means that, while *ANAE* benchmarks for F1 values arguably *can* be applied more or less at face value to the normalized output of FAVE, F2 benchmarks definitely *cannot* be: FAVE's reported normalized F2 values will average about 105 Hz larger than vowels of the same frontness under *ANAE* normalization. Thus, for example, when Gordon and Strelluf (2017) report that two of their thirteen FAVE-measured speakers satisfy the LOT F2 > 1450 Hz benchmark, they are actually using a much laxer standard than *ANAE*'s 1450-Hz criterion implies. To achieve the same degree of fronting represented by 1450 Hz in *ANAE* normalization, FAVE-rescaled Lobanov-normalized measurements would have to exceed 1560 Hz.

We will test the efficacy of this system for converting *ANAE* benchmarks into Lobanov-normalized benchmarks by applying it to *ANAE* data. Many benchmark values are used by *ANAE* to define dialect regions. Therefore, we can test the converted benchmarks by applying them to the dialect regions they are supposed to define and evaluating whether they characterize the regions at least as well as the original benchmarks do. This will be the focus of the next section of the paper.

4 Testing the Lobanov-transformed Benchmarks

ANAE defines three parameters that may be used to evaluate the quality of an isogloss (Labov et al. 2006:42–43). The **homogeneity** of a region with respect to a dialect feature is the percentage of points within the region that exhibit the feature; the **consistency** of a region is the percentage of points exhibiting the dialect feature that are located within the region; and the **leakage** of a region is the percentage of points *outside* the region that exhibit the feature. A well-drawn isogloss for a given feature will ideally have high homogeneity, high consistency, and low leakage, though of course in real-world cases of linguistic variation it is rare for all three parameters to have near-optimal values. As Stanley (2020) notes, many of the *ANAE* benchmark values were apparently chosen with the goal of optimizing isogloss homogeneity and consistency, rather than for theoretically-motivated reasons; therefore, it seems appropriate to use these isogloss parameters to evaluate the method defined above for converting the benchmarks to Lobanov values.

The benchmarks we will examine are those used in *ANAE* to draw isoglosses that define named dialect regions. Fruehwald (2010)'s packaging of the *ANAE* data codes each speaker as belonging to one of 22 dialect groups[2], and we will for the most part use these groups as the basis of our calculations of homogeneity, consistency, and leakage. Each speaker is coded as belonging to exactly one of these 22 regions, even when regions overlap or are subsets of each other; the Inland North is part of the North, but speakers from, for example, Chicago are coded only as belonging to

[2] Actually 23, but one of them is used only once and appears to be a typo.

the Inland North. When examining a benchmark that is used to define a region that contains subregions, such as the North, its subregions (such as the Inland North) will be included as part of the larger region for calculations of homogeneity, consistency, and leakage.

For example, *ANAE* defines a "southeastern super-region" characterized in part by fronting of the GOAT vowel to F2 > 1200 Hz. This "super-region" roughly includes the South, the Midland, the Mid-Atlantic, and Florida. The isogloss for the southeastern super-region drawn on *ANAE*'s Map 11.11 does not exactly correspond with the union of those regions: for example, a handful of transitional cities such as Washington, D.C. and Corpus Christi, Texas are within the isogloss but not counted in any of those regions. However, for the sake of simplicity—i.e., to avoid having to scrutinize the map in detail to determine exactly which communities are inside the isogloss—for the purpose of testing the GOAT F2 benchmark we will consider the southeastern super-region to consist of the set of communities coded as South, Inland South, Texas South, Midland, Mid-Atlantic, Florida, or Charleston.

By that definition, the homogeneity of the southeastern super-region with respect to the F2(GOAT) > 1200 Hz benchmark is 91.6%; its consistency is 58.7%; and its leakage is 39.8%. If we translate the value of 1200 Hz into Lobanov normalization in the manner described above, the benchmark becomes F2(GOAT) > –0.909. A total of 260 speakers in the corpus satisfy the Lobanov-transformed benchmark, of whom 155 are within the southeastern super-region. From this, we can calculate the homogeneity, consistency, and leakage of the super-region with respect to the Lobanov-transformed benchmark: respectively, 93.4%, 59.8%, and 39.0%. Each of these quantities is slightly better than the corresponding value for the original 1200-Hz benchmark. Therefore, even though the super-region was defined by a benchmark stated in terms of log-mean normalized formant measurements, the Lobanov-normalized benchmark actually does a slightly better job of characterizing the same region. These quantities are summarized in Table 3. Thus we can confirm that researchers using Lobanov-normalized data wishing to compare their data to *ANAE*'s standards for GOAT fronting in the southeastern super-region may use the F2 benchmark of –0.909.

The same benchmark is used in the opposite direction to define a different isogloss: one of the criteria *ANAE* uses to define the Northern dialect region is that F2 of GOAT is generally less than 1200 Hz (see *ANAE* map 11.8). Table 4 shows that the consistency and leakage of the broader North region are very slightly better under the Lobanov-transformed benchmark for GOAT backness than under the original *ANAE* benchmark, and its homogeneity is unchanged.

Southeastern super-region (cf. *ANAE* map 11.11)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		166	269			
ANAE benchmark	F2(GOAT) > 1200 Hz	152	107	**.916**	**.587**	**.398**
Lobanov-transformed	F2(GOAT) > –0.909	155	105	**.934**	**.596**	**.390**

Table 3. Comparison of benchmarks for GOAT F2 in the "southeastern super-region". Regions included: South, Inland South, Texas South, Midland, Mid-Atlantic, Florida, Charleston.

North (cf. *ANAE* map 11.8)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		117	318			
ANAE benchmark	F2(GOAT) < 1200 Hz	98	78	**.838**	**.557**	**.245**
Lobanov-transformed	F2(GOAT) < –0.909	98	77	**.838**	**.560**	**.242**

Table 4. Comparison of benchmarks for GOAT F2 in the North. Regions included: North, Inland North, Western New England, Providence.

We can apply this same approach to test Lobanov-transformed versions of other formant benchmarks used in *ANAE*. The Inland North is defined in part by the benchmarks F1(TRAP) < 700

and F2(LOT) > 1450 (see *ANAE* maps 14.4, 14.5); the TRAP isogloss also includes the St. Louis Corridor. Tables 5 and 6 demonstrate that the homogeneity, consistency, and leakage of the Inland North are all improved by the Lobanov-transformed versions of these benchmarks.

Inland North (cf. *ANAE* map 14.4)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		71	364			
ANAE benchmark	F1(TRAP) < 700 Hz	53	54	**.746**	**.495**	**.148**
Lobanov-transformed	F1(TRAP) < 0.329	60	37	**.845**	**.619**	**.102**

Table 5. Comparison of benchmarks for TRAP F1 in the Inland North (including St. Louis corridor).

Inland North (cf. *ANAE* map 14.5)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		62	373			
ANAE benchmark	F2(LOT) > 1450 Hz	48	38	**.774**	**.558**	**.102**
Lobanov-transformed	F2(LOT) > –0.334	50	37	**.806**	**.575**	**.099**

Table 6. Comparison of benchmarks for LOT F2 in the Inland North.

Eastern Corridor (cf. *ANAE* map 9.2)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		24	411			
ANAE benchmark	F1(THOUGHT) < 700 Hz	19	37	**.792**	**.339**	**.090**
Lobanov-transformed	F1(THOUGHT) < 0.329	19	30	**.792**	**.388**	**.073**

Table 7. Comparison of benchmarks for THOUGHT F1 in the "Eastern corridor". Regions included: New York City, Mid-Atlantic, Providence.

Canada (cf. *ANAE* map 15.4)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		24	411			
ANAE benchmark	F1(DRESS) > 660 Hz	23	203	**.958**	**.102**	**.494**
Lobanov-transformed	F1(DRESS) > 0.062	23	182	**.958**	**.112**	**.443**

Table 8. Comparison of benchmarks for DRESS F1 in Canada.

ANAE's Map. 9.2 uses a benchmark of 700 Hz (Labov et al. 2006:59) to define the raising of THOUGHT in a region roughly encompassing the Mid-Atlantic region, New York City, and Providence, plus a handful of Western New England data points in Connecticut; for the sake of simplicity we ignore the Western New England points so that the region can be expressed as a union of named regions. Table 7 shows that Lobanov transformation leaves the region's homogeneity unchanged with respect to THOUGHT-raising and slightly improves its consistency and leakage.

Map 15.4 in *ANAE* defines the Canadian dialect region in terms of the Canadian Shift, featuring lowered DRESS (F1 > 660 Hz), backed TRAP (F2 < 1825 Hz), and backed LOT (F2 < 1275). None of these features is unique to Canada, as discussed in depth in Becker (2019), and therefore all three

have very low consistency scores if measured as simply Canadian features. However, transforming these benchmarks to Lobanov values improves or maintains Canada's homogeneity and consistency for all three of them, and improves leakage for two of the three, as shown in Tables 8–10.

Canada (cf. *ANAE* map 15.4)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		24	411			
ANAE benchmark	F2(TRAP) < 1825 Hz	24	169	**1.00**	**.124**	**.411**
Lobanov-transformed	F2(TRAP) < 0.527	24	145	**1.00**	**.142**	**.353**

Table 9. Comparison of benchmarks for TRAP F2 in Canada.

Canada (cf. *ANAE* map 15.4)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		24	411			
ANAE benchmark	F2(LOT) < 1275 Hz	21	88	**.875**	**.193**	**.214**
Lobanov-transformed	F2(LOT) < –0.736	22	90	**.917**	**.196**	**.219**

Table 10. Comparison of benchmarks for LOT F2 in Canada.

Inland Canada (cf. *ANAE* map 15.7)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		9	23			
ANAE benchmark	F2(FACE) > 2200 Hz	7	11	**.778**	**.389**	**.478**
Lobanov-transformed	F2(FACE) > 1.39	7	9	**.778**	**.438**	**.391**

Table 11. Comparison of benchmarks for FACE F1 in Inland Canada vs. the rest of Canada. Cities included: Edmonton, Calgary, Saskatoon, Regina, Winnipeg, Thunder Bay.

Inland Canada (cf. *ANAE* map 15.7)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		9	23			
ANAE benchmark	F2(MOUTH) < 1550 Hz	6	3	**.667**	**.667**	**.130**
Lobanov-transformed	F2(MOUTH) < –0.105	5	6	**.556**	**.455**	**.261**

Table 12. Comparison of benchmarks for MOUTH F2 in Inland Canada vs. the rest of Canada. Cities included: Edmonton, Calgary, Saskatoon, Regina, Winnipeg, Thunder Bay.

ANAE also defines an "Inland Canada" subregion with conservatively back GOAT and MOUTH and front FACE. Since these isoglosses are "internal to Canada, with the intent of differentiating one Canadian region from another" (Labov et al. 2006:223), we evaluate the consistency and leakage of Inland Canada only in comparison to the rest of Canada. For the purposes of testing these benchmarks, Inland Canada will be taken to consist of all data points in Alberta, Saskatchewan, and Manitoba, plus Thunder Bay, Ontario: the narrowest definition of the region on *ANAE*'s Map 15.7.[3] FACE shows improved consistency and leakage with the Lobanov-transformed benchmarks, and

[3] If we use the separate isoglosses for each vowel on *ANAE*'s Map 15.7, we get generally the same results.

unchanged homogeneity, as shown in Table 11; MOUTH does worse on all three parameters (Table 12); and GOAT has improved homogeneity but worse consistency and leakage (Table 13).

Inland Canada (cf. *ANAE* map 15.7)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		9	23			
ANAE benchmark	F2(GOAT) < 1100 Hz	7	5	**.778**	**.583**	**.217**
Lobanov-transformed	F2(GOAT) < –1.14	8	8	**.889**	**.500**	**.348**

Table 13. Comparison of benchmarks for GOAT F2 in Inland Canada vs. the rest of Canada. Cities included: Edmonton, Calgary, Saskatoon, Regina, Winnipeg, Thunder Bay.

The Atlantic Provinces constitute a top-level dialect region separate from the rest of Canada in *ANAE*, with an isogloss based on the fronting of START to F2 > 1450 Hz, shown on *ANAE*'s Map 15.6. Labov et al. (2006:221)'s discussion of this feature suggests that the isogloss is drawn only with respect to the rest of Canada, not based on comparing the Atlantic Provinces to the full *ANAE* data set[4]; and so when we evaluate the isogloss parameters for this benchmark and its Lobanov-transformed version, we ignore the American data points, as we did for Inland Canada above. Table 14 shows that the Atlantic Provinces have better homogeneity under the original log-mean benchmark, but better consistency and leakage under the Lobanov-transformed benchmark.

Atlantic Provinces (cf. *ANAE* map 15.6)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		8	24			
ANAE benchmark	F2(START) > 1450 Hz	6	3	**.750**	**.667**	**.125**
Lobanov-transformed	F2(START) > –0.334	5	1	**.625**	**.833**	**.042**

Table 14. Comparison of benchmarks for START F2 in Atlantic Provinces vs. the rest of Canada.

North Central (cf. *ANAE* map 11.13)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		9	424			
ANAE benchmark	F2(GOOSE) < 1700 Hz	6	89	**0.667**	**.063**	**.210**
Lobanov-transformed	F2(GOOSE) < 0.240	6	85	**0.667**	**.066**	**.200**

Table 15. Comparison of benchmarks for GOOSE F2 after coronals in the North Central. Cities included: Brockway, Lemmon, Minot, Bismarck, Fargo, Bemidji, Chisholm, Superior, Marquette.

Map 11.13 of *ANAE* identifies a "North Central" region characterized by "limited fronting" of GOOSE. The communities included in this isogloss are Brockway, Montana; Lemmon, South Dakota; Minot, Bismarck, and Fargo, North Dakota; Bemidji and Chisholm, Minnesota; Superior, Wisconsin; and Marquette, Michigan. Labov et al. (2006:141–142) offer two benchmarks for GOOSE backness in this region: F2 is less than 1700 Hz after coronal consonants, and less than 1300 Hz after non-coronals. Tables 15 and 16 show that the coronal benchmark is very slightly improved by the Lobanov transformation, while the non-coronal benchmark produces better results in the log-mean normalization. Not all speakers in the *ANAE* data have calculable GOOSE means in both

[4] If we do compare the Atlantic Provinces to the entire remainder of the *ANAE* data set, all three parameters are better with the original log-mean normalized benchmark than with the Lobanov-transformed benchmark.

environments, so the total number of speakers adds up to less than 435 in these tables.

North Central (cf. *ANAE* map 11.13)		*n* inside region	*n* outside region	homogeneity	consistency	leakage
total speakers		9	407			
ANAE benchmark	F2(GOOSE) < 1300 Hz	9	138	**1.00**	**.061**	**.339**
Lobanov-transformed	F2(GOOSE) < –0.679	8	145	**0.889**	**.052**	**.356**

Table 16. Comparison of benchmarks for GOOSE F2 after non-coronals in the North Central. Cities included: Brockway, Lemmon, Minot, Bismarck, Fargo, Bemidji, Chisholm, Superior, Marquette.

5 Discussion and Conclusion

We have examined 14 benchmarks used in *ANAE* as part of the definition of named dialect regions. Although the dialect regions were drawn with an eye to maximizing isogloss homogeneity and consistency with respect to benchmarks defined in terms of log-mean normalized formants, transforming the benchmarks into values compatible with Lobanov normalization actually improves the isogloss parameters in the majority of cases; these results are summarized in Table 17. The few cases where the isogloss parameters are worse under Lobanov normalization are either worse by a very small margin, or in regions defined by a very small number of data points, which are likely to be the least reliable anyway. This suggests that, on the whole, defining benchmarks in terms of Lobanov-normalized formant values does at least as good a job of characterizing dialect regions as the log-mean-normalized benchmarks according to which the isoglosses were defined.

	better with Lobanov	same results	better with original
homogeneity	5	6	3
consistency	11	0	3
leakage	10	0	4

Table 17: Of the 14 isoglosses evaluated, did Lobanov-transformed benchmarks or the original *ANAE* benchmarks produce better isogloss parameters?

Apart from very small dialect regions, the benchmark that shows the most change under Lobanov transformation is TRAP-raising, for which homogeneity and consistency both improve by more than 10%. This suggests that *ANAE* contains quite a few speakers like Ernest P. (shown on the right side of Figure 1 above), with TRAP unraised but in a relatively "short" vowel space, so that 700 log-mean–normalized hertz is the F1 value of low rather than mid vowels. Since Lobanov-normalized formants measure a vowel's relative height or backness within the range of vowel space a speaker uses, Lobanov normalization is more effective at diagnosing whether a single phoneme such as TRAP is low, mid, or high within the overall structure of a speaker's vowel space. It may be the case that the reason the Northern Cities Shift shows greater improvements under Lobanov normalization than other dialect features do is that features like the Canadian Shift and the fronting of back vowels often involve multiple vowel phonemes changing in the same direction, thus changing the overall shape of the vowel space rather than just the position of one phoneme in it.

In any event, the goal of this paper is not to argue, contra Barreda and Nearey (2017) and Rankinen and de Jong (2021), that Lobanov normalization is superior to log-mean normalization for evaluating regional dialect features. The point is merely, given that many researchers do use Lobanov normalization, and many researchers use *ANAE* formant benchmarks, to provide a well-motivated and standardizable way to do both at the same time. Some researchers have overlooked the fact that Lobanov-normalized data is not necessarily comparable to *ANAE* benchmarks, and judged speakers' participation in vowel shifts by inappropriate standards. Others have recognized that fact and been forced to resort to elaborate additional computations, not necessarily replicable by other researchers, in order to compare their data to the benchmarks in a meaningful way. Clearly, as long as *ANAE* benchmarks continue to be used, it is desirable to have a standardized methodology

available for different researchers to adapt the benchmarks in the same way.

I have shown that converting benchmarks from *ANAE*'s hertz values to *z*-scores relative to the entire *ANAE* formant data set does, overall, about as good a job at distinguishing dialect regions as the original benchmarks do, if not better; and therefore I encourage researchers wishing to use such benchmarks for evaluating Lobanov-normalized data to convert them in this way. The relevant means and standard deviations for carrying out this calculation are shown in Table 2 above.

Researchers working with FAVE-extract output, in which Lobanov-normalized formants are rescaled to hertz, may use the *ANAE* benchmarks at face value for F1. However, they are strongly encouraged *not* to do so for F2, but rather to recalculate the benchmarks in terms of *z*-scores and FAVE's rescaling, using the FAVE conversion values on the right side of Table 2. For instance, an *ANAE* benchmark of 1200 Hz transforms to a Lobanov benchmark of –0.909 *z*-score units; FAVE sets mean F2 at 1700 Hz with standard deviation of 420 Hz, so a *z*-score benchmark of –0.909 is represented in FAVE output as 1318 Hz.

Since *ANAE* set the baseline for our current understanding of the dialectology of North America, it is valuable to be able to use the standards it set in order to contextualize and evaluate new data. Therefore, as methodological trends in the field change, it's necessary to find responsible ways of expressing *ANAE*'s standards in such a way that they can be meaningfully applied in newer research.

References

Barreda, Santiago. 2021. Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change* 33.1:27–53.

Barreda, Santiago and Terrance Nearey. 2017. A regression approach to vowel normalization for missing and unbalanced data. *Journal of the Acoustical Society of America* 144:500–520.

Becker, Kara (ed.). 2019. *The Low Back Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and Short Front Vowel Shifts across North America.* Publication of the American Dialect Society 104.

Brumbaugh, Susan and Christian Koops. 2017. Vowel variation in Albuquerque, New Mexico. In *Speech in the Western States, Volume 2: The Mountain West*, ed. V. Fridland, A. B. Wassink, T. Kendall, and B. E. Evans, Publication of the American Dialect Society 102:31–57.

Fruehwald, Josef. 2010. Compilation, coding, cleaning, and plotting of the ANAE data. Ms. and R workspace, University of Pennsylvania.

Gordon, Matthew J. and Christopher Strelluf. 2017. Working the early shift: Older Inland Northern speech and the beginnings of the Northern Cities Shift. *Journal of Linguistic Geography* 4:31–46.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English*. Berlin: Mouton/de Gruyter.

Lobanov, Boris M. 1971. Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49:606–608.

Nearey, Terrance Michael. 1978. *Phonetic Feature Systems for Vowels*. Bloomington, Ind.: Indiana University Linguistics Club.

Rankinen, Wil and Kenneth de Jong. 2021. The entanglement of dialectal variation and speaker normalization. *Language and Speech* 64:181–202.

Roeder, Rebecca. 2009. The effects of phonetic environment on English /æ/ among speakers of Mexican heritage in Michigan. In *Toronto Working Papers in Linguistics 31*, ed. R. Compton and M. Irimia.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite.

Stanley, Joseph A. 2020. Vowel Dynamics of the Elsewhere Shift: A Sociophonetic Analysis of English in Cowlitz County, Washington. Doctoral dissertation, University of Georgia.

Watt, Dominic and Anne Fabricius. 2002. Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 ~ F2 plane. In *Leeds Working Papers in Linguistics and Phonetics 9,* 159–73.

Department of Linguistics and Asian/Middle Eastern Languages
San Diego State University
5500 Campanile Drive
San Diego CA 92182-7727
adinkin@sdsu.edu