COMPARING VOWEL FORMANT NORMALIZATION METHODS

Nicholas Flynn & Paul Foulkes

Department of Language and Linguistics, University of York, UK
`nejf100@york.ac.uk; paul.foulkes@york.ac.uk`

ABSTRACT

Results from a large-scale comparative study of vowel formant normalization methods are presented. Effectiveness of methods was evaluated by their ability to improve the equalization and alignment of speaker vowel spaces over raw Hertz measurements. Vowel-intrinsic methods performed poorly, while vowel-extrinsic, formant-intrinsic methods performed the best overall.

Keywords: vowel formants, normalization, sociophonetics

1. INTRODUCTION

The process of normalizing vowel formant data to permit accurate cross-speaker comparisons is an issue that has grown in importance in recent years. The viability of normalizing has been opened to a greater number of researchers, thanks to the online normalization tool NORM [15]. The sheer number of available algorithms indicates a lack of consensus about how best to normalize formants.

Normalization methods have traditionally been categorized according to whether they are vowel intrinsic or extrinsic, formant intrinsic/extrinsic, speaker intrinsic/extrinsic, or a combination [18]. Classification of methods included in the study is included alongside the results in Tables 1 and 2.

The main goals of normalization are [15, 18]:
- to minimize inter-speaker variation due to physiological or anatomical differences;
- to preserve inter-speaker variation due to dialect or social differences, or sound change;
- to maintain phonemic differences;
- to model the cognitive processes that allow listeners to understand different speakers.

Of course, it is unlikely that any method can perfectly fulfill all these criteria [1, 15].

Several comparative studies have assessed the effectiveness of methods [1, 4, 5, 6, 9]. However, the range of methods tested and the nature of the data used varies considerably between studies. The present study offers a more complete and up-to-date picture. It presents results using a large dataset collected under non-laboratory conditions and includes a larger number of methods.

2. METHOD

2.1. Data

The data came from 20 speakers, 5 young (18-22) and 5 older (40-50) speakers of each sex from Nottingham, UK. F_1 and F_2 measurements were extracted from monophthongal word list items. Mono tracks sampled at 22,050Hz were used, and formant measurements were taken in Praat v5.0.42 using a script with manual correction. A minimum of three adjacent glottal pulses plotted by Praat's inbuilt formant tracker were averaged for each token. 3605 tokens were measured (mean 180 per speaker).

2.2. Details of the normalization methods

Twenty methods were compared, six vowel-intrinsic and fourteen vowel-extrinsic.

(1) and (2) show simple base 10 (henceforth *Log*) and natural (*Ln*) logarithmic transformations. (3), (4) and (5) show vowel-intrinsic methods from the literature using Mels [13], ERBs [8] and Barks [16].

(1) $$F_i^N = \log_{10}(F_i)$$

(2) $$F_i^N = \ln(F_i)$$

(3) $$F_i^N = 1127\ln\left(1 + \frac{F_i}{700}\right)$$

(4) $$F_i^N = 21.4\ln(0.00437F_i + 1)$$

(5) $$F_i^N = 26.81\left(\frac{F_i}{1960 + F_i}\right) - 0.53$$

Bladon [3] adapts the Bark scale method to normalize females' data relative to males', using (5) for males and (6) for females.

(6) $$F_i^N = 26.81\left(\frac{F_i}{1960 + F_i}\right) - 1.53$$

Bark-diff [14] using (7) is a further adaptation of Bark scale normalization. A slight modification was made in that $B_3 - B_1$ was substituted in place of $B_1 - B_0$ (B_i represents F_i in Barks) [15].

(7)　　$F_i^N = B_3 - B_i$

Nordström [12] scales females' formant values relative to males' using an estimation of the difference in vocal tract length. (8) is used for males and (9) for females. μ_{F_3} is defined as the mean F_3 across all vowel tokens where $F_1 > 600$Hz.

(8)　　$F_i^N = F_i$

(9)　　$F_i^N = \left(\dfrac{\mu_{F_3}^{male}}{\mu_{F_3}^{female}} \right) F_i$,

　　　and $\left(\dfrac{\mu_{F_3}^{male}}{\mu_{F_3}^{female}} \right) = 0.876$ for this dataset.

LCE [10] scales formant values as a proportion of a speaker's maximum for that formant (10). In effect, vowel spaces are aligned by anchoring them at the maximum values for individual formants.

(10)　　$F_i^N = \dfrac{F_i}{F_i^{max}}$

Gerstman [7] builds on *LCE* by aligning vowel spaces at both endpoints of their formant frequency range (11). Values are scaled so that the extremities are 0 and 999 rather than 0 and 1.

(11)　　$F_i^N = 999 \left(\dfrac{F_i - F_i^{min}}{F_i^{max} + F_i^{min}} \right)$

Lobanov [10] expresses values relative to the hypothetical centre of a speaker's vowel space. It uses a method similar to that used widely in statistics. A speaker's mean formant frequency is subtracted from a formant value and then divided by the standard deviation for that formant (12).

(12)　　$F_i^N = \dfrac{F_i - \mu_i}{\sigma_i}$

The Watt & Fabricius method expresses values relative to the centroid of a speaker's vowel space [17]. A vowel space is considered triangular, with the apices at points representing the minimum and maximum F_1 and F_2 for the speaker (13). [u′] is constructed so that $F_1[u'] = F_2[u'] = F_1[i]$. Formants are then expressed relative to the centroid (14).

Three formulations of Watt & Fabricius were included in the study. *origW&F* [17] followed the original method detailed above and using (13) for both F_1 and F_2. *1mW&F* [6] implemented a revised formula for calculating the *x*-coordinate of the centroid, following observations that the original formulation can skew values in the lower part of the vowel space [15]. Shown in (15), this modification excludes $F_2[a]$, thus placing *S* equidistant between [i] and [u′] on the F_2 axis.

(13)　　$S(F_i) = \dfrac{F_i[i] + F_i[a] + F_i[u']}{3}$

(14)　　$F_i^N = \dfrac{F_i}{S(F_i)}$

(15)　　$S(F_2) = \dfrac{F_2[i] + F_2[u']}{2}$

2mW&F constructs [u′] such that $F_2[u']$ is set equal to the lowest mean F_2 value of the point vowels, and $F_1[u']$ is set equal to the lowest F_1 mean value of the point vowels. This gives a more realistic placement of [u′]. (13) and (15) were then used to identify *S*. *Bigham* [2] is a further derivation of *origW&F,* where *S* is the centroid of a quadrilateral not a triangle (16). The apices are constructed at points representing minimum and maximum F_1 and F_2 frequencies for the speaker.

(16)　　$S(F_i) = \dfrac{F_i[i'] + F_i[a'] + F_i[o'] + F_i[u']}{4}$

Letter used actual mean values of a speaker's *lett*ER (i.e. schwa) vowel as the centroid, rather than a constructed *S* (17).

(17)　　$F_i^N = \dfrac{F_i}{F_{i[lettER]}} - 1$

The final methods were based on Nearey [11], which normalize by subtracting the mean of the log-transformed formant frequency across all vowels for the speaker from a log-transformed formant value. *Nearey1* is the formant-intrinsic formulation (18). *NeareyGM* is formant-extrinsic, calculating the mean using both F_1 and F_2 (19). Following [15], exponentials of *Nearey*-normalized values were also computed (20), (21).

(18)　　$F_i^N = \ln(F_i) - \mu_{\ln(F_i)}$

(19)　　$F_i^N = \ln(F_i) - \left(\dfrac{\mu_{\ln(F_1)} + \mu_{\ln(F_2)}}{2} \right)$

(20)　　$F_i^N = \exp\{\ln(F_i) - \mu_{\ln(F_i)}\}$

$$(21) \quad F_i^N = \exp\left\{\ln(F_i) - \left(\frac{\mu_{\ln(F_1)} + \mu_{\ln(F_2)}}{2}\right)\right\}$$

2.3. Methods of comparison

To assess the relative effectiveness of each method in normalizing the dataset, a series of evaluative tests were performed. Results for two of these tests are reported. The first assesses the ability to equalize and the second to align vowel spaces.

For each speaker, the vowel space was taken to be quadrilateral. Four points were defined using mean values taken from vowel categories with maximum and minimum F_2 and F_1. The lines connecting these four points were taken to represent the hypothetical limits of the vowel space. The general formula for calculating the area of a trapezium was then used to calculate the area of each vowel space.

Following [6], the equalization of vowel spaces was quantified by examining the reduction of variance in the speakers' vowel spaces. The squared coefficient of variance (SCV, 22) of each method was compared, as SCV is scale-invariant.

$$(22) \quad SCV = \left(\frac{\sigma}{\mu}\right)^2$$

A low SCV is indicative of a dataset having small variance. If a method's results yield a lower SCV than that of the raw Hertz data, it has reduced the variance of inter-speaker vowel spaces, and hence made different vowel spaces more similar.

The alignment and overlap of vowel spaces were quantified using Python v2.6.4 incorporating the Shapely v1.2.6 package. The area of the intersection of all 20 vowel spaces was divided by the area of the union of all 20 vowel spaces to give the percentage of area that overlapped. As the overlaps are percentages they can be compared directly: higher percentage shows better alignment.

3. RESULTS

3.1. Equalizing vowel spaces

Table 1 gives the SCV of speakers' hypothetical total vowel spaces under each normalization method, and the ranking of each method. The classification of methods as vowel (V), formant (F) and speaker (Sp) intrinsic or extrinsic is also given.

All 20 techniques showed improvement over raw Hertz values. *Gerstman* displayed the smallest SCV, so produced vowel spaces with the least

variance, and hence was the most effective at equalizing them. *1mW&F* outperformed the original *origW&F* formulation. *Nearey1* and *NeareyGM* yielded the same result, giving no reason to favor one over the other.

Nordström performed least well, with *exp{Nearey1}, exp{NGM}, Mel* and *Bark* all ranked low. Although the three worst-performing methods were vowel-extrinsic, the results show that overall, vowel-intrinsic scaling formulae performed less well than vowel-extrinsic formulae.

Table 1: SCVs of vowel spaces for each method.

Method	V	F	Sp	SCV	Rank
Hertz	N/A	N/A	N/A	0.06212	N/A
Gerstman	Extr	Intr	Intr	0.01020	1
LCE	Extr	Intr	Intr	0.01487	2
Lobanov	Extr	Intr	Intr	0.02032	3
Bigham	Extr	Intr	Intr	0.02556	4
1mW&F	Extr	Intr	Intr	0.02587	5
Letter	Extr	Intr	Intr	0.02637	6
origW&F	Extr	Intr	Intr	0.02671	7
2mW&F	Extr	Intr	Intr	0.02818	8
ERB	Intr	Intr	Intr	0.03233	9
Nearey1	Extr	Intr	Intr	0.03250	=10
NeareyGM	Extr	Extr	Intr	0.03250	=10
Log	Intr	Intr	Intr	0.03250	=10
Ln	Intr	Intr	Intr	0.03250	=10
Bladon	Intr	Intr	Intr	0.03409	=14
Bark	Intr	Intr	Intr	0.03409	=14
Bark-diff	Intr	Extr	Intr	0.03549	16
Mel	Intr	Intr	Intr	0.03583	17
exp{Nearey1}	Extr	Intr	Intr	0.03798	=18
exp{NGM}	Extr	Extr	Intr	0.03798	=18
Nordström	Extr	Extr	Extr	0.03977	20

3.2. Aligning vowel spaces

Table 2 presents the overlap percentages and corresponding rankings of each method based on their alignment of the vowel spaces. Impressionistically the results fall into three groups (indicated by thicker lines in Table 2). For visual illustration, see accompanying image files 1-3.

The three versions of *W&F* and *Bigham* (itself derived from *W&F*) were much the most effective. All four formulations of *Nearey* showed clear improvement over raw Hertz, with the exponential versions performing slightly better. Except for *Bladon*, results for the vowel-intrinsic scaling

methods were largely similar, with little if any improvement over raw Hertz.

Table 2: Overlap percentages for each method.

Method	V	F	Sp	% overlapping	Rank
Bigham	Extr	Intr	Intr	45.8%	1
2mW&F	Extr	Intr	Intr	43.8%	2
origW&F	Extr	Intr	Intr	43.4%	3
1mW&F	Extr	Intr	Intr	42.3%	4
Gerstman	Extr	Intr	Intr	30.0%	5
Lobanov	Extr	Intr	Intr	29.2%	6
Nordstrom	Extr	Extr	Extr	28.7%	7
exp{Nearey1}	Extr	Intr	Intr	27.6%	8
Nearey1	Extr	Intr	Intr	27.1%	9
exp{NGM}	Extr	Extr	Intr	26.9%	10
Bladon	Intr	Intr	Intr	25.9%	11
NeareyGM	Extr	Extr	Intr	25.7%	12
Letter	Extr	Intr	Intr	24.1%	13
LCE	Extr	Intr	Intr	23.1%	14
Bark-diff	Intr	Extr	Intr	13.5%	15
Bark	Intr	Intr	Intr	13.2%	16
Mel	Intr	Intr	Intr	13.1%	17
ERB	Intr	Intr	Intr	12.8%	18
Ln	Intr	Intr	Intr	12.2%	=19
Log	Intr	Intr	Intr	12.2%	=19
Hertz	N/A	N/A	N/A	12.6%	N/A

4. DISCUSSION AND CONCLUSION

The results of the comparative tests bring to light some striking patterns and conclusions, some in accordance with what other method comparisons have found, but others in apparent contrast.

The five vowel-intrinsic scaling transformations all performed poorly, as did *Bark-diff*. This result supports claims that vowel-intrinsic methods are inadequate, at least for sociophonetics [1, 4]. However, even the worst-performing methods were an improvement on the raw Hertz measures, suggesting that any normalization is better than none if the aim is to compare different speakers. Vowel-intrinsic normalization may have a role to play in studies of speech perception, but with respect to sociophonetic research other methods appear superior.

Letter and *LCE* performed well at equalizing vowel spaces, but comparatively poorly at aligning them, while *Nordström* was better at aligning than equalizing. These results demonstrate the possibility of methods performing to different levels of effectiveness depending on the method of comparison used, and suggest evaluation of

methods should ideally be based on a range of comparative tests. For both comparisons, the best methods were vowel-extrinsic, formant-intrinsic, speaker-intrinsic. This conclusion supports that of [1, 4, 6]: this type of method performs best for sociophonetic research.

5. REFERENCES

[1] Adank, P., Smits, R., van Hout, R. 2004. A comparison of vowel normalization methods for language variation research. *JASA* 116(5), 3099-3107.
[2] Bigham, D.S. 2008. *Dialect Contact and Accommodation among Emerging Adults in a University Setting*. PhD Dissertation, University of Texas at Austin.
[3] Bladon, R.A.W., Henton, C.G., Pickering, J.B. 1984. Towards an auditory theory of speaker normalization. *Language and Communication* 4(1), 59-69.
[4] Clopper, C.G. 2009. Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistic Compass* 3(6), 1430-1442.
[5] Disner, S.F. 1980. Evaluation of vowel normalization methods. *JASA* 67(1), 253-261.
[6] Fabricius, A.H., Watt, D.J.L., Johnson, D.E. 2009. A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* 21(3), 413-435.
[7] Gerstman, L. 1968. Classification of self-normalized vowels. *IEEE Transactions of Audio Electroacoustics AU-16*, 78-80.
[8] Glasberg, B.R., Moore, B.C.J. 1990. Derivation of auditory filter shapes from notched noise data. *Hearing Research* 47, 103-138.
[9] Hindle, D. 1978. Approaches to vowel normalization in the study of natural speech. In Sankoff, D. (ed.), *Linguistic Variation: Models and Methods*. New York: Academic Press, 161-171.
[10] Lobanov, B.M. 1971. Classification of Russian vowels spoken by different speakers. *JASA* 49(2), 606-608.
[11] Nearey, T.M. 1978. *Phonetic Feature Systems for Vowels*. PhD Dissertation, Indiana University.
[12] Nordström, P.E. 1977. Female and infant vocal tracts simulated from male area functions. *J. Phon.* 5(1), 81-92.
[13] Stevens, S.S., Volkmann, J. 1940. The relation of pitch to frequency: A revised scale. *American Journal of Psychology* 53(3), 329-353.
[14] Syrdal, A.K., Gopal, H.S. 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *JASA* 79(4), 1086-1100.
[15] Thomas, E.R., Kendall, T. 2007. *NORM: The Vowel Normalization and Plotting Suite*. Online Resource. http://ncslaap.lib.ncsu.edu/tools/norm, accessed:17/11/08.
[16] Traunmüller, H. 1990. Analytical expressions for the tonotopic sensory scale. *JASA* 88(1), 97-100.
[17] Watt, D.J.L., Fabricius, A.H. 2002. Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F_1-F_2 plane. *Leeds Working Papers in Linguistics and Phonetics* 9, 159-173.
[18] Watt, D.J.L., Fabricius, A.H., Kendall, T. 2010. More on vowels: plotting and normalization. In Di Paolo, M., Yaeger-Dror, M. (eds.), *Sociophonetics: A Student's Guide*. London: Routledge, 107-118.