The implementation of methods in Visible Vowels

Wilbert Heeringa & Hans Van de Velde

June 24, 2021

1 Introduction

Visible Vowels allows to convert and normalize vowel data and calculate some specific metrics. This document explains how these values are calculated. It is not written as a manual for Visible Vowels. Please contact us at `wjheeringa@gmail.com` if you detect any errors, or if you want to implement other metrics. Refer to this document as: Heeringa, W. & Van de Velde, H. (2018), "The implementation of methods in Visible Vowels," vignette of the R package 'visvow'.

In this document we focus on averaging and long-term formants (Section 2), scale conversion methods (Section 3), speaker normalization of formants (Section 4), speaker normalization of duration (Section 5), methods for the evaluation of speaker formant normalization methods (Section 6), methods for measuring vowel dynamics (Section 7) and exploratory methods (Section 8).

2 Averaging and long-term formants

2.1 Contours, dynamics, duration

Ideally, a data set contains the same number of realizations per vowel and per speaker. When this is not the case, both cross-vowel and cross-speaker averaging will produce confounding results, as illustrated in the following example. Assume you want to calculate the average duration for a particular dialect region. In the data set the dialect region is represented by speakers $s1$ and $s2$ who pronounced the vowels [ɪ] and [aː]. The durations that were measured in milliseconds are:

	[ɪ]	[aː]
$s1$	100, 105	160, 168, 171, 180
$s2$	110, 112, 113	175, 179, 182, 186, 192

By default the average duration of the dialect region is found by calculating the average of the 14 realizations:

$$\frac{100 + 105 + 160 + 168 + 171 + 180 + 110 + 112 + 113 + 175 + 179 + 182 + 186 + 192}{14} \quad (1)$$

which equals to 152ms.

However, as can be seen, both speakers pronounced vowel [aː] more often than vowel [ɪ]. Additionally, speaker $s2$ pronounced more realizations for both vowels than speaker $s1$. When averaging the measurements of all 14 realizations, both vowel [aː] and speaker $s2$ are weighed too heavily. This, however, can be corrected by calculating the average of the measurements of the realizations for each combination of speaker and vowel, and subsequently calculating the average of the (in our example four) averages:

$$duration_{mean} = \frac{\sum_{i=1}^{n_{vowels}} \sum_{j=1}^{n_{speakers}} \frac{\sum_{k=1}^{n_{realizations_{ij}}} duration_{ijk}}{n_{realizations_{ij}}}}{n_{vowels} \times n_{speakers}} \tag{2}$$

where $n_{speakers}$ is the number of speakers, n_{vowels} the number of vowels and $n_{tokens_{ij}}$ the number of realizations for each combination of speaker and vowel. When applying the formula to our example we get:

$$\frac{\left(\frac{100+105}{2}\right) + \left(\frac{160+168+171+180}{4}\right) + \left(\frac{110+112+113}{3}\right) + \left(\frac{175+179+182+186+192}{5}\right)}{4} \tag{3}$$

which equals to 142ms.

This solution works fine when each speaker has pronounced each vowel at least one time. But assume the situation that speaker $s1$ has not pronounced vowel [ɪ]:

	[ɪ]	[aː]
$s1$		160, 168, 171, 180
$s2$	110, 112, 113	175, 179, 182, 186, 192

In that case just three combinations of a speaker and a vowel type are given. When calculating the average duration as the average of the averages of the three combinations like in Formula 2, vowel type [aː] is weighed more heavily than vowel [ɪ].

This is solved by first calculating the average of the measurements of the realizations for each combination of speaker and vowel, then calculating the average duration per vowel, and subsequently calculating the average of the (in our example two) vowel averages. This is accomplished by the following formula:

$$duration_{mean} = \frac{\sum_{i=1}^{n_{vowels}} \frac{\sum_{j=1}^{n_{speakers}} \frac{\sum_{k=1}^{n_{realizations_{ij}}} duration_{ijk}}{n_{realizations_{ij}}}}{n_{speakers}}}{n_{vowels}} \tag{4}$$

When applying the formula to our example we get:

$$\frac{\frac{\left(\frac{(110+112+113)}{3}\right)}{1} + \frac{\left(\frac{160+168+171+180}{4}\right) + \left(\frac{175+179+182+186+192}{5}\right)}{2}}{2} \tag{5}$$

which equals to 144 ms.

Formula 4 is applied in the 'Contours' tab, the 'Dynamics' tab or the 'Duration' tab when the option 'average' is checked.

2.2 Formants

In the 'Formants tab' the options 'average' and 'long-term formants' are found. The options are related since both start by averaging the measurements of realizations of each vowel category per speaker. This assures that all vowel types and speakers are weighed the same (see Section 2.1).

However, in the next step, 'average' and 'long-term formants' are different. The two steps are reflected in the formulas (6) and (7). Using the option 'average' per vowel category the average of the speaker averages is calculated:

$$F_{f_{mean}} = \frac{\sum_{j=1}^{n_{speakers}} \frac{\sum_{k=1}^{n_{realization_{ij}}} F_{f_{ijk}}}{n_{realizations_{ij}}}}{n_{speakers}} \tag{6}$$

where f is equal to 1, 2 or 3. In the plot the average vowel positions are shown.

Using the option 'long-term formants' per speaker the average of the vowel averages is calculated:

$$F_{f_{mean}} = \frac{\sum_{i=1}^{n_{vowels}} \frac{\sum_{k=1}^{n_{realization_{ij}}} F_{f_{ijk}}}{n_{realizations_{ij}}}}{n_{vowels}} \tag{7}$$

where f is equal to 1, 2 or 3. In the plot the average speaker positions are shown.

When both 'average' and 'long-term formants' are checked, 'average' has no effect but is overruled by 'long-term formants'.

3 Scale conversion methods

Conversion methods aim to represent frequencies and frequency differences of pitch and/or formants in accordance with how they are perceived. Visible Vowels offers six scales under the 'f0' tab: Hz, bark, ERB, ln, mel and ST. The same scales are available under the 'Formants' tab, except the ST scale. For bark, ERB and mel multiple versions are provided. The measurements in the input table are supposed to be in Hz.

Scales are applied to f0, F1, F2 and F3 for every time point given in the input table (e.g. 20%, 50%, etc.). All the scales mentioned here are discussed in the following subsections.

3.1 bark

The bark scale is a psychoacoustical scale proposed by Zwicker (1961). Several formulas were approached in order to approach the bark scale as closely as possible. In Visible Vowels Hz values can be converted to bark values by three formulas:

- Schroeder et al. (1979):

$$F_i^{bark} = 7.0 \times ln\left(\frac{F_i^{Hz}}{650} + \sqrt{1 + \left(\frac{F_i^{Hz}}{650}\right)^2}\right) \tag{8}$$

where ln is the natural logarithm. This formula is also used in the function `hertzToBark` in PRAAT (Boersma and Weenink, 2017).

- Zwicker and Terhardt (1980):

$$F_i^{bark} = 13 \times arctan\left(\frac{0.76}{1000} \times F_i^{Hz}\right) + 3.5 \times arctan\left(\left(\frac{F_i^{Hz}}{7500}\right)^2\right) \tag{9}$$

- Traunmüller (1983):

$$F_i^{bark} = \frac{26.81 \times F_i^{Hz}}{1960 + F_i^{Hz}} - 0.53 \tag{10}$$

This formula is the closest approximation of the tabulated data originally published by Zwicker (1961) when all formant frequences are between 200 Hz and 6700 Hz. Traunmüller (1990) added corrections that are applied at both ends of the scale:

$$\begin{array}{ll} \text{if } F_i^{bark} < 2 & \text{add } 0.15 \times (2 - F_i^{bark}) \\ \text{if } F_i^{bark} > 20.1 & \text{add } 0.22 \times (F_i^{bark} - 20.1) \end{array}$$

In Visible Vowels we included these corrections.

3.2 ERB

The equivalent rectangular bandwidth or ERB was proposed by Moore and Glasberg (1983). In Visible Vowels Hz values can be converted to ERB values by three formulas:

- Greenword (1961):

$$F_i^{ERB} = 16.7 \times log_{10}\left(1 + \frac{F_i^{Hz}}{165.4}\right) \tag{11}$$

- Moore and Glasberg (1983):

$$F_i^{ERB} = 11.17 \times ln\left(\frac{F_i^{Hz} + 312}{F_i^{Hz} + 14675}\right) + 43 \tag{12}$$

Almost the same formula is used in the function `hertzToErb` in PRAAT. However the value '14675' is replaced by '14680'.

- Glasberg and Moore (1990):

$$F_i^{ERB} = 21.4 \times log_{10}\left(\left(\frac{4.37}{1000} \times F_i^{Hz}\right) + 1\right) \tag{13}$$

3.3 ln

Miller (1989) proposed to scale formant frequencies to a (natural) logarithmic scale which aligns better with perceived frequency differences. The natural logarithmic transform is calculated as:

$$F_i^{ln} = ln(F_i^{Hz}) \tag{14}$$

3.4 mel

The mel scale (from *melody*) is a perceptual scale of pitches judged by listeners to be equal in distance from one another. In Visible Vowels Hz values can be converted to mel values by two formulas:

- Fant (1968):

$$F_i^{mel} = \frac{1000}{log(2)} \times log\left(\frac{Hz}{1000} + 1\right) \tag{15}$$

  This formula yields that same results for any logarithm base.

- O'Shaughnessy (1987):

$$F_i^{mel} = 1127 \times ln\left(1 + \frac{F_i^{Hz}}{700}\right) \tag{16}$$

3.5 ST

Each musical octave is divided into twelve semitones. Semitones (ST) represent a pitch interval which is one twelfth of an octave. Nolan (2003) found in a provisional analysis that speakers' intuitions about equivalence of intonational span across speakers were best modelled by a psycho-acoustic pitch scale which is logarithmic (semitones) or near-logarithmic (ERB-rate) in the frequency range of interest.

Rietveld and Van Heuven (1997) give a formula that defines the semitone scale as a log-frequency display of f0 frequencies in terms of the departure from a reference of 50 Hz. Choosing 50 Hz as a reference value centers the semitone scale at 50 Hz = 0 semitones. Their formula is:

$$f0^{ST} = 39.87 \times log_{10}\left(\frac{f0^{Hz}}{50}\right) \tag{17}$$

which yields almost the same results as:

$$f0^{ST} = 12 \times log_2 \left(\frac{f0^{Hz}}{50} \right) \tag{18}$$

Fant et al. (2002) uses a semitone scale with a reference frequency of 100 Hz, using the following formula:

$$f0^{ST} = 12 \times log_2 \left(\frac{f0^{Hz}}{100} \right) \tag{19}$$

The same formula is used in the function `hertzToSemitones` in PRAAT.

It is likely that measurements of older male speakers are (partially) below 100 Hz. When converting them to semitones, negative values would be obtained which can cause calculation errors in subsequent analyses. This is less likely to happen with a reference value of 50 Hz.

In Visible Vowels the log_2 is used as in formulas (10) and (11), but the user can choose any reference value. The default value is set to 50 Hz.

4 Speaker normalization of formants

Under the 'Formants' tab Visible Vowels offers 16 normalization procedures. Most of them are described in at least one of the following publications: Adank (2003), Adank et al. (2004), Flynn (2011), Van der Harst (2011) and Esfandiaria and Alinezhadb (2014). With regard to the implementation of the normalization methods in Visible Vowels the following should be considered:

- Normalization is applied **per speaker** to F1 and F2 for all time points in the vowel interval (e.g. 20%, 50%, etc.).
- Some procedures also normalize f0.
- Some procedures also normalize F3. Normalization procedures that normalize only F1 and F2 are not available when one of the axes of a plot represents F3 (see Table 1).
- The procedure does **not** consider any **subsetting** according to any categorical, variable, i.e. all cases in the input table are involved in the normalization procedure.
- In Visible Vowels there are ten normalization procedures that can be applied to any scale (Hz, bark, etc.) There are six normalization methods in which a logarithmic transformation is included. The latter ones can be applied only to formant measurements in the Hz scale in order to avoid the measurements being converted twice (see Table 1).
- The procedures are not applied individually per time point, but to all time points at once. Applying normalization procedures individually per time point would erroneously eliminate differences between time points.

	applied to				requires		base	use
	f0	F1	F2	F3	f0	f3	scale	descr.
Formant-ratio normalization								
Peterson (1951)	x	x	x			x		
Sussman (1986)		x	x	x		x	Hz	
Syrdal & Gopal (1986)		x	x		x	x		
Miller (1989)		x	x	x	x		Hz	x
Thomas & Kendall (2007)	x	x	x			x		
Range normalization								
Gerstman (1968)	x	x	x	x				x
Centroid normalization								
Lobanov (1971)	x	x	x	x				x
Watt & Fabricius (2002)		x	x					x
Fabricius et al. (2009)		x	x					x
Bigham (2008)		x	x					x
Heeringa & Van de Velde (2021) I		x	x					x
Heeringa & Van de Velde (2021) II		x	x					x
Log-mean normalization								
Nearey (1978) I	x	x	x	x			Hz	x
Nearey (1978) II	x	x	x	x	x	x	Hz	x
Labov (2006) I		x	x				Hz	x
Labov (2006) II		x	x	x		x	Hz	x

Table 1: *Overview of formant normalization methods. When no scale is given under 'base scale' any scale is possible. An 'x' in de column 'use descr.' indicates that the procedure uses descriptives like minimum, mean, etc.*

- Some procedures require f0 and/or f3 for normalizing (see Table 1). When no values are given for the required variable(s) of a procedure, the procedure is not made available in Visible Vowels.

- Some normalization methods use descriptives like the minimum, maximum, mean or standard deviation (see Table 1). When using any of these methods, the user can choose which time points should be included when these descriptives are calculated. Below we refer to them as *descriptive time points*. Selecting points to be used for calculating descriptives is independent of selecting points to be plotted in graphs. The latter selection is done by means of a separate input. When the user does not choose any descriptive time points, Visible Vowels chooses the most central time point, i.e. given n time points, the $round(n/2)$-th point is chosen by default.

In Visible Vowels the 16 normalization methods are classified in four types: formant-ratio normalization, range normalization, centroid normalization and log-mean normalization. The four groups are discussed in Sections 4.1, 4.2, 4.3 and 4.4 respectively. Per section the methods are discussed in chronological order.

4.1 Formant-ratio normalization

All of the methods in this section are different formulations of the formant ratio hypothesis. The basic idea is that vowels are relative patterns, not absolute formant frequencies (Johnson, 2005).

4.1.1 Peterson (1951)

Peterson (1951) plotted F1/F3 ratios against F2/F3 ratios. Mohanan and Idsardi (2010) reintroduced the use of the same ratios and confirmed that the auditory cortex is sensitive to modulations of the F1/F3 ratio.

For time point t the measurements of the vowel realizations are normalized as follows:

$$F_{t1}^{Peterson} = \frac{F_{t1}}{F_{t3}} \tag{20}$$

$$F_{t2}^{Peterson} = \frac{F_{t2}}{F_{t3}} \tag{21}$$

Peterson (1951) calculated the ratios on the basis of mel-transformed measurements. In Visible Vowels the ratios can be calculated on the basis of any scale that is available in the app.

4.1.2 Sussman (1986)

Using the normalization method of Sussman (1986) each formant value is expressed relative to the geometric mean of F1, F2 and F3. For time point t and variable i the measurements of the vowel realizations are normalized as follows:

$$F_{ti}^{Sussman} = ln\left(\frac{F_{ti}}{\hat{F}_t}\right) \tag{22}$$

where $\hat{F}_t$ is the geometric mean, which is calculated from the values of the three formants of the vowel realiziation that is being normalized, and ln is the natural logarithm. In Visible Vowels this normalization procedure can be applied only to measurements in the Hz scale.

4.1.3 Syrdal & Gopal (1986)

The normalization procedure of Syrdal and Gopal (1986) is based on their observation that the distance between neighbouring formants is similar across speakers. In their model F1 minus F0 corresponds to the height dimension and F3 minus F2 to the front back dimension. F1 and F2 frequencies of vowel realizations are normalized at time point t by the following formulas:

$$F_{t1}^{S\&G} = F_{t1} - F_{t0} \tag{23}$$

$$F_{t2}^{S\&G} = F_{t3} - F_{t2} \tag{24}$$

Syrdal and Gopal (1986) applied the two formulas to frequencies that were scaled to bark and therefore this method is known as the "Bark-Distance Method". In Visible Vowels, however, the formulas can also be applied to frequencies in the other scales that are available in the app.

The transformation of the F2 frequencies causes that large F2 values become small, and small F2 values become large. In order to prevent the graph from being mirrored in the F2 dimension, the transformed F2 values are multiplied by -1.

4.1.4 Miller (1989)

Using the normalization method of Miller (1989) formants are compared with their lower neighbours, i.e. F3 with F2 and F2 with F1. The first formant is normalized against a sensory reference (SR). The Sensory Reference (SR) is calculated for each vowel realization using geometric mean f0 (μ_{f0}) that is corrected for a constant c.

In our implementation, the geometric mean μ_{f0} is calculated as follows. First the f0 values are averaged per combination of speaker, vowel type and time point, where time point is one of the time points that were chosen by the user to be considered when calculating descriptives (see introduction of Section 4). Using the averages in the data set thus obtained the geometric mean μ_{f0} per speaker and across the vowel categories and the time points is calculated.

The constant c is the geometric mean of the f0 average of the male speakers and the f0 average of the female speakers. Miller suggested to use $c = 168$ Hz, a value he found on the basis of f0 measurements in the Peterson and Barney database (Peterson and Barney, 1952). The f0 average of the male speakers was 125 Hz and f0 average of the female speakers was 225 Hz. The geometric mean is $\sqrt{(125 \times 225)} = 168$ Hz.

Rather than recalculating the constant c from the input data, we use $c = 168$ Hz. Thus the procedure can still be used when the input table contains measurements from only male speakers or only female speakers or when the number of speakers is small. Using this constant, SR is computed as:

$$SR = 168 \times \left(\frac{\mu_{f0}}{168}\right)^{\frac{1}{3}} \tag{25}$$

Now formant frequencies in Hz of all vowel realizations are normalized for each time point t by the following formulas:

$$F_{t1}^{Miller} = ln\left(\frac{F_{t1}}{SR}\right) \tag{26}$$

$$F_{t2}^{Miller} = ln\left(\frac{F_{t2}}{F_{t1}}\right) \tag{27}$$

$$F_{t3}^{Miller} = ln\left(\frac{F_{t3}}{F_{t2}}\right) \tag{28}$$

where ln is the natural logarithm.

4.1.5 Thomas & Kendall (2007)

The normalization procedure of Thomas and Kendall (2007) is similar to the procedure of Syrdal and Gopal (1986), but both F1 and F2 are normalized by subtracting them from F3. Formant frequencies of vowel realizations are normalized for a time point t as follows:

$$F_{t1}^{S\&G} = F_{t3} - F_{t1} \tag{29}$$

$$F_{t2}^{S\&G} = F_{t3} - F_{t2} \tag{30}$$

The transformation of both the F1 and the F2 frequencies causes that large formant values become small, and small formant values become large. In order to prevent the graph from being mirrored in both the F1 and the F2 dimension, the transformed values are multiplied by -1.

4.2 Range normalization

4.2.1 Gerstman (1968)

Gerstman (1968) normalizes the frequencies of a formant on the basis of the lowest and highest values found per speaker and across the selected descriptive time points. The frequencies are scaled so that they range from 0 to 999.

As a first step, for each of the variables F1 and F2 the values are averaged per combination of speaker, vowel type and descriptive time point which is a time point that was chosen by the user to be considered when calculating descriptives (see introduction of Section 4). This avoids that vowel types that occur frequently are weighed more heavily then those that occur less frequently.

Next, per speaker the formant values are averaged across the descriptive time points, and the minima and maxima for F1, F2, F3 are found. Then for the vowel realizations of speaker k, time point t and variable i we calculate:

$$F_{kti}^{Gerstmann} = 999 \times \frac{F_{kti} - F_{ki}^{min}}{F_{ki}^{max} - F_{ki}^{min}} \tag{31}$$

4.3 Centroid normalization

4.3.1 Lobanov (1971)

A normalization procedure that expresses values relative to the hypothetical centre of a speaker's vowel space is that developed by Lobanov (1971). Using this method a speaker's mean formant frequency is subtracted from a specific formant value and then divided by the standard deviation for that formant. In the normalized F1-F2 plot Lobanov's centroid lies at (0,0).

Visible Vowels first detects whether the number of different vowel categories is larger than 1. If this is not the case and all measurements represent realizations of the same vowel category, for each of the variables F1, F2 and F3 the mean and the standard deviation are calculated per speaker and across the realizations and the descriptive time points that were chosen by the user to be

considered when calculating descriptives (see introduction of Section 4). Using the mean and the standard deviation the formant frequencies of all vowel realizations of speaker k, time point t and variable i are normalized as follows:

$$F_{kti}^{Lobanov} = \frac{F_{kti} - \mu_{ki}}{\sigma_{ki}} \tag{32}$$

If the number of different vowels is larger than 1, the mean and the standard deviation are obtained in a slightly different way. We have to assure that vowel types are weighed equally rather than by the number of realizations each of them has. This solved by averaging the F1, F2 and F3 measurements per combination of speaker, vowel type and descriptive time point, where time point is one of the time points that were chosen by the user to be considered when calculating descriptives (see introduction of Section 4). Then per speaker the mean and standard deviation of these averaged values are calculated.

4.3.2 Watt & Fabricius (2002)

The procedure of Watt and Fabricius (2002) expresses frequency values relative to a constructed centroid that is based on points that represent the corners of the vowel envelope of a speaker's vowel space. After normalization the centroid lies at (1,1) in the F1-F2 plot.

As a first step, for each of the variables F1 and F2 the values are averaged per combination of speaker, vowel type and descriptive time point which is a time point that was chosen by the user to be considered when calculating descriptives (see introduction of Section 4). This avoids that vowel types that occur frequently are weighed more heavily then those that occur less frequently.

Next per speaker the formant values are averaged across the descriptive time points. Thus we get a data set that contains average F1 and F2 frequencies for each vowel type per speaker. Using this data set for each speaker we find the corners of the vowel envelope which we call [i], [a] and [u']. The coordinates of [i] are minimum F1 and the maximum F2. The coordinates of [a] are the maximum F1 and the F2 of the vowel type that has the maximum F1. The minimum F1 is also assigned to the F1 *and* the F2 of [u']. Now the coordinate of formant i of the centroid for speaker k is calculated as:

$$S_{ki} = \frac{F_{ki[i]} + F_{ki[a]} + F_{ki[u']}}{3} \tag{33}$$

Formant values of the vowel realizations of speaker k, formant i and time point t are normalized as follows:

$$F_{kti}^{W\&F} = \frac{F_{kti}}{S_{ki}} \tag{34}$$

4.3.3 Fabricius et al. (2009)

A weakness of the normalization method of Watt and Fabricius (2002) that was noticed by Thomas and Kendall (2007) is that the F2 of [a] might differ considerably from the mean value of the F2 of [i] and the F2 of [u'] and thus distort

the lower part of the vowel space. Therefore, Fabricius et al. (2009) proposed an alternative with a modified formula for calculation of the coordinates of the centroid:

$$S_{ki} = \begin{cases} \dfrac{F_{ki[i]} + F_{ki[a]} + F_{ki[u']}}{3} & , i = 1 \\ \\ \dfrac{F_{ki[i]} + F_{ki[u']}}{3} & , i = 2 \end{cases} \tag{35}$$

4.3.4 Bigham (2008)

When using the procedures of Watt and Fabricius (2002) and Fabricius et al. (2009) it is assumed that the vowel space has the shape of a triangle. The normalization method of Bigham (2008) is another derivation of the procedure of Watt and Fabricius (2002), but its centroid is obtained on the basis of the corners of a quadrilateral. As corners Bigham choose the American English vowels [ɪ], [u], [æ] and the average of [ɑ] and [ɔ], with tokens taken from word list items of the form /hVd/.

We implemented a modified version of Bigham's method that was proposed by Flynn (2011) (see also Flynn and Foulkes (2011)). When using this approach the centroid is obtained on the basis of the corners of the vowel envelope which are called [i'], [a'], [o'] and [u']. The coordinates of [i'] are minimum F1 and maximum F2. Minimum F1 is also assigned to [u']. Minimum F2 is assigned to [u'] and [o']. Maximum F1 is assigned to [o'] and [a'].

The F2 of [a'] was set equal to the F2 of the TRAP-vowel [ae]. In our implementation first we try to find the [ae] in the data set. If the vowel is not found, the procedure searches for [aeˑ], if that vowel is not found, the procedure searches for [aeː], then for [a], then for [aˑ], then for [aː], then for [ɛ], then for [ɛˑ], then for [ɛː].

Now the coordinate of formant i of the centroid for speaker k is calculated as:

$$S_{ki} = \frac{F_{ki[i']} + F_{ki[a']} + F_{ki[o']} + F_{ki[u']}}{4} \tag{36}$$

4.3.5 Heeringa & Van de Velde (2021) I

The assumption behind the procedures of Watt and Fabricius (2002) and Fabricius et al. (2009) is that vowel spaces have the shape of a triangle. When using the procedure of Bigham (2008) the vowel space is assumed to be a quadrilateral. In response to this Heeringa and Van de Velde (2021) developed a normalization method that does not assume a particular shape. It calculates the centroid on the basis of all points that constitute the convex hull that encloses the vowel space.

Just as for the procedures of Watt and Fabricius (2002), Fabricius et al. (2009) and Bigham (2008), we first calculate the average F1 and F2 per combination of speaker, vowel type and descriptive time point so that each vowel is

equally weighed in the normalization procedure. With 'descriptive time point' we mean a point that was chosen by the user to be considered when calculating descriptives (see introduction of Section 4).

Next per speaker the formant values are averaged across the descriptive time points. Thus we get a data set that contains average F1 and F2 frequencies for each vowel type per speaker. Using this data set the vowels are found that constitute the convex hull. To this end we use the R function `chull` from the `grDevices` package. This function uses an algorithm that is given by Eddy (1977). On the basis of the F1,F2 coordinates of the vowels that constitute the convex hull the coordinates of the centroid are found with the R function `poly_center` of the package `pracma`. This function calculates the centroid as the center (of mass) of the convex hull.

If S_{ki} is the coordinate of formant i of the centroid of the vowel space of speaker k, then the vowel realizations of speaker k, formant i and time point t are normalized as follows:

$$F_{kti}^{convex\ hull} = \frac{F_{kti}}{S_{ki}} \tag{37}$$

4.3.6 Heeringa & Van de Velde (2021) II

Heeringa and Van de Velde (2021) developed a second normalization method that can be considered as a variant of Lobanov's normalization method. Different from Lobanov's method their method does not depend on the distribution of the vowels within the vowel spaces of the speakers. When normalizing formant i of speaker k the μ in Lobanov's z-score formula is replaced by the centroid coordinate S_{ki} as calculated in Heeringa & Van de Velde I. The σ is calculated on the basis of the formant values i of the vowels that constitute the convex hull.

The vowels that constitute the convex hull may be irregular distributed. I.e. the Euclidean distances of pairs of consecutive vowels may vary (strongly). In order to solve this, the number of points on the convex hull is interpolated up to 1000 points. Next the points are classified in ten classes of equal width, both on the basis of F1 and F2. The authors found that by using ten classes there is an equilibrium between the even distribution of the points on the convex hull and providing sufficient detail. The ten F1 classes do not exactly correspond with the F2 classes since F1 and F2 differences of the pairs of two successive points do not exactly correlate. Therefore, points may have the same F1 class and different F2 classes, or the other way around. For each F1 class/F2 class combination points that had a F1 within the F1 class and a F2 within the F2 class are averaged. Then the number of points becomes equal to the number of F1 class/F2 class combinations. σ is calculated as the standard deviation of the formant values i of these points.

4.4 Log-mean normalization

4.4.1 Nearey (1978) I

The normalization methods of Nearey (1978) transform Hz measurements to logarithms, and subsequently subtract a reference value from the log-transformed frequencies. The reference value is a log-mean. In the version explained in this section the log-mean is calculated for each variable (F1, F2, F3) individually. Therefore, Van der Harst (2011) refers to this procedure as *Nearey's individual log-mean model*

In our implementation, we first calculate the natural logarithms of the formant values (F1, F2, F3) which should be given in Hz.

Next we calculate means for each combination of speaker, vowel type, descriptive time point and variable (F1, F2, F3), thus avoiding that vowel types that occur more frequently in the data are weighed more heavily than others. With 'descriptive time point' we mean a point that was chosen by the user to be considered when calculating descriptives (see introduction of Section 4).

Using these means, for each speaker k and variable i we calculate the average frequency μ_{ki}^{ln}. Then for each speaker k, each time point t and each variable i we calculate:

$$F_{kti}^{Nearey} = F_{kti}^{ln} - \mu_{ki}^{ln} \tag{38}$$

4.4.2 Nearey (1978) II

The method presented in this section is similar to the one explained in the previous section, except that the reference value is calculated by taking a speaker's mean of the log-means of the variables f0, F1, F2 and F3. Since the same reference value is used for normalizing F1, F2 and F3 frequencies, Van der Harst (2011) refers to this method as *Nearey's shared log-mean model*.

Normalized frequencies of the vowel realizations of speaker k, time point t and variable i are calculated as follows:

$$F_{kti}^{Nearey} = F_{kti}^{ln} - \frac{\mu_{ki}^{ln} + \mu_{ki}^{ln} + \mu_{ki}^{ln} + \mu_{ki}^{ln}}{4} \tag{39}$$

4.4.3 Labov's ANAE method 1

The normalization procedure of Labov et al. (2006) was designed for the *Atlas of North American English*.

First the natural logarithms of the formant values (F1, F2) given in Hz are calculated. Using these values the grand mean G is calculated. The grand mean G is the geometric mean of the logarithmically transformed values of both F1 and F2 of all speakers. The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers $x_1, x_2, ..., x_n$ the geometric mean is:

$$\sqrt[n]{x_1 \times x_2 \times ... \times x_n} \tag{40}$$

In order to avoid any bias towards vowel types that are more frequently found in the data, we generate a table that contains the geometric mean of the realizations for each combination of speaker, vowel type, descriptive time point and formant (F1, F2), where a descriptive time point is a point that was chosen by the user to be considered when calculating descriptives (see introduction of Section 4).

Given n_k speakers, n_v vowel types, n_t time points and n_i(=2) formants, we calculate G as the geometric mean of the $n_k \times n_v \times n_t \times n_i$ geometric means.

In addition, the mean S_k per speaker is calculated as the geometric mean of the $n_v \times n_t \times n_i$ geometric means of speaker k.

Subsequently the anti-log (with base e, i.e. the exponent) of the difference between the two means ($G - S_k$) is calculated, which results in a speaker-specific scaling factor d_k:

$$d_k = exp(G - S_k) \tag{41}$$

Next, the formant values of the vowel realizations of speaker k, time point t and formant i are multiplied by this scaling factor:

$$F_{kti}^{Labov} = d_k \times F_{kti} \tag{42}$$

Since this procedure uses only F1 and F2 measurements, in Visible Vowels only F1 and F2 formant frequencies are normalized when this procedure is chosen.

Labov et al. (2006) pointed out that G, using F1 and F2, stabilizes when the number of speakers exceeds 345. Thomas and Kendall (2007) mention that this may indicate "that this method (and perhaps speaker-extrinsic methods in general) are best only when a study has an exceptionally high subject count."

4.4.4 Labov's ANAE method 2

Labov's ANAE method 2 uses the same procedure as Labov's ANAE method 1, except that also F3 measurements are included when G and S_k are calculated. Therefore, when using this method in Visible Vowels F3 formant frequencies are normalized as well.

5 Speaker normalization of duration

Under the 'Duration' tab Visible Vowels offers one normalization procedure which is discussed in the subsection below.

The procedure does *not* consider any *subsetting* according to any categorical, variable, i.e. durations of all cases in the input table are involved in the normalization procedure.

5.1 Lobanov (1971)

Lobanov's z-score transformation (Lobanov, 1971) was used as a procedure for normalization duration by –among others– Wang and Van Heuven (2006), Kachkovskaia (2014) and Kocharov et al. (2015).

In our implementation it is first detected whether the number of different vowel types is larger than 1. If this is not the case and all measurements represent realizations of the same vowel, the mean duration and the standard deviation are calculated per speaker and across the vowel realizations. Using these descriptives the durations of the vowel realizations of speaker k are normalized as follows:

$$F_k^{Lobanov} = \frac{F_k - \mu_k}{\sigma_k} \tag{43}$$

If the number of different vowels is larger than 1, the mean and the standard deviation are obtained in a slightly different way. We have to assure that vowel types are weighed equally rather than by the number of realizations each of them has. Therefore, we first calculate average durations for each combination of speaker and vowel type. Using these average durations the mean and the standard deviation are calculated per speaker.

6 Evaluation of speaker formant normalization methods

The goal of the tab 'Evaluate' is to find the best combination of a scale conversion method and a speaker normalization method for the data set that was uploaded by the user. After having chosen the settings the user presses the Go! button. Running the evaluation procedure *may take some time*, depending on the size of the data set.

In case the speakers have pronounced different sets of vowels, the procedures are run on the basis of the set of vowels that are found across all speakers. The vowels thus excluded are printed.

6.1 Evaluate

After selecting 'Evaluate' under 'Choose' the evaluation methods of Adank et al. (2004) and Fabricius et al. (2009) become available.

The results are presented as a table where the columns represent the scale conversion methods and the rows the normalization procedures. Each score is shown on a background with a color somewhere between turquoise and yellow. The more yellow the background is, the better the result. Note that for some tests larger scores represent better results, and for other tests smaller scores represent better results.

For 'Hz' frequencies all normalization methods are given. In order to avoid double scaling, for the other scales no scores are given for normalization methods that implicitely scale frequencies themselves.

When f0 is checked, only those normalization procedures are evaluated that are able to normalize f0 scores. When F3 is checked, only those normalization procedures are evaluated that are able to normalize F3 scores. When f0 and/or F3 is checked, only results of the evaluation methods of Van der Harst (2011) are shown, since the evaluation methods of Fabricius et al. (2009) work only in F1/F2 space.

When f0 and/or F3 frequencies are not given in the data set, normalization procedures that use f0 and/or F3 for normalizing are left out in the results (see Table 1).

6.1.1 Adank et al. (2004)

Adank (2003), Adank et al. (2004) and Van der Harst (2011) evaluated vowel normalization methods by comparing them on how effectively they 1) preserve phonemic, 2) minimize anatomical/physiological and 3) preserve sociolinguistic information in acoustic representations of vowels. Whereas Adank (2003) and Adank et al. (2004) only tested the procedures for monophthongs, Van der Harst (2011) also took diphthongal vowels into account.

The normalization procedures were evaluated through Linear Discrimant Analysis (LDA), a standard pattern recognition technique. LDA assumes that the covariance matrices are equal across categories. Adank et al. (2004) mention that this assumption if often not met for vowel formant frequencies. The alternative is to use Quadratic discriminant analysis (QDA). Adank et al. (2004) mention that this method has the drawback that it requires much larger numbers of parameters to be estimated than LDA, thus risking overfitting. When comparing the use of LDA and QDA for establishing how well the normalization procedures preserved information about the vowel token's intended phonemic identity in the normalized acoustic variables, Adank (2003) and Adank et al. (2004) found that the percentages of correctly classified vowel tokens for QDA were only 1% to 2% higher than those for LDA. Given the parsimony of the LDA model relative to QDA, they used only LDA in the rest of their study. In our implementation only LDA is used as well.

In Visible Vowels multiple anatomic variables and multiple sociolinguistic variabels can be entered. When multiple anatomic variables are entered, they are 'fused'. Assume a variable with values 'A', 'B' en 'C', and another variable with values 'x' and 'y'. Then after fusing the following values are possible: 'A/x 'A/y', 'B/x', 'B/y', 'C/x' and 'C/y'. We refer to the fused variables as 'compound variable'. For sociolinguistic variables the same approach is used.

As input for the three procedures we use the F1, F2 and F3 (if chosen by the user and available in the data set) averaged per vowel, speaker, time point, compound anatomic variable value and compound sociolinguistic variable value.

Preserve phonemic variation
With the R function `lda` from the `MASS` package the vowel category is predicted by F1, F2 and (if chosen and available) F3. The percentage of correctly predicted vowel categories is returned for the combination of scale conversion method and speaker normalization method that is under consideration. The higher the percentage, the better the phonemic variation is preserved.

Minimize anatomic variation
When multiple time points are chosen by the user, the measurement at each time

point is considered as a separate vowel category. For example, when formants of [i], [a] and [u] are measured at the 25% point and the 75% point, we process them as six vowels: [i]25%, [i]75%, [u]25%, [u]75%, [a]25% and [a]75%.

Assume we consider two anatomic variables, one having the values 'male' and 'female and another having the values 'old' and 'young'. Then the compound anatomic variable has values 'male/old', 'male/young', 'female/old', 'female/young'.

Now for each speaker the value of the compound anatomical variable is predicted by the formant values of the vowels that were pronounced. Assume we consider F1, F2 and F3, then in our example there are 6 vowels $\times$ 3 formants is 18 predictors. If there are n speakers, n values are predicted, unless the compound anatomical variable defines different conditions under which the same speakers have pronounced vowels. If there are k conditions and n speakers, then $k \times n$ values are predicted.

The values of the compound anatomical variable are predicted by the R function `lda` from the `MASS` package. The percentage of correctly predicted values is returned for the combination of scale conversion method and speaker normalization method that is under consideration. The lower the percentage, the better anatomic differences are minimized.

Preserve sociolinguistic variation
When multiple time points are chosen by the user, the measurement at each time point is considered as a separate vowel category. For example, when formants of [i], [a] and [u] are measured at the 25% point and the 75% point, we process them as six vowels: [i]25%, [i]75%, [u]25%, [u]75%, [a]25% and [a]75%.

Assume we consider two sociolinguistic variables, one having the values 'north' and 'south' and another having the values 'rural' and 'urban'. Then the compound sociolinguistic variable has values 'north/rural', 'north/urban', 'south/rural', 'south/urban'.

Now per vowel the values of the compound sociolinguistic variable are predicted for all of the speakers using the `lda` function and the percentage of correctly predicted values is returned. If there are n speakers, n values are predicted, unless the compound sociolinguistic variable defines different conditions under which the same speakers have pronounced vowels. If there are k conditions and n speakers, then $k \times n$ values are predicted.

When v vowels are considered, v percentages are obtained, each percentage being obtained on the basis of n predictions. In our example $v = 6$. As a final result, the average of the v percentages is returned. The higher the average percentage, the better the sociolinguistic variation is preserved.

6.1.2 Fabricius et al. (2009)

Two evaluation methods that were proposed by Fabricius et al. (2009) are available. The first method assesses the ability to equalize vowel spaces and the second to align vowel spaces. The two methods were also used by Flynn (2011) and Flynn and Foulkes (2011). They evaluate normalization methods only on the basis of F1 and F2.

In case multiple vowels of the same vowel category are pronounced by the same speaker, their formant frequencies are averaged before using the two evaluation methods.

When multiple time points are chosen by the user, the evaluation results are produced *per time point* and subsequently *averaged across time points*.

Anatomical and sociolinguistic variables that the user may have entered are not considered when these two evaluation methods are used, unless they define different conditions under which the same speakers have pronounced vowels. In that case the number of vowels per speaker is the number of different vowel categories times the number of conditions.

Equalize vowel space areas

The idea behind the first method is to quantify the equalization of the areas of the vowel spaces by examining the reduction of variance in the speakers' vowel spaces. In order to calculate the area of a vowel space, Flynn and Foulkes (2011) assumed the vowel space to have the shape of a trapezium. Fabricius et al. (2009) calculated the area of a vowel space on the basis of its convex hull, which makes the procedure independent of any shape of the vowel space. In order to find the convex hull we used the R function `chull`. The area that is enclosed by the convex hull was calculated by the R function `polyarea` from the `pracma` package. Then the squared coefficient of variance (SCV) was calculated as:

$$SCV = \left(\frac{\sigma}{\mu}\right)^2 \tag{44}$$

Dividing σ by μ makes the SCV scale-invariant. Next, Fabricius et al. (2009) divided each method's SCV by the Hertz SCV, which gave the proportion of variance that remained after normalization. This proportion was subtracted from 1, resulting in the proportional reduction in variance.

Improve vowel space overlap

The second method proposed by Fabricius et al. (2009) was also used by Flynn (2011), Flynn and Foulkes (2011) and Esfandiaria and Alinezhadb (2014). When using this method the area of the intersection of the vowel spaces of the speakers is calculated and divided by the area of the union of the speaker's vowel spaces. This results in the proportion of area that overlaps. A higher proportion shows a better alignment. Again, the areas are found on the basis of their convex hulls, not assuming any particular shape a priori. However, different from what Fabricius et al. (2009) proposed, Flynn and Foulkes (2011) assumed a quadrilateral and Esfandiaria and Alinezhadb (2014) assumed a triangle.

Fabricius et al. (2009) calculated overlap for each pair of speakers. Following Flynn and Foulkes (2011) we divided the area of the intersection of the vowel spaces of all speakers by the area of the union of the vowel spaces of all speakers, thus obtaining one score for the complete set of speakers.

6.2 Compare

After having selected the option 'Compare' under 'Choose' one can choose to compare either scale conversion methods or speaker normalization methods.

The methods are compared on the basis of F1, F2 and F3 (if chosen by the user and available in the data set) averaged per vowel, speaker, time point, compound anatomic variable value and compound sociolinguistic variable value.

Scale normalization methods
When comparing the scale conversion methods, they are applied to the unnormalized formant measurements. Each scale conversion method is compared to each conversions method. Since there are 10 methods including 'no scaling' (i.e. retaining the original frequencies in Hz), the number of comparison pairs is $(10 \times (10 - 1))/2 = 45$. Two methods are compared by correlating the frequencies converted by the one method with the frequencies converted by the other method. The correlations are calculated individually for F1, F2 and (if chosen and available) for F3 frequencies. Then the three (or two) correlation coefficients are averaged. This average correlation is converted to a distance by calculating 1 - average correlation.

On the basis of these distances the ten methods are clustered with the R function `hclust` from the `stats` package. When using this function the option 'method' is set to 'average' which makes the function performing UPGMA clustering (Jain and Dubes, 1988).

Speaker normalization methods
When comparing the speaker normalization methods, they are applied to the raw Hz formant measurements. Each speaker normalization method is compared to each speaker normalization method. Since there are 16 methods including 'no normalization' (i.e. retaining the original frequencies in Hz), the number of comparison pairs is $(16 \times (16 - 1))/2 = 120$. Two methods are compared by correlating the frequencies normalized by the one method with the frequencies normalized by the other method. The correlations are calculated individually for F1, F2 and (if chosen and available) for F3 frequencies. Then the three (or two) correlation coefficients are averaged. This average correlation is converted to a distance by calculating 1 - average correlation.

On the basis of these distnaces the 16 methods are clustered by UPGMA clustering using the R function `hclust`.

7 Measuring vowel dynamics

In order to measure vowel dynamics, we implemented two methods that were introduced by Fox and Jacewicz (2009): *trajectory length* and *spectral rate of change*. Fox and Jacewicz (2009) used these methods on the basis of F1 and F2 measurements. In Visible Vowels the user can choose any subset of f0, F1, F2 and F3, including the individual variables or all of the variables.

7.1 Trajectory length

Fox and Jacewicz (2009) measured trajectory length as the sum of the lengths of the vectors a trajectory consists of, where the length of a vector is the Euclidean distance between the F1,F2 coordinates of the starting point and the end point respectively. Generalized to any subset of the variables mentioned above, the vector section length VSL between two consecutive temporal points i and $i+1$ is:

$$VSL_{i..i+1} = \sqrt{\sum_{v \in Vset} (var_{v,i} - var_{v,i+1})^2} \tag{45}$$

where $1 \leq i < n$, n being the number of time points for which the variables are measured in the vowel interval, and $Vset$ being the set of variables chosen by the user to be included in the calculation of VSL. The trajectory length TL is measured as:

$$TL = \sum_{i=1}^{n-1} VSL_{i..i+1} \tag{46}$$

7.2 Spectral rate of change

Fox and Jacewicz (2009) write that "differences in vowel dynamics are manifested in the way the spectral change varies across vowel's duration". To address this, they propose to measure the spectral **r**ate **o**f **c**hange TL_{roc}:

$$TL_{roc} = \frac{TL}{time_n - time_1} \tag{47}$$

where $time_1$ and $time_n$ represent the times (in seconds or milliseconds) of respectively the first and the last temporal point that was included in the calculation of TL.

This formula works correctly when the duration of any pair of consecutive temporal points is the same, as it was the case for the data used by the authors, but would not correctly measure the rate of speech when the sections are unequally sized. This problem is easily be solved by using $VSL_{roc_{i..i+1}}$ instead of $VSL_{i..i+1}$ in (46), hence:

$$VSL_{roc_{i..i+1}} = \frac{VSL_{i..i+1}}{time_{i+1} - time_i} \tag{48}$$

where $time_i$ and $time_{i+1}$ represent the times (in seconds or milliseconds) of respectively the i^{th} and $i+1^{th}$ temporal point that was included in the calculation of TL. This formula was also proposed by Fox and Jacewicz (2009), however only for calculating change in the individual sections of a trajectory. Including this formula in (46) we get:

$$TL_{roc} = \sum_{i=1}^{n-1} VSL_{roc_{i..i+1}} \tag{49}$$

which is used in Visble Vowels in order to measure the overall rate of change.

8 Exploratory methods

In the 'Explore' panel distances between speakers and between groups of speakers can be calculated, where the groups are defined according to one or more categorical variables.

In Visible Vowels two measures are available. The first one we refer to as the Euclidean distance. The second one is the ACCDIST metric of Huckvale (2004). Calculating Euclidean distances takes less computation time than calculating ACCDIST distances. The ACCDIST distance, however, is usefull when the user wants to consider the relative mutual relationships of vowels within the speaker's vowel spaces regardless the sizes of their vowel spaces.

8.1 Choice of parameters

The user can choose the vowels to be considered. However, the procedure requires that speakers are compared to each other on the basis of the same set of vowels. If the set of vowels is not the same for each speaker, the user can choose only those vowels that are shared by all speakers. The vowels that are excluded by the procedure are shown the first time the user enters the 'Explore' panel after having loaded the input table in the 'Load file' panel.

The user can choose the time points to be included. Given the user's selection of vowels and time points Visible Vowels calculates the average F1, F2 and F3 for any combination of speaker, time point and vowel.

The user can also choose which formants should be considered: F1, F2, F3, F1 & F2, F1 & F3, F2 & F3, F1 & F2 & F3.

8.2 Euclidean distance

Using the average F1 and/or F2 and/or F3 for any combination of the included speakers, time points and vowels, a distance is calculated for any pair of speakers.

In case formant frequencies are measured for n_v vowels and at n_t time points per vowel, we consider this as a set of $n_v \times n_t$ different vowels. For example, assume vowels [a] and [i] with measurements at the 20%, 50% and 80% time point. We consider them as six different vowels: [$a_{20\%}$], [$a_{50\%}$], [$a_{80\%}$], [$i_{20\%}$], [$i_{50\%}$] and [$i_{80\%}$]. We refer to the number of vowel/time point combinations as n_{vt}.

Given n_{vt} vowel/time point combinations and $Fset$ being the set of formants chosen by the user to be included in the distance measurements, the Euclidean distance between two speakers i and j is calculated as follows:

$$dist_k[i,j] = \sqrt{\sum_{c=1}^{n_{vt}} \sum_{f \in Fset} (F_{cfi} - F_{cfj})^2}$$

8.3 ACCDIST distance

The ACCDIST metric compares speakers on the basis of their vowel systems Huckvale (2004). We extended Huckvale's method by offering the possibility to

include F3 as well (in addition to F1 and F2).

8.3.1 Inter-vowel distances

Using the average F1 and/or F2 and/or F3 for any combination of the included speakers, time points and vowels, for each speaker the inter-vowel distances are calculated.

In case formant frequencies are measured for n_v vowels and at n_t time points per vowel, we consider this as a set of $n_v \times n_t$ different vowels among which distances are calculated for each speaker. For example, assume vowels [a] and [i] with measurements at the 20%, 50% and 80% time point. We consider them as six different vowels: [$a_{20\%}$], [$a_{50\%}$], [$a_{80\%}$], [$i_{20\%}$], [$i_{50\%}$] and [$i_{80\%}$]. We refer to the number of vowel/time point combinations as n_{vt}.

Given n_{vt} vowel/time point combinations and $Fset$ being the set of formants chosen by the user to be included in the distance measurements, the Euclidean inter-vowel distances of a speaker k are calculated as follows:

```
for (i in 2:n_vt)
{
  for (j in 1:(i-1))
  {
    dist_k[i,j] = sqrt( sum_{f in Fset} (F_fi - F_fj)^2 )
  }
}
```

where the Euclidean inter-vowel distances are stored in matrix $dist_k$.

8.3.2 Inter-speaker distances

Given n_s speakers the *similarities* between the speakers are calculated as:

```
for (i in 2:n_s)
{
  for (j in 1:(i-1))
  {
    DIST[i,j] = cor(dist_i, dist_j)
  }
}
```

In this formula the function *cor* calculates the Pearson correlation coefficient. The inter-speaker similarities are stored in the matrix $DIST$. When a user downloads the table in the 'Explore' panel, s/he obtains the matrix $DIST$.

In order to be able to apply cluster analysis and multidimensional scaling to the measurements, each similarity measurement r is converted to a distance by calculating $1 - r$.

8.4 Distances among speaker groups

When the option 'summarize' is checked, groups of speakers are compared to each other, where the groups are defined according to the categorical variables that are chosen under 'Sel. variable' and by the categories of those variables that are selected under 'Sel. categories'. Assume group 1 with speakers A, B and C, and group 2 with speakers X and Y, than similarity (and subsequently distance) is calculated as the average similarity of the speaker pairs AX, AY, BX, BY, CX and CY.

8.5 Cluster analysis

Huckvale (2004) uses cluster analysis in order to visualize the relationships between the speakers (and their accents). In Visible Vowels the user can choose from give cluster methods: Single-linkage, complete-linkage, UPGMA (or: group average), WPGMA (or: McQuitty) and the Ward's method. For more information about these methods see Jain and Dubes (1988). There exist two versions of the Ward's method: 'Ward1' and 'Ward2'. Murtagh and Legendre (2014) showed that 'Ward2' correctly implements Ward Jr. (1963)'s clustering criterion, therefore this version is used in Visible Vowels.

When using any cluster method the amount of variance in the distance matrix explained by the dendrogram is given. The explained variance is calculated as the squared cophenetic correlation coefficient. The cophenetic correlation coefficient is a measure of the agreement between the distances as implied by the dendrogram –the cophenetic distances– and those of the original distance matrix (Sokal and Rohlf, 1962). For finding the cophenetic distance between objects i and j we have to find the least signifcant (smallest) cluster in which both objects are first present. The cophenetic distance between i and j is equal to the distance between the subclusters of this cluster.

8.6 Multidimensional scaling

The ACCDIST distances can also be visualized by means of multidimensional scaling (MDS) procedures. These procedures visualize the multidimensional data by giving each speaker (or groups or speakers when the option 'summarize' is checked) a location in a two-dimensional map. There are four procedures: classical MDS (Torgerson, 1952, 1958), Kruskal's Non-metric MDS (Shepard, 1962; Kruskal, 1964; Kruskal and Wish, 1978), Sammon's Non-Linear Mapping (Sammon, 1969) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008; Van der Maaten, 2014).

As to the latter procedure, Van der Maaten and Hinton (2008) write that "t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales" and that t-SNE "produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map" which we found a justification for adding this method to the other –and maybe more well known– procedures.

Visible Vowels uses the function `Rtsne` of the R package `Rtsne` (Krijthe, 2015). This function has a perplexity parameter which is related to the number of nearest neighbors considered when placing each data point. When using lower values more attention is paid to the local aspects of the data, and the use of higher values increases the impact of more distant neighbors and global structure.

Van der Maaten and Hinton (2008) write that "typical values are between 5 and 50". The perplexity value is set to 30 by default. When the number of speakers (or groups of speakers) is smaller than 91, this would result in an error message: "Perplexity is too large". In order to get the perplexity value as close as possible to the default value when the number of speakers is smaller than 91, we calculate:

$$perplexity = (n - 1)\ DIV\ 3 \tag{50}$$

where n is the number of speakers or speaker groups.

For multidimensional scaling methods the explained variance is calculated as the squared correlation between the Euclidean inter-point distances of the two-dimensional plot and the distances of the original distance matrix.

8.7 Nota bene

When the same speaker pronounces vowels of different languages or under different conditions, different speaker labels should be used for each language or condition. For example: if a speaker X pronounces vowels of Dutch, Frisian and German, use label X_{Du} for the Dutch vowels, label X_{Fr} for the Frisian vowels and label X_{Ge} for the German vowels.

Distances between speakers are calculated when the user selects at least three vowels and five speakers, either by selecting five individual speakers or by one or more groups (defined by one or more categorical variables chosen under 'Sel. variable') that include at least five speakers.

References

Adank, P. (2003). *Vowel Normalization: a Perceptual-acoustic Study of Dutch Vowels*. PhD thesis, University of Nijmegen, Nijmegen.

Adank, P., Van Hout, R., and Smiths, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5):3099–3107. https://doi.org/10.1121/1.1795335.

Bigham, D. S. (2008). *Dialect Contact and Accommodation among Emerging Adults in a University Setting*. PhD thesis, University of Texas at Austin, Austin.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [computer program]. Version 6.0.36, retrieved 11 November 2017 from http://www.praat.org/.

Eddy, W. F. (1977). A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, 3(4):398–403 and 411–412.

Esfandiaria, N. and Alinezhadb, B. (2014). Evaluating normalization procedures on reducing the effect of gender in Persian vowel space. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 13(2):303–316.

Fabricius, A. H., Watt, D., and Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalisation algorithms for sociophonetics. *Language Variation and Change*, 21(3):413–435.

Fant, G. (1968). Analysis and synthesis of speech processes. In Malmberg, B., editor, *Manual of phonetics*, pages 173–177. North-Holland Publishing Comp., Amsterdam.

Fant, G., Kruckenberg, A., Gustafson, K., and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish. In *Speech Prosody 2002, International Conference; Aix-en-Provence, France, April 11-13, 2002*, pages 283–286.

Flynn, N. (2011). Comparing vowel formant normalisation procedures. *York Papers in Linguistics*, 2(11):1–28.

Flynn, N. and Foulkes, P. (2011). Comparing vowel formant normalization methods. In Lee, W. and Zee, E., editors, *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII): August 17-21, 2011*, pages 683–686. City University of Hong Kong.

Fox, R. A. and Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *Journal of the Acoustical Society of America*, 126(5):2603–2618. https://doi.org/10.1121/1.3212921.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions of Audio Electroacoustics*, AU-16:78–80.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103–138.

Greenword, D. D. (1961). Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustical Society of America*, 33(10):1344–1356. https://doi.org/10.1121/1.1908437.

Heeringa, W. and Van de Velde, H. (2021). A new vowel normalization for sociophonetics. In *Proceedings of the 22th Annual Conference of the International Speech Communication Association (Interspeech 2021), August 30-September 3, 2021, Brno, Czech Republic.*

Huckvale, M. (2004). ACCDIST: a Metric for Comparing Speakers' Accents. In *Proceedings of the International Conference on Spoken Language Processing, Jeju, Korea, Oct 2004*, pages 29–32.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey.

Johnson, . (2005). Speaker normalization in speech perception. In Pisoni, D. B. and Remez, R. E., editors, *The handbook of speech perception*, pages 363–389. Blackwell Publishers, Oxford.

Kachkovskaia, T. (2014). Phrase-final lengthening in russian: Pre-boundayr or pre-pausal? In Ronzhin, A., Potapova, R., and Delic, V., editors, *Speech and Computer; 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5–9, 2014. Proceedings*, volume LNAI 8773, pages 353–359. Springer.

Kocharov, D., Kachkovskaia, T., and Skrelin, P. (2015). Position-dependent vowel reduction in russian. In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of the 18th International Congress of Phonetic Sciences*, number 0402, Glasgow UK. University of Glasgow.

Krijthe, J. H. (2015). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. https://cran.r-project.org/web/packages/Rtsne/. R package version 0.13.

Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis. *Psychometrika*, 29:1–28.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Number 07–011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change. A Multimedia Reference Tool, Volume 1*. Mouton de Gruyter, Berlin.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49(2):606–608. https://doi.org/10.1121/1.1912396.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85(5):2114–2134. https://doi.org/10.1121/1.397862.

Mohanan, P. J. and Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and cognitive processes*, 25(6):808–839. https://doi.org/10.1080/01690965.2010.490047.

Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753. https://doi.org/10.1121/1.389861.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31(3):274–295. https://doi.org/10.1007/s00357–014–9161–z.

Nearey, T. (1978). *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club, Bloomington IN.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In Solé, M. J., Recasens, D., and Romero, J., editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, volume 39, pages 771–774, Barcelona. Causal Productions Pty Ltd.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine*. Addison-Wesley Publishing Company, Reading, Massachusetts.

Peterson, G. E. (1951). The phonetic value of vowels. *Language*, pages 541–553. https://doi.org/10.2307/410041.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184. https://doi.org/10.1121/1.1906875.

Rietveld, A. C. M. and Van Heuven, V. J. (1997). *Algemene fonetiek*. Coutinho, Bussum.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C 18:401–409.

Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66(6):1647–1652. https://doi.org/10.1121/1.383662.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125–140, 219–246.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11:33–40.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, 28(1):12–23. https://doi.org/10.1016/0093–934X(86)90087–8.

Syrdal, A. K. and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of american english vowels. *Journal of the Acoustical Society of America*, 79(4):1086–1100. https://doi.org/10.1121/1.393381.

Thomas, E. R. and Kendall, T. (2007). NORM: The Vowel Normalisation and Plotting Suite. Accessed 21 January 2018 at http://lingtools.uoregon.edu/norm/.

Torgerson, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika*, 17.4:401–419.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

Traunmüller, H. (1983). *On Vowels. Perception of Spectral Features, Related Aspects of Production and Sociophonetic Dilllensions*. PhD thesis, University of Stockholm, Stockholm.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88:97–100. https://doi.org/10.1121/1.399849.

Van der Harst, S. (2011). *The Vowel Space Paradox: A Sociophonetic Study on Dutch*. PhD thesis, Radboud University, Nijmegen.

Van der Maaten, L. J. P. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(Oct.):3221–3245.

Van der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov.):2579–2605.

Wang, H. and Van Heuven, V. J. (2006). Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. In van de Weijer, J. and Los, B., editors, *Linguistics in the Netherlands 2006*, volume 23, pages 237–248. https://doi.org/10.1075/avt.23.23wan. John Benjamins, Amsterdam.

Ward Jr., J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845.

Watt, D. J. L. and Fabricius, A. H. (2002). Evaluation of a technique for improving the mapping of multiple speakers. *Leeds Working Papers in Linguistics and Phonetics*, 9:159–173.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33(2):248–248. https://doi.org/10.1121/1.1908630.

Zwicker, E. and Terhardt, E. (1980). Analytical expressions for criticalband rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68(5):1523–1525. https://doi.org/10.1121/1.385079.